11-1-2016

# Vol. 15, No. 2 (Full Issue) 

JMASM Editors

Follow this and additional works at: http:// digitalcommons.wayne.edu/jmasm
Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

## Recommended Citation

Editors, JMASM (2016) "Vol. 15, No. 2 (Full Issue)," Journal of Modern Applied Statistical Methods: Vol. 15: Iss. 2, Article 1.
DOI: 10.22237/jmasm/1478004480
Available at: http://digitalcommons.wayne.edu/jmasm/vol15/iss2/1

```
do i1 = 1,4
    j(1) = i1
        do i2 = 1,4
            j(2) = i2
            do i3=1,4
                j(3) = i3
                    do i4=1,4
                        j(4) = i4
                                if (j(1) .eq. j(2) .or. j(1) .eq. j(3) .or. j(1) .eq. j(4)) cycle
                                if (j(2) .eq. j(3).or. j(2) .eq. j(4)) cycle
                                if (j(3).eq. j(4)) cycle
                            print*.j(1).j(2).j(3).j(4)
                end do
            end do
        end do
end do
```


# Journal of Modern Applied Statistical Methods 

## Invited Articles

C. R. Rao, Miodrag Lovric, Rand R. Wilcox and Timothy Hayes
with response from
Bruce D. Zumbo, Edward Kroc, and Shlomo S. Sawilowsky

Vol. 15, No. 2 • November, 2016

# Journal of <br> Modern Applied Statistical Methods 

Shlomo S. Sawilowsky<br>SENIOR EDITOR<br>College of Education<br>Wayne State University

Jack Sawilowsky<br>EDITOR<br>Reason Statistical Consulting

Harvey Keselman<br>ASSOCIATE EDITOR EMERITUS

Department of Psychology
University of Manitoba

## Alan Klockars <br> ASSISTANT EDITOR EMERITUS

Educational Psychology
University of Washington

Bruno D. Zumbo<br>ASSOCIATE EDITOR<br>Measurement, Evaluation, \& Research Methodology<br>University of British Columbia

Clayton Hayes<br>EDITORIAL ASSISTANCE

Vance W. Berger<br>ASSISTANT EDITOR<br>Biometry Research Group<br>National Cancer Institute

Joshua Neds-Fox<br>Editorial assistance

Todd C. Headrick ASSISTANT EDITOR<br>Educational Psychology \& Special Education So. Illinois UniversityCarbondale

Heather Marie Perrone<br>EDITORIAL ASSISTANCE

$J M A S M$ (ISSN 1538-9472, http://digitalcommons.wayne.edu/jmasm) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. JMASM seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computerintensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo- random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

# Journal of Modern Applied Statistical Methods 

Vol. 15, No. 2
$\approx$ November 2016 な
Table of Contents

## Invited Articles

| 2-21 | C. R. RAO <br> M. M. LOVRIC | Testing Point Null Hypothesis of a Normal Mean and the Truth: $21^{\text {st }}$ Century Perspective |
| :---: | :---: | :---: |
| 22-32 | S. S. SAWILOWSKY | Rao-Lovric and the Triwizard Point Null Hypothesis Tournament |
| $33-40$ | B. D. ZUMBO <br> E. KROC | Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: $21^{\text {st }}$ Century Perspective |
| 41-52 | R. R. WILCOX <br> T. HAYES | Within Groups ANOVA When Using a Robust Multivariate Measure of Location |
| Regular Articles |  |  |
| 53-66 | J. STEPHENS | Longitudinal Stability of Effect Sizes in Education Research |
| 67-89 | W. K.. LIM <br> A. W. LIM | A Comparison of Usual $t$-Test Statistic and Modified $t$-Test Statistics on Skewed Distribution Functions |
| 90-111 | A. I. AL-OMARI <br> C. N. BOUZA <br> D. COVARRUBIAS <br> R. PAL | A New Estimator of the Population Mean: An Application to Bioleaching Studies |
| $112-134$ | $\begin{aligned} & \text { T. N. SINDHU } \\ & \text { N. FEROZE } \\ & \text { M. ASLAM } \end{aligned}$ | Study of the Left Censored Data from the Gumbel Type II Distribution under a Bayesian Approach |

Journal of Modern Applied Statistical Methods November 2016, Vol. 15, No. 2, ii-vii doi: 10.22237/jmasm/1478001600

| $135-148$ | B. LANTZ |
| :--- | :--- |
|  | R. ANDERSSON |
|  | P. MANFREDSSON |

$149-170$
$171-186$

187 - 206

207-214

215-223 D. RAHARDJA

224-237 P. MEYVISCH

238-272 P. WANG
$273-292$

293-321

322-349

350-361
G. O. KADILAR
Y. YANG
Z. ZHANG
P. SA
A. K. SHARMA G. N. SINGH
R. MAJI
A. BANDYOPADHYAY G. N. SINGH
G. N. SINGH
A. K. SINGH
A. K. SHARMA
M. VOSSOUGHI S. SHAHVALI E. SADEGHI
I. PAPAGEORGIOU
T. N. SINDHU
N. FEROZE
M. ASLAM
O. E. OLUBUSOYE
O. S. YAYA
O. O. OJO

Preliminary Tests of Normality When Comparing Three Independent Samples

Estimation of Population Mean on Recent Occasion under Non-Response in $h$-Occasion Successive Sampling

Efficient and Unbiased Estimation Procedure of Population Mean in Two-Phase Sampling

An Improved Generalized Estimation Procedure of Current Population Mean in Two-Occasion Successive Sampling

A New Exponential Type Estimator for the Population Mean in Simple Random Sampling

A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects

Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood

A New Test for Correlation on Bivariate Nonnormal Distributions

Comparison of Some Multivariate Nonparametric Tests in Profile Analysis to Repeated Measurements

Optimal Estimation and Sampling Allocation in Survey Sampling Under a General Correlated Superpopulation Model

Doubly Censored Data from Two-Component Mixture of Inverse Weibull Distributions: Theory and Applications

Misspecification of Variants of Autoregressive GARCH models and Effect on In-Sample Forecasting

Journal of Modern Applied Statistical Methods November 2016, Vol. 15, No. 2, ii-vii doi: 10.22237/jmasm/1478001600

362-381 A. V. DORUGADE

382-399 E. C. NWOGU I. S. IWUEZE V. U. NLEBEDIM
$\begin{array}{ll}400-410 & \text { G. KHALAF } \\ & \text { M. IGUERNANE }\end{array}$

411-427 C. G. UDOMBOSO G. N. AMAHIA I. K. DONTWI

428-454 R. R. L. KANTAM M. S. RAVIKUMAR
$455-474$
R. W. WALKER
$475-494$
$495-509$
$510-525$
$526-535$
K. A. RAO
J. G. D'CUNHA
$536-562$
G. M. JAYCNA
S. L. ROSEN

563-579 H. S. RODRIGO
C. P. TSOKOS
T. SHARAF
$580-626$
J. N. DYER

Copyright © 2016 JMASM, Inc. ISSN 1538-9472

Improved Ridge Estimator in Linear Regression with Multicollinearity, Heteroscedastic Errors and Outliers

Some Tests for Seasonality in Time Series Data

Multicollinearity and a Ridge Parameter Estimation Approach

An Adjusted Network Information Criterion for Model Selection in Statistical Neural Network Models

Limited Failure Censored Life Test Sampling Plan in Burr Type X Distribution

On Generalizing Cumulative Ordered Regression Models

Estimation of Parameters of Misclassified Size Biased Borel Distribution

Estimating the Parameter of Exponential Distribution under Type II Censoring From Fuzzy Data

The Application of Legendre Multiwavelet Functions in Image Compression

Bayesian Inference for Median of the Lognormal Distribution

Developing Bayesian-based Confidence Bounds for Non-identically Distributed Observations using the Lyapunov Condition

Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach

Monte Carlo Simulation Design for Evaluating Normal-Based Control Chart Properties

Journal of Modern Applied Statistical Methods November 2016, Vol. 15, No. 2, ii-vii doi: 10.22237/jmasm/1478001600

| 627-642 | L. QIN <br> L. WEISSFELD <br> M. LEVINE <br> M. MARCUS <br> F. DAI | Latent Variable Model for Weight Gain Prevention Data with Informative Intermittent Missingness |
| :---: | :---: | :---: |
| 643-655 | P. NASIRI <br> H. ESFANDYARIFAR | E-Bayesian Estimation of the Parameter of the Logarithmic Series Distribution |
| 656-670 | $\begin{aligned} & \text { S. NAQASH } \\ & \text { S. P. AHMAD } \\ & \text { A. AHMED } \end{aligned}$ | Bayesian Analysis of Generalized Exponential Distribution |
| 671-695 | D. R. BAROT <br> M. N. PATEL | Hierarchical Bayes Estimation of Reliability Indexes of Cold Standby Series System under General Progressive Type II Censoring Scheme |
| 696-702 | A. HOSSIANZADEH K. ZARE | Bayesian Analysis of Discrete Skewed Laplace Distribution |
| $703-722$ | K. K. SURESH <br> S. UMAMAHESWARI | Designing of Bayesian Skip Lot Sampling Plan under Destructive Testing |
| $723-736$ | G. KRSTIĆ <br> N. S. KRSTIĆ <br> M. ZAMBRANO-BIGI | The br2-weighting Method for Estimating the Effects of Air Pollution on Population Health |

Algorithms and Code

| $737-742$ | S. A. ROSE <br> B. MARKMAN | Monte Carlo Simulations for Structural <br> Equation Modelling (Revolution R) |
| :--- | :--- | :--- |
|  | W. M. A. W. AHMAD | An Alternative Method for Multiple Linear |
| $743-754$ | M. A. A. NAWI | Model Regression Modeling, a Technical <br> Combining of Robust, Bootstrap and |
|  | N. A. ALENG | Fuzzy Approach |
|  | M. SHAFIQ |  |
|  | S. O. OGUNDELE | An Alternative Algorithm and R Programming |
|  | J. I. MBEGBU | Implementation for Least Absolute Deviation |
|  | C. R. NWOSU | Estimator of the Linear Regression Models |

## Statistical Software Applications \& Review

| $768-787$ | S. SU | Fitting Flexible Parametric Regression Models <br> with GLDreg in R |
| :--- | :--- | :--- |

## Emerging Scholars

| $788-820$ | M. ALMHEIDAT <br> C. LEE | A Generalization of the Weibull Distribution <br> with Applications |
| :--- | :--- | :--- |
|  | F. FAMOYE |  |

## Letter to the Editor

| $821-824$ | G. L. BAIRD | Reflections Concerning Recent Ban on NHST |
| :--- | :--- | :--- |
|  | S. R. DUERR | and Confidence Intervals |

## Invited Articles

# Testing Point Null Hypothesis of a Normal Mean and the Truth: $21^{\text {st }}$ Century Perspective 

Calyampudi Radhakrishna Rao<br>Penn State University<br>State College, PA

Miodrag M. Lovric<br>University of Kragujevac<br>Kragujevac, Serbia

Testing a point (sharp) null hypothesis is arguably the most widely used statistical inferential procedure in many fields of scientific research, nevertheless, the most controversial, and misapprehended. Since 1935 when Buchanan-Wollaston raised the first criticism against hypothesis testing, this foundational field of statistics has drawn increasingly active and stronger opposition, including draconian suggestions that statistical significance testing should be abandoned or even banned. Statisticians should stop ignoring these accumulated and significant anomalies within the current point-null hypotheses paradigm and rebuild healthy foundations of statistical science. The foundation for a paradigm shift in testing statistical hypotheses is suggested, which is testing interval null hypotheses based on implications of the Zero probability paradox. It states that in a real-world research point-null hypothesis of a normal mean has zero probability. This implies that formulated point-null hypothesis of a mean in the context of the simple normal model is almost surely false. Thus, Zero probability paradox points to the root cause of so-called large $n$ problem in significance testing. It discloses that there is no point in searching for a cure under the current point-null paradigm.

Keywords: zero-probability paradox, point null hypothesis, Lebesgue measure, rational numbers, algebraic numbers, almost sure false null hypothesis, inexactification, paradigm shift in testing statistical hypotheses.
> "It cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved"

Fisher (1922)

[^0]
## RAO \& LOVRIC

## Introduction

Following Fisher's foundational contribution to significance tests, and Neyman and Pearson to hypothesis tests, statistical testing has become widely adopted by researchers as the most common statistical inferential approach in almost all different branches of science. However, there has been a steadily growing dissatisfaction in the scientific community with traditional tests of the point (sharp, precise) null hypothesis. Since Buchanan-Wollaston (1935) raised the first criticism against significance testing, their application has been debated extensively, and numerous objections and severe complaints have been leveled against their utility. Critics also accentuated statistical tests are not only overused, but are often misunderstood and misused. Nickerson (2000) provided a summary of common misconceptions, and criticisms as well as arguments in support of null hypothesis testing, from a non-statistician viewpoint.

The most trenchant critics requested significance tests should be abandoned, banned or deinstitutionalized (e.g., Lindley, 1975; Hunter, 1997; Armstrong, 2007; Orlitzky, 2012). The editors of the American Journal of Public Health imposed a ban, although it only lasted two years. Similarly, in 1997 the officers of the American Psychological Association (APA) created a task force to make recommendations about appropriate statistical practice and to consider banning significance testing. The proposal was regarded as too extreme and was rejected (Wilkinson, 1999). More recently, in 2015, the editors of Basic and Applied Social Psychology journal enforced a ban on significance testing (as well as confidence intervals). On behalf of the ASA Board of Directors, Wasserstein \& Lazar (2016) formulated six principles regarding the usage of p -values, hoping that the ASA statement would open a fresh discussion with regards to the use of statistical inference.

The ASA's statement should be praised as the first organized reply from statistics community to the abovementioned issues. However, it did not address the fundamental problems and did not provide a new perspective on statistical testing.

Critics advocated reform of statistical inference and statistics education. They recommended less emphasis should be placed on reporting of $p$ values, cynically termed "harvest of asterisks" (Cohen, 1990). The reformers, mainly non-statisticians, argued attention should be shifted to effect size, point estimation, confidence interval, information theoretic approaches (e.g., Akaike Information Criterion), graphical methods, and progressively more on the communication of results using Bayesian inference.

Consider two of the most important criticisms of significance testing: (1) point null hypotheses are unlikely to be true, and (2) a statistical significant result is always obtainable with a sufficiently large sample. The scope of this paper is limited to the problem of testing the mean of a normal distribution, although this problem is of substantial importance because of its widespread application in statistical theory and practice. The primary objective is to prove that in the realworld research when testing the mean of a normal distribution using a point-null hypothesis, the probability of that hypothesis is zero. We call this result the Zero probability paradox. This paradox undoubtedly reveals logical deficiency of a point-null hypothesis of a normal mean: in reality, its testing is actually a procedure that unequivocally will lead (with sufficiently large sample) to a foregone conclusion that formulated null hypothesis is almost surely false. The logical name for this procedure in which a sharp null hypothesis is ultimately being rejected should be "inexactification," rather than testing (Good, 1994, p. 241).

## The Existence of Point Null Hypothesis: History and Overview

Testing a point null hypothesis is arguably the most widely used and at the same time the most controversial, misapprehended and severely criticized statistical procedure in many fields of scientific research. Focus on one of the most common criticisms, that point null hypotheses are not realistic. The Zero probability paradox, presented here, evolved as a result of persuasive and accumulated ideas of statisticians, and non-statisticians referred to in this section.

There is a vast amount of references in statistics and non-statistics literature with the claim that, in reality, point null hypotheses are almost always false. Critics, however, supported this statement only by intuitive arguments, empirical evidence, and common sense. One of the early critics, L. J. Savage (1954, p. 254), disproved the validity of tests "in which the null hypothesis is such that it would not really be accepted by anyone." I. R. Savage, (1957, p. 332-333) asserted the "null hypotheses of no difference are usually known to be false before the data are collected...when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science." Nunnally (1960, p. 642) expressed a similar assertion, but admitted he agreed although he cannot prove it directly. However, he argued it is supported both by common sense and by practical experience. Likewise, Meehl, (1967, p. 108) pointed out there is "universal agreement that the old point-null hypothesis...is

## RAO \& LOVRIC

[quasi-] always false in biological and social science." His opinion was based on the result that in "psychological and sociological investigations involving very large numbers of subjects, it is regularly found that almost all correlations or differences between means are statistically significant" (p. 109). Meehl illustrated this by providing an example of a large sample of over 55,000 Minnesota high school seniors that revealed $91 \%$ significant associations among a collection of 45 variables.

In the same way, Cohen (1990, p. 1308) stated the null hypothesis "taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null is always false, what's the big deal about rejecting it?"

There is near consensus in the literature that exactly true point null hypotheses are extremely rare in reality. This is exemplified by the following by Kadane (1987, p. 347): "For the last 15 or so years I have been looking for applied cases in which I might have some serious belief in a null hypothesis. In that time I found only one [testing an astrologer claim that on the bases of peoples birthdays it is possible to predict who is likely to have a drug problem]... I do not expect to test a precise hypothesis as a serious statistical calculation."

In a similar manner, there was a quest for an existence of a realistic case for which a null hypothesis cannot be regarded beforehand as false. As a result of this pursuit, a commonly given example is found, that there is no extrasensory effect in a parapsychological experiment. Good (1994, p. 241) argued there is at least one example of a precisely sharp null hypothesis: precognition is impossible. Similarly, Ghosh et al. (2006, p. 45) suggested astrology cannot predict the future. Berger and Delampady (1987, p. 320), although admitting that it is perhaps impossible to have a null hypothesis that can be exactly modeled as $\theta=\theta_{0}$, noted talking to plants has no effect on their growth. Nevertheless, they admitted minor biases in the design of the experiments may produce statistical significance. They also argued that point null hypotheses are reasonable approximations to fuzzy precise (small interval) nulls. However, as pointed out by Bernardo (1999, p. 102) "this approximation always breaks down for sufficiently large samples." Likewise, Rousseau (2007) showed for large samples the Bayes factor associated with point null hypotheses is a poor approximation of Bayes factors of interval null hypotheses unless the intervals are extremely small.

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

In contrast, Zellner (1987, p. 339) emphasized many realistic examples of point null hypotheses can be given in testing well-formulated physical laws, such as $\mathrm{s}=.5 \mathrm{gt}^{2}$ and $\mathrm{E}=\mathrm{mc}^{2}$. Kass and Raftery (1995, p. 788) argued although "one rarely believes a scientific law in an absolute sense, it is a great convenience to speak and to act as if laws are valid. When one says that a certain theory is correct, one means that deviations from it are sufficiently minor to be irrelevant for all practical purposes at hand."

Based on the above arguments, a natural question arises: why are we testing point null hypotheses at all, when it is known in advance they are almost never exactly true in the real world? Sprenger (2013) argued these hypotheses often give useful idealization of reality. He considered this originated in the Popperian philosophy of science: "only a highly testable or improbable theory is worth testing and is actually (and not only potentially) satisfactory if it withstands severe tests." (Popper, 1963, p. 219-220)

According to Cox (2006, p. 31) null hypothesis refers to a probability model, and this implies idealization. He argued it would be absurd to think that a mathematical model could be an exact representation of a real system. Thus, null hypotheses are postulated within a system that is untrue.

Good (1956, p. 254) remarked a null hypothesis is tested, although it is known in advance it cannot be exactly true, because "we wish to test whether the hypothesis is in some sense approximately true, or whether it is rejectable on the sort of size of sample that we intend to take." Kruskal (1968) indicated the need is to test whether the mean is near $\mu_{0}$, meaning as near as makes no substantive difference. He stated this will be achieved as long as the sample sizes and significance levels are reasonable and the power is at least moderately large for alternatives interestingly different from the null hypothesis.

Edwards, et al. (1963) presented a Bayesian view on the sharp null hypothesis problem. They acknowledged in usual applications the null hypothesis is known to be false from the outset, because realistically the null hypothesis cannot be infinitely sharp. From a Bayesian perspective, a sharp null hypothesis is likely to be appropriate only when it deserves special initial credence. They also highlighted in Bayesian analysis the null hypothesis is "a hazily defined small region rather than a point [italicized by authors]" (p. 235).

Finally, consider Krueger's (2001) attempt to explain why all null hypotheses are false. He started from the premise that in statistics populations are mathematical abstractions that contain infinite possible observations. "This implies an infinite number of possible states of the population, and each of these states may be a distinct hypothesis. With an infinite number of hypotheses, no
individual hypothesis can be true with any calculable probability" (p. 17). It is, however, clear that his arguments on the survival of the flawed significance testing are themselves flawed. It is erroneous to claim that one-sided and interval null hypotheses are always false.

It can be concluded existing literature does not offer proof of the extraordinary statement that all point null hypotheses are false.

## The Nature of a Point Null Hypothesis

Before exposing the Zero probability paradox, it is of fundamental importance to clarify some misconceptions about the nature of the point null hypothesis.

Suppose that a random sample of size $n, X=\left(X_{1}, X_{2}, \ldots, X_{n}\right)$, is selected from the normal population $N\left(\theta, \sigma^{2}\right)$, where $\theta$ is an unknown mean assuming values in a parameter space $\Theta \subset \mathbb{R}^{1}$. Suppose also that the variance, $\sigma^{2}>0$ is known. It is required to test the null hypothesis $H_{0}: \theta=\theta_{0}$ versus an unspecified alternative hypothesis $H_{1}: \theta \neq \theta_{0}$. Regard this sharp or point null hypothesis as a numerically exact statement, that is free of vagueness and ambiguity, namely as an assertion that exactly specifies a single value of a parameter $\theta_{0}$. In other words, it is obvious that $\theta_{0}$ as a crisp number, not a fuzzy number.

It is well known that to every real number there corresponds a unique point on the number line and vice versa. Obviously, point hypothetical value $\theta_{0}$ corresponds to a distinctive point on the real number line, not to an interval. As Euclid gave an intuitive definition in the first sentence from his Elements book 1, "a point is that which has no part, or which has no magnitude." In the contemporary notion, this is tantamount to saying that a point is a dimensionless entity that has only a location. It also naturally implies that "every point is unextended" (Playfair, 1819, p. 289).

Claims that there are different kinds of sharp hypotheses, some fuzzy sharp and some infinitely sharp, in other words, that equal sign can be perceived in infinitely different ways, are unconvincing. If testing "hazily defined small region" is considered a null hypothesis in a scientific, non-subjective way, then it is a sine qua non to formulate that hypothesis accurately, for example, as $H_{0}:\left|\theta-\theta_{0}\right| \leq \delta$ or using fuzzy set theory as $H_{0}: \tilde{\theta}=\tilde{\theta}_{0}$, where $\tilde{\theta}$ is the unknown fuzzy parameter and $\tilde{\theta}_{0}$ a known fuzzy number. However, in the traditional point null hypothesis $H_{0}: \theta=\theta_{0}$, in practice, (since the pioneering work of Arbuthnott (1710)) $\theta_{0}$ has always been formulated as a crisp rational number, never as a fuzzy number $\tilde{\theta}$.

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

A fuzzy number, $\tilde{\theta}$, in contrast, is a distinctly different entity. It is defined as a fuzzy set in $\mathbb{R}$ with a normal, fuzzy convex and a continuous membership function of bounded support. Note also that, in the fuzzy set framework, the possible values of the parameter of interest are expressed as linguistic variables, and that the data are observations of a normal fuzzy random sample. In conclusion, $\theta_{0}=\tilde{\theta}_{0}$, that is, (Crisp number $=$ Fuzzy number $)$, is nothing else but a selfdeception.

## Zero Probability Paradox

In a real-world research, the probability of an exact point-null hypothesis of the mean of a normally distributed population is zero. Let $\mathbb{Q}$ be the set of all rational real numbers, that is $\mathbb{Q}=\{m / n ; m, n \in \mathbb{Z}, n \neq 0\}$, where $\mathbb{Z}$ stands for the set of all integers. Suppose, as in the previous section, that a random sample of size $n$, $X=\left(X_{1}, X_{2}, \ldots, X_{n}\right)$, is selected from the normal population $N\left(\theta, \sigma^{2}\right)$, where $\theta$ is an unknown mean assuming values in a parameter space $\Theta \subset \mathbb{R}^{1}$. Divide parameter space into two disjoints sets $\Theta_{\mathbb{Q}}$ and $\Theta_{\mathbb{R} \mid \mathbb{Q}}$ that are mutually exclusive $\left(\Theta_{\mathbb{Q}} \cap \Theta_{\mathbb{R} \backslash \mathbb{Q}}=\varnothing\right)$ and exhaustive $\left(\Theta=\Theta_{\mathbb{Q}} \cup \Theta_{\mathbb{R} \backslash \mathbb{Q}}\right)$. Suppose further that the set $\Theta_{\mathbb{Q}}$ is equivalent to the set of all rational numbers $\mathbb{Q}$ and that $\Theta_{\mathbb{R} \mathbb{Q}}$ is equivalent to the set of all irrational numbers $\mathbb{R} \backslash \mathbb{Q}$.

It is desired to test the traditional null hypothesis

$$
\begin{equation*}
H_{0}: \theta=\theta_{0} \tag{1}
\end{equation*}
$$

versus an unspecified alternative hypothesis

$$
H_{1}: \theta \neq \theta_{0},
$$

where $\theta_{0}$ is a rational number, i.e. $\theta_{0} \in \Theta_{\mathbb{Q}}$

Point-null zero probability paradox (Zero Probability paradox). Probability of the null hypothesis (1) is equal to zero:

$$
\left.P\left(H_{0} \mid \forall \theta \in \Theta_{\mathbb{Q}}\right\}\right)=0 .
$$

## RAO \& LOVRIC

This is tantamount to saying that probability of the null hypothesis

$$
\begin{aligned}
& P\left(H_{0} \mid \theta \in\{\text { All rational numbers }\}\right)=0, \text { and } \\
& P\left(H_{1} \mid \theta \in\{\text { All irrational numbers }\}\right)=1 .
\end{aligned}
$$

Here, regard rationals on the number line as indicators of the means of corresponding normal distributions that have rational numbers as their means.

## Proof:

A) In scientific research and statistical practice, any point null hypothesis of the normal population is almost always stated as a single rational number.
B) As proved by Cantor in 1873, rational numbers are countable-that is, there is in one-one correspondence between the rational numbers and the natural numbers (see, for example, Calkin and Wilf, 2000, for a binary tree argument). Because the rational numbers, $q_{i}$, are countable, enumerate them as a sequence $\left\{q_{i}\right\}$, or $\mathbb{Q}=\bigcup_{i=1}^{\infty}\left\{q_{i}\right\}$. Hence, the set of all hypothetical null values of the point-null hypotheses that could be expressed using rational numbers, $\Theta_{\mathbb{Q}}$, is also countable. In other words, this set has a bijective correspondence to the set of rational numbers.
C) The Lebesgue measure of any singleton set, $\{x\}$, is zero (where singleton means the smallest possible nonempty set). Every countable set has Lebesgue measure zero (see, for example, Adams and Guillemin, 1996, p. 9). Therefore, Lebesgue measure of the set of all rational numbers is also zero, that is $\lambda(\mathbb{Q})=\lambda\left(\bigcup_{i=1}^{\infty}\left\{q_{i}\right\}\right)=\sum_{i=1}^{\infty} \lambda\left(\left\{q_{i}\right\}=0\right.$.

In light of this fact, Lebesgue measure of the set of all hypothetical null values of the point-null hypotheses that could be expressed using rational numbers ( $H_{0}: \theta \in \Theta_{\mathbb{Q}}$ ) is also zero because this set is countable, $\lambda\left(\Theta_{\mathbb{Q}}\right)=0$.
D) Normal distribution is absolutely continuous with respect to the Lebesgue measure $\lambda$. This signifies that all sets which have zero Lebesgue measure must also have zero probability under probability

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

measure; i.e., for all events $\mathrm{A} \in \mathrm{R}$ such that $\mu(\mathrm{A})=0 \rightarrow \mathrm{P}_{X}(\mathrm{~A})=0$. As Borovkov (2013, p. 39) has nicely exemplified "for an absolutely continuous distribution, the probability of hitting a set of zero Lebesgue measure is zero."
E) Because for an absolutely continuous distribution, a countably infinite set of all rational numbers has Lebesgue measure zero, conclude their probability measure is also zero.
F) Therefore, probability measure of a set of all possible hypothetical null rational values of the point-null hypotheses in testing a normal mean is also zero, $P\left(\left\{H_{0} \mid \forall \theta \in \Theta_{\mathbb{Q}}\right\}\right)=0$. This unequivocally amounts to the deduction that any single-point null hypothesis about the normal mean has also probability zero, that is,

P (Point-null hypothesis formulated as a rational number | Normal distribution) $=0$.

## Quod erat demonstrandum.

Subsequently, the probability of point null formulated as an irrational number is one. Figuratively speaking, rationals occupy zero length on a real line and the set of irrationals is uncountably infinite.

The scope of the Zero probability paradox can be further extended to the even more general set of all point null hypotheses asserted as real algebraic numbers, that is, the roots of single variable polynomial equations whose coefficients are all integers. This set includes rational numbers, Gaussian integers, golden ratio, constructible numbers, some irrational numbers such as $\sqrt{3}$, etc. Because this set is countable, as also proved by Cantor in 1874, (see, for example, Kaplansky, 2001, Paradox 4, p. 23) it has Lebesgue measure zero and therefore under Gaussian distribution its probability is zero. The cardinality (a measure of the "number of elements of the set") of the algebraic numbers is $\aleph_{0}$ (aleph-naught), the same as the natural numbers and rational numbers. However, the cardinality of the set of transcendental numbers is the same as that of the set of real numbers $|\mathbb{R}|$, the cardinality of the continuum. Almost all real numbers are transcendental, but we are familiar with almost none of them (except, for example, $\pi, e$, Liouville numbers, Champernowne constant, etc.).

It is important to emphasize that the Zero probability paradox applies both in the case when population variance is known and unknown.

It might be objected that a point-null hypothesis that the mean of the errors made in astronomical observations is equal to zero is reasonable and that its probability could be larger than zero. Karl Pearson (1935a, p. 296) replied, "I have never found a normal curve fit anything if there are enough observations! The astronomical data provided to prove that errors of observation follow normal curves are pitiably scanty, and if proper tests are applied usually show that they do not!"

## Conclusion

## Zero Probability and Impossibility.

Before discussing some of the implications of the Zero probability paradox, it is of considerable interest to clarify the difference between zero probability and impossibility. The most common and persistent misconception in the literature about probability is the interpretation that zero probability implies that an event is impossible. This is equally shared by many applied statistics textbooks writers (for example, Everitt, 1999, p. 14; de Muth, 2014, p. 20; Burns \& Burns, 2008, p. 164; Sharma, 2010, p. 191) and non-statisticians (for example, Poole \& Mackworth, 2010, p. 296; Finlayson \& McMahon, 2004, p. 360; Yoe, 2012, p. 305; Quinn and Keough, 2002, p. 7). This does not come as a surprise since many notable scholars held the same false impression in the past.

As reported by Finetti (2008, p. 49), Borel used to say "let us consider the probabilities $10^{-3}, 10^{-10}, 10^{-100}, 10^{-1000}$. A probability of $10^{-1000}$ is roughly equal to the probability of picking by chance a particular atom in the entire universe." Indeed, Borel (1962, p. 3), one of the founding fathers of measure theory, proposed in a book for the non-scientists published in 1943 "the single law of chance," or Borel's law. It states "Events with a sufficiently small probability never occur; or at least, we must act, in all circumstances, as if they were impossible." Similar interpretations were given by many other eminent scientists who tried to relate probabilities to the physical world. For example, Bernoulli (1713, pp. 211-212) stated in the first chapter of Part IV of his Ars Conjectandi that "if one thing is considered morally certain which has 999/1000 certainty, another thing will be morally impossible, which has only $1 / 1000$ certainty." Cournot (1843, p. 78) also tried to build a bridge from probability theory to the physical world by stating that "a physically impossible event is one whose probability is infinitely small." Likewise, Popper (2002, p. 195) pointed out that

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

"the rule that extreme improbabilities have to be neglected...agrees with the demand for scientific objectivity."

However, today, there is an almost general agreement among statisticians that probability zero means "almost surely impossible" or extremely unlikely. In other words, an event of zero probability will almost never happen but there may be exceptions. For example, Kolmogorov (1956, p. 5) emphasized that " $\mathrm{P}(\mathrm{A})=0$ does not imply the impossibility of A...all we can assert is that...event A is practically impossible." According to Hand (2014, p. 6), "extremely improbable events are commonplace. It's a consequence of more fundamental laws, which all tie together to lead inevitably and inexorably to the occurrence of such extraordinarily unlikely events." Although we approve of Hand's position that events of vanishingly small probability will ultimately happen, we strongly disagree with establishing statistical tests on point-null hypotheses and expecting for coincidences and miracles to happen.

In light of the previous discussion, we restate the Zero probability paradox in the following, more comprehensible way: in practice, when testing a mean of the normal distribution using a point-null hypothesis, the probability of that hypothesis is zero. This does not imply that it is "absolutely" impossible to state a true point-null hypothesis, but that formulated point-nulls in the context of the simple normal model are almost surely false.

## Some Implications of the Zero Probability Paradox.

Fisher's illuminating words (1922) are more relevant today than in 1922:

It cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved...This anomalous state of statistical science...the obscurity which envelops the theoretical bases of statistical methods may perhaps be ascribed to two considerations. In the first place, it appears to be widely thought, or rather felt, that in a subject in which all results are liable to greater or smaller errors, precise definition of ideas or concepts is, if not impossible, at least not a practical necessity. In the second place, it has happened that in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems. (p. 311-312)

## RAO \& LOVRIC

We argue that the Zero probability paradox has a specific power to shed new light on some fundamental problems in the foundations of statistical science that have been ignored, and help us to resolve some accumulated anomalies related to the point-null hypothesis testing, including so-called large $n$ problem in significance testing, and the Jeffreys-Lindley paradox. It can also elucidate the notion of the Bayes factor, mixed prior distribution advocated by Jeffreys, "irreconcilability of $p$-values and evidence" (Berger \& Sellke, 1987), and Cromwell's rule (Lindley, 1991, p. 104), among others.

However, detailed consideration of the implications of the Zero probability paradox for the Fisherian significance testing, Neyman-Pearson hypothesis testing, and Bayesian testing are beyond the scope of this paper. We confine ourselves, therefore, only to some general implications. Berkson (1938) was the first to notice dependence of significance testing on the sample size. He objected that it is possible to obtain a statistically significant chi-square test merely by increasing sample size:

I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small... we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all!" [italicized for emphasis]

Berkson failed to recognize that the same deficiency (sensitivity to sample size) is also shared by other significance tests based on point-null hypotheses and continuous data. Today this is well known as the large $n$ problem. As argued by Mayo (2006, p. 809): "for any discrepancy from the null, however small, one can find a sample size such as there is a high probability (as high as one likes) that the test will yield a statistically significant result (for any $p$-value one wishes)." She claims that the large $n$ problem is the basis for the famous Jeffreys-Lindley paradox (Lindley, 1957), probably the most quoted divergence between the frequentist and Bayesian approaches to inference. A number of suggestions have been proposed to alleviate this problem, including adjustment of $p$-values to a
fixed sample size (Good, 1988, p. 391), rules of thumb for decreasing $\alpha$ as $n$ increases, and indicated effect size.

Karl Pearson (1935b, p. 550) opined "there is only one case in which an hypothesis can be definitely rejected, namely when its probability is zero." Relating this to the Zero probability paradox leads to the following conclusion. Focusing on the inferential aspects of the problem (not on the decision-making approach) permits rejecting the point-null hypothesis a priori, before seeing data. To paraphrase Berkson, because the result of the significance tests are known, they are no test at all. Term testing is a misnomer in this case and should be replaced by inexactification. These tests are merely procedures that ask researchers to waste their time and financial resources, to collect enough data, and when ultimately reject their point nulls to confirm what they knew beforehand, that their point nulls were almost surely false.

Zero probability paradox points to the root cause of the large $n$ problem and discloses that there is no cure for it under the current point-null paradigm. Because classical significance tests ( $Z$ and $t$-test) are consistent, as the sample size increase, they will become extremely sensitive and therefore, detect even the tiniest discrepancy from the crisp hypothetical (almost surely false) null hypothesis. In other words, classical test statistic converges almost surely to $\infty$ and therefore, gives the asymptotically correct result (see, for example, DasGupta, 2008, p. 337, or Lehman and Romano, 2005, p. 462). Again, this means that in the real world testing any sharp null hypothesis of the normal mean will be ultimately, almost surely, rejected with large enough sample size.

This significant logical inconsistency of the significance testing was not an overwhelming issue in the first half of the past century when Gosset was "'naughtily' playing about with absurdly small numbers" (Eagon Pearson, 1939, p. 217). However, if Efron's view (2010, p. VII) is embraced that in the 21st century, statisticians will deal with large data sets and complex questions, it is clear that the current point-null paradigm is inadequate. Van der Laan and Rose (2010), for example, indicated that next generation of statisticians must construct new tools for massive data sets since the current ones are severely limited. Similarly, Hand (1998, p. 113) insisted in data mining instead of "statistical significance, consider more carefully substantive significance: is the effect important or valuable or not?"

To rephrase Box (1979): the only question of interest is "Is the normal model based on point-null hypothesis illuminating and useful?" The answer must be "No".

So, what should we do? This article is an initial contribution to making a paradigm shift in testing statistical hypotheses. Instead of testing highly
problematic and almost surely false point null hypotheses, as a natural replacement, test a negligible null hypothesis:

$$
\begin{aligned}
& H_{0}:\left|\theta-\theta_{0}\right| \leq \delta \text { (Effect size is negligible) against } \\
& H_{1}:\left|\theta-\theta_{0}\right|>\delta \text { (Effect size is practically meaningful). }
\end{aligned}
$$

We propose naming this avant-garde proposal the "Hodges-Lehmann paradigm". Hodges and Lehmann (1954) were among first statisticians who had noted deficiencies of the point null hypothesis and formulated testing of "material significance" in their path-breaking paper "Testing the approximate validity of statistical hypotheses". We do not regard the Hodges-Lehmann paradigm as deus ex machine, nor as a magic alternative to the traditional point-null testing. However, we argue that it will substantially improve scientific research based on statistical testing. The argument that point nulls are mathematically more tractable is obsolete and belongs to the pre-MCMC era.

We regard statistics as the grammar of science. Thus, we are responsible for providing unambiguous rules of that grammar. We should not feel proud if nonstatisticians are trying to make reform in statistical inference and statistics education. We, statisticians, are accountable to provide researchers in other sciences non-conflicting, coherent, and consistent concepts of testing the statistical hypotheses. Otherwise, significance tests "can actually impede scientific progress." (Kirk, 2003, p. 100) and even harm "development of scientific knowledge" (Armstrong, 2007, p. 321). Researchers and scientists will feel confused and deceived by statistics and statisticians. As pointed out by Cousins (2014, p. 35): "More than a half century after Lindley drew attention to the different dependence of $p$-values and Bayes factors on sample size $n$ (described two decades previously by Jeffreys), there is still no consensus on how best to communicate results of testing scientific hypotheses."

Presumably, we all agree on the point that overcoming of accumulated inconsistencies is always a crucial method in science. As pointed out by Good (1982, p. 489), "a Bayes/non-Bayes compromise or synthesis is necessary for human reasoning." We argue that this compromise is impossible to reach within the point null-hypothesis testing paradigm, as Jeffreys-Lindley paradox evidently testifies.

In sharp contrast to the current point-nulls model, we argue that it is possible to harmonize inferential results of frequentist and Bayesian testing within the new framework. In other words, frequentist and Bayesian inference will become, in

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

principle, compatible and would (or at least could) lead to the similar conclusions in (a) one-sided testing, (b) two-sided testing, and (c) interval estimation.

However, to make this proposal fully justifiable it is necessary to obtain a proof that point nulls are also almost always false in the case of two samples. The initial clue is given by Tukey (1991, p. 100):
"Statisticians classically asked the wrong questions-and were willing to answer with a lie, one that was often downright lie. They asked "Are the effects of A and B different?" and they were willing to answer "no." All we know about the world teaches us that the effects of A and B are always different-in some decimal place-for any A and B. Thus asking 'Are the effects different' is foolish."

Only then, we can set as one of the fundamental rules of the $21^{\text {st }}$ century Statistical Science Decalogue: Hypotheses exactas non fingo!

## References

Adams, M. \& Guillemin, V. (1996). Measure theory and probability. Basel, Switzerland: Birkhäuser.

Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. Philosophical Transactions of the Royal Society of London, 27(325-336), 186-190.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. International Journal of Forecasting, 23(2), 321-327. doi:
10.1016/j.ijforecast.2007.03.004

Berger, J. O. \& Delampady, M. (1987). Testing precise hypotheses. Statistical Science, 2(3), 317-335. doi: 10.1214/ss/1177013238

Berger, J. O. \& Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p values and evidence, Journal of the American Statistical Association, 82(397), 112-122. doi: 10.1080/01621459.1987.10478397

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33(203), 526-542. doi: 10.1080/01621459.1938.10502329

Bernardo, J. M. (1999). Nested hypothesis testing: the Bayesian reference criterion. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Eds.

## RAO \& LOVRIC

Bayesian Statistics (6 ${ }^{\text {th }}$ Ed.). Oxford: Oxford University Press, 101-130 (with discussion).

Bernoulli, J. (1713). Ars conjectandi. (Opus posthumum). Impensis Thurnisiorum, Fratrum.

Borel, E. (1962). Probabilities and life. Mineola, New York: Dover Publications.

Borovkov, A. (2013). Probability theory. NY: Springer.
Box, G. E. P. (1979). Robustness in the strategy of scientific model building, In R. L. Launer and G. N. Wilkinson, Eds. Robustness in Statistics. Cambridge, MA: Academic Press, pp. 201-236.

Buchanan-Wollaston, H. J. (1935). Statistical tests. Nature, 136(3431), 182183. doi: 10.1038/136182b0

Burns, R. P. \& Burns, R. (2008). Business research methods and statistics using SPSS. London: Sage Publications.

Calkin, N. \& Wilf, H. (2000). Recounting the rationals. American Mathematical Monthly, 107(4), 360-363. doi: 10.2307/2589182

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312. doi: 10.1037/0003-066x.45.12.1304

Cousins, R. D. (2014). The Jeffreys-Lindley paradox and discovery criteria in high energy physics. Synthese, 1-38. doi: 10.1007/s11229-014-0525-z

Cournot, A. A. (1843). Exposition de la théorie des chances et des probabilities. Paris: Hachette.

Cox, D. R. (2006). Principles of statistical inference. Cambridge, UK: Cambridge University Press.

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability. NY: Springer.
de Finetti, B. (2008). Philosophical lectures on probability: collected, edited, and annotated by Alberto Mura. NY: Springer.

De Muth, J. E. (2014). Basic statistics and pharmaceutical statistical applications ( $3^{\text {rd }}$ Ed.). Boca Raton, FL: CRC Press.

Edwards, W., Lindman, H. \& Savage, L. J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70(3), 193-242. doi: 10.1037/h0044139

Efron, B. (2010). The Future of Statistics. In M. Lovric (Ed.). International Encyclopedia of Statistical Science. NY: Springer, pp. VII-X.

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

Everitt, B. (1999) Chance rules: an informal guide to probability, risk and statistics. NY: Springer-Verlag.

Finlayson, B. L. \& McMahon, T. A. (2004). Stream hydrology: an introduction for ecologists. Chichester, UK: John Wiley \& Sons.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, Series A, 222(594-604), 309-368. doi: 10.1098/rsta.1922.0009

Ghosh, J. K, Delampady M. \& Tapas, S. (2006). An introduction to Bayesian analysis: theory and methods. NY: Springer.

Good, I. J. (1994). The existence of sharp null hypothesis. Journal of Statistical Computation and Simulation, 49(3-4), 241-242. doi: 10.1080/00949659408811587

Good, I. J. (1956). Which comes first, probability or statistics? Journal of the Institute of Actuaries, 82(2), 249-255. doi: 10.1017/S0020268100046448

Good. I. J. (1988). The interface between statistics and philosophy of science. Statistical Science, 3(4), 386-397. doi: 10.1214/ss/1177012754

Hand, D. (1998). Data mining: statistics and more? The American Statistician, 52(2), 112-118. doi: 10.1080/00031305.1998.10480549

Hand, D. J. (2014). The improbability principle: why coincidences, miracles, and rare events happen every day. NY: Scientific American / Farrar, Straus and Giroux.

Hodges, J. L. \& Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. Journal of the Royal Statistical Society. Series B, 16(2), 262-268.

Hunter, J. E. (1997). Needed: A ban on the significance test. Psychological Science, 8(1), 3-7. doi: 10.1111/j.1467-9280.1997.tb00534.x

Kadane, J. B. (1987). [Testing precise hypotheses]: Comment. Statistical Science, 2(3), 347-348. doi: 10.1214/ss/1177013244

Kaplansky, I. (2001). Set Theory and Metric Spaces (2 ${ }^{\text {nd }}$ Ed.). Providence, RI: AMS Chelsea Publishing.

Kass, R. E. \& Raftery, A. E. (1995). Bayes Factors. Journal of the American Statistical Association, 90(430), 773-795. doi: 10.1080/01621459.1995.10476572

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), Handbook of research methods in experimental psychology. Malden, MA:
Blackwell, pp. 83-105

## RAO \& LOVRIC

Kolmogorov, A. N. (1956). Foundations of the theory of probability (2 ${ }^{\text {nd }}$ English Ed.). Providence, RI: AMS Chelsea Publishing.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. American Psychologist, 56(1), 16-26. doi: 10.1037//0003066x.56.1.16

Kruskal, W. H. (1968). Significance, tests of. In: The international encyclopedia of the social sciences, 14. NY: The Macmillan Co. and the Free Press, pp. 238-250.

Lehmann, E. L. \& Romano, J. P. (2005). Testing statistical hypotheses, $3^{\text {rd }}$ Ed. NY: Springer.

Lindley, D. V. (1957) A statistical paradox. Biometrika, 44(1-2), 187-192. doi: 10.1093/biomet/44.1-2.187

Lindley, D. V. (1975). The future of statistics: a Bayesian 21 st century. Advances in Applied Probability, 7(Supplement: Proceedings of the Conference on Directions for Mathematical Statistics), 106-115. doi: 10.2307/1426315

Lindley, D. V. (1991). Making decisions (2 ${ }^{\text {nd }}$ Ed.). NY: John Wiley \& Sons.
Mayo, D. 2006. Philosophy of statistics. In: S. Sarkar \& J. Pfeifer, Eds. The philosophy of science: an encyclopedia. London: Routledge, pp. 802-815.

Meehl, P. E. (1967). Theory testing in psychology and physics: a methodological paradox. Philosophy of Science, 34(2), 103-115 (doi:
10.1086/288135). Reprinted in The significance test controversy - a reader, D. E. Morrison and R. E. Henkel, Eds. 1970. London: Aldine Publishing Company (Butterworth Group).

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. Psychological Methods, 5(2), 241-301. doi: 10.1037//1082-989x.5.2.241

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20(4), 641-650. doi: 10.1177/001316446002000401

Orlitzky, M. (2012). How can significance tests be deinstitutionalized?
Organizational Research Methods, 15(2), 199-228. doi:
10.1177/1094428111428356

Pearson, E. S. (1939). "Student" as statistician. Biometrika, 30(3-4), 210250. doi: 10.2307/2332648

Pearson, K. (1935a). Statistical tests. Nature, 136(3434), 296-297. doi: 10.1038/136296a0

## TESTING POINT NULL HYPOTHESIS OF A NORMAL MEAN

Pearson, K. (1935b). Statistical tests. Nature, 136(3440), 550. doi: 10.1038/136550a0

Playfair, J. (1819). Elements of geometry: containing the first six books of Euclid, with a supplement on the quadrature of the circle and the geometry of solids; to which are added, Elements of plane and spherical trigonometry. NY: G. Long.

Poole, D. L. \& Mackworth, A. K. (2010). Artificial intelligence:
foundations of computational agents. Cambridge, UK: Cambridge University Press.

Popper, K. R. (2002).The logic of scientific discovery, $2^{\text {nd }} E d$. NY: Routledge.

Popper, K. R. (1963). Conjectures and refutations: the growth of scientific knowledge. New York: Harper.

Quinn, G. P. \& Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge, UK: Cambridge University Press.

Rousseau, J. (2007). Approximating interval hypothesis: p-values and Bayes factors. In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Eds. Bayesian Statistics (8 ${ }^{\text {th }}$ Ed.). Oxford: University Press, pp. 417-452.

Savage, I. R. (1957). Nonparametric Statistics. Journal of the American Statistical Association, 52(279), 331-344. doi: 10.1080/01621459.1957.10501392

Savage, L. J. (1954). The foundations of statistics. New York: John Wiley and Sons.

Sharma, J. K. (2010). Fundamentals of business statistics. Noida, India: Dorling Kindersley.

Sprenger, J. (2013). Testing a precise null hypothesis: the case of Lindley's paradox. Philosophy of Science, 80(5), 733-744. doi: 10.1086/673730

Trafimow, D. \& Marks, M. (2015). Editorial. Basic and Applied Social Psychology, 37(1), 1-2. doi: 10.1080/01973533.2015.1012991.

Tukey, J. W. (1991). The philosophy of multiple comparisons. Statistical Science, 6(1), 100-116. doi: 10.1214/ss/1177011945
van der Laan, M. J. \& Rose, S. (2010). Statistics ready for a revolution: next generation of statisticians must build tools for massive data sets. Amstat News, 399, 38-39.

Wasserstein, R. L. \& Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. American Statistician. 70(2), 129-133. doi: 10.1080/00031305.2016.1154108

## RAO \& LOVRIC

Wilkinson, L. (1999). Statistical methods in psychology journals: guidelines and explanations. American Psychologist. 54(8), 594-604. doi: 10.1037/0003066x.54.8.594

Yoe, C. (2012). Principles of risk analysis: decision making under uncertainty. Boca Raton, FL: CRC Press.

Zellner, A. (1987). [Testing precise hypotheses]: Comment. Statistical Science, 2(3), 339-341. doi: 10.1214/ss/1177013241

# Rao-Lovric and the Triwizard Point Null Hypothesis Tournament 

Shlomo S. Sawilowsky<br>Wayne State University Detroit, MI

The debate if the point null hypothesis is ever literally true cannot be resolved, because there are three competing statistical systems claiming ownership of the construct. The local resolution depends on personal acclimatization to a Fisherian, Frequentist, or Bayesian orientation (or an unexpected fourth champion if decision theory is allowed to compete). Implications of Rao and Lovric's proposed Hodges-Lehman paradigm are discussed in the Appendix.

Keywords: true null hypothesis, Rao-Lovric, Hodges-Lehman.

In their historical reviews of experimental design, Cochran (1977) and Frank Yates posited the first planned controlled experiment was conducted by Daniel ( $7^{\text {th }}-6^{\text {th }}$ century BCE), who employed a ten day treatment vs comparison group post-test only trial. The purpose was to demonstrate the efficacy of a Kosher diet of high protein, low fat, dried legume seeds and water on soldiering skills vs Nebuchadnezzar's army's royal comestible of non-Kosher wine and meat (Daniel 1:3-16). In Contra Celsus (1:15), Origen of Alexandria (153-253 CE) cited Hermippus ( $5^{\text {th }}$ century BCE) and Hecatæus ( $4^{\text {th }}$ century, BCE, presumably of Abdera) who opined subsequent development of analytical analyses of experimental principles by the Jews influenced, if not culminated in, Pythagoras' philosophy of mathematical sciences. Subsequently, Tana Kama (Mishna Gittin 7:1; Talmud Gittin 67b) underscored the importance of co-variables and the minimum number of repetitions for a reliable single subject study design. Shimon ben Chalafta also invoked experimental replications to test claims (e.g., Talmud Chulin 57b).

In the middle of the $2^{\text {nd }}$ century CE, Galen (Aelius/Claudius Galenus) mused how much credence should be given, if any, to a $50^{\text {th }}$ medical study if the previous 49 replications were of no significance. In the early $11^{\text {th }}$ century CE,

[^1]
## SHLOMO S. SAWILOWSKY

Avicenna (Abu ibn Sina) reacted to haphazard methods in the conduct and analysis of experiments and presented seven governing rules. In 1266 CE, Roger Bacon systematized observation of empirical data in controlled experiments. Arthur Young (1771, Figure 1) published a course on experimental agriculture, wherein comparative designs employing standardized methods and analyses were proposed. The analysis of the hypothesis "every year there shall be born more males than females" (1710-1712, p. 188) by John Arbuthnott (un-admittedly inspired by Sir William Petty \& John Graunt) is considered the origin of the nonparametric Sign Test, although it predates more formal origins of empirical probability captured in the treatises on the doctrines of conjecture and chance by Jacob Bernoulli (1713), Abraham de Moivre (1718) and Thomas Bayes (Price, 1763, p. 370).

In the early part of the $20^{\text {th }}$ century CE, Sir Ronald Fisher (influenced by Pierre-Simon Laplace, Carl Gauss, Joseph Jastrow, Sir Francis Galton, Karl Pearson, G. Udny Yule, William Gosset, and certainly others; perhaps later also with Andrey Kolmogorov \& E. J. G. Pittman) defined the null hypothesis, the fundamental building block of modern hypothesis testing, as being true unless there is evidence from the sample (randomly obtained or data at hand) to the contrary. His innovations regarding blocking variables and factorial layouts were pioneering developments in the design of experiments.

Following the logic of experimentation by C. S. Peirce in late $19^{\text {th }}$ century, the Frequentist lemma by Jerzy Neyman and Egon Pearson developed in the 1930s-1940s violated the Fisherian cannon with the introduction of the alternative hypothesis. It was indeed irrefragable blasphemy, because Frequentists must admit the choice and magnitude of the alternative are subjective and independent of both the null hypothesis and the sample. Other $20^{\text {th }}$ century developments in experimental design included orthogonal arrays by my esteemed colleague Professor C. R. Rao, sequential experiments by Abraham Wald and later Herman Chernoff, and the quality control designs of Genichi Taguchi.

Nevertheless, the Frequentists had the advantage, because in the Fisherian system the lack of an alternative obviated the desired notion of fixed comparative statistical power, and by extension, stable effect size. These two modern approaches to statistics are antipodal. Many misunderstandings in hypothesis testing are due to their intrinsic incompatibility, starting with Sir Fisher's "lapsus linguae" (Neyman, 1941, p. 129) fiducial argument (see Sawilowsky, 2003).


Figure 1. Arthur Young (1801), Annals of Agriculture and other Useful Arts, Vol 37. London: Rackham \& Hill. (From the JMASM Archives.)

## SHLOMO S. SAWILOWSKY

This struggle provided the segue for a Bayesian resurrection from Fisher's epithet, "From a purely historical standpoint it is worth noting that the ideas and nomenclature for which I am responsible were developed only after I had inured myself to the absolute rejection of the postulate of Inverse Probability" (1937a, p. 151; see also 1937b, 1939). Although also receiving a boost from C. S. Peirce's logic, Bayesian analysis during Sir Fisher's reign was conducted without benefit of his development of degrees of freedom. The initial inability to replicate Fisherian/Frequentist numerical results was a serious setback to the modern Bayesian paradigm (Sawilowsky, 2002, 2003). Although they have since recovered and inverse probability is currently quite popular, unless there are documented informative prior probabilities available, such as baseball batting averages, Fisher's inurement prevails.

Now comes the debate on certifying the literal truth of the null hypothesis. Original Fisherians needs no proof, because postulation of the putative null was the pivotal theoretic spanning well over two millennia in the science of discovering truth. Frequentists, however, can never accept any proof. The most that can be said is based on the current sample there is no evidence to support the alternative. (This should not be considered an open invitation to collecting potentially endless (a) random samples, known as the quest for a Type I error and its attendant rewards of publishing and tenure or (b) data sets at hand, known as non-representative findings never interpreted with caution to support situational truths with its attendant rewards of political fodder, ill-begotten relief from the court, financial returns based on false advertising, etc.) Moreover, it wouldn't matter even if the null hypothesis is always literally false, because it must be false to an a priori specified magnitude to be rejected.

The Frequentist nomenclature, failure to reject the null hypothesis, was just the ticket in the social and behavioral sciences, where politically correct thinking of the 1960s had begun to take control of those in charge of the keys to situational truths. At best, near-null, near-nil, and the like, were approved substitutes. Philosophically, the yellow submarine is a closed system, so at some decimal of the mantissa there must be a non-Zero value.

The various Frequentist counterproofs were flawed attempts to make something out of nothing by incorrectly preserving the post hoc effect size even when the statistical test was not significant. For example, in the two sample layout, the $t$ statistic is a test of difference between two means. If the $p$ value is above the a priori selected nominal $\alpha$ level, it means the observed difference is not real and should be read as zero. Based on the sample, assumed to be random for generalization purposes, there is no evidence that the populations from which they
were drawn differed in terms of location. Just as the observed difference in means can be safely ignored, the effect size was not statistically significantly different from zero, and can be safely ignored.

This means regardless of the magnitude of the obtained value (e.g., Cohen's $d, 1962,1969,1977,1988)$ in the two sample layout [from very small ( 0.01 ; Sawilowksy, 2009) to small (0.2; Cohen, 1988) to moderate ( 0.5 ; Cohen, 1988) to large ( 0.8 ; Cohen, 1988) to very large (1.2; Sawilowsky, 2009) to huge (2.0; Sawilowsky, 2009)], it should be read and interpreted as zero. Hence, the point null hypothesis, to the Fisherian, is indeed considered to be literally true regardless of the magnitude of Cohen's $d$ when the $p$ value is greater than nominal $\alpha$.

In the antecedent article, colleagues C. R. Rao and M. Lovric (http://digitalcommons.wayne.edu/jmasm/vol15/iss2/2), cited Cohen (1990) who wrote the null hypothesis can only be true "in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false)" (p. 1, 308). Based on my letters with him, documented elsewhere, Cohen's statement was not surprising.

Subsequently, this was discussed conceptually in Knapp and Sawilowsky (2001, p. 71-74; for expanded commentary relative to the debate see Harlow, et al., 1997; Imbens \& Rubin, 2015). I included Meehl's (1990) recapitulation that he initially referred only to quasi-experiments and surveys (Meehl, 1978), but later admitted the null hypothesis can be literally true in an "experimental study" (Meehl, 1990, p. 204). (Carol H. Ammons, the co-Editor of Psychological Reports where it was published, sent me a reprint of Meehl (1990) soon after its publication. In our subsequent conversation, I was supportive of Meehl's recapitulation, and I remain so today.) Similarly, in Knapp and Sawilowksy (2001) I also included Hagen's (1997, p. 20) imputed recapitulation of Cohen (1994).

A simple demonstration of the algorithm I presented in Knapp and Sawilowsky (2001) is coded in R in Figure 2. When executed, it creates two groups, $x$ and $y$, and populates them with scores randomly selected from the standard normal curve. Although a Monte Carlo is unnecessary when underlying assumptions are met, it is employed to facilitate the demonstration. The two independent samples pooled variance $t$ test is conducted on the data, and if the $p$ value is less than nominal $\alpha=0.05$, a counter is incremented. The process is repeated 100,000 times. The final value of the counter is divided by the number of repetitions to produce the Type I error rate.

## SHLOMO S. SAWILOWSKY

The code will produce the same result on any computer platform and operating system, because the seed number is set for the pseudo-random number generator. That result is 0.04919 . Rejections occurred across the 100,000 repetitions, but they were known false positives. The point null hypothesis was indeed literally true, because it was programmed to be so. The collection of false positives that give rise to the notion the point null is never literally true were simply the constituent figments of imagination that sum to the Type I error rate.

```
set.seed (123457)
to5 <- NULL
rep <- 100000
rejt05 <- numeric(length=rep)
ss <- 30
for (i in 1:rep) {
    x1 <- rnorm(ss)
    x <- x1+0.0
    y<- rnorm(ss)
    tp <- t.test(x,y,var=TRUE)[["p.value"]]
    rejt05[i] <- ifelse (tp < 0.05,1,0)
    }
t05 <- sum(rejt05)/rep
```

Figure 2. Monte Carlo $t$ Test in R Code

The rejection rate obtained from the code will approach 0.05 as (a) the sample size, set to 30 per group in this example, increases, (b) the number of repetitions of the experiment increases, or (c) possibly even with the current study parameters if a different initial seed number is selected (Hill \& Sawilowsky, 2011). For example, if the number of repetitions is increased to $1,000,000$, the Type I error improves to 0.049858 .

A non-null condition can be created by replacing the 0.0 with a non-zero number (positive or negative) in the line $x<-x 1+0.0$. For example, to model a very small effect size of 0.01 (Sawilowsky, 2009), replace the 0.0 in this code segment with a constant $c=0.01$ (representing $0.01 * \sigma$; where $\sigma$ refers to the standard deviation of the normal curve $=1$ ). The constant $c$ is added to each member of the $x$ group and shifts its location by that magnitude, while leaving the
scale unaffected. The resulting rejection rate is known as statistical power (not Type I error rate). With 100,000 repetitions it amounts to 0.04923 , a nuanced but detectable difference of 0.00004 above nominal $\alpha$ for this sample size and data pseudo-randomly sampled from the standard normal curve.

If the effect size is increased to 0.05 the power yield increases to .05342 , and for an effect size of 0.1 the power increases to 0.06542 . For Cohen's (1988) small effect size of 0.2 , the power increases further to 0.11611 . As the effect size approaches infinity (and depending on the distribution and sample size, the effect size may not need to increase past a small fraction or multiple of its $\sigma$ ) the power approaches 1 .

Random numbers represent a literally true null condition. This R code proves that when the point null is literally true, the $t$ test (if all conditions are met, i.e., normality, homoscedasticity, independence) will retain the null hypothesis to the nominal $\alpha$ level. Hence, in real world applications of a true randomized experimental design, if there is no difference between $\bar{x}$ and $\bar{y}$ (the two sample means) the $t$ test will testify to that fact.

Execution of the R code demonstrates increasing the sample size and/or number of repetitions of the experiment to $\infty$ will not lead to a rejection rate of the null hypothesis different from nominal $\alpha$, which is the answer to Cohen's speculation of what might happen in the bowels of a Monte Carlo study. Moreover, despite the current fascination with big data (and hopefully its ardent fans are able to recognize and deprecate its often hidden or embedded stepwise methods), Gosset noted many in applied disciplines we are forced to work with small samples. This was aptly captured in Sir Fisher's revelation to Samuel Stouffer regarding the inspiration for deriving a certain postulate: something had to be done when rabbits got into the garden and ate a lot of the degrees of freedom.

To the Fisherian, QED. To the Frequentist, the discussion is much ado about (something that can never be literally) nothing. To the Bayesian, add noninformative priors to the perils of non-normality, heteroscedasticity, and nonindependence; and then choose sides.

## References

Arbuthnott, J. (1710-1712). An argument for divine providence, taken from constant regularity observ'd in the births of both sexes. Philosophical transactions (1683-1775), 27 (1710-1712), 186-190.

## SHLOMO S. SAWILOWSKY

Bernoulli, J. (1713). Ars Conjectandi Opus Posthumum. Basileæ:
Thurnisiorum.
Cochran, W. G. (1977). Sampling techniques. $3^{\text {rd }}$ ed. NY: Wiley.
Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal Social Psychology, 65(3), 145-153. doi: 10.1037/h0045186

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. San Diego, CA: Academic Press.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences, Rev. Ed. San Diego, CA: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312. doi: 10.1037/0003-066x.45.12.1304

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997-1003. doi: 10.1037/0003-066x.49.12.997

De Moivre, A. (1718). The doctrine of chances: Or, a method for calculating the probabilities of events in play. London: W. Pearson.

Fisher, R. A. (1937a). Editorial note. Annals of Eugenics, 7, 146-151.
Fisher, R. A. (1937b). On a point raised by M. S. Bartlett on fiducial probability. Annals of Eugenics, 7(4), 370-375. doi: 10.1111/j.14691809.1937.tb02154.x

Fisher, R. A. (1939). A note on fiducial inference. The Annals of Mathematical Statistics, 10(4), 383-388. doi: 10.1214/aoms/1177732151

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52(1), 15-24. doi: 10.1037/0003-066x.52.1.15

Harlow, L. L., Mulaik, S. A., \& Steiger, J. H. (Eds.). (1997). What if there were no significance tests? Mahwah, NJ: Erlbaum.

Hodges, J. L., \& Lehmann, E. L. (1963). Estimates of location based on rank tests. Annals of Mathematical Statistics, 34(2), 598-611. doi:
10.1214/aoms/1177704172

Hill, J. C., \& Sawilowsky, S. S. (2011). Bias in Monte Carlo simulations due to pseudo-random number generator initial seed selection. Journal of Modern Applied Statistical Methods, 10(1), 29-50.

Imbens, G. W., \& Rubin, D. B. (2015). Fisher's exact P-values for completely randomized experiments. In G. W. Imbens \& D. B. Rubin, Eds.

## THE TRIWIZARD POINT NULL HYPOTHESIS TOURNAMENT

Causal inference in statistics, social, and biomedical sciences. NY: Cambridge University Press. doi: 10.1017/CBO9781139025751.006

Knapp, T. R., \& Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. The Journal of Experimental Education, 70(1), 65-79. doi: 10.1080/00220970109599498

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46(4), 806-834. doi: 10.1037/0022-006x.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. Psychological Reports, 1(2), 108-141. doi: 10.1207/s15327965pli0102_1

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. Biometrika, 32(2), 128-150. doi: 10.1093/biomet/32.2.128

Price. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R. S., communicated by Mr. Price, in a letter to John Canton, A.M. F.R.S. Philosophical Transactions, 53, 370-418.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_{1}{ }^{2} \neq \sigma_{2}{ }^{2}$. Journal of Modern Applied Statistical Methods, 1(2), 461-472.

Sawilowsky, S. S. (2003). Deconstructing the arguments from the case against hypothesis testing. Journal of Modern Applied Statistical Methods, 2(2), 467-474.

Sawilowsky, S. S. (2009). New effect size rules of thumb. Journal of Modern Applied Statistical Methods, 8(2), 597 - 599.

Young, A. (1771). A course on experimental agriculture; containing an exact register of all the business transacted during five years on near three hundred acres of various soils; including a variety of experiments on the cultivation of all sorts of grain and pulse, both in the old and new methods. Dublin: Exshaw, Saunders, Chamberlaine, Sleater, Potts, Hoey, Williams, Porter, Moncrieffe, and Walter.

## SHLOMO S. SAWILOWSKY

## Appendix

In Knapp and Sawilowsky (2001), I presented rebuttals to "the following propositions:

- The null hypothesis is always false.
- A sufficiently large enough sample guarantees rejection of the null hypothesis.
- Statistical tests are of no use because the results do not address practical importance.
- Testing a near-nil null hypothesis is better than testing a null hypothesis.
- Hypothesis testing does not lead to scientific discoveries.
- Confidence intervals are superior to hypothesis testing.
- Effect sizes should be reported regardless of the outcome of hypothesis testing." (p. 71).

The subjectivity of defining a near-nil null hypothesis will also have a deleterious effect on equivalence testing, and could be added to the above list.

With regard to testing a near-nil null instead of a null hypothesis, Rao and Lovric, in the antecedent article, proposed a paradigm shift to testing the negligible null hypothesis:
$H_{0}:\left|\theta-\theta_{0}\right| \leq \delta$ (Effect size is negligible) against
$H_{1}:\left|\theta-\theta_{0}\right|>\delta$ (Effect size is practically meaningful).

They aptly named it the "Hodges-Lehmann paradigm," a nomenclature well known in other contexts. In R-measures of location, for example, the inversion of signed ranks can lead to the Hodges-Lehmann estimator, a robust (median unbiased) pseudo- $\theta$ point estimator of symmetry (Hodges \& Lehmann, 1963). In bracketed (see Sawilowsky, 2003, p. 128) intervals, the Hodge-Lehmann treatment alternative is modeled by a systematic progression from pseudo- $\theta$, although no expertise is called on to determine negligible or practical meaningfulness.

Regarding near-nill null hypotheses within the context of hypothesis testing, I’ve opined (Knapp \& Sawilowsky, 2001),

This remedy's attendant difficulties are obvious considering the chaos that would arise from the infinite number of near-nils that might be chosen. (Eventually, we speculate, some common near-nils would emerge and evolve into a universally accepted traditional near-nil, completing the circle.) Moreover, the near-nil weakens the Fisherian logic regarding the null hypothesis, which is indirect proof by contradiction. If the probability associated with sample data obtained from a designed study is so remote, the null hypothesis or the model that generated it is contradicted. Rejecting a null hypothesis should be more compelling than rejecting an arbitrarily chosen near-nil hypothesis. Also, in the social and behavioral sciences for cases in which treatment effects or naturally occurring differences are often tiny, using the near-nil hypothesis when investigating interventions with potentially subtle differences may hide a treatment effect. Similarly, as the magnitude of the near-nil increases, the sample size necessary to detect a false near-nil null hypothesis increases in the treatment versus control group and related designs, which would be highly undesirable. (p. 73).

# Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective' 

Bruno D. Zumbo<br>University of British Columbia<br>Vancouver, BC, Canada

Edward Kroc<br>University of British Columbia<br>Vancouver, BC, Canada

Although we have much to agree with in Rao and Lovric's important discussion of the test of point null hypotheses, it stirred us to provide a way out of their apparent Zero probability paradox and cast the Hodges-Lehmann paradigm from a Serlin-Lapsley approach. We close our remarks with an eye toward a broad perspective.

Keywords: Hypothesis testing, point null, statistical practice

Statistical methods and the testing of hypotheses play a pivotal role in day-to-day practical science, but not always an enlightened one. There are several wellknown criticisms of testing a point null hypothesis in the statistical literature that go back at least to Berkson (1938, 1942) and Hodges and Lehmann (1954). Debates about the role of statistical hypothesis testing, its uses, misinterpretations, and abuses as well as adjacent discussions of interpretations and abuses of confidence intervals, effect sizes, and statistical power continue unabated in the methodological, statistical and substantive literatures. Through all of this, however, conventional significance tests, point null hypotheses, and p-values continue to be used in nearly all experimental publications in the social, behavioral, natural, and health sciences to dichotomize claims from statistical hypotheses in to significant versus nonsignificant findings. The use of significance tests of point null hypotheses, as a kind of ritualistic cultural behaviour, continues unabated because these statistical techniques appear (at least to practicing scientists) to be objective and exact, they are easily and readily available in statistical software packages and on web applets, students are taught to use them, and journal reviewers and editors demand them.

[^2]
## SOME REMARKS ON RAO AND LOVRIC

Rao and Lovric's (2016, this issue) recent paper rests in this backdrop of our discipline's longstanding ineffective critical obsession to challenge and repurpose our most sacred of empirical methodological cows: the testing of point null hypotheses via significance testing. Rao and Lovric are to be most warmly thanked for bringing this fundamentally important issue to the attention of readers of the Journal of Modern Applied Statistical Methods and initiating an important conversation. Their recent contribution to the literature gives us much to agree with, but also stirs us to critically reflect on some of their claims and observations. This is distinctly a sign of good scholarship.

We have arranged our remarks in to three categories. In what follows we: (i) reserve the majority of our remarks for Rao and Lovric's point null Zero probability paradox and the matter of events of vanishingly small probability ultimately happening, a key point in the on-going controversy surrounding testing point null hypotheses, (ii) bring what we call the Serlin-Lapsley perspective to the Hodges-Lehmann paradigm briefly attending to its strengths and limitations, and (iii) close with some remarks that aim to move us to a broader perspective.

## Rao and Lovric's point null Zero probability paradox, and on the event of vanishingly small probability ultimately happening

As Rao and Lovric remind us, Hand (2014, p. 6) stated: "extremely improbable events are commonplace. It's a consequence of more fundamental laws, which all tie together to lead inevitably and inexorably to the occurrence of such extraordinarily unlikely events". We are in agreement, per Kolmogorov (1956), that the probability of an event A being zero does not imply that the event A is impossible. Indeed, it is the support of a probability measure that separates the possible from the impossible, not the value of the measure on its support. It is true, for example, that the probability of observing the event $\{X=1\}$ is zero when $X$ is an exponential random variable, but that the event $\{X=1\}$ should not be considered impossible, since the measure is well-defined and nonzero on any open set containing this event. However, the event $\{X=-1\}$ is indeed ontologically impossible when $X$ is an exponential random variable; this event is simply not in the support of $X$.

We must disagree though with Hand's claim that "events of vanishingly small probability will ultimately happen." This is not true in general, at least, not if an event of "vanishingly small probability" is to be interpreted as an event that is almost surely null; i.e., an event whose probability is equal to zero. Broadly

## ZUMBO \& KROC

speaking, most random variables of practical interest fall into one of two categories: they are defined by measures that are either (1) absolutely continuous with respect to Lebesgue measure, usually defined on the real line or half line; or (2) absolutely continuous with respect to counting measure on some countable set, usually the positive or nonnegative integers. We remind the reader that a measure $\mu$ is absolutely continuous with respect to another measure $\lambda$ if every $\lambda$-null set is also a $\mu$-null set.

We note that absolute continuity implies the existence of a probability density function, case (1), or a probability mass function, case (2), by the classical Radon-Nikodym Theorem; indeed, this is precisely how these objects are formally defined. We also note that this definition presupposes the specification of a legitimate $\sigma$-algebra, and it suffices to take the Borel sets on the real line, or the power set on the integers respectively. With this terminology in mind, we will see that Hand's statement is false when our probability measure is absolutely continuous with respect to Lebesgue measure, and unnecessary when our measure is absolutely continuous with respect to counting measure.

Let $X$ be distributed according to a probability measure, $\operatorname{Pr}$, that is absolutely continuous with respect to Lebesgue measure. Choose any real number $a$. What is the probability that we eventually sample $\{a\}$ ? Formally, if we let $X_{i}$ denote the $i^{\text {th }}$ (independent) sampling, we wish to calculate the probability of the union of events $\left\{X_{i}=a\right\}$ over all $i>0$. Apply countable subadditivity of the measure (a defining property of measures) to bound this probability by the sum of $\operatorname{Pr}\left(\left\{X_{i}=a\right\}\right)$. Each of these is identically zero (by absolute continuity), therefore the probability of their union is as well. Thus, we can sample infinitely often and we will in fact never sample the singleton $\{a\}$, almost surely. This argument immediately generalizes to any countable set, which is automatically a Pr-null set by absolute continuity. So, for example, the probability that we ever observe any rational number in infinitely many samples of $X$ is zero. The argument can be fully extended to apply to any Lebesgue-null event, including those containing certain uncountable sets of reals, such as Cantor or various other fractal sets.

Two key points are noteworthy about the above argument. First, we take as definition that any sampling scheme must consist of a countable number of steps. That is, we do not allow the possibility of drawing uncountably many samples. Theoretically, this kind of uncountable sampling scheme is not impossible, but it would be completely meaningless in practice: any mechanistic process requires countability of its steps.

The second point to note is that the theoretical argument above relies on the infinite precision of our sample, and this is where the crux of the matter lies. A
careful reading of the above argument will reveal an apparent paradox: the probability of ever observing any rational number under a probability measure, absolutely continuous with respect to Lebesgue measure, is identically zero; yet, in practice, every singleton sample that we draw from such a distribution will be a rational number. This is simply another, equivalent instantiation of what Rao and Lovric term the Zero probability paradox. Any practical measuring device will demand that a sampled point is drawn to only a finite level of precision; i.e., we can only observe real numbers with finite decimal expansion in practice. The way out of this apparent paradox is to realize that all probability measures, in practice, are only supported on a finite set. The size of this set is dictated by the precision of our measurements, but we know that this precision must always be of finite detail. Consequently, if we choose any real number $a$ in the support of our practical probability measure, $\operatorname{Pr}^{*}$, we have $\operatorname{Pr}^{*}(X=a)>0$; this follows since any finite set of real numbers, under the classic topology, is nowhere dense. Revisiting our generic sampling scheme from before, we now calculate:

$$
\begin{aligned}
\operatorname{Pr} *\left(\bigcup_{i=1}^{\infty}\left\{X_{i}=a\right\}\right) & =1-\operatorname{Pr}^{*}\left(\bigcap_{i=1}^{\infty}\left\{X_{i} \neq a\right\}\right) \\
& =1-\prod_{i=1}^{\infty} \operatorname{Pr}^{*}\left(X_{i} \neq a\right) \\
& =1-\lim _{N \rightarrow \infty}\left[1-\operatorname{Pr}^{*}\left(X_{i}=a\right)\right]^{N}
\end{aligned}
$$

The limit goes to zero since $\operatorname{Pr}^{*}\left(X_{i}=a\right)>0$ for all $i$, so the probability that we eventually observe the singleton $\{a\}$ is exactly 1 , almost surely. The same reasoning applies to any subset of the practical probability space.

This is the distinction between probability in practice, the ultimate subject of statistics, and the platonic structure of the mathematical objects that we use to conveniently describe that practice. These descriptions are nearly always approximations: we simplify our practical probability spaces by smudging them into theoretical ones. This has undeniably proven to be an extremely fruitful tactic, but it has also given rise to several conundrums and apparent paradoxes like the ones discussed here. Point null hypotheses may be almost surely false in the platonic sense, but this is only a reflection of the disconnect between the literal structure of the objects we study and the approximations, like the various scaledLebesgue measures, that we use to conveniently describe them mathematically. It is meaningful effects that we truly care about, relative to the precision of our

## ZUMBO \& KROC

measurements and the object of our research, and in this regard we are very much in agreement with Rao and Lovric.

Finally, we note that this entire discussion is unnecessary when considering random variables that are absolutely continuous with respect to counting measure (all the standard "discrete" distributions, for example). By definition, such a corresponding probability space contains no nontrivial null sets in the support of the measure; thus, there are no events of "vanishingly small probability" to speak of. Just as in the resolution of our apparent paradox on a practical probability space above, every event will eventually happen almost surely.

## Hodges-Lehmann Paradigm with a Serlin-Lapsley Twist

We would agree with Rao and Lovric that the Hodges-Lehmann method is not a magical alternative to the traditional point null testing but that it may provide a useful paradigm for the practicing empirical scientist. However, we would contend that in its day-to-day use among empirical researchers the HodgesLehmann paradigm still suffers from some of the same issues as the point null. In particular, the magic choice of "delta", in the Hodges-Lehmann or Serlin-Lapsley senses, remains arbitrary - or necessarily defined subjectively by the researcher, contingent on precision, etc., as before. Furthermore, a key to the widespread adoption of the Hodges-Lehmann paradigm is what we will refer to as the SerlinLapsley approach to statistical science that incorporates a 'good enough' principle and embodies Imre Lakatos's view of science. Our message is the same as Rao and Lovric's but from a different framework.

Efforts to facilitate testing what may be called 'range nulls', which require assumptions about the distribution of a statistic when the null is false, have been made by Serlin and Lapsley (1985, 1993). In short, this approach involves incorporating an external criterion or statistic, such as an effect size, into the hypothesis test via a range-null hypothesis approach. The kernel of the 'range-null hypothesis approach' specifies a range of values under the null hypothesis for which a rejection implies a meaningful result. One tests against a negligibly small or trivial effect. If one rejects the null range hypothesis this implies not only that, for example, the mean of the experimental group is different than control group, but the difference is of large enough magnitude to be meaningful. As Serlin and Lapsley (1985) note, minimum effects testing can test more realistic hypotheses, rather than the "straw man" zero effect (p. 74). The important difference in terms of scientific practice is that Serlin and Lapsley's (1985) framework focuses on testing one's own theory as the null, along with using what they call a "good-
enough belt" around a "complex null hypothesis" (p. 79). Central to this paradigm is the principle that this "good-enough belt" be defined subject to the analyst's particular research question, knowledge from the previous literature, and the precision of measurement. One machinery for applying Serlin and Lapsley's framework is the Hodges-Lehmann paradigm described in Rao and Lovric or one could, as Serlin and Lapsley (1985) do, construct a test criterion by directly computing the percentile of the noncentral distribution involved in the test of the hypothesis to set the critical value - for example, the noncentral F distribution in a situation similar to the one described in Rao and Lovric.

## Closing Remarks

In closing, Rao and Lovric's paper highlights for us the continuing need for dialogue on conceptual and foundational matters in statistics. The statistical significance test is the most widely known point in empirical science wherein probabilities and probability models enter the scientific process as either a platonic structure of the mathematical objects or a practical mathematical model. In this light, the nature, use and misuse of significance tests have been widely discussed in both statistical and non-statistical circles. Clearly for significance tests to be of much use to empirical researchers they must focus on sensible and interesting null hypotheses. As is widely discussed in the methodological literature and growing in importance in the statistical literature, there are clear distinctions between statistical significance and the more important notions of practical, clinical or biological significance. Likewise, we need to move beyond the conventional language of Type I and Type II error rates and also consider errors that are directly related to day-to-day statistical practice such as Type S (sign) and Type M (magnitude) errors, which Gelman and Tuerlinckx (2000) describe as relating to the probability that claims with confidence have the wrong sign or are far in magnitude from underlying effect sizes. These errors speak more directly to quantifying our subjective assumptions about what matters and what does not.

Highest among our concerns is that there is a misunderstanding among some experimental researchers that statistical theories of hypothesis testing (be they of the Fisherian, Neyman-Pearson, or some blended approach of the two frameworks) are intended to give an automated and (naively) objective support to an empirical claim. This misunderstanding reflects a lack of alignment of statistical and scientific reasoning. Cox (1982) stated the matter best:

## ZUMBO \& KROC

"Failure to achieve an interesting level of significance in a study does not mean that the topic should be abandoned. Significance tests are not intended to inhibit the free judgment of investigators. Rather they may on the one hand warn that the data alone do not establish an effect, and hence guard against over interpretation and unwarranted claims, and on the other hand show that an effect is reasonably firmly proved." (Cox, 1982, p. 327).

## References

Berkson, J. (1938). Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. Journal of the American Statistical Association, 33(203), 526-542. doi: 10.2307/2279690

Berkson, J. (1942). Tests of Significance Considered as Evidence. Journal of the American Statistical Association, 37(219), 325-335. doi: 10.1080/01621459.1942.10501760

Cox, D. R. (1982). Statistical Significance Tests. British Journal of Clinical Pharmacology, 14(3), 325-331. doi: 10.1111/j.1365-2125.1982.tb01987.x

Gelman, A., \& Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics, 15, 373-390. doi: 10.1007/s0018000000040

Hand, D. (2014). The improbability principle: why coincidences, miracles, and rare events happen every day. NY: Scientific American / Farrar, Straus and Giroux.

Hodges, J. L. \& Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. Journal of the Royal Statistical Society. Series B (Methodological), 16(2), 261-268.

Kolmogorov, A. N. (1956). Foundations of the theory of probability (Second English edition). NY: Chelsea.

Rao, C. R., \& Lovric, M. M. (2016). Testing Point Null Hypothesis of a Normal Mean and the Truth: 21 st Century Perspective. Journal of Modern Applied Statistical Methods, 15(2), 2-21. doi: 10.22237/jmasm/1478001660

Serlin, R. C., \& Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40(1), 73-83. doi: 10.1037//0003-066x.40.1.73

## SOME REMARKS ON RAO AND LOVRIC

Serlin, R. C., \& Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren \& C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 199-228). Hillsdale, NJ: Erlbaum.

# Within Groups ANOVA When Using a Robust Multivariate Measure of Location 

Rand Wilcox<br>University of Southern California<br>Los Angeles, California

Timothy Hayes<br>University of Southern California<br>Los Angeles, California

For robust measures of location associated with $J$ dependent groups, various methods have been proposed that are aimed at testing the global hypothesis of a common measure of location applied to the marginal distributions. A criticism of these methods is that they do not deal with outliers in a manner that takes into account the overall structure of the data. Location estimators have been derived that deal with outliers in this manner, but evidently there are no simulation results regarding how well they perform when the goal is to test the some global hypothesis. The paper compares four bootstrap methods in terms of their ability to control the Type I error probability when the sample size is small, two of which were found to perform poorly. The choice of location estimator was found to be important as well. Indeed, for several of the estimators considered here, control over the Type I error probability was very poor. Only one estimator performed well when using the first of two general approaches that might be used. It is based on a variation of the (affine equivariant) Donoho-Gasko trimmed mean. For the second general approach, only a skipped estimator performed reasonably well. (It removes outliers via a projection method and averages the remaining data.) Only one bootstrap method was found to perform well when using the first approach. A different bootstrap method is recommended when using the second approach.

Keywords: Bootstrap methods, outliers, skipped estimator, Donoho-Gasko trimmed mean

## Introduction

Methods for comparing dependent groups, based on the usual sample mean, are not robust under general conditions. A fundamental concern with any inferential technique based on the mean is that it can result in relatively low power when dealing with heavy-tailed distributions (e.g., Marrona, Martin, \& Yohai, 2006; Staudte \& Sheather, 1990; Wilcox, 2012). Roughly, heavy-tailed distributions are

[^3]
## WITHIN GROUPS ANOVA

characterized by outliers that inflate the standard error of the sample mean. Even an arbitrarily small departure from normality can result in poor power. Another concern is that the breakdown point of the sample mean is only $1 / n$, where n is the sample size. That is, the minimum proportion of points that must be altered to completely destroy the sample mean (make it arbitrarily large or small) is $1 / n$.

Various methods for comparing $J \geq 2$ dependent groups have been derived and studied that are based on replacing the marginal means with some robust estimator (e.g., Wilcox, 2012, Ch. 8). That is, if $X_{i j}(i=1, \ldots, n ; j=1, \ldots, J)$ is a random sample of $n$ vectors from some $J$-variate distribution, for each $j$, a robust measure of location is computed. These methods deal with outliers among the marginal distributions, but they do not deal with outliers in a manner that takes into account the overall structure of the data. As a simple example of what this means, it is not unusual to be young, it is not unusual to have heart disease, but it is very unusual to be both young and have heart disease.

Situations are encountered where there are no outliers among the marginal distributions based on, for example, a boxplot or the MAD-median rule, yet there are outliers when using a multivariate outlier detection technique that takes into account the overall structure (e.g., Wilcox, 2012).

Another possible criticism of applying a robust estimator to each of the marginal distributions is that the resulting measure of location is not affine equivariant (e.g., Rousseeuw \& Leroy, 1987). To elaborate, note that a basic requirement for $\hat{\theta}_{j}$ to qualify as a location estimator is that it be both scale and location equivariant. That is, if $\hat{\theta}_{j}=T\left(X_{i j}, \ldots, X_{n j}\right)$ is some estimate of $\theta_{j}$, then for $\hat{\theta}_{j}$ to qualify as a location estimator, it should be the case that for constants $a$ and $b$,

$$
T\left(a X_{1}+b, \ldots, X_{n} A+b\right)=a T\left(X_{1}, \ldots, X_{n}\right)+b .
$$

In the multivariate case, a generalization of this requirement, affine equivariance, is that for a $J$-by- $J$ nonsingular matrix $\mathbf{A}$ and vector $\mathbf{b}$ having length $J$,

$$
\begin{equation*}
T\left(\mathbf{X}_{1} \mathbf{A}+\mathbf{b}, \ldots, \mathbf{X}_{n} \mathbf{A}+\mathbf{b}\right)=T\left(\mathbf{X}_{1}, \ldots, \mathbf{X}_{n}\right) \mathbf{A}+\mathbf{b} . \tag{1}
\end{equation*}
$$

In particular, the estimate is transformed properly under rotations of the data as well as changes in location and scale.

## WILCOX \& HAYES

The goal in this paper is to report simulation results on several methods for comparing dependent groups with an emphasis on situations where the sample size is small. Several multivariate estimators were considered that take into account the overall structure of the data when dealing with outliers. All of them are location and scale equivariant, but one is not affine equivariant.

Here, two types of global hypotheses are considered. To describe them, let $\hat{\Theta}(\mathbf{X})$ represent one of the multivariate location estimators to be considered. Letting $\Theta=\left(\theta_{1}, \ldots, \theta_{j}\right)$ represent the estimand associated with $\hat{\Theta}(\mathbf{X})$ (the population analog of $\hat{\Theta}(\mathbf{X})$, the first global hypothesis is

$$
\begin{equation*}
H_{0}: \theta_{1}=\cdots=\theta_{j} \tag{2}
\end{equation*}
$$

To describe the second hypothesis, let $D_{i j k}=X_{i j}-X_{i k}, j<k$, and let $\hat{\Theta}(D)$ be some multivariate location estimator based on the $D_{i j k}$ values. There are $L=\left(J^{2}-J\right) / 2$ parameters being estimated, which are labeled $\Delta=\left(\delta_{1}, \ldots, \delta_{L}\right)$. Now the goal is to test

$$
\begin{equation*}
H_{0}: \delta_{1}=\cdots=\delta_{L}=0 . \tag{3}
\end{equation*}
$$

From basic principles, when dealing with means, there is no distinction between (2) and (3). But under general conditions, this is not the case when using a robust estimator. (It is readily verified, for example, that the difference between the marginal medians is not necessarily equal to the median of the difference scores.)

Two bootstrap methods for testing (2) were considered here, and another two methods were considered when testing (3). As will be seen, the choice of estimator, as well as the bootstrap method that is used, is crucial in terms of controlling the Type I error probability, at least when the sample size is small.

## Description of the Methods

## The Location Estimators

The first estimator is based on a particular variation of an affine equivariant estimator derived by Donoho and Gasko (1992), which will be labeled the DG estimator henceforth. Roughly, it begins by quantifying how deeply each point is

## WITHIN GROUPS ANOVA

nested within the cloud of points. Here, this is done using a projection-type method, which provides an approximation of half-space depth (Wilcox, 2012, section 6.2.5). To elaborate, let $\hat{\tau}$ be some initial affine equivariant location estimator. Here, the (fast) minimum covariance determinant estimator (MCD) is used (e.g., Wilcox, 2012, section 6.3.2). Briefly, the MCD estimator searches for a subset of half the data that minimizes the generalized variance. The mean of this subset is the MCD measure of location. Let

$$
\begin{aligned}
\mathbf{U}_{i} & =\mathbf{X}_{i}-\hat{\tau} \\
B_{i} & =\mathbf{U}_{i} \mathbf{U}_{i}^{\prime}
\end{aligned}
$$

$(i=1, \ldots, n)$ and for any $j(j=1, \ldots, n)$, let

$$
W_{i j}=\sum_{k=1}^{j} U_{i j} U_{j k}
$$

and

$$
\begin{equation*}
T_{i j}=\frac{W_{i j}}{B_{i}}\left(U_{i 1}, \ldots, U_{i j}\right) \tag{4}
\end{equation*}
$$

The distance between $\hat{\theta}$ and the projection of $X_{\mathrm{j}}$ (when projecting onto the line connecting $\mathbf{X}_{i}$ and $\hat{\tau}$ ) is

$$
H_{i j}=\operatorname{sign}\left(W_{i j}\right)\left\|T_{i j}\right\|
$$

where $\left\|T_{i j}\right\|$ is the Euclidean norm associated with the vector $T_{i j}$.
Let $d_{i j}$ be the depth of $X_{j}$ when projecting points onto the line connecting $X_{i}$ and $\hat{\theta}$. That is, for fixed $i$ and $j$, the depth of the projected value of $X_{j}$ is

$$
d_{i j}=\min \left(\#\left\{H_{i j} \leq H_{i k}\right\}, \#\left\{H_{i j} \geq H_{i k}\right\}\right),
$$

Where $\#\left\{H_{i j} \leq H_{i k}\right\}$ indicates how many $H_{i k}(k=1, \ldots, n)$ values satisfy $H_{i j} \leq H_{i k}$. The depth of $X_{j}$ is taken to be $L_{j}=\min d_{i j}$, the minimum being taken over all $i=1, \ldots, n$.

## WILCOX \& HAYES

The Donoho-Gasko (DG) $\gamma$ trimmed mean associated with the $X_{i j}$ values is the average of all points that are at least $\gamma$ deep in the sample. That is, points having depth less than $\gamma$ are trimmed and the mean of the remaining points is computed. If the maximum depth among all $n$ points is at least $\gamma$, the breakdown point of the DG estimator is $\gamma /(1+\gamma)$, where the breakdown point refers to the minimum proportion of points that must be altered to completely destroy an estimator. Here, $\gamma=.2$ is used.

The other estimator considered here, which performed well in simulations when testing (3), is a skipped estimator based on a projection method for detecting outliers, which will be labeled the SP estimator. Fix $i$, and for the point $\mathbf{X}_{i}$ let

$$
\begin{gathered}
\mathbf{A}_{i}^{\prime}=\mathbf{X}_{i}-\hat{\tau}, \\
\mathbf{B}_{j}^{\prime}=\mathbf{X}_{i}-\hat{\tau} \\
\mathbf{C}_{j}=\frac{A_{i} B_{j}}{B_{j} B_{j}^{\prime}} \mathbf{B}_{j},
\end{gathered}
$$

$j=1, \ldots, n$. Then when projecting the points onto the line between $\mathbf{X}_{i}$ and $\hat{\tau}$, the distance of the $j^{\text {th }}$ point from $\hat{\tau}$ is

$$
V_{i j}=\left\|C_{j}\right\| .
$$

The $j^{\text {th }}$ point is declared an outlier if

$$
\begin{equation*}
V_{i j}>M_{V}+c\left(q_{2}-q_{1}\right) \tag{5}
\end{equation*}
$$

Where $M_{V}, q_{1}$ and $q_{2}$ are the usual sample median and estimates of the lower and upper quartiles, respectively, based on the $V_{i 1}, \ldots, V_{i n}$ values, and $c$ is the .95 quantile of a chi-squared distribution with $J$ degrees of freedom. (Here, the quartiles are estimated via the ideal fourths; see Frigge, Hoaglin, \& Iglewicz, 1989.)

The process just described is for a single projection. Repeating this process for each $i(i=1, \ldots, n), \mathbf{X}_{j}$ is declared an outlier if for any of these projections, $V_{i j}$ satisfies (5). Removing any points declared an outlier, the mean of the remaining

## WITHIN GROUPS ANOVA

data is taken to be the SP estimator of location. Its small-sample efficiency compares well to the DG estimator (Wilcox, 2012). Note that the estimate of interquartile range, $q_{2}-q_{1}$, based on the ideal fourths, has a breakdown point of .25 indicating that the breakdown point of the SP estimator is .25 as well. The small-sample efficiency of the SP estimator compares well to several other robust estimators that have been derived ( Ng \& Wilcox, 2010).

Several other affine equivariant estimators were considered but which performed poorly in simulations in terms of controlling the Type I error probability. So computational details related to these other estimators are not provided. They included the minimum volume ellipsoid (MVE) estimator (Rousseeuw \& van Zomeren, 1990), the minimum covariance determinant (MCD) estimator (Rousseeuw \& Van Driessen, 1999), the translated-biweight S-estimator (Rocke, 1996), the median ball algorithm (Olive, 2004) and the orthogonal Gnanadesikan-Kettenring (OGK) estimator (Maronna \& Zamar, 2002).

## Testing (2) and (3)

Two bootstrap methods for testing (2), as well as (3), were considered. The first, which is designed to test (2) and corresponds to the RMPB3 in Wilcox (2012, section 8.2.5), is applied as follows. Compute the test statistic

$$
Q=\sum\left(\hat{\theta}_{j}-\bar{\theta}\right)^{2},
$$

Where $\bar{\theta}=\sum \hat{\theta} / J$. An appropriate critical value is estimated by first setting $Z_{i j}=X_{i j}-\hat{\theta}_{j}$. That is, shift the empirical distributions so that the null hypothesis is true. Next, a bootstrap sample is obtained by resampling, with replacement, $n$ rows from the matrix $\mathbf{Z}$ yielding $Z_{i j}^{*}(i=1, \ldots, n ; j=1, \ldots, J)$. Compute the measure of location that is of interest based on this bootstrap sample yielding $\hat{\theta}_{j}^{*}$ and test statistic $Q^{*}$. Repeat this process $B$ times yielding $Q_{1}^{*}, \ldots, Q_{B}^{*}$. Put these $B$ values in ascending order yielding $Q_{(1)}^{*} \leq \cdots \leq Q_{(B)}^{*}$. Then reject the hypothesis of equal measures of location at the $\alpha$ level if $Q>Q_{(u)}^{*}$, where $u=(1-\alpha) B$ rounded to the nearest integer.

The second method for testing (2) is based in part on bootstrap samples obtained from the $X_{i j}$ values rather than the $Z_{i j}$ values. The strategy is based on determining how deeply the grand mean is nested within the resulting bootstrap

## WILCOX \& HAYES

cloud. Details about this strategy can be found in Wilcox (2012, pp. 392-393). Because this approach performed poorly for the situation at hand, no details are provided.

The two bootstrap methods for testing (3) can be roughly described as follows. Take $B$ bootstrap samples by resampling with replacement from the matrix $\mathbf{X}$, compute a measure of location based on the resulting difference scores and determine how deeply the null vector 0 is nested within the bootstrap cloud. Here, two methods were used to measure the depth of a point in data cloud: Mahalanobis distance and projection distance. In general this approach did not perform well. But when coupled with the DG estimator, it did perform reasonably well when testing (3).

To provide more details, let $\hat{\Delta}_{b}^{*}(b=1, \ldots, B)$ indicate the location estimate of $\Delta$ based on the $b^{\text {th }}$ bootstrap sample and for convenience let $\hat{\Delta}_{0}^{*}$ denote the null vector. Let $P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)$ be the projection distance of $\hat{\Delta}_{b}^{*}$ based on the $B+1$ points $\hat{\Delta}_{0}^{*}, \ldots, \hat{\Delta}_{B}^{*}$. So $P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)$ reflects how far the null vector is from the center of the bootstrap cloud. Then, from general theoretical results in Liu and Singh (1997), a $p$-value is

$$
1-\frac{1}{B} \sum_{b=1}^{B} I\left(P_{d}^{*}\left(\hat{\Delta}_{0}^{*}\right) \geq P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)\right)
$$

where the indicator function $I\left(P_{d}^{*}\left(\hat{\Delta}_{0}^{*}\right) \geq P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)\right)=1$ if $P_{d}^{*}\left(\hat{\Delta}_{0}^{*}\right) \geq P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)$; otherwise $I\left(P_{d}^{*}\left(\hat{\Delta}_{0}^{*}\right) \geq P_{d}^{*}\left(\hat{\Delta}_{b}^{*}\right)\right)=0$. This will be called method D-P. When the projection distance is replaced by Mahalanobis distance, this will be called method D-M.

## Simulation

Simulations were used to study the small-sample properties of the methods described in the previous section. The simulations were run using the software R, with much of the code written in $\mathrm{C}++$. In addition, the R functions took advantage of a multi-core processor via the R package parallel. Despite this, execution time was relatively high, particularly when using the DG estimator in conjunction with method D-P. Consequently, estimated Type I error probabilities were based on 2000 replications. Four types of distributions were used: normal, symmetric and

## WITHIN GROUPS ANOVA

heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, the marginal distributions were taken to be one of four $g$-and- $h$ distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. (The R function rmul, in Wilcox, 2012, was used to generate observations.) If $Z$ has a standard normal distribution and $g>0$, then

$$
W=\frac{\exp (g Z)-1}{g} \exp \left(h Z^{2} / 2\right)
$$

has a $g$-and- $h$ distribution where $g$ and $h$ are parameters that determine the first four moments. If $g=0$, this last equation is taken to be

$$
W=Z \exp \left(h Z^{2} / 2\right)
$$

The four distributions used here were the standard normal ( $g=h=0.0$ ), asymmetric heavy-tailed distribution ( $h=0.2, g=0.0$ ), an asymmetric distribution with relatively light tails ( $h=0.0, g=0.2$ ), and an asymmetric distribution with heavy tails ( $g=h=0.2$ ). Table 1 shows the skewness ( $\kappa_{1}$ ) and kurtosis ( $\kappa_{2}$ ) for each distribution. Additional properties of the $g$-and-h distribution are summarized by Hoaglin (1985). The number of bootstrap samples was taken to be $B=500$. This choice generally seems to perform well in other settings, in terms of controlling the Type I error probability (Wilcox, 2012). But a possibility is that a larger choice for $B$ might yield more power (e.g., Racine \& MacKinnon, 2000). The correlation among the variables was taken to be $\rho=0$ or $\rho=.5$.

Table 1. Some properties of the $g$-and- $h$ distribution.

| $g$ | $h$ | $K_{1}$ | $\kappa_{2}$ |
| ---: | ---: | ---: | ---: |
| 0.0 | 0.0 | 0.00 | 3.0 |
| 0.0 | 0.2 | 0.00 | 21.46 |
| 0.2 | 0.0 | 0.61 | 3.68 |
| 0.2 | 0.2 | 2.81 | 155.98 |

As a partial check on the impact of heteroscedasticity on the Type I error probability, the $X_{i j}$ values were taken to be $\lambda X_{i j}(i=1, \ldots, n)$. The two choices for $\lambda$ were 1 and 4 . For symmetric $g$-and- $h$ distributions ( $g=0$ ), all of the measures of location considered here are equal to zero, so for $\lambda=4$ the null hypothesis remains

## WILCOX \& HAYES

true. But when dealing with skewed distributions ( $g>0$ ), this is not the case. To deal with this, the expected value of an estimator was determined by generating 4000 samples of size $n$ from a specified $g$-and- $h$ distribution (with $\lambda=1$ ) and then averaging the resulting estimates. So with $p=4$, in essence 16,000 estimates are being used. Then the marginal distributions were shifted so that, based on the expected value of an estimator, the null hypothesis is true.

Shown in Table 2 are the results when using the SP estimator with methods D-M and D-P to test (3). Although the seriousness of a Type I error depends on the situation, Bradley (1978) has suggested that as a general guide, when testing at the .05 level, at a minimum the actual level should be between .025 and .075 . As can be seen, this criterion is generally met when using D-M. But under normality, with $\rho=.5$, is this not the case, the largest estimate being .098 . In contrast, when using D-P, the largest estimate is .075 .

Table 2. Estimated Type I error probabilities when testing (3), $n=20, \alpha=.05$ using the SP estimator.

|  |  | D-M |  |  |  | D-P |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\lambda=1$ |  | $\lambda=4$ |  | $\lambda=1$ |  | $\lambda=4$ |  |
| $g$ | $h$ | $\rho=.0$ | $\rho=.5$ | $\rho=.0$ | $\rho=.5$ | $\rho=.0$ | $\rho=.5$ | $\rho=.0$ | $\rho=.5$ |
| 0.0 | 0.0 | . 069 | . 065 | . 096 | . 083 | . 055 | . 063 | . 075 | . 065 |
| 0.0 | 0.2 | . 052 | . 047 | . 055 | . 049 | . 033 | . 042 | . 041 | . 043 |
| 0.2 | 0.0 | . 070 | . 071 | . 039 | . 046 | . 054 | . 070 | . 054 | . 056 |
| 0.2 | 0.2 | . 044 | . 044 | . 030 | . 040 | . 035 | . 039 | . 028 | . 040 |

Reported in Table 3 are simulation results when using method $Q$ to test (2) with the DG estimator and $n=30$. For $n=20$, estimated Type I error probabilities exceed .075. But as indicated in Table 3, with $n=30$, the estimates ranged between .025 and .061 when testing at the .05 level. When testing (2) instead via methods D-M or D-P, control over the Type I error probability was poor.

Table 3. Estimated Type I error probabilities, $n=30, \alpha=.05$ using method $Q$ to test (2) with the DG estimator

|  | $\lambda=1$ |  |  |  | $\rho=.5$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $g$ | $h$ | $\rho=.0$ | $\rho=.5=.0$ | $\rho=.5$ |  |
| 0.0 | 0.0 | .056 | .057 | .053 | .060 |
| 0.0 | 0.2 | .031 | .034 | .040 | .041 |
| 0.2 | 0.0 | .054 | .060 | .057 | .061 |
| 0.2 | 0.2 | .026 | .025 | .038 | .040 |

## WITHIN GROUPS ANOVA

## Conclusion

When using a location estimator that takes into account the overall structure of data when dealing outliers, finding a method for testing (2) and (3) appears to be nontrivial when the sample size is small. The bulk of the methods considered here performed poorly in terms of controlling the Type I error probability, particularly when using an affine equivariant estimator.
Only one method performed well in simulations when testing (2) and an affine equivariant estimator is used: method $Q$ in conjunction with the DG estimator. No method based on an affine equivariant estimator was found to perform reasonably well when testing (3). Moreover, several bootstrap methods that perform reasonably well using a robust estimator applied to each of marginal distributions did not perform well for the situations considered here. However, the skipped estimator studied here, which is location and scale equivariant, was found to perform reasonably well when testing (3) via a percentile bootstrap method that measures the depth of null vector using projection distances. Another possible appeal of the SP estimator over the DG estimator is that for light-tailed distributions, including normal distributions, the DG estimator has relatively poor efficiency (e.g., Massé \& Plante, 2003; Wilcox, 2012, p. 251). In contrast, the SP estimator performs nearly as well as the usual sample mean.

R functions are available for applying the methods that performed well in the simulations. The R function bd1GLOB tests (2). The DG estimator can be used by setting the argument est=dmean. Setting the argument MC=TRUE takes advantage of multi-core processor, if multiple cores are available, via the $R$ package parallel, which can be installed via R command install.packages. The R function rmdzD applies method D-P in conjunction with the SP estimator. Again, setting the argument MC=TRUE will take advantage of a multi-core processor if one is available and the R package parallel has been installed. These functions can be installed with the R command install.packages("WRS",repos="http:R-Forge.R-project.org"). They are also stored in the file Rallfun-v24, which can be downloaded from the first author's web page.

## References

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

## WILCOX \& HAYES

Donoho, D. L. \& Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. Annals of Statistics, 20(4), 1803-1827. doi: 10.1214/aos/1176348890

Frigge, M., Hoaglin, D. C., \& Iglewicz, B. (1989). Some implementations of the boxplot. American Statistician, 43(1), 50-54.
doi: 10.1080/00031305.1989.10475612
Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. C. Hoaglin, F. Mosteller \& J. W. Tukey (Eds.) Exploring data tables trends and shapes (pp. 461-515). Hoboken, NJ: John Wiley \& Sons, Inc.

Liu, R. Y. \& Singh, K. (1997). Notions of limiting $p$ values based on data depth and bootstrap. Journal of the American Statistical Association, 92(437), 266-277. doi: 10.1080/01621459.1997.10473624

Maronna, R. A., Martin, D. R., \& Yohai, V. J. (2006). Robust statistics: Theory and methods. New York, NY: Wiley.

Maronna, R. A., \& Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional data sets. Technometrics, 44(4), 307-317. doi: 10.1198/004017002188618509

Massé, J.-C., \& Plante, J.-F. (2003). A Monte Carlo study of the accuracy and robustness of ten bivariate location estimators. Computational Statistics \& Data Analysis, 42(1/2), 1-26. doi: 10.1016/S0167-9473(02)00103-2

Ng, M., \& Wilcox, R. R. (2010). The small-sample efficiency of some recently proposed multivariate measures of location. Journal of Modern and Applied Statistical Methods, 9(1), 28-42. Available at: http://digitalcommons.wayne.edu/jmasm/vol9/iss1/5

Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. Computational Statistics \& Data Analysis, 46(1), 93-102. doi: 10.1016/S0167-9473(03)00119-1

Racine, J. S., \& MacKinnon, J. G. (2007). Simulation-based tests that can use any number of simulations. Communications in Statistics - Simulation and Computation, 36(2), 357-365. doi: 10.1080/03610910601161256

Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. Annals of Statistics, 24(3), 1327-1345. doi: 10.1214/aos/1032526972

Rousseeuw, P. J., \& Leroy, A. M. (1987). Robust regression and outlier detection. New York, NY: John Wiley \& Sons, Inc.

## WITHIN GROUPS ANOVA

Rousseeuw, P. J., \& Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212-223. doi: 10.1080/00401706.1999.10485670

Rousseeuw, P. J., \& van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411), 633-639. doi: 10.2307/2289995

Staudte, R. G., \& Sheather, S. J. (1990). Robust estimation and testing. New York, NY: Wiley.

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing ( $3^{\text {rd }}$ ed.). San Diego, CA: Academic Press.

## Regular Articles

# Longitudinal Stability of Effect Sizes in Education Research 

Joshua Stephens<br>Cleveland State University<br>Cleveland, OH

Educators use meta-analyses to decide best practices. It has been suggested that effect sizes have declined over time due to various biases. This study applies an established methodological framework to educational meta-analyses and finds that effect sizes have increased from 1970-present. Potential causes for this phenomenon are discussed.

Keywords: Effect sizes, meta-analysis, research methodology, publication bias.

## Introduction

Effect sizes are commonly used in conducting meta-analyses, such as in educational research. Jennions and Moller (2002) suggested reliance on effect sizes has declined somewhat due to various sources of bias. The primary concern of this study is with the application of increased rigor to educational literature. It is important that educators and educational policy-makers use practices and policies based on the strongest empirical evidence. Because public school funding is a limited resource, it is important for that funding to be spent wisely and on effective innovations. This applies to other fields as well, such as social work (Shlonsky, Noonan, Littell, \& Montgomery, 2011).

## Meta-analysis

Effect sizes describe the magnitude difference between the null and alternative hypothesis. Effect sizes are calculated for each study, weighted by sample size and study quality, and then averaged to produce an overall effect size (Littell, Corcoran, \& Pillai, 2008). Although typical data analysis uses multiple observations of a phenomenon as data points, meta-analysis uses multiple studies as data points (Wolf, 1986; Littell et al., 2008). The resulting literature synthesis

[^4]
## LONGITUDINAL STABILITY OF EFFECT SIZES

may become stronger than that provided in a qualitative or narrative fashion (Asher, 1990).

## Unstable Effect Size

Ecologists discovered several examples of diminishing effect sizes (Alatalo, Mappes, \& Elgar, 1997; Gontard-Danek \& Moller, 1999; Poulin, 2000; Simmons, Tomkins, Kotiaho, 1999). An interpretation of why effect sizes apparently diminish over time has not emerged. The following are possible explanations. Alatalo et al. (1997) attributed diminishing effect sizes to changing belief systems. Palmer (2000) attributed the phenomenon to fads. Tregenza and Wedell (1997) attributed it to biased study design. Alatalo et al. (1997) suggested submitting findings for publication that support previously held ideas makes it easier to get published. Simmons et al. (1999) suggested that it is easier to publish confirmatory findings during early stages of research in a particular field, but it becomes more difficult as critique of that field narrows. This may be particularly emphasized in the social sciences, where it takes longer to publish non-significant results (Stern \& Simes, 1997).

Social science researchers who study the phenomenon of diminishing effect sizes cite two primary potential causes: dissemination bias and citation bias.

Dissemination bias is a broad term encompassing many different sorts of biases related to the publication and dissemination process, including bias related to date of publication, language, multiple publication bias, selective reference citation, database index bias, media attributed bias, selective publication bias, familiarity of techniques, and the cost of research reports (Rothstein, Sutton \& Bornstein, 2005; Song, Eastwood, Gilbody, Duley, \& Sutton, 2000). "Dissemination bias occurs when the dissemination profile of a study's results depends on the direction or strength of its findings" (Song et al., 2000, p. 17). It refers to the notion that a given literature review does not represent a random sampling of all studies in a given field, and therefore is a type of non-random sampling error similar to that found when conducting primary research (Song et al., 2000).

Both indirect and direct evidence support the existence of dissemination bias (Sohn, 1996). Examples of indirect evidence include disproportionately high percentage of positive findings in journals, or larger effect sizes in small studies relative to large studies. Small studies are more vulnerable to dissemination biases, as the results of these studies will be more widely spread around the true results owing to greater random error (Begg \& Berlin, 1988). Direct evidence includes

## JOSHUA STEPHENS

such things as admissions by investigators and publishers and comparison of results from published and unpublished studies (Song et al., 2000). Rotton, Foos, VanMeek, and Levitt (1995) found that the most significant reason given by authors for not submitting their work for publication was the failure to find statistical significance.

The strongest evidence supporting the existence of dissemination bias comes from comparisons between published and unpublished studies (Song et al., 2000). Simes (1986) performed meta-analyses on both published and unpublished studies of a cancer treatment regimen and discovered that the published findings found that the treatment was effective, but when the published and unpublished studies were analyzed together, the treatment effect was not found.

There are specific types of dissemination bias. Biases in addition to those mentioned earlier include positive results bias, hot stuff bias, time-lag bias, grey literature bias, full publication bias, place of publication bias, outcome reporting bias, and retrieval bias (Song et al., 2000). These forms of bias may be prevalent in many disciplines and may account for observed decline in effect sizes in ecology and other fields.

## Methodology

The purpose of this study is to analyze whether meta-analytically derived results are longitudinally stable in education research. To accomplish this task, a process similar to that used by Jennions and Moller (2002) will be invoked.

## Study Selection

First, a set of meta-analyses, based on $\mathrm{K}-12$ classroom interventions from the years 1970 to 2011, was selected from the EBSCOHost databases. Studies were included if they specifically provide effect size results based on meta-analytical techniques and provide a comprehensive list of studies used to generate effect sizes.

Table 1. Descriptive statistics of included studies

| N | Year of Publication <br> Range | Mean Year <br> of Publication | Mean Number of <br> Reported Effect Sizes <br> Per Meta-Analysis |
| :---: | :---: | :---: | :---: |
| 60 | $1984-2010$ | 2002.3 | 42.7 |

## LONGITUDINAL STABILITY OF EFFECT SIZES

The study involved a broad search for literature, which was then winnowed down through a rigorous paring process, resulting in a final set of 60 studies that were analyzed. Descriptive statistics of these studies are shown in Table 1. After final literature was selected, data analysis was initiated.

## Statistical Procedure

The 60 selected studies were then analyzed using a process outlined by Jennions and Moller (2002), involving the use of four Spearman's $\rho$ (rho) analyses on two levels. The first set of analyses dealt with the effect sizes reported in the selected studies. This will hereafter be known as the "study level" of analysis. The second set of analyses were conducted on the meta-analyses themselves. This is hereafter known as the "meta-analysis level."

On both the study level and the meta-analysis level, four relationships were analyzed: (i) the relationship between effect size and year of publication; (ii) the relationship between effect size and sample size; (iii) the relationship between standardized effect size and sample size; and (iv) the relationship between effect size and year of publication, after weighting for variation in sampling effort. The first three relationships were conducted using a Spearman's $\rho$ (rho) test and were performed in SPSS.

The fourth relationship was conducted using MetaWin 2.0. This relationship was estimated by creating a random-effects continuous model meta-analysis with year of publication as the independent variable and the inverse of sampling variance as the weighting factor. Random-effects meta-analysis was selected over a fixed-effects model, as fixed-effects models become problematic when some studies have very large sample sizes. These studies then dominate the analysis, and the results from the studies with smaller sample sizes are largely ignored (Helfenstein, 2002).

MetaWin 2.0 was used to obtain a one-tailed $\rho$-value for year of publication generated by a randomization method with 999 replicates. A one-tailed $\rho$-value was chosen because the Jennions and Moller (2002) study used a one-tailed test, since they postulated that a declining effect size was more likely. The effect size generated by the meta-analysis was converted to a Spearman's $\rho$ - (rho-) value so that all results were reported in a uniform manner. The formula to do this is as follows:

$$
\rho=\sqrt{\frac{d^{2}}{d^{2}+4}}
$$

## JOSHUA STEPHENS

All Spearman's $\rho$-values were then converted to standard normal deviates (Zscores), using the formula:

$$
\rho=\frac{\sqrt{Z^{2}}}{n}
$$

This was done so that all results were normalized, thus diminishing the effects of outliers and providing a more robust answer to the research question.

## Results

Results regarding the possibility of effect sizes diminishing over time are compiled in Table 2.

Table 2. Relationships ( $\rho$ ) between effect size, standardized effect size, year of publication, and sample size.

|  | Method of Calculation |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Weighted meta- <br> analysis of: | Year v. <br> Effect Size | $\boldsymbol{n} \mathbf{v}$. <br> Effect Size | $\boldsymbol{n}$ v. <br> Standard Effect | Year v. Effect Size (after <br> weighing for sampling variance) |
| Datasets | $0.105^{\star}$ | $-0.073^{* *}$ | $-0.073^{\star *}$ | $0.440^{\star}$ |
| Original Meta- <br> Analyses | $0.317^{* *}$ | -0.148 | -0.148 | $0.333^{\star}$ |

Note: * Significant at the <0.001 level; **Significant at the <0.01 level

Beginning at the study level, these results indicate that there is a statistically significant positive relationship between year of publication and effect size ( $\rho=0.105, p<0.001, n=1167$ ). However, there was also a significant relationship between sample size and both effect size and standardized effect size, so the relationship was re-assessed after accounting for sampling variance. Still, however, a statistically significant positive relationship was observed ( $\rho=0.440$, $p<0.001, n=1167$ ). Figures $1-4$ show scatterplots of these four relationships.

## LONGITUDINAL STABILITY OF EFFECT SIZES



Figure 1. Publication year compared to effect size $(g)$ at the study level


Figure 3. Sample size compared to standardized effect size ( $z$-transformed) at the study level


Figure 2. Sample size compared to effect size $(g)$ at the study level


Figure 4. Year of publication compared to effect size ( $g$ ) after weighting for sample size at the study level

A similar observation is found at the meta-analysis level. These results indicate that there is a statistically significant positive relationship between year of publication and effect size ( $\rho=0.317, p<0.009, n=60$ ). However, there was not a significant relationship between sample size and both effect size and standardized effect size. Still, however, a statistically significant, positive relationship was observed ( $\rho=0.333, p<0.001, n=60$ ) after accounting for sampling variance. Figures $5-8$ below show scatterplots of the relationships from the meta-analysis level.

## JOSHUA STEPHENS



Figure 5. Publication year compared to effect size ( $g$ ) at the meta-analysis level


Figure 7. Sample size compared to standardized effect size ( $z$-transformed) at the meta-analysis level

It is notable that effect sizes increase at both the study and meta-analysis levels. Data were parsed out to show mean effect sizes by decade to allow for simpler understanding of how effect sizes have increased over time. Table 3 shows this descriptive information.

## LONGITUDINAL STABILITY OF EFFECT SIZES

Table 3. Mean effect sizes by decade.

|  | N | Mean effect size <br> $(\boldsymbol{g})$ | Range | Standard <br> Deviation |
| :---: | :---: | :---: | :---: | :---: |
| 1970s / 1980s | 2 | 0.100 | $-0.20-0.40$ | 0.424 |
| 1990s | 21 | 0.424 | $-0.09-1.61$ | 0.329 |
| 2000s | 31 | 0.509 | $-0.75-1.40$ | 0.506 |
| $2010 s$ | 6 | 0.595 | $0.33-0.91$ | 0.276 |

## Conclusion

It was found that education meta-analyses do not appear to follow the pattern seen in the natural sciences, because the effect sizes on which they are based did not decline. On the contrary, for the sample included in this study, they tended to increase over time.

This finding bears some consideration. If no statistically significant relationships had been observed between effect sizes and year of publication, then it could be assumed that meta-analysis provides a longitudinally stable measure, and a strong argument could have been made for wider use of this analytical technique. However, as measured effect sizes tend to increase over the time period 1970 - 2012, either there is some persistent set of biases that are impacting the conduct or publication of educational research, or effect sizes are indeed increasing over time as the field of education develops into a more complex and sophisticated science and leaves behind ineffective educational practices.

## Persistent Bias in Educational Research

One explanation for the observed phenomenon of longitudinally increasing effect sizes is publication bias. Given the findings of this study, it seems reasonable to conclude that it is possible that some of these forms of bias may be more active than others. In particular, the following forms of publication bias are possible explanations for the findings of this study: positive results bias; hot stuff bias; grey literature bias; and confirmation bias.

Positive results bias Positive results bias refers to the tendency of authors to submit-and for editors to publish—positive or significant research results while ignoring non-significant results (Song et al., 2000). This seems to be a likely cause of increasing effect sizes. Since researchers generally will find statistically significant results when they are searching for literature to use to

## JOSHUA STEPHENS

conduct meta-analyses, they will find ever-increasing effect sizes across time. Then this effect becomes multiplied, as other researchers use published metaanalyses to generate effect sizes for new research and duplicate biases from past research.

Hot stuff bias
Another form of bias that could account for the phenomenon of increasing effect sizes is hot stuff bias. This refers to the phenomenon of journal publishers tending to publish topics that are timely or popular but which may only have relatively weak results (Sackett, 1979). This seems to be a likely form of publication bias in education where fads and trends dominate pedagogical practice. These trends may be pushed by textbook publishers looking to profit from a product, or politicians who make educational policy with little understanding of educational systems and processes.

Hot stuff bias may account for increasing effect sizes through publishers choosing articles to publish based on what they believe will promote their journal's readership. Publishers choose articles that may be methodologically unsound; these articles are then indexed in electronic indexes and used to conduct meta-analyses, thereby creating the appearance of increasing effect sizes over time. When the particular timely trend ends, no researcher bothers to fully repudiate it or no journal chooses to publish these repudiations, so it appears that these effect sizes are significant and increasing over time.

Grey literature bias Grey literature refers to things such as conference presentations, dissertations, working papers, and other pieces of literature that are difficult to obtain as they are not electronically indexed in any systematic manner (Auger, 1998). Grey literature bias refers to the notion that these pieces of literature tend to show non-significant or statistically weaker results and that excluding these from meta-analyses produces an artificially high effect size (Song et al., 2000). McAuley et al. (1999) sampled 135 meta-analyses, 38 of which included grey literature, found that those meta-analyses that included grey literature showed a diminished effect size of approximately $12 \%$.

Grey literature bias would appear to be a significant problem in the field of educational research where many universities have large numbers of master's and doctoral students who are producing volumes of research that is never published. While it is difficult to quantify specifically how much research is conducted and never included in any sort of meta-analysis, it is safe to assume it must be a large amount every year. When one includes classroom research done by practicing teachers, the amount of grey literature skyrockets. While not all of this research

## LONGITUDINAL STABILITY OF EFFECT SIZES

would meet methodological criteria for publication or for inclusion in properly conducted meta-analyses, some certainly would. The exclusion of this grey literature could be a significant factor in the observed phenomenon of increasing effect sizes. If established researchers get their statistically significant findings published while student researchers or others who find non-significance do not, then effect sizes would tend to increase over time as no one individual or organization reputes earlier findings.

Confirmation bias Confirmation bias refers to the psychological phenomenon whereby humans tend to subconsciously look for ideas and information that confirms their earlier beliefs. This information tends to be more readily assimilated and utilized than does information that contradicts what an individual believes (Bushman \& Wells, 2001).

Confirmation bias seems like a likely cause of increasing effect sizes. As researchers look for studies to help them build the case for their study, they will naturally begin by searching for studies that confirm what they already believe. As they find increasing numbers of these studies, it seems that the results of the study are a foregone conclusion. This may lead researchers to discount or ignore studies that may disagree with what they believe is true about a research question. In a meta-analysis, this may take the form of a researcher applying more stringent selection criteria to studies that don't confirm his or her hypothesis, leading to effect sizes that increase across time.

## Increasing effect sizes represent educational reality

There is another explanation for the phenomenon of longitudinally increasing effect sizes in educational research: it is possible that effect sizes seem to be increasing because they actually are. This is a hopeful notion that as educational researchers have begun to more rigorously conduct research and educational practitioners have received better training in the utilization of research-based educational techniques, that educational practices have become more effective. This would be supported by the fact that, over the past 40 years in the sample considered in this study, many states have implemented tougher teacher training and licensure laws, and departments of education at universities have taken a more rigorously quantitative approach. However, when the outcomes of largescale assessments of student learning are observed across this time period, no similarly significant gains are apparent. It is beyond the scope of this study to adequately assess the growth of students in comparison to the perceived growth of

## JOSHUA STEPHENS

teacher effectiveness. However, it does seem less likely that this is the case and more likely that the correct explanation for the phenomenon of longitudinally increasing effect sizes is publication bias.

## Potential solutions for addressing increasing effect sizes

If, as the results of this study suggest, effect sizes are in fact increasing over time, then this potentially indicates that there is a problem in the publication process that should be corrected by researchers and publishers. Failure to do so may cause misperceptions regarding the efficacy of a host of educational interventions that may diminish the impact of schooling for students which is a patently undesirable outcome.

Educational researchers should strive to conduct meta-analyses and other research in the most methodologically sound manner possible. Narrative literature reviews should be only used when a research question is either very limited in scope or is so new that very little literature is available such that it would be possible for a researcher to adequately summarize findings from the literature base without quantitative methods. It may also be useful to provide narrative literature reviews as an element of a meta-analysis. Meta-analytic techniques should be included in most literature reviews and these techniques should follow the guidelines set forth by the Cochrane and Campbell Collaborations (Pfeffer \& Sutton, 2006). These organizations have initiated programming to assist researchers with developing the most accurate summarizations of literature possible. Following their recommendations globally would create a less biased body of educational literature that would be more useful to practitioners and researchers alike.

To further ameliorate this phenomenon, there would need to be a change in the way education research is published. First and foremost, there must be a journal dedicated to publishing only null or statistically insignificant findings. This journal must be indexed properly in major educational research databases and should draw from as many countries and languages as possible. By doing so, researchers who wish to properly conduct meta-analyses will be able to more readily access these results and then conduct a more methodologically sound and less biased meta-analysis.

Additionally, a comprehensive effort should be made to index the wide body of grey literature that is generated globally each year. Conference presentations, dissertations, theses, working papers, action research and other forms of grey literature may provide important insight into research questions and should not be

## LONGITUDINAL STABILITY OF EFFECT SIZES

ignored. Moreover, publishers should be conservative when announcing special issues or accepting papers on topics that are very new. Although this is difficult to do and may not always be advisable, this would help alleviate the problems associated with hot stuff bias, as described above.

## Limitations of the present study

There are two limitations in this study that require comment. First, component studies came from a limited subset of education studies. Hence, a more inclusive literature search may invalidate or temper the results found here. Second, it has been opined that meta-analysis be conducted using a team of reviewers who make decisions regarding which studies to include. Presumably, that process creates a less biased set of inclusion criteria. It is possible that, had this research been conducted utilizing a team of researchers or assistants to help determine which studies should be included, the results of this project may have been different.

The larger question remains as to the cause of the observed phenomenon. Is it caused by pervasive publication biases that should be immediately addressed and remedied, or have effect sizes increased because educators have become better at their jobs over the past 40 years? This causal question is truly vexing and should be a primary focus of future research. In general, publication biases are not widely studied in education, and should be a source of concern for the community of educational researchers and for those who utilize that research.

## Acknowledgments

This paper was developed in part from the author's dissertation (Stephens, 2013).

## References

Alatalo, R. V., Mappes, J. \& Elgar, M. (1997). Heritabilities and paradigm shifts. Nature, 385, 402 - 403. doi: 10.1038/385402a0

Asher, W. (1990). Education psychology, research methodology, and metaanalysis. Educational Psychologist, 25(2), 143. doi: 10.1207/s15326985ep2502_3

Auger, C. P. (1998). Information sources in grey literature, 4th ed. London: Bowker Saur.

Begg, C. B. \& Berlin, J. A. (1988). Publication bias: a problem in interpreting medical data. Journal of the Royal Statistical Society. Series A (Statistics in Society), 151(3), 419 - 463. doi: 10.2307/2982993

## JOSHUA STEPHENS

Bushman, B. J. \& Wells, G. L. (2001). Narrative impressions of literature: The availability of bias and corrective properties of eta-analytic approaches. Personality and Social Psychology Bulletin, 27(9), 1123 - 1130. doi: 10.1177/0146167201279005

Gontard-Danek, M. C. \& Moller, A. P. (1999). The strength of sexual selection: a meta-analysis of bird studies. Behavioral Ecology, 10(5), 476-486. doi: 10.1093/beheco/10.5.476

Helfenstein, U. (2002). Data and models determine treatment proposals - an illustration from meta-analysis. Postgraduate Medical Journal, 78(917), 131 134. doi: 10.1136/pmj.78.917.131

Jennions, M. \& Moller, A. (2002). Relationships fade with time: a metaanalysis of temporal trends in publication in ecology and evolution. Proceedings of the Royal Society of London B: Biological Sciences, 269(1486), 43-48. doi: 10.1098/rspb.2001.1832

Littell, J. H., Corcoran, J., \& Pillai, V. (2008). Systematic reviews and metaanalysis. New York, NY: Oxford University Press.

McAuley, L., Moher, D., Pham, B., Tugwell, P. (1999). Evaluation of the impact of grey literature in meta-analysis. In: Proceedings of the 7th Cochrane Colloquium; Oct 5-9; Rome. Rome: Universitas Tommaso D’ Aquino: 17.

Palmer, A. R. (2000). Quasireplication and the contract of error: Lessons from sex ratios, heritabilities and fluctuating asymmetry. Annual Review of Ecology and Systematics, 31, 441 - 480. doi: 10.1146/annurev.ecolsys.31.1.441

Pfeffer, J. \& Sutton, R. I. (2006). Demanding proof: engineers are poised to lead an evidence-based management movement. Industrial Engineer, 38(6), 43-47.

Poulin, R. (2000). Manipulation of host behaviour by parasites: a weakening paradigm? Proceedings of the Royal Society of London B: Biological Sciences, 267(1445), 787 - 792. doi: 10.1098/rspb.2000.1072

Rosenberg, M., Adams, D., \& Gurevitch, J. (2000). Meta-win: statistical software for meta-analysis, v. 2.0. Sunderland, MA: Sinauer.

Rothstein, H., Sutton, A. J., \& Bornstein, M. (Eds.). (2005). Publication bias in meta-analysis: Prevention, assessment, and adjustments. Chichester, UK: Wiley.

Rotton, J., Foos, P. W., Vanmeek, L., \& Levitt, M. (1995). Publication practices and the file drawer problem - a survey of published authors. Journal of Social Behavior and Personality, 10(1), 1-13.

## LONGITUDINAL STABILITY OF EFFECT SIZES

Sackett, D. L. (1979). Bias in analytic research. Journal of Chronic Disease, 32(1-2), 51 - 53. doi: 10.1016/0021-9681(79)90012-2

Shlonsky, A., Noonan, E., Littell, J., \& Montgomery, P. (2011). The role of systematic reviews and the Campbell Collaboration in the realization of evidenceinformed practice. Clinical Social Work Journal, 39(4), 362-368. doi: 10.1007/s10615-010-0307-0

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. Journal of Clinical Oncology, 4(10), 1529-1541.

Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. \& Hunt, J. (1999).
Fluctuating paradigm. Proceedings of the Royal Society of London B: Biological Sciences, 266(1419), 593 - 595. doi: 10.1098/rspb.1999.0677

Sohn, D. (1996). Publications bias and the evaluation of psychotherapy efficacy in reviews of the research literature. Clinical Psychology Review, 16(2), 147 - 156. doi: 10.1016/0272-7358(96)00005-0

Song, F., Eastwood, A. J., Gilbody, S., Duley, L. \& Sutton, A. J. (2000). Publication and related biases. Health Technology Assessment, 4(10), 1 - 125. doi: 10.3310/hta4100

Stephens, J. (2013). Longitudinal stability of effect sizes in educational research (Doctoral dissertation). Retrieved from EngagedScholarship@CSU (http://engagedscholarship.csuohio.edu/etdarchive/281/)

Stern, J. M. \& Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. British Medical Journal, 315(7109), 640-645.

Tregenza, T. \& Wedell, N. (1997). Natural selection bias. Nature, 386, 234. doi: $10.1038 / 386234 \mathrm{~b} 0$

Wolf, F. (1986). "Meta-analysis." Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills: Sage Publications.

# A Comparison of Usual $\boldsymbol{t}$-Test Statistic and Modified $t$-Test Statistics on Skewed Distribution Functions 

Wooi K. Lim<br>William Paterson University<br>Wayne, New Jersey

Alice W. Lim<br>University of California at Los Angeles<br>Los Angeles, California


#### Abstract

When the sample size $n$ is small, the random variable $T=\sqrt{n}(\bar{X}-\mu) / S$ is said to follow a central $t$ distribution with degrees of freedom $(n-1)$, where $\bar{X}$ is the sample mean and $S$ is the sample standard deviation, provided that the data $X \sim N\left(\mu, \sigma^{2}\right)$. The random variable $T$ can be used as a test statistic to hypothesize the population mean $\mu$. Some argue that the $t$-test statistic is robust against the normality of the distribution and claim that the normality assumption is not necessary. In this article we will use simulation to study whether the $t$-test is really robust if the population distribution is not normally distributed. In particular, we will study how the skewness of a probability distribution will affect the confidence interval as well as the $t$-test statistic.


Keywords: $\quad$ Skewness, $t$-test, confidence interval, Edgeworth expansion

## Introduction

The effect of skewness, denoted by $\gamma$ from here on, of a random variable $X$ on $t$-test have been investigated by Johnson (1978), Hall (1992), Abramovitch and Singh (1985) and many others; but, those are more on the theoretic investigation and concentrated on the $t$-test. Very little has been studied on the confidence interval. Two independent samples $t$-test are studied by Sawilowsky and Blair (1992). Their studies are based on several skewed distributions and various sample sizes. Their simulation results show that the proportions of rejection in the upper tail or lower tail are affected by the skewness of the distribution when samples sizes are small. Blair and Sawilowsky (1993) comparing the performance

Wooi K. Lim is an Associate Professor of Statistics in the Department of Mathematics at William Paterson University. Email him at: limw@ wpunj.edu. Alice W. Lim is a student majoring in Mathematics at UCLA. Email her at: canalice @gmail.com.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

usual independent samples $t$-test and modified $t$-tests under different distribution functions and various samples sizes. For further details on the performance and recommendation of which $t$-test under various distributions, see Blair and Sawilowsky (1993).

Consider the one sample $t$-test. Based on simulation studies, skewness of the distribution does not affect the $t$-test as much as the confidence interval. It can be shown that the coverage error is larger than the pre-determined coverage error, $\alpha$, if the data follow a skewed distribution function.

Intuitively, if $X$ is a random variable with mean $\mu$ but is positively skewed, $\gamma>0$, then the population median is less than the population mean $\mu$. A sample of size $n$ from $X$ is likely to have more than $50 \%$ of values to be less than $\mu$; hence most likely $(\bar{X}-\mu)<0$. If $\gamma>0$, then a $(1-\alpha) \times 100 \%$ confidence interval for $\mu$

$$
\begin{equation*}
\left(\bar{X}-t_{\alpha / 2} \frac{S}{\sqrt{n}}, \bar{X}+t_{\alpha / 2} \frac{S}{\sqrt{n}}\right) \tag{1}
\end{equation*}
$$

will miss the mean $\mu$ more on the upper side than the lower side. This effect is reported by Boos and Hughes-Oliver (2000). Define the missed right and missed left as given in Boos and Hughes-Oliver (2000, p. 122), where miss right occurs when the population mean $\mu$ is above the upper confidence limit, i.e.,

$$
\mu>\bar{X}+t_{\alpha / 2} \frac{S}{\sqrt{n}},
$$

and miss left occurs when the population mean $\mu$ is below the lower confidence limit, i.e.,

$$
\mu<\bar{X}-t_{\alpha / 2} \frac{S}{\sqrt{n}}
$$

and miss $=($ miss right + miss left $)$. Tables 1,2,3 and 4 are the simulated results of missed right, missed left and missed of usual confidence interval given in equation (1). Four types of population distributions, namely normal distribution $(\gamma=0)$, Laplace distribution $(\gamma=0)$, Gamma distribution $(\gamma>0)$ and Gumbel distribution $(\gamma<0)$ were selected for the simulation study.

## LIM \& LIM

Table 1. Table of miss right, miss left and miss of $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Normal}(1,2), n=10$ and skewness $=0.0$.

| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| miss right | 0.005 | 0.011 | 0.015 | 0.021 | 0.025 | 0.029 | 0.033 | 0.041 | 0.043 | 0.052 |
| miss left | 0.004 | 0.009 | 0.017 | 0.022 | 0.023 | 0.033 | 0.034 | 0.038 | 0.049 | 0.06 |
| miss | 0.009 | 0.02 | 0.032 | 0.044 | 0.048 | 0.061 | 0.067 | 0.079 | 0.092 | 0.111 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss right | 0.053 | 0.057 | 0.067 | 0.073 | 0.07 | 0.082 | 0.088 | 0.1 | 0.098 | 0.101 |
| miss left | 0.056 | 0.061 | 0.063 | 0.071 | 0.082 | 0.082 | 0.087 | 0.093 | 0.107 | 0.101 |
| miss | 0.109 | 0.118 | 0.129 | 0.144 | 0.152 | 0.164 | 0.175 | 0.192 | 0.205 | 0.201 |

Table 2. Table of miss right, miss left and miss of $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Laplace}(1,2), n=10$ and skewness $=0.0$.

| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| miss right | 0.003 | 0.007 | 0.012 | 0.015 | 0.02 | 0.025 | 0.03 | 0.037 | 0.041 | 0.047 |
| miss left | 0.004 | 0.005 | 0.01 | 0.014 | 0.02 | 0.026 | 0.031 | 0.035 | 0.039 | 0.047 |
| miss | 0.007 | 0.012 | 0.022 | 0.029 | 0.04 | 0.051 | 0.061 | 0.072 | 0.08 | 0.094 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss right | 0.052 | 0.059 | 0.058 | 0.07 | 0.075 | 0.081 | 0.086 | 0.095 | 0.095 | 0.101 |
| miss left | 0.054 | 0.058 | 0.07 | 0.07 | 0.079 | 0.079 | 0.083 | 0.089 | 0.1 | 0.1 |
| miss | 0.106 | 0.117 | 0.128 | 0.14 | 0.154 | 0.16 | 0.169 | 0.184 | 0.195 | 0.201 |

Table 3. Table of miss right, miss left and miss of $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gamma}(1,2), n=10$ and skewness $=2$.

| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| miss right | 0.051 | 0.061 | 0.08 | 0.093 | 0.101 | 0.108 | 0.121 | 0.122 | 0.125 | 0.13 |
| miss left | 0.0 | 0.001 | 0.001 | 0.002 | 0.005 | 0.004 | 0.007 | 0.009 | 0.009 | 0.015 |
| miss | 0.051 | 0.062 | 0.081 | 0.095 | 0.106 | 0.112 | 0.128 | 0.132 | 0.134 | 0.145 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss right | 0.144 | 0.149 | 0.155 | 0.157 | 0.167 | 0.174 | 0.177 | 0.175 | 0.185 | 0.192 |
| miss left | 0.015 | 0.017 | 0.02 | 0.025 | 0.03 | 0.029 | 0.037 | 0.044 | 0.047 | 0.047 |
| miss | 0.159 | 0.166 | 0.175 | 0.182 | 0.197 | 0.203 | 0.215 | 0.219 | 0.232 | 0.239 |

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

Table 4. Table of miss right, miss left and miss of $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gumbel}(1,2), n=10$ and skewness $=-1.14$.

| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| miss right | 0.001 | 0.003 | 0.005 | 0.008 | 0.01 | 0.014 | 0.016 | 0.02 | 0.021 | 0.029 |
| miss left | 0.019 | 0.03 | 0.036 | 0.044 | 0.054 | 0.066 | 0.068 | 0.082 | 0.084 | 0.089 |
| miss | 0.02 | 0.033 | 0.041 | 0.052 | 0.064 | 0.08 | 0.084 | 0.102 | 0.105 | 0.118 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss right | 0.031 | 0.031 | 0.039 | 0.042 | 0.047 | 0.049 | 0.055 | 0.062 | 0.063 | 0.063 |
| Miss left | 0.099 | 0.094 | 0.103 | 0.117 | 0.118 | 0.118 | 0.129 | 0.137 | 0.142 | 0.149 |
| miss | 0.13 | 0.125 | 0.142 | 0.159 | 0.165 | 0.167 | 0.184 | 0.199 | 0.206 | 0.212 |

It is shown in Table 1 if $X$ is normally distributed, the nominal coverage error $\alpha$ is close to the simulated missed coverage error. Results in Tables 1 and 2 show that if the probability distributions are symmetrically distributed, then the missed left $\approx$ missed right. Tables 3 and 4 show that if $X$ is skewed, such as in the Gamma distribution (Table 3) or Gumbel distribution (Table 4), the missed coverage error is more than the nominal coverage error $\alpha$. Interestingly, the results show that for a right skewed distribution, the missed right coverage errors are substantially greater than the missed left coverage errors (see Table 3). The opposite is true for the left skewed, Gumbel, distribution (see Table 4).

It is well known that the random variable $T=\sqrt{n}(\bar{X}-\mu) / S$ is a ratio of the normal random variable and $\chi_{(n-1)}$ random variable with $(\bar{X}-\mu)$ and $S$ statistically independent. Will the random variable $T=\sqrt{n}(\bar{X}-\mu) / S$ be affected by the skewness of the probability distribution? Simulated empirical distribution of $T$ for the same four chosen population distributions are under studied. Our results are summarized in Figures 1, 2, 3 and 4. In this simulation, a sample of $n=10$ is drawn from the population distribution with replications of $M=5000$. Each figure contains figures (a), (b) and (c), with the exception of Figure 1. Figures (a) are histograms of the $t$-test statistics, $t^{*}=\sqrt{n}\left(\bar{X}-\mu_{0}\right) / S$, under the assumption that $H_{0}: \mu=\mu_{0}$ is true. Figures (b) are the plots of $\bar{X}-\mu_{0}$ versus $S$.

In Figure $1, X_{1}, X_{2}, \ldots, X_{10} \stackrel{i i d}{\sim} N(1,2)$ were sampled. The histogram in Figure 1(a) is an empirical distribution of $t^{*}$ under the assumption of $H_{0}: \mu=1$. The histogram is quite symmetric and the plot of $\bar{X}-\mu_{0}$ versus $S$ does not seem
to have any correlation. This is what we expected from a $t$-test statistic. What happens if $X$ is not normal?

Figure 2(a) is the distribution of $t^{*}$ with $X_{1}, X_{2}, \ldots, X_{10} \sim$ idaplace $(1,2)$. The histogram shows that $t^{*}$ is symmetric. The plot of $\left(\bar{X}-\mu_{0}\right)$, where $\mu_{0}=1$, versus $S$, see Figure 2(b), does not show any correlation. Figure 2(c) is the empirical distribution of $t^{*}$ based on $X \sim N(1,2)$ versus $X \sim$ Laplace $(1,2)$. It can be seen that the distribution of $t^{*}$ based on $X \sim$ Laplace $(1,2)$ has shorter tails than the $t^{*}$ computed from $X \sim N(1,2)$. Clearly the variability of $\left(\bar{X}-\mu_{0}\right)$ plays a role in the distribution of $t^{*}$. This may suggest that $t^{*}$ generated from $X$ following a Laplace distribution may not be as sensitive as the $t^{*}$ obtained from a normal distribution.


Figure 1. (a) histogram of $t^{*}=\sqrt{n}\left(\bar{X}-\mu_{0}\right) / S$, where $\mu_{0}=1$,
$X_{1}, X_{2}, \ldots, X_{10} \sim \operatorname{Normal}(1,2)$. (b) plot of $\left(\bar{X}-\mu_{0}\right)$ versus $S$.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION



Figure 2. (a) histogram of $t^{*}=\sqrt{n}\left(\bar{X}-\mu_{0}\right) / S$, where $\mu_{0}=1$,
$X_{1}, X_{2}, \ldots, X_{10} \sim \operatorname{Laplace}(1,2)$. (b) plot of $\left(\bar{X}-\mu_{0}\right)$ versus $S$. (c) Fig. 2a over Fig. 1a


Figure 3. (a) histogram of $t^{*}=\sqrt{n}\left(\bar{X}-\mu_{0}\right) / S$, where $\mu_{0}=2$,
$X_{1}, X_{2}, \ldots, X_{10} \sim \operatorname{Gamma}(1,2)$. (b) plot of $\left(\bar{X}-\mu_{0}\right)$ versus $S$. (c) Fig 3a over Fig 1a

## T-TEST ON SKEWED DISTRIBUTION FUNCTION



Figure 4. (a) histogram of $t^{*}=\sqrt{n}\left(\bar{X}-\mu_{0}\right) / S$, where $\mu_{0} \approx-0.1544$, $X_{1}, X_{2}, \ldots, X_{10} \sim \operatorname{Gumbel}(1,2)$. (b) plot of $\left(\bar{X}-\mu_{0}\right)$ versus $S$. (c) Fig 4a over Fig 1a

Figures in 3 and 4 are simulation results from a skewed probability distributions. Figure 3(a) is the distribution of $t^{*}$ with $X_{1}, X_{2}, \ldots, X_{10}{ }^{\text {iid }} \sim \operatorname{Gamma}(1,2)$. Interestingly, Gamma distribution is a right skewed distribution but the distribution of $t^{*}$ is left skewed. One can see in Figure 3(b) that there is a positive correlation between $\left(\bar{X}-\mu_{0}\right)$, where $\mu_{0}=2$, and $S$.

Further, Figure 3(c) shows that $t^{*}$ s constructed from $X \sim \operatorname{Gamma}(1,2)$ lie below 0 more often than fall above 0 . One can see in Figure 3(b) that $\left(\bar{X}-\mu_{0}\right)$ versus $S$ is more disperse when $\left(\bar{X}-\mu_{0}\right)>0$ while it is less varied when $\left(\bar{X}-\mu_{0}\right)<0$. Thus, when $\left(\bar{X}-\mu_{0}\right)>0$ and large, it tends to counter by large $S$ making $t^{*}$ more concentrated on the right hand side. On the other hand, when $\left(\bar{X}-\mu_{0}\right)<0$ and $S$ is small, $t^{*}$ tends to stretch further towards the negative side making $t^{*}$ skewed negatively. Similar arguments can explain why left skewed distributions will have $\bar{X}$ overestimate $\mu$ more often and making the distribution of $t^{*}$ positively skewed, see Figure 4 where $\mu_{0} \approx-0.1544$. In the next section we will compare the two transformation methods, proposed by Hall (1992), with the usual test statistics $T$.

## Correction and Transformation

Johnson (1978) and others noticed some undesired effects on skewed distributions on the $t$-test. Hall (1992) proposed to modify the $t$-test statistic $T$, say $g(T)$, so that $g(T)$ is less skew and less bias. The transformed test statistic $g(T)$ must be invertible to obtain a unique modified confidence interval for $\mu$. He suggested $g$ been a monotonic function to achieve the invertibility. The two monotonic transformations of $T$ proposed by Hall (1992) are:

$$
\begin{equation*}
T_{1}=T+\frac{a}{\sqrt{n}} \hat{\gamma} T^{2}+\frac{1}{3 n} a^{2} \hat{\gamma}^{2} T^{3}+\frac{1}{n} b \hat{\gamma} \tag{2}
\end{equation*}
$$

and

$$
\begin{equation*}
T_{2}=\frac{\sqrt{n}}{2 a \hat{\gamma}}\left(\operatorname{Exp}\left(\frac{2 a \hat{\gamma} T}{\sqrt{n}}\right)-1\right)+\frac{1}{n} b \hat{\gamma} \tag{3}
\end{equation*}
$$

where $\hat{\gamma}=\left\{\sum_{i=1}^{n}\left(X_{i}-\bar{X}\right)^{3} / n\right\} / S^{3}$ is an estimate of $\gamma$. Note that as the estimated skewness $\hat{\gamma} \rightarrow 0$ both $T_{1} \rightarrow 0$ and $T_{2} \rightarrow 0$. The test statistic $T_{1}$ is a direct consequence of the Edgeworth expansion of $T$ given below, see for example, A. DasGupta (2008) page 191.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

Theorem 1. Let $X_{1}, X_{2}, \ldots, X_{n}$ be iid with CDF $F$ having mean $\mu$, variance $\sigma^{2}$, and $E\left(X_{1}-\mu\right)^{4}<\infty$. Assume that $F$ satisfies Cramër's condition. Let $\gamma=E\left(X_{1}-\mu\right)^{3} / \sigma^{3}, \quad P_{1}(x)=\left(2 x^{2}+1\right) / 6, \quad$ then the CDF of $t$-statistic $T=\sqrt{n}(\bar{X}-\mu) / S$ admits the expansion

$$
\begin{equation*}
P(T \leq x)=\Phi(x)+\frac{\gamma P_{1}(x)}{\sqrt{n}} \phi(x)+O\left(n^{-1}\right) \tag{4}
\end{equation*}
$$

uniformly in $x$, where $\Phi(x)$ and $\phi(x)$ are standard normal distribution and density function, respectively.

From the above theorem, the skewness of the distribution $F$ has significant effect on $T$ especially when the sample size $n$ is small. One term Edgeworth expansion for $T$ is (see Hall 1987)

$$
\begin{equation*}
P(T \leq x)=\Phi(x)+\frac{\gamma}{6 \sqrt{n}}\left(2 x^{2}+1\right) \phi(x)+O\left(n^{-1}\right) \tag{5}
\end{equation*}
$$

From (5) a modified test statistic is

$$
\begin{equation*}
T_{0}=T+\frac{\hat{\gamma}}{3 \sqrt{n}} T^{2}+\frac{\hat{\gamma}}{6 \sqrt{n}}, \tag{6}
\end{equation*}
$$

which may be used to correct the skewness of $T$. One may use $a=1 / 3$ and $b=1 / 6$ in equations (2) and (3). As indicated by Hall (1992), $T_{0}$ is not a monotonic function and hence is not invertible to construct a confidence interval for $\mu$. Hall (1992) modified $T_{0}$ to $T_{1}$ as given in (2) so that it can be inverted to construct the confidence interval as well as to correct the bias and skewness. We are not sure why the last term of (2) and (3) is $(b \hat{\gamma}) / n$ rather than $(b \hat{\gamma}) / \sqrt{n}$. Zhou and Gao (2000) uses

$$
\begin{equation*}
T_{1}=T+\frac{\hat{\gamma}}{\sqrt{n}}\left(a T^{2}+b\right)+\frac{1}{3 n} a^{2} \hat{\gamma}^{2} T^{3} \tag{7}
\end{equation*}
$$

which is slightly different from Hall's $T_{1}$ and we will called (7) the $T_{1}$ from now on.

## LIM \& LIM

The simulation will justify the $P_{1}(x)$ in the one term of Edgeworth expansion of $T$ is indeed a polynomial function of order 2 . Consider the Edgeworth expansion of $T$,

$$
P\left(T \leq t_{\alpha}\right)=\Phi\left(t_{\alpha}\right)+\frac{\gamma P_{1}\left(t_{\alpha}\right)}{\sqrt{n}} \phi\left(t_{\alpha}\right)+O\left(n^{-1}\right)
$$

and one can show that

$$
\begin{equation*}
P(\text { miss left })=P\left(\mu<\bar{X}-t_{\alpha / 2} \frac{S}{\sqrt{n}}\right)=\frac{\alpha}{2}-\frac{\gamma}{\sqrt{n}} P_{1}\left(t_{\alpha / 2}\right) \phi\left(t_{\alpha / 2}\right)+O\left(n^{-1}\right) . \tag{8}
\end{equation*}
$$

Similarly,

$$
\begin{equation*}
P(\text { miss right })=P\left(\mu>\bar{X}+t_{\alpha / 2} \frac{S}{\sqrt{n}}\right)=\frac{\alpha}{2}+\frac{\gamma}{\sqrt{n}} P_{1}\left(-t_{\alpha / 2}\right) \phi\left(t_{\alpha / 2}\right)+O\left(n^{-1}\right) . \tag{9}
\end{equation*}
$$

If $P_{1}(t)>0$ for all $t$, one can see then, a positively skewed distribution $(\gamma>0)$ $P($ miss right $)>P($ miss left $)$. It can be seen in (9) that, with $\gamma>0$ and $P_{1}(t)>0$, $P($ miss right $)>\frac{\alpha}{2}$ and in equation (8) one obtains $P($ miss left $)<\frac{\alpha}{2}$. The opposite is true for a negatively skewed distribution. Let

$$
\begin{align*}
P(\text { miss }) & =P(\text { miss left })+P(\text { miss right }) \\
& =\alpha+\frac{\gamma}{\sqrt{n}}\left\{P_{1}\left(-t_{\alpha / 2}\right)-P_{1}\left(t_{\alpha / 2}\right)\right\} \phi\left(t_{\alpha / 2}\right)+O\left(n^{-1}\right) . \tag{10}
\end{align*}
$$

Let $k(\alpha)=(P($ miss $)-\alpha)$ and $g\left(t_{\alpha / 2}\right)=P_{1}\left(-t_{\alpha / 2}\right)-P_{1}\left(t_{\alpha / 2}\right)$. Rewrite Equation (10) as

$$
k(\alpha)=\frac{\gamma}{\sqrt{n}} g\left(t_{\alpha / 2}\right) \phi\left(t_{\alpha / 2}\right)+O\left(n^{-1}\right)
$$

and

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

$$
\frac{\sqrt{n} k(\alpha)}{\gamma \phi\left(t_{\alpha / 2}\right)} \approx g\left(t_{\alpha / 2}\right) .
$$

A plot of $\sqrt{n} k(\alpha) / \hat{\gamma} \phi\left(t_{\alpha / 2}\right)$ versus $t_{\alpha / 2}$ will review the structure of $P_{1}(x)$ if the random variable $X$ is skewed.


Figure 5. Plot of $\sqrt{n} k(\alpha) / \hat{\gamma} \phi\left(t_{\alpha / 2}\right)$ versus $t_{\alpha / 2}$ (a) $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Normal}(1,2)$ and (b) $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Laplace}(1,2)$.

Finding the structure of $g$ is a matter of regression. However, we are interested in whether the structure of $g$ agrees with the quadratic function given in Theorem 1. Note that $P_{1}(x)=2 x^{2}+1$ is an even function. Hence, $P\left(-t_{\alpha / 2}\right)-P_{1}\left(t_{\alpha / 2}\right)=0$. However, if $X$ is a right skewed distributed function, we have seen in Table 3 that $P$ (miss right) $>P($ miss left $)$; we expect $g\left(t_{\alpha / 2}\right)>0$ (see Figure 6(a)). Similarly, one can see that $g\left(t_{\alpha / 2}\right)<0$ when $X$ has left skewed distribution (see Figure 6(b)). If $X$ is a symmetrically distributed function with skewness $\gamma=0$, then the plot of $\sqrt{n} k(\alpha) / \hat{\gamma} \phi\left(t_{\alpha / 2}\right)$ versus $t_{\alpha / 2}$ does not show any pattern as seen in Figures 5 (a) and (b). Figures 6 (a) and (b) show that $g\left(t_{\alpha / 2}\right)$ resembles a quadratic function, confirming that the use of the second term in (5) is necessary if skewness appeared in the data. From the simulation and the

## LIM \& LIM

equation (10) one can see that if $X$ is skewed, then $P$ (miss) $>\alpha$; because $\left\{\frac{\hat{\gamma}}{\sqrt{n}} g\left(t_{\alpha / 2}\right) \phi\left(t_{\alpha / 2}\right)\right\}>0$. Thus, it explains that when $X$ is a skewed distribution the coverage error will be larger than the nominal coverage error $\alpha$.


Figure 6: Plot of $\sqrt{n} k(\alpha) / \hat{\gamma} \phi\left(t_{\alpha / 2}\right)$ versus $t_{\alpha / 2}$ (a) $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gamma}(1,2)$ and (b) $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gumbel}(1,2)$.

## Comparison of $\boldsymbol{T}, \boldsymbol{T}_{\mathbf{1}}$ and $\boldsymbol{T}_{\mathbf{2}}$

The objective is to compare the test statistics $T, T_{1}$, and $T_{2}$. The modified test statistic $T_{2}$ given in Hall (1992) has not been paid attention to as far as we know. As mentioned earlier, $\lim _{\hat{\gamma} \rightarrow 0} T_{2}=T$ and if $X$ is positively skewed, most likely $(\bar{X}-\mu)<0$. We modify the $T$ in $T_{2}$ by adding the term $b \bar{\gamma} / \sqrt{n}$ to $\bar{X}$ so that it shifts to the correct direction, i.e.,

$$
\begin{equation*}
T_{b}=\sqrt{n}\left(\bar{X}+\frac{b \gamma}{\sqrt{n}}-\mu\right) / S . \tag{11}
\end{equation*}
$$

Unlike $T_{2}$ in (3), our modified $T_{2}$ is

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

$$
\begin{equation*}
T_{2}^{*}=\frac{\sqrt{n}}{2 a \hat{\gamma}}\left(\operatorname{Exp}\left(\frac{2 a \hat{\gamma} T_{b}}{\sqrt{n}}\right)-1\right)+(b \hat{\gamma}) / \sqrt{n} . \tag{12}
\end{equation*}
$$

The simulation study is repeated on the four chosen distributions but this time we compare the empirical distributions of three test statistics, i.e., $T, T_{1}$ and $T_{2}^{*}$. Both $T_{1}$ and $T_{2}^{*}$ are less skewed than $T$ if $X$ is simulated from a skewed distribution, which are shown in Figures 9 and 10. Figures 7 and 8 show that if $X$ is a symmetric distribution, the distributions of $T_{1}$ and $T_{2}^{*}$ remain symmetric.

## Confidence Interval

A simulation study of confidence intervals derived from $T, T_{1}$ and $T_{2}^{*}$ was conducted. The $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ derived from $T_{1}$ is

$$
\begin{equation*}
\left(\bar{X}-\frac{S}{\sqrt{n}} h_{1}^{-1}\left(-t_{\alpha / 2}\right), \bar{X}+\frac{S}{\sqrt{n}} h_{1}^{-1}\left(t_{\alpha / 2}\right)\right), \tag{13}
\end{equation*}
$$

where $h_{1}^{-1}(t)=\sqrt{n}(a \hat{\gamma})^{-1}\left(1+\frac{3 a \hat{\gamma}}{\sqrt{n}}\left(t-\frac{b \hat{\gamma}}{\sqrt{n}}\right)\right)^{1 / 3}-\sqrt{n}(a \hat{\gamma})^{-1}$. One can see that $h_{1}^{-1}(t)$ may produce complex values for some $\hat{\gamma}$ and $t$. If the first 4 terms of $h_{1}^{-1}(t)$ are expanded and the expansion is simplified,

$$
\begin{equation*}
h_{1}^{-1}(t) \approx\left(t-\frac{b \hat{\gamma}}{\sqrt{n}}\right)-\frac{a \hat{\gamma}}{\sqrt{n}}\left(t-\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{2}+\frac{5}{3} \frac{(a \hat{\gamma})^{2}}{n}\left(t-\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{3}=h_{1^{*}}^{-1}(t) \tag{14}
\end{equation*}
$$

Then, replace $h_{1}^{-1}(t)$ by $h_{1^{*}}^{-1}(t)$ in (13), the approximation confidence interval of $T_{1}$, called $T_{1}^{*}$, will guarantee to produce a real valued confidence interval. The confidence interval of $T_{1}^{*}$ is

$$
\begin{equation*}
\left(\bar{X}-\frac{S}{\sqrt{n}} h_{1^{*}}^{-1}\left(-t_{\alpha / 2}\right), \bar{X}+\frac{S}{\sqrt{n}} h_{1^{*}}^{-1}\left(t_{\alpha / 2}\right)\right), \tag{15}
\end{equation*}
$$

LIM \& LIM


Figure 7. Left figure is the histogram of $T=\sqrt{n}(\bar{X}-\mu) / S, X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Normal}(1,2)$. The graph in the middle is the histogram of $T_{1}$ and the histogram on the right is the empirical distribution of $T_{2}^{*}$.


Figure 8. Left figure is the histogram of $T=\sqrt{n}(\bar{X}-\mu) / S, X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Laplace}(1,2)$. The graph in the middle is the histogram of $T_{1}$ and the histogram on the right is the empirical distribution of $T_{2}^{*}$.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION



Figure 9. Left figure is the histogram of $T=\sqrt{n}(\bar{X}-\mu) / S, X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gamma}(1,2)$. The graph in the middle is the histogram of $T_{1}$ and the histogram on the right is the empirical distribution of $T_{2}^{*}$.




Figure 10. Left figure is the histogram of $T=\sqrt{n}(\bar{X}-\mu) / S, X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gumbel}(1,2)$. The graph in the middle is the histogram of $T_{1}$ and the histogram on the right is the empirical distribution of $T_{2}^{*}$.

## LIM \& LIM

where $h_{1^{*}}^{-1}(t)$ is given in (14). The $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ derived from $T_{2}^{*}$ is

$$
\left[\begin{array}{l}
\left(\bar{X}+\frac{b \gamma}{\sqrt{n}}\right)-\frac{S}{\sqrt{n}}\left\{\frac{\sqrt{n}}{2 a \hat{\gamma}} \ln \left(1-\frac{2 a \hat{\gamma}}{\sqrt{n}}\left(t_{\alpha / 2}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)\right)\right\}, \\
\left(\bar{X}+\frac{b \gamma}{\sqrt{n}}\right)-\frac{S}{\sqrt{n}}\left\{\frac{\sqrt{n}}{2 a \hat{\gamma}} \ln \left(1+\frac{2 a \hat{\gamma}}{\sqrt{n}}\left(t_{\alpha / 2}-\frac{b \hat{\gamma}}{\sqrt{n}}\right)\right)\right\}
\end{array}\right] .
$$

It is not surprising that the logarithm function may produce a complex number. Expand the logarithm function and keep the first 3 terms of the Taylor series expansions, the approximation confidence interval for $\mu$ is

$$
\left[\begin{array}{l}
\left(\bar{X}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)-\frac{S}{\sqrt{n}}\left\{\left(t_{\alpha / 2}-\frac{b \hat{\gamma}}{\sqrt{n}}\right)-\frac{a \hat{\gamma}}{\sqrt{n}}\left(t_{\alpha / 2}-\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{2}+\frac{4}{3} \frac{(a \hat{\gamma})^{2}}{\sqrt{n}}\left(t_{\alpha / 2}-\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{3}\right\},  \tag{16}\\
\left(\bar{X}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)+\frac{S}{\sqrt{n}}\left\{\left(t_{\alpha / 2}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)+\frac{a \hat{\gamma}}{\sqrt{n}}\left(t_{\alpha / 2}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{2}+\frac{4}{3} \frac{(a \hat{\gamma})^{2}}{\sqrt{n}}\left(t_{\alpha / 2}+\frac{b \hat{\gamma}}{\sqrt{n}}\right)^{3}\right\}
\end{array}\right] .
$$

The above confidence interval (16) may be called a confidence interval from $T_{2}^{* *}$. The confidence interval in (16) is different from that of in (15). The confidence interval in (15) subtract $(b \hat{\gamma}) / \sqrt{n}$ from $t_{\alpha / 2}$ on upper and lower confidence limit. Unlike the confidence interval in (15), the confidence interval in (16) tends to subtract $(b \hat{\gamma}) / \sqrt{n}$ from $t_{\alpha / 2}$ on the lower confidence limit but add $(b \hat{\gamma}) / \sqrt{n}$ on the upper confidence limit.

It can be seen in Table 7 that if $X$ is severely skewed, the modified confidence intervals $T_{1}^{*}$ and $T_{2}^{* *}$ perform substantially better than the usual confidence interval derived from $T$. If the skewness is not severe, $T$ performs better than the modified $T$.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

Table 5. Table of nominal coverage error $\alpha$ and the simulated missed for $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Normal}(1,2)$.

| $n=10$ | $M=10000$ |  | Normal $(1,2)$ |  | skewness $=0$ |  |  |  | $a=1 / 3$ | $b=1 / 6$ |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |  |
| miss $T$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.049 | 0.06 | 0.069 | 0.079 | 0.089 | 0.101 |  |
| miss $T_{1}^{*}$ | 0.011 | 0.022 | 0.031 | 0.04 | 0.05 | 0.06 | 0.069 | 0.08 | 0.09 | 0.101 |  |
| miss $T_{2}^{* * *}$ | 0.013 | 0.025 | 0.035 | 0.045 | 0.055 | 0.064 | 0.074 | 0.084 | 0.094 | 0.104 |  |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |  |
| miss $T$ | 0.109 | 0.122 | 0.13 | 0.14 | 0.15 | 0.16 | 0.168 | 0.177 | 0.189 | 0.202 |  |
| miss $T_{1}^{*}$ | 0.109 | 0.12 | 0.128 | 0.138 | 0.149 | 0.158 | 0.167 | 0.176 | 0.187 | 0.2 |  |
| miss $T_{2}^{\text {me* }}$ | 0.114 | 0.125 | 0.132 | 0.142 | 0.152 | 0.162 | 0.171 | 0.181 | 0.191 | 0.203 |  |

Table 6. Table of nominal coverage error $\alpha$ and the simulated missed for $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Laplace}(1,2)$.

| $n=10$ | $M=10000$ |  | Laplace $(1,2)$ |  | skewness $=0$ |  | $a=1 / 3$, | $b=1 / 6$ |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| miss $T$ | 0.005 | 0.014 | 0.023 | 0.032 | 0.041 | 0.051 | 0.061 | 0.073 | 0.084 | 0.095 |
| miss $T_{1}^{*}$ | 0.01 | 0.023 | 0.037 | 0.049 | 0.061 | 0.074 | 0.086 | 0.1 | 0.113 | 0.125 |
| miss $T_{2}^{* *}$ | 0.015 | 0.032 | 0.049 | 0.064 | 0.076 | 0.092 | 0.104 | 0.119 | 0.132 | 0.144 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss $T$ | 0.105 | 0.116 | 0.127 | 0.14 | 0.151 | 0.161 | 0.171 | 0.182 | 0.194 | 0.202 |
| miss $T_{1}^{*}$ | 0.135 | 0.148 | 0.159 | 0.172 | 0.186 | 0.194 | 0.205 | 0.216 | 0.23 | 0.237 |
| miss $T_{2}^{* *}$ | 0.154 | 0.167 | 0.178 | 0.191 | 0.204 | 0.213 | 0.224 | 0.236 | 0.249 | 0.256 |

Table 7. Table of nominal coverage error $\alpha$ and the simulated missed for $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gamma}(1,2)$.

| $n=10$ | $M=10000$ | Gamma (1,2) |  |  |  |  |  |  | skewness $=2$ |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| miss $T$ | 0.047 | 0.062 | 0.077 | 0.089 | 0.1 | 0.109 | 0.119 | 0.13 | 0.139 | 0.148 |
| miss $T_{1}^{*}$ | 0.024 | 0.035 | 0.048 | 0.058 | 0.068 | 0.077 | 0.088 | 0.098 | 0.108 | 0.118 |
| miss $T_{2}^{* *}$ | 0.022 | 0.033 | 0.046 | 0.056 | 0.065 | 0.074 | 0.086 | 0.097 | 0.107 | 0.117 |
| $\alpha$ | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss $T$ | 0.156 | 0.166 | 0.176 | 0.186 | 0.19 | 0.198 | 0.206 | 0.22 | 0.229 | 0.238 |
| miss $T_{1}^{*}$ | 0.128 | 0.138 | 0.149 | 0.16 | 0.167 | 0.178 | 0.186 | 0.2 | 0.21 | 0.221 |
| miss $T_{2}^{* *}$ | 0.127 | 0.137 | 0.149 | 0.159 | 0.168 | 0.178 | 0.186 | 0.201 | 0.211 | 0.223 |

## LIM \& LIM

Table 8. Table of nominal coverage error $\alpha$ and the simulated missed for $(1-\alpha) \times 100 \%$ confidence interval for $\mu$ with $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gumbel}(1,2)$.

| $n=10$ | $M=10000$ |  | Gumbel (1,2) |  | skewness $\approx-1.14, \quad a=1 / 3, \quad b=1 / 6$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\alpha$ | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 |
| miss $T$ | 0.019 | 0.033 | 0.043 | 0.055 | 0.064 | 0.075 | 0.085 | 0.094 | 0.105 | 0.115 |
| miss $T_{1}^{*}$ | 0.015 | 0.027 | 0.037 | 0.048 | 0.058 | 0.068 | 0.079 | 0.088 | 0.099 | 0.109 |
| miss $T_{2}^{* *}$ | 0.017 | 0.028 | 0.039 | 0.050 | 0.061 | 0.071 | 0.081 | 0.091 | 0.101 | 0.111 |
| a | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.19 | 0.2 |
| miss $T$ | T 0.125 | 0.134 | 0.142 | 0.153 | 0.162 | 0.173 | 0.182 | 0.193 | 0.201 | 0.212 |
| miss $T_{1}^{*}$ | 0.12 | 0.129 | 0.138 | 0.149 | 0.158 | 0.169 | 0.18 | 0.19 | 0.199 | 0.208 |
| miss $T_{2}^{\text {+**}}$ | 0.121 | 0.133 | 0.141 | 0.152 | 0.163 | 0.174 | 0.182 | 0.191 | 0.200 | 0.210 |

## Hypothesis Testing

The three test statistics $T, T_{1}$ and $T_{2}^{*}$ are compared in terms of the power of their tests. A Computer Approach Technique (CAT), given in Pal, Lim and Ling (2007), will be used. For a normal distribution all three test statistics perform relatively well. However, $T_{1}$ and $T_{2}^{*}$ lost some power on the Laplace distribution, more on the $T_{1}$ than $T_{2}^{*}$. If $X$ is a positively skewed distribution, such as Gamma, $T_{1}$ and $T_{2}^{*}$ perform slightly better than $T$ on the right side of $\mu_{0}$ while $T$ performs better than other two on the left side of $\mu_{0}$. The opposite is true for negatively skewed distribution. In terms of modified test statistics, $T_{2}^{*}$ performs slightly better than $T_{1}$ from the point of view of power of the test. The simulation results for the power of the tests are summarized in Figures 11, 12, 13 and 14.

## T-TEST ON SKEWED DISTRIBUTION FUNCTION



Figure 11. Plot of $\eta$ versus power of the test for where $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Normal}(\eta, 2)$, with $n=10$. The hypothesis testing is $H_{0}: \eta=4(\mu=4)$ versus $H_{1}: \eta \neq 4(\mu \neq 4)$


Figure 12. Plot of $\eta$ versus power of the test for where $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Laplace}(\eta, 2)$, with $n=10$. The hypothesis testing is $H_{0}: \eta=4(\mu=4)$ versus $H_{1}: \eta \neq 4(\mu \neq 4)$

## LIM \& LIM



Figure 13. Plot of $\eta$ versus power of the test for where $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gamma}(\eta, 2)$, with $n=10$. The hypothesis testing is $H_{0}: \eta=4(\mu=8)$ versus $H_{1}: \eta \neq 4(\mu \neq 8)$


Figure 14. Plot of $\eta$ versus power of the test for where $X_{1}, X_{2}, \ldots, X_{n} \sim \operatorname{Gumbel}(\eta, 2)$, with $n=10$. The hypothesis testing is $H_{0}: \eta=4(\mu=2.84557)$ versus $H_{1}: \eta \neq 4(\mu \neq 2.84557)$

## Conclusion

Based on these results, it appears that the usual t -test statistic, $T$, is quite robust regardless of the skewness of the distribution. The modified t-test statistics $T_{1}$ and

## T-TEST ON SKEWED DISTRIBUTION FUNCTION

$T_{2}^{*}$ can improve the power on one side of the $\mu=\mu_{0}$ only, but not on both sides. From the results, it appears the modified confidence intervals perform much better than the usual confidence interval derived from $T$ when $X$ is simulated from a skewed distribution.

## Acknowledgements

The authors thank Eswar Phadia, Professor (Ret'd), William Paterson University of New Jersey for many helpful suggestions. Also, the authors thank the editor and referees for suggestions to improve the presentation of this article.

## References

Abramovitch, L. \& Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. The Annals of Statistics, 13(1), 116-132. doi: 10.1214/aos/1176346580

Blair, R. C. \& Sawilowsky, S. S. (1993). Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. Statistics in Medicine, 12(23), 2233-2243. doi: 10.1002/sim. 4780122308

Boos, D. D. \& Hughes-Oliver, J. M. (2000). How large does $n$ have to be for $Z$ and $t$ intervals. The American Statistician, 54(2), 121-128.
doi: 10.1080/00031305.2000.10474524
DasGupta, A. (2008). Asymptotic theory of statistics and probability. Springer Science \& Business Media.

Hall, P. (1987). Edgeworth expansion for student's $t$ statistic under minimal moment conditions. Annals of Probability, 15(3), 920-931.
doi: 10.1214/aop/1176992073
Hall, P. (1992). On the removal of skewness by transformation. Journal of Royal Statistical Society, Series B (Methodological), 54(1), 221-228.

Johnson, N. J. (1978). Modified $t$-test and confidence intervals for asymmetrical population. Journal of the American Statistical Association, 73(363), 536-544. doi: 10.1080/01621459.1978.10480051

Pal, N., Lim, W. K., \& Ling, C. (2007). A computational approach to statistical inferences. Journal of Applied Probability \& Statistics, 2(1), 13-35.

Sawilowsky, S. S. \& Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the $t$-test to departures from population

## LIM \& LIM

normality. Psychological Bulletin, 111(2), 352-360. doi: 10.1037/0033-2909.111.2.352

Zhou, X.-H. \& Gao S. (2000). One-sided confidence intervals for mean of positively skewed distributions. The American Statistician, 54(2), 100-104. doi: 10.1080/00031305.2000.10474519

# A New Estimator of the Population Mean: An Application to Bioleaching Studies 

Amer I. Al-Omari<br>Al al-Bayt University Mafraq, Jordan

Carlos N. Bouza<br>Universidad de La Habana<br>Havana, Cuba

Dante Covarrubias<br>Universidad Autónoma de<br>Guerrero, México

Roma Pal<br>Bhat-Sarkhar Eng. Design \& QC Advisors, India

The multistage balanced groups ranked set samples (MBGRSS) method is considered for estimating the population mean for samples of size $m=3 k$ where $k$ is a positive real integer. It is compared with the simple random sampling (SRS) and ranked set sampling (RSS) schemes. For the symmetric distributions considered in this study, the MBGRSS estimator is an unbiased estimator of the population mean and it is more efficient than SRS and RSS methods based on the same number of measured units. Its efficiency is increasing in $s$ for fixed value of the sample size, where $s$ is the number of stages. For non symmetric distributions considered in this paper, the MBGRSS estimator is biased. The method is applied in a study of bioleaching.

Keywords: Ranked set sampling, simple random sampling, multistage balanced groups, ranked set samples, symmetric and asymmetric distribution

## Introduction

Ranked set sampling is a sampling procedure, which is a less costly as compared to the widely used simple random sampling in cases where visual ranking of a set of observations can be easily done, while the exact measurement of observations is not easy and cost. The RSS mean was considered by McIntyre (1952) as an estimator of the population mean. The RSS mean estimator was considered more efficient than the SRS counterpart.

Takahasi and Wakimoto (1968) introduced the mathematical theory of ranked set sampling. Al-Saleh and Al-Kadiri (2000) suggested double RSS method in order to estimate the population mean. Al-Saleh and Al-Omari (2002) suggested multistage RSS method to increase the efficiency of estimating the

[^5]
## AL-OMARI ET AL.

mean for fixed value of the sample size. Jemain and Al-Omari (2006a, 2006b) considered double percentile RSS and multistage median RSS methods, respectively, for the mean estimation. They found that both methods are more efficient than the SRS based on the same sample size.

Jemain, Al-Omari, and Ibrahim (2008) investigated balanced groups RSS method for estimating the population mean. Jemain, Al-Omari, and Ibrahim (2007) suggested multistage extreme ranked set sampling method for estimating the population mean. Al-Hadhrami and Al-Omari (2009) considered the Bayesian inference of the variance of the normal distribution using moving extreme ranked set sampling. Ozturk (2011) used the RSS for parametric inference about the parameters of the location-scale family of distributions. Dong and Cui (2011) investigated the optimal sign test for quantiles in ranked set samples. Al-Omari, Ibrahim, Jemain, and Al-Hadhrami (2009) proposed multistage balanced groups ranked set samples for estimating the population median. For more details about RSS see Herrera and Al-Omari (2011), Al-Omari (2011), Vock and Balakrishnan (2011), and Drikvandi, Modarres, and Jalilian (2011).

Let $X_{1}, X_{2}, \ldots, X_{m}$ be a SRS of size $m$ from cdf $F(x)$. The $i^{\text {th }}$ order statistic $X_{(i: m)}$ has the probability density function (pdf) and the cumulative distribution function (cdf), $f_{(i: m)}(x)$ and $F_{(i: m)}(x)$, respectively, given by

$$
\begin{gather*}
f_{(i: m)}(x)=\frac{1}{B[i, m-i+1]}[F(x)]^{i-1}[1-F(x)]^{m-i} f(x)  \tag{1}\\
F_{(i: m)}(x)=\int_{-\infty}^{x} \frac{1}{B[i, m-i+1]}[F(t)]^{i-1}[1-f(t)]^{m-i} f(t) d t, \quad-\infty<x<\infty \tag{2}
\end{gather*}
$$

where $B[\alpha, \beta]=\int_{0}^{1} t^{\alpha-1}[1-t]^{\beta-1} d t$ is the complete beta function. The mean and the variance of $X_{(i: m)}$ are given by $\mu_{(i: m)}=\int_{-\infty}^{\infty} x f_{(i: m)}(x) d x \quad$ and $\sigma_{(i: m)}^{2}=\int_{-\infty}^{\infty}\left[x-\mu_{(i: m)}\right]^{2} f_{(i: m)}(x) d x$, respectively.

## A NEW ESTIMATOR OF THE POPULATION MEAN

## Multistage Balanced Groups Ranked Set Samples

The RSS can be described as: randomly select $m^{2}$ units from the target population. Allocate these units into $m$ sets, each of size $m$. Rank the $m$ units within each set visually or by any cheap method with respect to the characteristic of interest. From the $i^{\text {th }}$ set select the $i^{\text {th }}$ ranked unit for $i=1,2, \ldots, m$. The process can be repeated $n$ cycles to obtain a set of size $m n$ from the initial $m^{2} n$ units.

The MBGRSS as suggested by Al-Omari et al. (2009) consists from the following steps:

Step 1: Randomly select $(3 k)^{s+1}$ for $k=1,2,3, \ldots$ units from the target population, and then allocate them into $(3 k)^{s}$ sets, each of size $3 k$.

Step 2: The $3 k$ units of each set are ranked based on professional judgment or by any cheap method in terms of the variable of interest. Then the $(3 k)^{s}$ sets are divided into three groups, each of $3^{s-1} k^{s}$ sets.

Step 3: From each set in the first group, the smallest ranked unit is selected; from each set in the second group; the median ranked unit is selected, and from each set in the third group, the largest ranked unit is selected. This step yields $(3 k)^{s-1}$ sets, $3^{s-2} k^{s-1}$ sets in each group.

Step 4: Without doing any actual measurement, from the $3^{s-2} k^{s-1}$ sets in the first group the smallest ranked unit is selected, from the $3^{s-2} k^{s-1}$ sets in the second group the median ranked unit is selected, and from the $3^{s-2} k^{s-1}$ sets in the third group the largest ranked unit is selected. This step yields $(3 k)^{s-2}$ sets, each group of $3^{s-3} k^{s-2}$ sets of size $3 k$.

Step 5: The process is continued using Steps (3) and (4) until we end up with one $s^{\text {th }}$ stage balanced groups RSS of size $3 k$.

The procedure can be repeated $n$ times if needed to obtain a sample of size $3 k n$ from the initial $(3 k)^{s+1} n$ units.

Al-Omari et al. (2009) introduced an example to illustrate the MBGRSS when $m=3$. In this paper we will illustrate the MBGRSS in estimating the population mean using $m=9$.

## AL-OMARI ET AL.

## Example

Let $s=3$ and $k=3$, then $m=9$. Therefore, we have to select 6561 units, say $X_{1}, X_{2}, \ldots, X_{6561}$. Allocate the 6561 selected units into 729 sets each of size 9 . The 9 observations of each set are ranked with respect to the study variable as follows: $\left\{X_{i(1: 9)}, X_{i(2: 9)}, \ldots, X_{i(9: 9}\right\}$, for $i=1,2, \ldots, 729$. Now, allocate the 729 sets into 3 groups, each of 243 sets as:

$$
\begin{array}{ll}
\text { 1st Group: } & \left\{X_{i(1: 99}, X_{i(2: 9)}, \ldots, X_{i(9: 9}\right\}, \text { for } i=1,2, \ldots, 243, \\
\text { 2nd Group: } & \left\{X_{i(1: 9),}, X_{i(2: 9)}, \ldots, X_{i(9: 9\}} \text {, for } i=244,245, \ldots, 486,\right. \\
\text { 3rd Group: } & \left\{X_{i(1: 9)}, X_{i(2: 9)}, \ldots, X_{i(9: 9}\right\}, \text { for } i=487,488, \ldots, 729 .
\end{array}
$$

For $s=1$, select the smallest ranked unit, $X_{i(19) 9}^{(1)}$ for $i=1,2, \ldots, 243$ from each set in the first group, and the median ranked unit, $X_{i(5.9)}^{(1)}$ for $i=244,245, \ldots, 486$ from each set in the second group, and finally, the largest ranked unit, $X_{i(9: 9)}^{(1)}$ for $i=487,488, \ldots, 729$ from each set in the third group. This step yields 729 units, which are $X_{1(199)}^{(1)}, X_{2(199)}^{(1)}, \ldots, X_{243(1: 9)}^{(1)}, X_{244(5: 9)}^{(1)}, X_{245(5: 9)}^{(1)}, \ldots$, $X_{486(5: 9)}^{(1)}, X_{487(9: 9)}^{(1)}, X_{488(9: 9)}^{(1)}, \ldots, X_{729(9: 9)}^{(1)}$. Allocate these units into 81 sets, 27 sets in each group as follows:

1st Group: $\quad\left\{X_{9(i-1)+1(1: 9)}^{(1)}, X_{9(i-1)+2(1: 9)}^{(1)}, \ldots, X_{9(i-1)+9(1: 9)}^{(1)}\right\}$, for $i=1,2, \ldots 27$,
2nd Group: $\quad\left\{X_{9(i-1)+1(5: 9)}^{(1)}, X_{9(i-1)+2(5: 9)}^{(1)}, \ldots, X_{9(i-1)+9(5: 9)}^{(1)}\right\}$, for $i=28,29, \ldots, 54$,
3rd Group: $\quad\left\{X_{9(i-1)+1(9 \cdot 9)}^{(1)}, X_{9(i-1)+2(9 \cdot 9)}^{(1)}, \ldots, X_{9(i-1)+9(9 \cdot 9)}^{(1)}\right\}$, for $i=55,56, \ldots, 81$.

Now, for $s=2$, rank the units within each set in all the three groups and then select the smallest ranked unit, $X_{i(1 \cdot 9)}^{(2)}$ for $i=1,2, \ldots 27$ from each set in the 1 st group, and the median ranked unit, $X_{i(5: 9)}^{(2)}$ for $i=28,29, \ldots, 54$ from each set in the 2nd group, and the largest ranked unit, $X_{i(9 \cdot 9)}^{(2)}$ for $i=55,56, \ldots, 81$ from each set in the 3 rd group. This step yields 81 units, which are $X_{1(199)}^{(2)}, X_{2(19)}^{(2)}, \ldots$, $X_{27(199)}^{(2)}, X_{28(5: 9)}^{(2)}, X_{29(5: 9)}^{(2)}, \ldots, X_{54(5: 9)}^{(2)}, X_{55(9: 9)}^{(2)}, X_{56(9: 9)}^{(2)}, \ldots, X_{81(9: 9)}^{(2)}$. Allocate them into 9 sets, 3 sets in each group as follows:

## A NEW ESTIMATOR OF THE POPULATION MEAN

1st Group: $\quad\left\{X_{9(i-1)+1(1: 9)}^{(2)}, X_{9(i-1)+2(1: 9)}^{(2)}, \ldots, X_{9(i-1)+9(1: 9)}^{(2)}\right\}$, for $i=1,2,3$,
2nd Group: $\quad\left\{X_{9(i-1)+1(5: 9)}^{(2)}, X_{9(i-1)+2(5: 9)}^{(2)}, \ldots, X_{9(i-1)+9(5: 9)}^{(2)}\right\}$, for $i=4,5,6$,
3rd Group: $\quad\left\{X_{9(i-1)+(9 \cdot 9)}^{(2)}, X_{9(i-1)+2(9 \cdot 9)}^{(2)}, \ldots, X_{9(i-1)+9(9 \cdot 9)}^{(2)}\right\}$, for $i=7,8,9$.

Next, for $s=3$ rank the units within each set in each group, then select the smallest ranked unit, $X_{i(1.9)}^{(3)}$ for $i=1,2,3$ from each set in the 1st group, the median ranked unit, $X_{i(5: 9)}^{(3)}$ for $i=4,5,6$ from each set in the 2 nd group, and the largest ranked unit, $X_{i(9,9)}^{(3)}$ for $i=7,8,9$ from each set in the 3rd group. This step yields 9 units, which are $X_{1(1: 9)}^{(3)}, X_{2(1: 9)}^{(3)}, X_{3(1: 9)}^{(3)}, X_{4(5: 9)}^{(3)}, X_{5(5: 9)}^{(3)}, X_{6(5: 9)}^{(3)}, X_{7(9: 9)}^{(3)}, X_{8(9: 9)}^{(3)}$, $X_{9(99)}^{(3)}$ to be a MBGRSS of size 9 . The mean of these units is considered as an estimator of the population mean.

It is of interest to note here that the RSS and the MBGRSS are equivalent when $m=3$ for $s=1$.

## Estimation of the Population Mean

Assume that $X_{1}, X_{2}, \ldots, X_{m}$ is a random sample from the cdf $F(x)$ with a finite mean $\mu$ and variance $\sigma^{2}$ Also, assume that $X_{11 h}, X_{12 h}, \ldots, X_{1 m h} ; X_{21 h}, X_{22 h}, \ldots, X_{2 m h}$; $X_{m 1 h}, X_{m 2 h}, \ldots, X_{m m h}$ are $m$ independent SRS of size $m$ each in the $h^{\text {th }}$ cycle for $h=1,2, \ldots, n$. If $X_{i(1: m) h}, X_{i(2: m) h}, \ldots, X_{i(m: m) h}$ are the order statistics of the $i^{\text {th }}$ sample $X_{i 1 h}, X_{i 2 h}, \ldots, X_{i m h}$, for $i=1,2, \ldots, m$. Then, the measured RSS units are $X_{1(1: m) h}$, $X_{2(2: m) h}, \ldots, X_{m(m: m) h}$.

The SRS estimator of the population mean based on a sample of size $m$ is defined as

$$
\begin{equation*}
\bar{X}_{S R S}=\frac{1}{m n} \sum_{h=1}^{n} \sum_{i=1}^{m} X_{i h}, \tag{3}
\end{equation*}
$$

with variance

$$
\begin{equation*}
\operatorname{Var}\left(\bar{X}_{S R S}\right)=\frac{\sigma^{2}}{m n} . \tag{4}
\end{equation*}
$$

## AL-OMARI ET AL.

The RSS estimator of the population mean (see McIntyre (1952)) is given by

$$
\begin{equation*}
\bar{X}_{R S S}=\frac{1}{m n} \sum_{h=1}^{n} \sum_{i=1}^{m} X_{i(i: m) h}, \tag{5}
\end{equation*}
$$

with variance

$$
\begin{equation*}
\operatorname{Var}\left(\bar{X}_{R S S}\right)=\frac{1}{m^{2} n} \sum_{h=1}^{n} \sum_{i=1}^{m} \operatorname{Var}\left(X_{i(i: m) h}\right)=\frac{\sigma^{2}}{m n}-\frac{1}{m^{2} n} \sum_{i=1}^{m}\left[\mu_{(i: m)}-\mu\right]^{2} \tag{6}
\end{equation*}
$$

If $m$ is odd, in the $h^{\text {th }}$ cycle $(h=1,2, \ldots, n)$, let $X_{i(1: m) h}^{(s)}$ be the smallest ranked observation of the $i^{\text {th }}$ sample for $i=1,2, \ldots, k, X_{i\left(\frac{m+1}{2}: m\right) h}^{(s)}$ be the median ranked observation of the $i^{\text {th }}$ sample for $i=k+1, k+2, \ldots, 2 k$, and $X_{i(m \cdot m) h}^{(s)}$ be the largest ranked observation of the $i^{\text {th }}$ sample when $i=2 k+1,2 k+2, \ldots, 3 k$. Therefore, when $m$ is odd, the measured units $X_{1(1: m) h}^{(s)}, X_{2(1: m) h}^{(s)}, \ldots X_{k(1: m) h}^{(s)}, X_{k+1\left(\frac{m+1}{2}: m\right) h}^{(s)}, \ldots$, $X_{2 k\left(\frac{m+1}{2}: m\right) h}^{(s)}, X_{2 k+1(m: m) h}^{(s)}, \ldots, X_{3 k(m: m) h}^{(s)}$ will be denoted by MBGRSSO. It is of interest to mention here that the measured units within each group are identically independent (iid) but all units are independent but not identically distributed.

The suggested estimator of the population mean based on MBGRSSO is given by

$$
\begin{equation*}
\bar{X}_{\text {MBGRSSO }}^{(s)}=\frac{1}{3 k n} \sum_{h=1}^{n}\left(\sum_{i=1}^{k} X_{i(1: m) h}^{(s)}+\sum_{i=k+1}^{2 k} X_{i\left(\frac{m+1}{2}: m\right)^{h}}^{(s)}+\sum_{i=2 k+1}^{3 k} X_{i(m: m) h}^{(s)}\right) \tag{7}
\end{equation*}
$$

with variance

$$
\begin{equation*}
\operatorname{Var}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}\right)=\frac{1}{9 k^{2} n}\left[\sum_{i=1}^{k} \operatorname{Var}\left(X_{i(1: m)}^{(s)}\right)+\sum_{i=k+1}^{2 k} \operatorname{Var}\left(X_{i\left(\frac{m+1}{2}: m\right.}^{(s)}\right)+\sum_{i=2 k+1}^{3 k} \operatorname{Var}\left(X_{i(m: m)}^{(s)}\right)\right] \tag{8}
\end{equation*}
$$

## A NEW ESTIMATOR OF THE POPULATION MEAN

For even sample size, let $X_{i(1: m) h}^{(s)}$ be the smallest ranked observation of the $i^{\text {th }}$ sample for $i=1,2, \ldots, k, \frac{1}{2}\left(X_{i\left(\frac{m}{2}: m\right) h}^{(s)}+X_{i\left(\frac{m+2}{2}: m\right) h}^{(s)}\right)$ be the median ranked observation of the $i^{\text {th }}$ sample for $i=k+1, k+2, \ldots, 2 k$, and $X_{i(m: m) h}^{(s)}$ be the largest ranked observation of the $i^{\text {th }}$ sample for $i=2 k+1,2 k+2, \ldots, 3 k$. However, the measured observations $\quad X_{1(1: m) h}^{(s)}, X_{2(1: m) h}^{(s)}, \ldots, \quad X_{k(1: m) h}^{(s)}, \frac{1}{2}\left(X_{k+1\left(\frac{m}{2}: m\right) h}^{(s)}+X_{k+1\left(\frac{m+2}{2}: m\right) h}^{(s)}\right), \ldots$, $\frac{1}{2}\left(X_{2 k\left(\frac{m}{2} \cdot m\right) h}^{(s)}+X_{2 k\left(\frac{m+2}{2}: m\right) h}^{(s)}\right), X_{2 k+1(m: m) h}^{(s)}, \ldots, X_{3 k(m: m) h}^{(s)}$ will be denoted as MBGRSSE. The suggested MBGRSSE estimator of the population mean is defined as

$$
\begin{equation*}
\bar{X}_{M B G R S E}^{(s)}=\frac{1}{3 k n} \sum_{h=1}^{n}\left\{\sum_{i=1}^{k} X_{i(1: m) h}^{(s)}+\sum_{i=k+1}^{2 k}\left[\frac{1}{2}\left(X_{i\left(\frac{m}{2}: m\right)^{h}}^{(s)}+X_{i\left(\frac{m+2}{2}: m\right)^{h}}^{(s)}\right)\right]+\sum_{i=2 k+1}^{3 k} X_{i(m: m) h}^{(s)}\right\} \tag{9}
\end{equation*}
$$

with variance

$$
\left.\operatorname{Var}\left(\bar{X}_{M B G R S S E}^{(s)}\right)=\frac{1}{9 k^{2} n}\left\{\begin{array}{l}
\frac{1}{4} \sum_{i=k+1}^{2 k}\left[\begin{array}{l}
\operatorname{Var}\left(X_{i\left(\frac{m}{2}: m\right.}^{(s)}\right.
\end{array}\right)+\operatorname{Var}\left(X_{i\left(\frac{m+2}{2}: m\right)}^{(s)}\right)  \tag{10}\\
\left.+2 \operatorname{Cov}\left(X_{i\left(\frac{m}{2}: m\right)}^{(s)}\right) X_{i\left(\frac{m+2}{2}: m\right)}^{(s)}\right)
\end{array}\right]\right\}
$$

Define the following notations. For $i=1,2, \ldots, m$ in the $h^{\text {th }}$ cycle, $h=1,2, \ldots, n, \quad$ let $\quad \mu_{(j: m)}^{(s)}=E\left(X_{i(j: m) h}^{(s)}\right), \quad \sigma_{(j: m)}^{2(s)}=\operatorname{Var}\left(X_{i(j: m) h}^{(s)}\right), \quad$ where $j=1, \frac{m}{2}, \frac{m+2}{2}, \frac{m+1}{2}, m$. Whether the sample size is even or odd the measured units $X_{1(1: m) h}^{(s)}, X_{2(1: m) h}^{(s)}, \ldots, X_{k(1: m) h}^{(s)}$ are iid, and also $X_{2 k+1(m: m) h}^{(s)}, X_{2 k+2(m: m) h}^{(s)}, \ldots$, $X_{3 k(m: m) h}^{(s)}$ are iid. Also, when the sample size is odd, $X_{k+1\left(\frac{m+1}{2}: m\right) h}^{(s)}, X_{k+2\left(\frac{m+1}{2}: m\right) h}^{(s)}, \ldots$, $X_{2 k\left(\frac{m+1}{2}: m\right) h}^{(s)}$ are iid. Hence, Equations (8) and (10), respectively, can be written as

## AL-OMARI ET AL.

$$
\begin{gather*}
\operatorname{Var}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}\right)=\frac{1}{9 k n}\left(\sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}+\sigma_{(m: m)}^{2(s)}\right)  \tag{11}\\
\operatorname{Var}\left(\bar{X}_{\text {MBGRSSE }}^{(s)}\right)=\frac{1}{9 k n}\left[\sigma_{(1: m)}^{2(s)}+\frac{1}{4}\left(\sigma_{\left(\frac{m}{2}: m\right)}^{2(s)}+\sigma_{\left(\frac{m+2}{2}: m\right)}^{2(s)}+2 \sigma_{\left(\frac{m}{2}: m\right),\left(\frac{m+2}{2}: m\right)}^{(s)}\right)+\sigma_{(m: m)}^{2(s)}\right] \tag{12}
\end{gather*}
$$

If the parent distribution is symmetric about its mean $\mu$, then $X_{(i: m)}^{(s)}=X_{(m-i+1: m)}^{(s)}$ in distribution and then $\operatorname{Var}\left(X_{(i: m)}^{(s)}\right)=\operatorname{Var}\left(X_{(m-i+1: m)}^{(s)}\right)$ for $i=1,2, \ldots, m$ (David \& Nagaraja, 2003). Therefore, we have

$$
\begin{equation*}
\operatorname{Var}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}\right)=\frac{1}{9 k n}\left(2 \sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}\right) \tag{13}
\end{equation*}
$$

and

$$
\begin{equation*}
\operatorname{Var}\left(\bar{X}_{\text {MBGRSSE }}^{(s)}\right)=\frac{1}{9 k n}\left[2 \sigma_{(1: m)}^{2(s)}+\frac{1}{2}\left(\sigma_{\left(\frac{m}{2}: m\right)}^{2(s)}+\sigma_{\left(\frac{m}{2}: m\right) \cdot\left(\frac{m+2}{2}: m\right)}^{(s)}\right)\right] \tag{14}
\end{equation*}
$$

Lemma 3.1. If the population of study is symmetric about its mean $\mu$, then $\bar{X}_{\text {MBGRSSO }}^{(s)}$ and $\bar{X}_{\text {MBGRSSE }}^{(s)}$ are unbiased estimators of the population mean.

Proof: When the sample size is odd, the expectation of (7) is

$$
\begin{aligned}
E\left(\bar{X}_{M B G S S O}^{(s)}\right) & =\frac{1}{3 k n} \sum_{h=1}^{n}\left[\sum_{i=1}^{k} E\left(X_{i(1: m) h}^{(s)}\right)+\sum_{i=k+1}^{2 k} E\left(X_{i\left(\frac{m+1}{2}: m\right)^{h}}^{(s)}\right)+\sum_{i=2 k+1}^{3 k} E\left(X_{i(m: m) h}^{(s)}\right)\right] \\
& =\frac{1}{3 k n} \sum_{h=1}^{n}\left(\sum_{i=1}^{k} \mu_{(1: m) h}^{(s)}+\sum_{i=k+1}^{2 k} \mu_{\left(\frac{m+1}{2}: m\right)^{h}}^{(s)}+\sum_{i=2 k+1}^{3 k} \mu_{(m: m) h}^{(s)}\right) \\
& =\frac{1}{3}\left(\mu_{(1: m)}^{(s)}+\mu_{\left(\frac{m+1}{2}: m\right)}^{(s)}+\mu_{(m: m)}^{(s)}\right)
\end{aligned}
$$

## A NEW ESTIMATOR OF THE POPULATION MEAN

Because the distribution is symmetric about $\mu$, then we have $\mu_{(1: m)}^{(s)}+\mu_{(m: m)}^{(s)}=2 \mu$ and $\mu_{\left(\frac{m m}{2}: m\right)}^{(s)}=\mu$. Therefore, $E\left(\bar{X}_{M B G R S S O}^{(s)}\right)=\frac{1}{3}(2 \mu+\mu)=\mu$.
Also, the expectation of (9) is

$$
\begin{aligned}
E\left(\bar{X}_{M B G R S S E}^{(s)}\right) & =\frac{1}{3 k n} \sum_{h=1}^{n}\left\{\sum_{i=1}^{k} E\left(X_{i(1: m) h}^{(s)}\right)+\frac{1}{2} \sum_{i=k+1}^{2 k}\left[E\left(X_{i\left(\frac{m}{2}: m\right)^{\prime} h}^{(s)}\right)+E\left(X_{i\left(\frac{m+2}{2}: m\right) h}^{(s)}\right)\right]+\sum_{i=2 k+1}^{3 k} E\left(X_{i(m: m) h}^{s s}\right)\right\} \\
& =\frac{1}{3 k n} \sum_{h=1}^{n}\left[\sum_{i=1}^{k} \mu_{(1: m) h}^{(s)}+\sum_{i=k+1}^{2 k}\left[\frac{1}{2}\left(\mu_{\left(\frac{m}{2}: m\right)^{(s)}}^{(s)}+\mu_{\left.\left(\frac{m+2}{2}: m\right)^{(s)}\right)}^{(s)}\right)\right]+\sum_{i=2 k+1}^{3 k} \mu_{(m: m) h}^{(s)}\right] \\
& =\frac{1}{3}\left[\mu_{(1: m)}^{(s)}+\frac{1}{2}\left(\mu_{\left(\frac{m}{2}: m\right)}^{(s)}+\mu_{\left(\frac{m+2}{2}: m\right)}^{(s)}\right)+\mu_{(m: m)}^{(s)}\right] \\
& =\frac{1}{3}\left[\left(\mu_{(1: m)}^{(s)}+\mu_{(m: m)}^{(s)}\right)+\frac{1}{2}\left(\mu_{\left(\frac{m}{2}: m\right)}^{(s)}+\mu_{\left(\frac{m+2}{2}: m\right)}^{(s)}\right)\right]
\end{aligned}
$$

Because the distribution is symmetric about $\mu$, then we have $\mu_{(1: m)}^{(s)}+\mu_{(m: m)}^{(s)}=2 \mu$ and $\mu_{\left(\frac{m}{2}: m\right)}^{(s)}+\mu_{\left(\frac{m+2}{2}: m\right)}^{(s)}=2 \mu$. Therefore, $E\left(\bar{X}_{\text {MBGRSSE }}^{(s)}\right)=\frac{1}{3}\left[2 \mu+\frac{1}{2}(2 \mu)\right]=\mu$.

## Theorem 3.2:

1) $\quad \bar{X}_{\text {MBGRSSO }}^{(s)}$ is more efficient than $\bar{X}_{S R S}$ if $2 \sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}<3 \sigma^{2}$.
2) $\bar{X}_{\text {MBGRSSE }}^{(s)}$ is more efficient than $\bar{X}_{\text {SRS }}$ if $4 \sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m}{2}: m\right)}^{2(s)}+\sigma_{\left(\frac{m}{2}: m\right)\left(\frac{m+2}{2}: m\right)}^{(s)}<6 \sigma^{2}$.

Proof: The proof is directly using the MSE equations of the MBGRSS estimators with that of SRS method.

## AL-OMARI ET AL.

## Simulation Study

The suggested MBGRSS estimators of the population mean will be compared with their competitors using RSS and SRS schemes. Six probability distribution functions are investigated for the populations: uniform, normal, beta, exponential, gamma and Weibull. The averages of 60,000 samples estimates using $k=1,2,3$ corresponding to the sample sizes $m=3,6,9$ are compared. Assume that the cycle is repeated once. The efficiency of RSS relative to SRS is defined as

$$
\begin{equation*}
e f f\left(\bar{X}_{R S S}, \bar{X}_{S R S}\right)=\frac{\operatorname{Var}\left(\bar{X}_{S R S}\right)}{\operatorname{Var}\left(\bar{X}_{R S S}\right)}=1-\frac{\sum_{i=1}^{m}\left[\mu_{(i: m)}-\mu\right]^{2}}{m \sigma^{2}} \tag{15}
\end{equation*}
$$

If the distribution is symmetric, the efficiency of MBGRSSO and MBGRSSE relative to SRS are defined as:

$$
\begin{align*}
\operatorname{eff}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}, \bar{X}_{\text {SRS }}\right)= & \frac{\operatorname{Var}\left(\bar{X}_{\text {SRS }}\right)}{\operatorname{Var}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}\right)}=\frac{3 \sigma^{2}}{\sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}+\sigma_{(m: m)}^{2(s)}} \\
\operatorname{eff}\left(\bar{X}_{\text {MBGRSSE }}^{(s)}, \bar{X}_{\text {SRS }}\right)= & \frac{\operatorname{Var}\left(\bar{X}_{\text {SRS }}\right)}{\operatorname{Var}\left(\bar{X}_{\text {MBGRSSE }}^{(s)}\right)} \\
& =\frac{3 \sigma^{2}}{\left.\sigma_{(1: m)}^{2(s)}+\frac{1}{4}\left[\sigma_{\left(\frac{m}{2}: m\right)}^{2(s)}+\sigma_{\left(\frac{m+2}{2}: m\right)}^{2(s)}+2 \sigma_{\left(\frac{m}{2}: m\right)}^{(s)}\right)\left(\frac{m+2: m}{2}\right)\right]+\sigma_{(m: m)}^{2(s)}} . \tag{16}
\end{align*}
$$

The mean square errors of $\bar{X}_{\text {MBGRSSO }}^{(s)}$ and $\bar{X}_{\text {MBGRSSE }}^{(s)}$ are defined as

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{X}_{M B G R S S O}^{(s)}\right)=\frac{1}{9 k n}\left(\sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}+\sigma_{(m: m)}^{2(s)}\right)+\left[E\left(\bar{X}_{M B G R S S O}^{(s)}\right)-\mu\right]^{2} \tag{17}
\end{equation*}
$$

and

## A NEW ESTIMATOR OF THE POPULATION MEAN

$$
\left.\left.\begin{array}{rl}
\operatorname{MSE}\left(\bar{X}_{M B G R S S E}^{(s)}\right)= & \frac{1}{9 k n}\left[\sigma_{(1: m)}^{2(s)}+\frac{1}{4}\left(\sigma_{\left(\frac{m}{2}: m\right)}^{2(r)}+\sigma_{\left(\frac{m+2}{2}: m\right)}^{2(s)}+2 \sigma_{\left(\frac{m}{2}: m\right)}^{(s)}\right)\left(\frac{m+2}{2}: m\right)\right. \tag{18}
\end{array}\right)+\sigma_{(m: m)}^{2(s)}\right] \quad \text {. }
$$

If the distribution is asymmetric, the efficiency is defined as:

$$
\begin{align*}
\operatorname{eff}\left(\bar{X}_{\text {MBGRSSO }}^{(s)}, \bar{X}_{\text {SRS }}\right) & =\frac{\operatorname{Var}\left(\bar{X}_{\text {SRS }}\right)}{\operatorname{MSE}\left(\bar{X}_{M B G R S S O}^{(s)}\right)} \\
& =\frac{3 \sigma^{2}}{\sigma_{(1: m)}^{2(s)}+\sigma_{\left(\frac{m+1}{2}: m\right)}^{2(s)}+\sigma_{(m: m)}^{2(s)}+9 k n\left[E\left(\bar{X}_{\text {MBGRSSO}}^{(s)}\right)-\mu\right]^{2}}, \tag{19}
\end{align*}
$$

and

$$
\begin{align*}
\operatorname{eff}\left(\bar{X}_{\text {MBGRSSE }}^{(s)}, \bar{X}_{\text {SRS }}\right) & =\frac{\operatorname{Var}\left(\bar{X}_{\text {SRS }}\right)}{\operatorname{MSE}\left(\bar{X}_{\text {MBGRSE }}^{(s)}\right)} \\
& =\frac{3 \sigma^{2}}{\sigma_{(1: m)}^{2(s)}+\frac{1}{4}\left[\begin{array}{c}
\left.\sigma_{\left(\frac{m}{2}: m\right)}^{2(s)}\right)+\sigma_{\left(\frac{m+2}{2}: m\right)}^{2(s)} \\
\left.+2 \sigma_{\left(\frac{m}{2}: m\right)}^{(s)}\right)\left(\frac{m+2: m}{2}: m\right)
\end{array}\right]+\sigma_{(m: m)}^{2(s)}+9 k n\left[E\left(\bar{X}_{\text {MBGRSSE }}^{(s)}\right)-\mu\right]^{2}} . \tag{20}
\end{align*}
$$

In terms of the efficiency and bias values, the results are summarized in Tables $1-3$ with $m=3,6,9$, respectively for several values of $s$.

## AL-OMARI ET AL.

Table 1. The efficiency of RSS and MBGRSSO for estimating the population mean with $m=3$ and $1 \leq s \leq 5$

| Distribution | RSS | MBGRSSO |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $s=1$ | $s=2$ | $s=3$ | $s=4$ | $s=5$ |
| Uniform (0,1) | 2.000 | Eff | 2.000 | 5.746 | 16.148 | 38.257 | 89.134 |
| Normal (0,1) | 1.914 | Eff | 1.914 | 3.288 | 5.010 | 6.937 | 9.046 |
| Beta (4,4) | 1.989 | Eff | 1.989 | 4.249 | 8.763 | 17.514 | 35.017 |
| Exponential (1) | 1.636 | Eff | 1.636 | 1.355 | 0.683 | 0.331 | 0.185 |
|  |  | Bias | 0.000 | 0.232 | 0.546 | 0.900 | 1.263 |
| Gamma (2,1) | 1.767 | Eff | 1.767 | 1.764 | 1.100 | 0.565 | 0.321 |
|  |  | Bias | 0.000 | 0.243 | 0.587 | 0.966 | 1.355 |
| Weibull (1,3) | 1.802 | Eff | 1.802 | 1.402 | 0.683 | 0.336 | 0.190 |
|  |  | Bias | 0.000 | 0.688 | 1.650 | 2.697 | 3.765 |

Table 2. The efficiency of RSS and MBGRSSE for estimating the population mean with $m=6$ and $1 \leq s \leq 3$

| Distribution | RSS | MBGRSSE |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
|  |  | $s=1$ | $s=2$ | $s=3$ |  |
| Uniform (0,1) | 3.500 | Eff | 4.258 | 32.067 | 89.541 |
| Normal (0,1) | 3.226 | Eff | 2.880 | 5.906 | 8.472 |
| Beta (4,4) | 3.319 | Eff | 3.287 | 11.647 | 28.294 |
| Exponential (1) | 2.460 | Eff | 1.497 | 0.335 | 0.154 |
|  |  | Bias | 0.135 | 0.647 | 0.996 |
| Gamma (2,1) | 2.725 | Eff | 1.916 | 0.557 | 0.268 |
|  |  | Bias | 0.141 | 0.688 | 1.057 |
| Weibull (1,3) 2.424 | Eff | 1.472 | 0.324 | 0.154 |  |
|  |  | Bias | 0.408 | 1.943 | 2.984 |

Table 3. The efficiency of RSS and MBGRSSO for estimating the population mean with $m=9$ and $1 \leq s \leq 3$

| Distribution | RSS | MBGRSSO |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
|  |  |  | $s=1$ | $s=2$ | $s=3$ |
| Uniform (0,1) | 5.000 | Eff | 6.395 | 53.838 | 342.878 |
| Normal (0,1) | 4.442 |  | Eff | 3.467 | 7.283 |
| Beta (4,4) | 4.726 |  | Eff | 4.278 | 18.386 |
| Exponential (1) | 3.251 | Eff | 0.967 | 0.130 | 71.316 |
|  |  | Bias | 0.230 | 0.896 | 1.621 |
| Gamma (2,1) 3.610 | Eff | 1.449 | 0.222 | 0.071 |  |
|  |  | Bias | 0.242 | 0.960 | 1.747 |
| Weibull (1,3) | 3.162 | Eff | 0.971 | 0.128 | 0.041 |
|  |  | Bias | 0.685 | 2.696 | 4.867 |

## A NEW ESTIMATOR OF THE POPULATION MEAN

Based on the results in Tables 1-3, we can conclude the following:
(1) When the parent distribution is symmetric about its mean we have:
a. MBGRSS method is more efficient than the usual SRS. For example, for $m=9$ and $s=2$, the efficiency of MBGRSSO is 18.740 for estimating the mean of beta $(4,4)$.
b. MBGRSS estimators are unbiased of the population mean.
c. $\quad$ The efficiency of MBGRSS is increasing in $s$ for specific value of the sample size. For example, for $m=6$, the efficiency values for $s=1,2,3$ are 4.258, 32.067, and 89.541 respectively for estimating the mean of the uniform distribution.
d. The efficiency of MBGRSS estimators is increasing as the sample size increasing. As an example, for the standard normal distribution, for $s=2$ and $m=3,6,9$ the efficiency values are $3.288,5.906$, and 7.283 , respectively.
(2) When the underlying distribution is asymmetric about the population mean we have:
a. MBGRSS estimators are biased of the population mean. For example, with $m=9$ and $s=1$, the efficiency of MBGRSSO is 0.971 with bias 0.685 when estimating the mean of the Weibull distribution with parameters 1 and 3 .
b. The efficiency is decreasing in $s$ for specific value of the sample size. For example, for $m=6$ and $s=1,2,3$, the efficiency values of MBGRSSE are 1.497, 0.335 and 0.154 , respectively for estimating the mean of exponential distribution with parameter 1.
c. The bias of MBGRSS estimators is increasing in $s$. For example, if the parent distribution is gamma with parameters 2 and 1 , then for $m=3$ and $s=1,2,3,4$, the bias values are 0 , $0.243,0.587$ and 0.966 respectively.
(3) For $m=3$ and $s=1$, MBGRSSO is the same as RSS. Otherwise, when $s>1$ and for any $m$ the MERSSO is found to be more efficient.

## AL-OMARI ET AL.

## Application to Bioleaching Studies

Generally, RSS is more efficient than SRSWR. In practice, the interest is in estimating confidence intervals (CI). When the distribution is not known using resampling seems to be a good approach for evaluating the efficiency of RSS and for performing inferences. Bootstrap has proven to be good general resampling method for deriving the sampling distribution of statistics in SRS. Chen et al. (2004) considered a bootstrapping procedure for RSS re-sampling row-wise. Hui et al. (2005) proposed bootstrapping as a method to obtain confidence interval for estimation. We are going to use their proposals for deriving estimations of the sampling errors and CI's.

A Bootstrap procedure for RSS, BRSSR, is instrumented by the following algorithm.

## BRSSR algorithm:

1. Assign to each element of the $r^{\text {th }}$ row a probability the same probability of being selected and select $m$ units randomly from $F_{(r), m}$ with replacement to get

$$
\begin{aligned}
& X_{1(1: m) h}^{*(s)}, X_{2(1: m) h}^{*(s)}, \ldots, X_{k(1: m) h}^{*(s)} X_{k+1}^{*(s)}\left(\frac{m+1}{2}: m\right) h \\
& \ldots, X_{2 k\left(\frac{m+1}{2}: m\right)}^{*(s)}, X_{2 k+1(m: m) h}^{*(s)}, \ldots, X_{3 k\left(\frac{m+1}{2}: m\right) h}^{*(s)}
\end{aligned}
$$

2. Perform Step 1 for $r=1,2, \ldots, k$ to get a bootstrap ranked set samples
3. Define the Bootstrap distributions

$$
F_{(1) m}^{*}(t)=\frac{1}{k} \sum_{h=1}^{k} I\left(X_{1(1: m) h}^{*(s)} \leq t\right), F_{\left(\frac{m+1}{2}\right)^{m}}^{*}(t)=\frac{1}{k} \sum_{h=1}^{k} I\left(X_{k+1\left(\frac{m+1}{2}: m\right)^{h}}^{*(s)} \leq t\right)
$$

and

$$
F_{(m)^{m}}^{*}(t)=\frac{1}{k} \sum_{h=1}^{k} I\left(X_{k+1(m: m) h}^{*(s)} \leq t\right)
$$

## A NEW ESTIMATOR OF THE POPULATION MEAN

The BRSSR scheme of Bootstrap resamples from each $F_{(\mathrm{r}), m}(t)$ independently and then combine to have a Bootstrap sample.

Denote by $F(2)$ a collection of distribution functions having finite second moments, $H_{n, F}$ as the sampling distribution of $T_{n}=\frac{\sum_{r=1}^{k}\left[\bar{X}_{(r)}-\mu_{(r: m)}\right] \sqrt{m_{r}}}{\sqrt{k}}, H_{n, F n}$ as the sampling distribution of the corresponding BRSSR replica $T_{n}^{*}$, and $\rho(2)(G, H)=\inf _{\tau_{X, Y}} \sqrt{E|X-Y|^{2}}$, where $\tau_{X, Y}$ is the collection of all possible joint distributions of the pairs $(X, Y)$ whose marginal distributions are $G$ and $H$, respectively. An important result is the following proposition:

Proposition 5.1: (Modarres et al. 2006).
If $F \in F(2)$ the statistics $T_{n}=\frac{\sum_{r=1}^{k}\left[\bar{X}_{(r)}-\mu_{(r: m)}\right] \sqrt{m_{r}}}{\sqrt{k}}, T_{n}^{*}$ as the corresponding
BRSSR replica. Then, $(2)\left(H_{n, F_{n}}, H_{n}\right) \xrightarrow{\text { a.s }} 0$.
Once a number of Bootstrap samples $B$ is fixed the Monte Carlo approximation of $H_{n, F_{n}}(t)$ is defined as

$$
\hat{H}_{n, F_{n}}(t)=\frac{\sum_{b=1}^{B} I\left(\bar{X}_{M B G R S S O b}^{*(s)} \leq t\right)}{B} .
$$

Because $F_{n}$ is completely specified, we can make $\hat{H}_{n, F_{n}}(t)$ arbitrarily close to $H_{n, F_{n}}(t)$ by taking a sufficiently large $B$. Now, we can estimate the moments of MBGRSSA, $A=O, E$ using

$$
\hat{\mu}_{n}^{q}=\frac{\sum_{b=1}^{B}\left(\bar{X}_{M B G R S S A b}^{*(s)}\right)^{q}}{B} .
$$

These estimators allow estimating the variance of the estimator using

$$
\hat{V}_{B}\left(\bar{X}_{\text {MBGRSSA }}^{* * s}\right)=\hat{\mu}_{n}^{2}-\left(\hat{\mu}_{n}\right)^{2}
$$

## AL-OMARI ET AL.

Approximate Bootstrap confidence intervals can also be determined computing the needed quantiles

$$
I C_{N P}(\mu)=\left(\bar{X}_{M B G R S A L}^{*}, X_{M B G R S S A U}^{*(s)}\right)
$$

such that

$$
P\left(\bar{X}_{M B G R S S A L}^{*(s)}<\mu\right) \leq \frac{\alpha}{2}, P\left(X_{M B G R S S A U}^{*(s)}<\mu\right) \leq 1-\frac{\alpha}{2}
$$

or the $T$-Student approximation:

$$
I C_{P}(\mu)=\left(\bar{X}_{M B G R S S A}^{*(s)}-t_{\left(B-1,1-\frac{\alpha}{2}\right)} \sqrt{\hat{V}_{B}\left(\bar{X}_{M B G R S S A}^{*(s)}\right)}, \bar{X}_{M B G R S S A}^{*(s)}+t_{\left(B-1,1-\frac{\alpha}{2}\right)} \sqrt{\hat{V}_{B}\left(\bar{X}_{M B G R S S A}^{*(s)}\right)}\right)
$$

Bioleaching is increasingly being used because of its economical and environmental advantages. A bioleaching is the most acceptable manner of processing of ores since it does not require elaboration of mining complexes and allows increasing the source of raw materials along with providing integrated approach to metals extraction. In terms of economy and environmental protection, biotechnological methods are more sufficient than chemical methods used for processing of ores. It consists of the acid leaching of the mineral enhanced by bacteria.


Dump ore leaching on a slope


In situ leaching in a mine

Figure 1. Two leaching procedures

## A NEW ESTIMATOR OF THE POPULATION MEAN

Figure 1 is a sketch of engineering procedures and Figure 2 of the mechanism. Direct molecular analysis of DNA has greatly enhanced the ability to assess the diversity of microorganisms growing in an ecosystem. The samples were collected in the agglomeration of mineral ores of a combination of nickel and cobalt. It is of interest to grant that the observations cover small, medium and large concentrations after bioleaching.


Figure 2 Bioleaching mechanism

The geophysics evaluate the contents of the mineral samples by a cheap method periodically. They are interested in evaluating the mean contents of cobalt in the ore. We considered the use of MBGRSS. The parameters of the example developed previously in which $s=1,2, \ldots, 5, k=1,2,3, m=3,6,9$. The sample units were taken from the existent data base compiled in the last 5 years. We computed the estimation of the variances as well as the estimation of $\operatorname{Var}\left(\bar{X}_{\text {RSS }}\right)$ using the Bootstrap estimator

$$
\hat{V}_{B}\left(\bar{X}_{R S S}^{*}\right)=\frac{\sum_{b=1}^{B}\left(\bar{X}_{R S S b}^{*}\right)^{2}}{B}-\left(\frac{\sum_{b=1}^{B} \bar{X}_{R S S b}^{*}}{B}\right)^{2}
$$

The variance of the SRSWR mean, $\operatorname{Var}\left(\bar{X}_{\text {RSS }}\right)$, was estimated computing the usual estimator of $\sigma^{2}$ as

$$
\hat{\sigma}^{2}=\frac{\sum_{i=1}^{n} \sum_{h=1}^{m} X_{i h}}{m n-1}
$$

## AL-OMARI ET AL.

The estimated efficiencies were computed as follows

$$
\begin{aligned}
& \operatorname{eff}\left(\bar{X}_{B_{\text {MBGRSSA }}^{*}(s)}^{*}, \bar{X}_{S R S}\right)=\frac{\hat{V}_{B}\left(\bar{X}_{M B G R S S A}^{*(s)}\right)}{\operatorname{Var}\left(\bar{X}_{S R S}\right)}, A=O, E \\
& \operatorname{eff}\left(\bar{X}_{B_{\text {MBGRSSA }}(s)}^{*(s)}, \bar{X}_{\text {RSS }}\right)=\frac{\hat{V}_{B}\left(\bar{X}_{M B G R S S A}^{*(s)}\right)}{\operatorname{Var}\left(\bar{X}_{\text {RSS }}\right)}, A=O, E
\end{aligned}
$$

The behavior of the proposed model was studied in the variable "Estimated lengths in nucleotides of the 16-23S intergenic spacer region in strains." The data were collected on:

$$
\begin{aligned}
& X(1)=\text { T. ferrooxidans } 530545 \\
& X(2)=\text { T. thiooxidans } 480555 \\
& X(3)=\text { L. ferrooxidans } 495505
\end{aligned}
$$

The ranking variable was a consideration on the concentrations reported by the engineers associated with each sample send to the laboratory. An R-code was developed for selecting the multistage RSS sample and performing the Bootstrap samples selections and the needed calculations. The results are presented in next tables.

Table 4. Estimated efficiency of RSS and MBGRSSO estimators for estimating the population mean $m=3$ and $1 \leq s \leq 5$

|  | $R S S$ | $s=1$ | $s=2$ | $s=3$ | $s=4$ | $s=5$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $X(1)$ | 2.391 | 2.391 | 5.019 | 5.158 | 7.541 | 8.400 |
| $X(2)$ | 1.634 | 1.634 | 4.402 | 8.460 | 10.020 | 13.066 |
| $X(3)$ | 2.703 | 2.703 | 5.969 | 8.709 | 9.522 | 12.893 |

Table 5. Estimated efficiency of RSS and MBGRSSE estimators for estimating the population mean for $m=6$ and $1 \leq s \leq 3$

|  | $R S S$ | $s=1$ | $s=2$ | $s=3$ |
| ---: | ---: | ---: | ---: | ---: |
| $X(1)$ | 2.013 | 2.013 | 5.096 | 5.222 |
| $X(2)$ | 1.900 | 1.900 | 5.158 | 8.467 |
| $X(3)$ | 2.198 | 2.198 | 5.202 | 9.741 |

## A NEW ESTIMATOR OF THE POPULATION MEAN

Table 6. Estimated efficiency of RSS and MBGRSSO estimators for estimating the population mean for $m=9$ and $1 \leq s \leq 3$

|  | RSS | $s=1$ | $s=2$ | $s=3$ |
| :--- | :--- | :--- | :--- | :--- |
| $X(1)$ | 2.223 | 2.223 | 7.771 | 8.138 |
| $X(2)$ | 2.650 | 8.467 | 10.332 | 10.910 |
| $X(3)$ | 2.073 | 9.741 | 10.779 | 12.189 |

Based on Tables 4-6, we conclude the following:
a) MBGRSS method is more efficient than the SRS and RSS methods.
b) MBGRSSO estimators with $m=3$ and $s=5$ obtains the best results in terms of efficiency for all the variables.
c) The efficiency of MBGRSS estimators is increasing as $s$ increasing.

## Conclusion

Based on MBGRSS, it can be conclude that

1) If the underlying distribution is symmetric about the population mean $\mu$, then

- The MBGRSS estimators are unbiased of the population mean.
- $\quad \operatorname{Var}\left(\bar{X}_{M B G R S S}^{(s)}\right)<\operatorname{Var}\left(\bar{X}_{S R S}\right)$,
- $\quad \operatorname{Var}\left(\bar{X}_{M B G R S}^{(s)}\right)<\operatorname{Var}\left(\bar{X}_{\text {RSS }}\right)$ for $s>1$, and $s \geq 1$ for the uniform distribution.
- The efficiency of MBGRSS estimators is increasing in $s$.

2) If the parent distribution is asymmetric about $\mu$, then

- $\quad \bar{X}_{M B G R S S}^{(s)}$ is biased.
- For $m=3,6$ and $s=1$, the MSE of $\bar{X}_{\text {MBGRSS }}^{(s)}$ is less than the variance of $\bar{X}_{S R S}$,

3) It seems that MBGRSS should be preferred in bioleaching studies to RSS and SRS.

## AL-OMARI ET AL

It is recommended that the MBGRSS be used to estimate the population mean of symmetric distribution.

## Acknowledgements

The authors are grateful to the Department of Microbiology and Biochemistry of Bhat-Sarkhar Engineering Design \& Quality Control Advisors for providing access to the databases on bioleaching that allowed to develop the study of the efficiency of MBGRSS and to CONACYT, Mexico for the support of the research.

## References

Al-Hadhrami, S. A. \& Al-Omari, A. I. (2009). Bayesian inference on the variance of normal distribution using moving extremes ranked set sampling. Journal of Modern Applied Statistical Methods, 8(1), 273-281. Available at http://digitalcommons.wayne.edu/jmasm/vol8/iss1/25

Al-Omari, A. I. (2011). Estimation of mean based on modified robust extreme ranked set sampling. Journal of Statistical Computation and Simulation, 81(8), 1055-1066. doi: 10.1080/00949651003649161

Al-Omari, A. I., Ibrahim, K., Jemain, A. A. \& Al-Hadhrami, S. A. (2009). Multistage balanced groups ranked set samples for estimating the population median. Statistics in Transition-new series, 10(2), 223-233.

Al-Saleh, M. F. \& Al-Kadiri, M. A. (2000). Double-ranked set sampling. Statistics and Probability Letters, 48(2), 205-212. doi: 10.1016/S0167-7152(99)00206-0

Al-Saleh, M. F. \& Al-Omari, A. I. (2002). Multistage ranked set sampling. Journal of Statistical Planning and Inference, 102(2), 273-286. doi: 10.1016/S0378-3758(01)00086-6

Chen, Z, Bai, Z. \& Sinha, B. (2003). Ranked set sampling: Theory and applications. New York, NY: Springer.

Dong, X. \& Cui, L. (2010). Optimal sign test for quantiles in ranked set samples. Journal of Statistical Planning and Inference, 140(11), 2943-2951. doi: 10.1016/j.jspi.2010.05.005

David, H. A. \& Nagaraja, H. N. (2003). Order Statistics (3 ${ }^{\text {rd }}$ ed.). Hoboken, NJ: John Wiley \& Sons, Inc.

## A NEW ESTIMATOR OF THE POPULATION MEAN

Drikvandi, R., Modarres, R., \& Jalilian, A. H. (2011). A bootstrap test for symmetry based on ranked set samples. Computational Statistics \& Data Analysis, 55(4), 1807-1814. doi: 10.1016/j.csda.2010.11.012

Herrera, C. N. B. \& Al-Omari, A. I. (2011). Ranked set estimation with imputation of the missing observations: The median estimator. Revista Investigación Operacional, 32(1), 30-37.

Harvey T. J., Van Der Merwe, W. \& Afewu, K. (2002). The application of the GeoBiotics GEOCOAT ${ }^{\circledR}$ biooxidation technology for the treatment of sphalerite at Kumba resources Rosh Pinah mine. Minerals Engineering, 15(11), 823-829. doi: 10.1016/S0892-6875(02)00132-2

Hui, T. P., Modarres R. \& Zheng G. (2005). Bootstrap confidence interval estimation of mean via ranked set sampling linear regression, Journal of Statistical Computation and Simulation, 75(7), 543-553. doi: 10.1080/00949650412331286124

Jemain, A. A. \& Al-Omari, A. I. (2006a). Double percentile ranked set samples for estimating the population mean. Advances and Applications in Statistics, 6(3), 261-276.

Jemain, A. A. \& Al-Omari, A. I. (2006b). Multistage median ranked set samples for estimating the population mean. Pakistan Journal of Statistics, 22(3), 195-207.

Jemain, A. A., Al-Omari, A. I. \& Ibrahim, K. (2007). Multistage extreme ranked set sampling for estimating the population mean. Journal of Statistical Theory and Applications, 6(4), 456-471.

Jemain, A. A., Al-Omari, A. I. \& Ibrahim, K. (2008). Some variations of ranked set sampling. Electronic Journal of Applied Statistical Analysis, 1(1), 1-15. doi: 10.1285/i20705948v1n1p1

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. Australian Journal of Agricultural Research, 3(4), 385-390. doi: 10.1071/AR9520385

Modarres, R., Hui, T. P., \& Zheng G. (2006). Resampling methods for ranked set samples. Computational Statistics \& Data Analysis, 51(2), 1039-1050. doi: 10.1016/j.csda.2005.10.010

Ozturk, O. (2011). Parametric estimation of location and scale parameters in ranked set sampling. Journal of Statistical Planning and Inference, 141(4), 1616-1622. doi: 10.1016/j.jspi.2010.11.019

## AL-OMARI ET AL.

Rohwerder, T., Gehrke, T., Kinzler, K, \& Sand, W. (2003). Bioleaching review part A: Progress in bioleaching: Fundamentals and mechanisms of bacterial metal sulfide oxidation. Applied Microbiology and Biotechnology, 63(3), 239-248. doi: 10.1007/s00253-003-1448-7

Takahasi, K. \& Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. Annals of the Institute Statistical Mathematics, 20(1), 1-31. doi: 10.1007/BF02911622

Vock, M. \& Balakrishnan, N. (2011). A Jonckheere-Terpstra-type test for perfect ranking in balanced ranked set sampling. Journal of Statistical Planning and Inference, 141 (2), 624-630. doi: 10.1016/j.jspi.2010.07.005

# Study of the Left Censored Data from the Gumbel Type II Distribution under a Bayesian Approach 

Tabassum Naz Sindhu
Quaid-i-Azam University Islamabad, Pakistan

Navid Feroze<br>Ripha International University<br>Islamabad, Pakistan

Muhammad Aslam<br>Ripha International University Islamabad, Pakistan

Based on left type II censored samples from a Gumbel type II distribution, the Bayes estimators and corresponding risks of the unknown parameter were obtained under different asymmetric loss functions, assuming different informative and non-informative priors. Elicitation of hyper-parameters through prior predictive approach has also been discussed. The expressions for the credible intervals and posterior predictive distributions have been derived. Comparisons of these estimators are made through simulation study using numerical and graphical methods.

Keywords: Left censoring, loss functions, credible intervals, posterior predictive distributions

## Introduction

Gumbel type II distribution is very useful in life testing. Kotz and Nadarajah (2000) have given a brief characterization of the Gumbel type II distribution. Corsini, Gini, Greco, and Verrazzani (2002) studied the maximum likelihood (ML) algorithms and Cramer-Rao (CR) bounds for the location and scale parameters of the Gumbel distribution. Mousa, Jaheen, and Ahmad (2002) considered the Bayesian estimation to analyze both parameters of the Gumbel distribution based on record values.

The probability density function of the Gumbel distribution of the second kind is given by

$$
\begin{equation*}
f(x)=\tau v x^{-(v+1)} \exp \left[-\tau x^{-\nu}\right], \quad x>0, \tau, v>0 . \tag{1}
\end{equation*}
$$

[^6]
## SINDHU ET AL.

The corresponding cumulative distribution function is:

$$
\begin{equation*}
F(x)=1-\exp \left[-\tau x^{-v}\right], \quad x>0, \tau, v>0 . \tag{2}
\end{equation*}
$$

The parameter $v$ (being known) is a shape parameter of the model, and $\tau$ is the scale parameter.

The use of a Bayesian approach allows both sample and prior information to be incorporated into the statistical analysis, which will improve the quality of the inferences and permit a reduction in sample size. The decision-theoretic viewpoint takes into account additional information concerning the possible consequences of decisions (quantified by a loss function). The aim of this is to consider the statistical analysis of the unknown parameters when the data are left censored from the Gumbel distribution of the second kind. There is a widespread application and use of left-censoring or left-censored data in survival analysis and reliability theory. For example, in medical studies patients are subject to regular examinations. Discovery of a condition only tells us that the onset of sickness fell in the period since the previous examination and nothing about the exact date of the attack. Thus the time elapsed since onset has been left censored. Similarly, consider left-censored data when estimating functions of exact policy duration without knowing the exact date of policy entry; or when estimating functions of exact age without knowing the exact date of birth. Coburn, McBride and Ziller (2002) faced this problem due to the incidence of a higher proportion of rural children whose spells were left censored (i.e., those children who entered the sample uninsured), and who remained uninsured throughout the sample. As another example, job duration might be incomplete because the beginning of the job spells is not observed, which is an incidence of left censoring (Bagger, 2005).

## Likelihood Function and Posterior Distribution

Let $X_{(r+1)}, \ldots, X_{(n)}$ be the last $n-r$ order statistics from a random sample of size $n$ following Gumbel type II distribution. Then the joint probability density function of $X_{(r+1)}, \ldots, X_{(n)}$ is given by

$$
\begin{align*}
& f\left(x_{(r+1)}, \ldots, x_{(n)} ; \tau, \nu\right)=\frac{n!}{r!}\left(F\left(x_{(r+1)}\right)\right)^{r} f\left(x_{(r+1)}\right) \ldots f\left(x_{(n)}\right) \\
& \propto \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \tau^{s} \exp \left[-\tau \varsigma\left(x_{(i)}\right)\right], \tag{3}
\end{align*}
$$

## LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION

where $s=n-r$, and

$$
\varsigma\left(x_{(i)}\right)=\exp \left[-\tau\left\{\sum_{i=r+1}^{n} x_{(i)}^{-\nu}+k x_{(r+1)}^{-\nu}\right\}\right] .
$$

## Prior and Posterior Distributions

The uniform prior is assumed to be

$$
\begin{equation*}
p(\tau) \propto k, \tau>0 . \tag{4}
\end{equation*}
$$

The posterior distribution under the uniform prior for the left censored data is:

$$
\begin{equation*}
p(\tau \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k}(\tau)^{s} \exp \left[-\tau \varsigma\left(x_{(i)}\right)\right]}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}, \tau>0 \tag{5}
\end{equation*}
$$

The informative prior for the parameter $\tau$ is assumed to be exponential distribution:

$$
\begin{equation*}
p(\tau)=w e^{-\tau w}, \quad w>0, \quad \tau>0 \tag{6}
\end{equation*}
$$

The posterior distribution under the assumption of exponential prior is:

$$
\begin{equation*}
p(\tau \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k}(\tau)^{s} \exp \left[-\tau\left\{w+\varsigma\left(x_{(i)}\right)\right\}\right]}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}, \tau>0 \tag{7}
\end{equation*}
$$

The informative prior for the parameter $\tau$ is assumed to be gamma distribution:

## SINDHU ET AL.

$$
\begin{equation*}
p(\tau)=\frac{b^{a}}{\Gamma(a)} \tau^{a-1} e^{-b \tau}, a, b, \tau>0 \tag{8}
\end{equation*}
$$

The posterior distribution under the assumption of gamma prior for the left censored data is:

$$
\begin{equation*}
p(\tau \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k}(\tau)^{s+a-1} \exp \left[-\tau\left\{b+\varsigma\left(x_{(i)}\right)\right\}\right]}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+a)}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a)}}}, \tau>0 . \tag{9}
\end{equation*}
$$

The informative prior for the parameter $\tau$ is assumed to be inverse Levy distribution:

$$
\begin{equation*}
p(\tau)=\sqrt{\frac{c}{2 \pi}} \tau^{-\frac{1}{2}} e^{-\left(\frac{c \tau}{2}\right)}, c, \tau>0 \tag{10}
\end{equation*}
$$

The posterior distribution under the inverse Levy prior for the left censored data is:

$$
\begin{equation*}
p(\tau \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k}(\tau)^{s+\frac{1}{2}-1} \exp \left[-\tau\left\{\frac{c}{2}+\varsigma\left(x_{(i)}\right)\right\}\right]}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma\left(s+\frac{1}{2}\right)}{\left\{\frac{c}{2}+\varsigma\left(x_{(i)}\right)\right\}^{\left(s+\frac{1}{2}\right)}}}, \tau>0 \tag{11}
\end{equation*}
$$

## Bayes Estimators and Posterior Risks under Different Loss Functions

Consider the derivation of the Bayes estimator and corresponding posterior risks under different loss functions. The Bayes estimators are evaluated under precautionary loss function (PLF), weighted squared error loss function (WSELF), squared-log error loss function (SLELF), and entropy loss function (ELF). The

## LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION

Bayes estimator and corresponding posterior risks under different loss functions are given in the Table 1.

Table 1. Bayes estimator and posterior risks under different loss functions

| Loss Function $=L(\tau, \hat{\tau})$ | Bayes Estimator | Posterior Risk |
| :---: | :---: | ---: |
| PLF: | $\frac{(\tau-\hat{\tau})^{2}}{\hat{\tau}}$ | $\sqrt{E\left(\tau^{2} \mid \mathbf{x}\right)}$ |
| WSELF: | $2\left\{\sqrt{E\left(\tau^{2} \mid x\right)}-E(\tau \mid x)\right\}$ |  |
| SLELF: | $(\ln \hat{\tau}-\ln \tau)^{2}$ | $\exp \{E(\ln \tau \mid x)\}$ |
| ELF: $\quad\left\{\left(\frac{(\tau)}{\tau}\right)-\ln \left(\frac{\hat{\tau}}{\tau}\right)-1\right\}$ | $\{E(\ln \tau \mid x)\}^{2}-\{E(\ln \tau \mid x)\}^{2}$ |  |
|  | $\left\{E\left(\tau^{-1} \mid x\right)\right\}^{-1}$ | $\ln \left\{E\left(\tau^{-1} \mid x\right)\right\}+E(\ln \tau)$ |

The Bayes estimators and posterior risks under uniform prior are:

$$
\left.\left.\begin{array}{c}
\hat{\tau}_{P L F}=\sqrt{\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+3)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+3)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}}
\end{array}\right] \sqrt{\rho\left(\hat{\tau}_{P L F}\right)=2\left[\sqrt{\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+3)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+3)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+2)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}} \frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{}}\right]} \text { ] }\right]
$$

## SINDHU ET AL.

$$
\begin{aligned}
& \hat{\tau}_{\text {WSELF }}=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s)}}}, \\
& \rho\left(\hat{\tau}_{\text {WSELF }}\right)=\left[\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+2)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}-\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s)}}}\right] . \\
& \hat{\tau}_{\text {SLELF }}=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1) \exp \{\psi(s+1)\}}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}, \\
& \rho\left(\hat{\tau}_{\text {SLELF }}\right)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1) \psi^{\prime}(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}} . \\
& \hat{\tau}_{E L F}=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s)}}},
\end{aligned}
$$

$$
\rho\left(\hat{\tau}_{E L F}\right)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{\Gamma(s+1)}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}+\psi(s+1)-\ln \left[\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}\right] .
$$

The Bayes estimators and posterior risks under the rest of priors can be obtained in a similar manner.

## Bayes Credible Interval for the Left Censored Data

The Bayesian credible intervals for type II left censored data under informative and non-informative priors, as discussed by Saleem and Aslam (2009) are presented in the following. The credible intervals for type II left censored data under all priors are:

$$
\begin{aligned}
& \frac{\chi_{2(s+1)\left(\frac{\alpha}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}<\tau_{\text {Uniform }}<\frac{\chi_{2(s+1)\left(1-\frac{\alpha}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}} \\
& \frac{\chi_{2(s+1)\left(\frac{a}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}<\tau_{\text {Exponential }}<\frac{\chi_{2(s+1)\left(1-\frac{\pi}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+2)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}} \\
& \frac{\chi_{2(s+1)\left(\frac{a}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a+1)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a)}}}<\tau_{\text {Gamma }}<\frac{\chi_{2(s+1)\left(1-\frac{a}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a+1)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a)}}}
\end{aligned}
$$

## SINDHU ET AL.

$$
\frac{\chi_{2(s+1)\left(\frac{c}{2}\right)}^{r} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{c / 2+\varsigma\left(x_{(i)}\right)\right\}^{\left(s+\frac{3}{2}\right)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{c / 2+\varsigma\left(x_{(i)}\right)\right\}^{(s+1 / 2)}}}<\tau_{\operatorname{ln-Levy}}<\frac{\chi_{2(s+1)\left(1-\frac{\alpha}{2}\right)}^{2} \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{c / 2+\varsigma\left(x_{(i)}\right)\right\}^{(s+3 / 2)}}}{2 \sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{c / 2+\varsigma\left(x_{(i)}\right)\right\}^{(s+1 / 2)}}}
$$

## Elicitation

Consider a probability elicitation method known as prior predictive elicitation. Predictive elicitation is a method for estimating hyper-parameters of prior distributions by inverting corresponding prior predictive distributions. Elicitation of hyper-parameter from the prior $p(\tau)$ is conceptually difficult task because we first have to identify prior distribution and then its hyper-parameters. The prior predictive distribution is used for the elicitation of the hyper-parameters which is compared with the experts' judgment about this distribution and then the hyperparameters are chosen in such a way so as to make the judgment agree closely as possible with the given distribution (see Grimshaw, 1993; Kadane, 1980; O'Hagan et al., 2006; Grimshaw, Collings, Larsen, \& Hurt, 2001; Jenkinson, 2005; and León, Vázquez-Polo, \& González, 2003).

According to Aslam (2003), the method of assessment is to compare the predictive distribution with experts' assessment about this distribution and then to choose the hyper-parameters that make the assessment agree closely with the member of the family. He discusses three important methods to elicit the hyperparameters: (i) via the prior predictive probabilities (ii) via elicitation of the confidence levels (iii) via the predictive mode and confidence level. We will use the prior predictive approach by Aslam (2003).

## Prior predictive distribution

The prior predictive distribution is:

$$
\begin{equation*}
p(y)=\int_{0}^{\infty} p(y \mid \tau) p(\tau) d \tau \tag{12}
\end{equation*}
$$

The predictive distribution under exponential prior is:

$$
\begin{equation*}
p(y)=\int_{0}^{\infty} w v \tau y^{-(v+1)} \exp \left\{-\tau\left(y^{-v}+w\right)\right\} d \tau \tag{13}
\end{equation*}
$$

After some simplification it reduces as

$$
\begin{equation*}
p(y)=\frac{v w}{y^{(v+1)}\left\{w+y^{-v}\right\}^{2}}, \quad y>0 \tag{14}
\end{equation*}
$$

The predictive distribution under gamma prior is:

$$
\begin{gather*}
p(y)=\frac{v a b^{a}}{y^{(v+1)}\left\{b+y^{-v}\right\}^{a+1}}, \quad 0<y<\infty .  \tag{15}\\
p(y)=\frac{v \sqrt{c}}{2^{3 / 2} y^{(v+1)}\left\{c / 2+y^{-\nu}\right\}^{3 / 2}}, \quad 0<y<\infty . \tag{16}
\end{gather*}
$$

By using the method of elicitation defined by Aslam (2003), we obtain the following hyper-parameters $w=0.798566, a=0.152109, b=6.523695$ and $c=15.985795$.

## Posterior Predictive Distribution

The predictive distribution contains the information about the independent future random observation given preceding observations. The reader desire more details can see Bansal (2007).

The posterior predictive distribution of the future observation $y=x_{n+1}$ is

$$
\begin{equation*}
p(y \mid \mathbf{x})=\int_{0}^{\infty} p(\tau \mid \mathbf{x}) p(y \mid \tau) d \tau \tag{17}
\end{equation*}
$$

## SINDHU ET AL.

Where $p(y)=\tau v x^{-(\nu+1)} \exp \left[-\tau x^{-\nu}\right]$, is the future observation density and $p(\tau \mid \mathbf{x})$ is the posterior distribution obtained by incorporating the likelihood with the respective prior distributions.

The posterior predictive distribution of the future observation $y=x_{n+1}$ under uniform prior is

$$
\begin{equation*}
p(y \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{s+1}{y^{(v+1)}\left\{\varsigma\left(x_{(i)}\right)+y^{-\nu}\right\}^{(s+2)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}, y>0 . \tag{18}
\end{equation*}
$$

The posterior predictive distribution of the future observation $y=x_{n+1}$ under exponential prior is

$$
\begin{equation*}
p(y \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{s+1}{y^{(v+1)}\left\{w+\varsigma\left(x_{(i)}\right)+y^{-\nu}\right\}^{(s+2)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{w+\varsigma\left(x_{(i)}\right)\right\}^{(s+1)}}}, y>0 \tag{19}
\end{equation*}
$$

The posterior predictive distribution of the future observation $y=x_{n+1}$ under gamma prior is

$$
\begin{equation*}
p(y \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{s+a}{y^{(v+1)}\left\{b+\varsigma\left(x_{(i)}\right)+y^{-v}\right\}^{(s+a+1)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{b+\varsigma\left(x_{(i)}\right)\right\}^{(s+a)}}}, y>0 \tag{20}
\end{equation*}
$$

The posterior predictive distribution of the future observation $y=x_{n+1}$ under Inverse-Levy prior is

## LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION

$$
\begin{equation*}
p(y \mid x)=\frac{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{s+1 / 2}{y^{(v+1)}\left\{c / 2+\varsigma\left(x_{(i)}\right)+y^{-\nu}\right\}^{(s+3 / 2)}}}{\sum_{k=0}^{r}(-1)^{k}\binom{r}{k} \frac{1}{\left\{c / 2+\varsigma\left(x_{(i)}\right)\right\}^{(s+1 / 2)}}}, y>0 . \tag{21}
\end{equation*}
$$

## Simulation Study

Simulations can be helpful and an illuminating way to approach problems in Bayesian analysis. Bayesian problems of updating estimates can be handled easily and straight forwardly with simulation. Because the distribution function of the Gumbel type II distribution can be expressed, as well as its inverse in closed form, the inversion method of simulation is straightforward to implement. The study was carried out for different values of $(n, r)$ using $\tau \in 2.5$ and $v=0.5$. Censoring rates are assumed to be $5 \%$ and $10 \%$.

Sample size is varied to observe the effect of small and large samples on the estimators. Changes in the estimators and their risks have been determined when changing the loss function and the prior distribution of $\tau$ while keeping the sample size fixed. All these results are based on 5,000 repetitions. Tables 2-6 give the estimated value of the parameter, posterior risks and $95 \%$ confidence limits (Lower Confidence Limit (LCL) and Upper Confidence Limit (UCL)) for the parameter. The results are summarized in the following Tables and Figures 1-8. The amounts of posterior risks have been presented in the parenthesis in the tables.

## SINDHU ET AL.

Table 2. Bayes estimates and the posterior risks under PLF for $\tau \in 2.5$.

| n | Uniform Prior |  |  |
| :---: | :---: | :---: | :---: |
|  | No Censoring | 5\% Censoring | 10\% Censoring |
| 20 | $\begin{array}{r} 2.737920 \\ (0.125898) \end{array}$ | $\begin{array}{r} 3.35045 \\ (0.157935) \end{array}$ | $\begin{array}{r} 3.77639 \\ (0.181710) \end{array}$ |
| 40 | $\begin{array}{r} 2.677940 \\ (0.064145) \end{array}$ | $\begin{array}{r} 3.15159 \\ (0.077609) \end{array}$ | $\begin{array}{r} 3.64915 \\ (0.097539) \end{array}$ |
| 60 | $\begin{array}{r} 2.62145 \\ (0.042453) \end{array}$ | $\begin{array}{r} 3.09163 \\ (0.051534) \end{array}$ | $\begin{array}{r} 3.54489 \\ (0.060447) \end{array}$ |
| 80 | $\begin{array}{r} 2.57594 \\ (0.031510) \end{array}$ | $\begin{array}{r} 3.04116 \\ (0.038311) \end{array}$ | $\begin{array}{r} 3.50579 \\ (0.045182) \end{array}$ |
| 100 | $\begin{array}{r} 2.56138 \\ (0.025173) \end{array}$ | $\begin{array}{r} 3.03806 \\ (0.030759) \end{array}$ | $\begin{array}{r} 3.47670 \\ (0.036015) \end{array}$ |
| n | Exponential Prior |  |  |
| 20 | $\begin{array}{r} 2.58014 \\ (0.118643) \end{array}$ | $\begin{array}{r} 2.96201 \\ (0.138226) \end{array}$ | $\begin{array}{r} 3.38135 \\ (0.156758) \end{array}$ |
| 40 | $\begin{array}{r} 2.52198 \\ (0.060409) \end{array}$ | $\begin{array}{r} 2.95898 \\ (0.072220) \end{array}$ | $\begin{array}{r} 3.36035 \\ (0.084258) \end{array}$ |
| 60 | $\begin{array}{r} 2.52440 \\ (0.040720) \end{array}$ | $\begin{array}{r} 2.95009 \\ (0.049112) \end{array}$ | $\begin{array}{r} 3.35418 \\ (0.057015) \end{array}$ |
| 80 | $\begin{array}{r} 2.52171 \\ (0.030847) \end{array}$ | $\begin{array}{r} 2.94949 \\ (0.037501) \end{array}$ | $\begin{array}{r} 3.33655 \\ (0.043241) \end{array}$ |
| 100 | $\begin{array}{r} 2.50779 \\ (0.024647) \\ \hline \end{array}$ | $\begin{array}{r} 2.92773 \\ (0.030070) \\ \hline \end{array}$ | $\begin{array}{r} 3.30688 \\ (0.035032) \\ \hline \end{array}$ |
| n | Gamma Prior |  |  |
| 20 | $\begin{array}{r} 1.43895 \\ (0.068852) \end{array}$ | $\begin{array}{r} 1.55700 \\ (0.075152) \end{array}$ | $\begin{array}{r} 1.64308 \\ (0.079688) \end{array}$ |
| 40 | $\begin{array}{r} 1.82853 \\ (0.044707) \end{array}$ | $\begin{array}{r} 2.04504 \\ (0.050801) \end{array}$ | $\begin{array}{r} 2.21285 \\ (0.055460) \end{array}$ |
| 60 | $\begin{array}{r} 2.00816 \\ (0.032974) \end{array}$ | $\begin{array}{r} 2.26658 \\ (0.037962) \end{array}$ | $\begin{array}{r} 2.49874 \\ (0.042352) \end{array}$ |
| 80 | $\begin{array}{r} 2.11237 \\ (0.026111) \end{array}$ | $\begin{array}{r} 2.41150 \\ (0.030475) \end{array}$ | $\begin{array}{r} 2.67252 \\ (0.034264) \end{array}$ |
| 100 | $\begin{array}{r} 2.218482 \\ (0.021653) \\ \hline \end{array}$ | $\begin{array}{r} 2.51014 \\ (0.025478) \\ \hline \end{array}$ | $\begin{array}{r} 2.79600 \\ (0.028819) \\ \hline \end{array}$ |
| n | Inverse Levy Prior |  |  |
| 20 | $\begin{array}{r} 1.32737 \\ (0.062473) \end{array}$ | $\begin{array}{r} 1.43304 \\ (0.067927) \end{array}$ | $\begin{array}{r} 1.49803 \\ (0.071294) \end{array}$ |
| 40 | $\begin{array}{r} 1.72182 \\ (0.041743) \end{array}$ | $\begin{array}{r} 1.91963 \\ (0.047193) \end{array}$ | $\begin{array}{r} 2.05833 \\ (0.051005) \end{array}$ |
| 60 | $\begin{array}{r} 1.93203 \\ (0.031544) \end{array}$ | $\begin{array}{r} 2.16662 \\ (0.036031) \end{array}$ | $\begin{array}{r} 2.37030 \\ (0.039845) \end{array}$ |
| 80 | $\begin{array}{r} 2.04177 \\ (0.025129) \end{array}$ | $\begin{array}{r} 2.32593 \\ (0.029234) \end{array}$ | $\begin{array}{r} 2.55092 \\ (0.032477) \end{array}$ |
| 100 | $\begin{array}{r} 2.12131 \\ (0.020951) \\ \hline \end{array}$ | $\begin{array}{r} 2.41626 \\ (0.024413) \\ \hline \end{array}$ | $\begin{array}{r} 2.68807 \\ (0.027552) \\ \hline \end{array}$ |

## LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION

Table 3. Bayes estimates and the posterior risks under WSELF for $\tau \in 2.5$.

| n | Uniform Prior |  |  |
| :---: | :---: | :---: | :---: |
|  | No Censoring | 5\% Censoring | 10\% Censoring |
| 20 | $\begin{array}{r} 2.66809 \\ (0.133405) \end{array}$ | $\begin{array}{r} 3.08160 \\ (0.157976) \end{array}$ | $\begin{array}{r} 3.54947 \\ (0.186003) \end{array}$ |
| 40 | $\begin{array}{r} 2.55583 \\ (0.063896) \end{array}$ | $\begin{array}{r} 3.05530 \\ (0.078578) \end{array}$ | $\begin{array}{r} 3.43934 \\ (0.090409) \end{array}$ |
| 60 | $\begin{array}{r} 2.55213 \\ (0.042536) \end{array}$ | $\begin{array}{r} 3.02388 \\ (0.051901) \end{array}$ | $\begin{array}{r} 3.42741 \\ (0.060168) \end{array}$ |
| 80 | $\begin{array}{r} 2.53489 \\ (0.031686) \end{array}$ | $\begin{array}{r} 3.01692 \\ (0.038842) \end{array}$ | $\begin{array}{r} 3.41996 \\ (0.04506) \end{array}$ |
| 100 | $\begin{array}{r} 2.51670 \\ (0.025167) \end{array}$ | $\begin{array}{r} 3.00774 \\ (0.030991) \end{array}$ | $\begin{array}{r} 3.40597 \\ (0.035925) \end{array}$ |
| n | Exponential Prior |  |  |
| 20 | $\begin{array}{r} 2.37956 \\ (0.118978) \end{array}$ | $\begin{array}{r} 2.93114 \\ (0.139567) \end{array}$ | $\begin{array}{r} 3.35007 \\ (0.158471) \end{array}$ |
| 40 | $\begin{array}{r} 2.42840 \\ (0.060710) \end{array}$ | $\begin{array}{r} 2.87664 \\ (0.073818) \end{array}$ | $\begin{array}{r} 3.27245 \\ (0.085679) \end{array}$ |
| 60 | $\begin{array}{r} 2.46768 \\ (0.041128) \end{array}$ | $\begin{array}{r} 2.85571 \\ (0.049693) \end{array}$ | $\begin{array}{r} 3.270610 \\ (0.057314) \end{array}$ |
| 80 | $\begin{array}{r} 2.47487 \\ (0.030936) \end{array}$ | $\begin{array}{r} 2.72288 \\ (0.037589) \end{array}$ | $\begin{array}{r} 3.134120 \\ (0.043824) \end{array}$ |
| 100 | $\begin{array}{r} 2.48550 \\ (0.024855) \\ \hline \end{array}$ | $\begin{array}{r} 2.624320 \\ (0.030108) \\ \hline \end{array}$ | $\begin{array}{r} 3.02926 \\ (0.035046) \\ \hline \end{array}$ |
| n | Gamma Prior |  |  |
| 20 | $\begin{array}{r} 1.33348 \\ (0.069626) \end{array}$ | $\begin{array}{r} 1.44368 \\ (0.075839) \end{array}$ | $\begin{array}{r} 1.51586 \\ (0.080755) \end{array}$ |
| 40 | $\begin{array}{r} 1.75474 \\ (0.044819) \end{array}$ | $\begin{array}{r} 1.98012 \\ (0.050968) \end{array}$ | $\begin{array}{r} 2.12591 \\ (0.055810) \end{array}$ |
| 60 | $\begin{array}{r} 1.95524 \\ (0.03306) \end{array}$ | $\begin{array}{r} 2.25507 \\ (0.038435) \end{array}$ | $\begin{array}{r} 2.44299 \\ (0.042656) \end{array}$ |
| 80 | $\begin{array}{r} 2.07625 \\ (0.026231) \end{array}$ | $\begin{array}{r} 2.40362 \\ (0.030624) \end{array}$ | $\begin{array}{r} 2.63342 \\ (0.034421) \end{array}$ |
| 100 | $\begin{array}{r} 2.244640 \\ (0.021630) \\ \hline \end{array}$ | $\begin{array}{r} 2.50664 \\ (0.025501) \\ \hline \end{array}$ | $\begin{array}{r} 2.77998 \\ (0.029085) \\ \hline \end{array}$ |
| n | Inverse Levy Prior |  |  |
| 20 | $\begin{array}{r} 1.24650 \\ (0.063923) \end{array}$ | $\begin{array}{r} 1.31807 \\ (0.068090) \end{array}$ | $\begin{array}{r} 1.38627 \\ (0.071871) \end{array}$ |
| 40 | $\begin{array}{r} 1.665110 \\ (0.042155) \end{array}$ | $\begin{array}{r} 1.74892 \\ (0.044659) \end{array}$ | $\begin{array}{r} 1.84547 \\ (0.047385) \end{array}$ |
| 60 | $\begin{array}{r} 1.86831 \\ (0.031400) \end{array}$ | $\begin{array}{r} 2.10212 \\ (0.035987) \end{array}$ | $\begin{array}{r} 2.32167 \\ (0.040176) \end{array}$ |
| 80 | $\begin{array}{r} 1.99783 \\ (0.02513) \end{array}$ | $\begin{array}{r} 2.33427 \\ (0.030086) \end{array}$ | $\begin{array}{r} 2.50929 \\ (0.032640) \end{array}$ |
| 100 | $\begin{array}{r} 2.18089 \\ (0.020913) \end{array}$ | $\begin{array}{r} 2.40249 \\ (0.024701) \end{array}$ | $\begin{array}{r} 2.64028 \\ (0.027546) \end{array}$ |

## SINDHU ET AL.

Table 4. Bayes estimates and the posterior risks under SLELF for $\tau \in 2.5$.

| n | Uniform Prior |  |  |
| :---: | :---: | :---: | :---: |
|  | No Censoring | 5\% Censoring | 10\% Censoring |
| 20 | $\begin{array}{r} 2.70493 \\ (0.048771) \end{array}$ | $\begin{array}{r} 3.16249 \\ (0.051271) \end{array}$ | $\begin{array}{r} 3.67867 \\ (0.054041) \end{array}$ |
| 40 | $\begin{array}{r} 2.60860 \\ (0.024690) \end{array}$ | $\begin{array}{r} 3.08320 \\ (0.025973) \end{array}$ | $\begin{array}{r} 3.52510 \\ (0.027396) \end{array}$ |
| 60 | $\begin{array}{r} 2.548760 \\ (0.016529) \end{array}$ | $\begin{array}{r} 3.04864 \\ (0.017391) \end{array}$ | $\begin{array}{r} 3.48125 \\ (0.018348) \end{array}$ |
| 80 | $\begin{array}{r} 2.53947 \\ (0.012422) \end{array}$ | $\begin{array}{r} 3.02895 \\ (0.013072) \end{array}$ | $\begin{array}{r} 3.46749 \\ (0.013793) \end{array}$ |
| 100 | $\begin{array}{r} 2.53070 \\ (0.009950) \\ \hline \end{array}$ | $\begin{array}{r} 3.019810 \\ (0.010471) \\ \hline \end{array}$ | $\begin{array}{r} 3.24692 \\ (0.011050) \\ \hline \end{array}$ |
| n | Exponential Prior |  |  |
| 20 | $\begin{array}{r} 2.42262 \\ (0.048771) \end{array}$ | $\begin{array}{r} 2.89396 \\ (0.051271) \end{array}$ | $\begin{array}{r} 3.13621 \\ (0.054041) \end{array}$ |
| 40 | $\begin{array}{r} 2.46614 \\ (0.024690) \end{array}$ | $\begin{array}{r} 2.87997 \\ (0.025973) \end{array}$ | $\begin{array}{r} 3.11318 \\ (0.027396) \end{array}$ |
| 60 | $\begin{array}{r} 2.47732 \\ (0.016529) \end{array}$ | $\begin{array}{r} 2.79474 \\ (0.017391) \end{array}$ | $\begin{array}{r} 3.01411 \\ (0.018348) \end{array}$ |
| 80 | $\begin{array}{r} 2.48808 \\ (0.012422) \end{array}$ | $\begin{array}{r} 2.64583 \\ (0.013072) \end{array}$ | $\begin{array}{r} 3.006108 \\ (0.013793) \end{array}$ |
| 100 | $\begin{array}{r} 2.497560 \\ (0.009950) \end{array}$ | $\begin{array}{r} 2.60852 \\ (0.010471) \end{array}$ | $\begin{array}{r} 2.985631 \\ (0.011050) \end{array}$ |
| n | Gamma Prior |  |  |
| 20 | $\begin{array}{r} 1.37081 \\ (0.050874) \end{array}$ | $\begin{array}{r} 1.48503 \\ (0.0536004) \end{array}$ | $\begin{array}{r} 1.56354 \\ (0.056635) \end{array}$ |
| 40 | $\begin{array}{r} 1.78940 \\ (0.025218) \end{array}$ | $\begin{array}{r} 1.98832 \\ (0.026557) \end{array}$ | $\begin{array}{r} 2.15504 \\ (0.028047) \end{array}$ |
| 60 | $\begin{array}{r} 1.98230 \\ (0.016764) \end{array}$ | $\begin{array}{r} 2.23221 \\ (0.017651) \end{array}$ | $\begin{array}{r} 2.45581 \\ (0.018638) \end{array}$ |
| 80 | $\begin{array}{r} 2.081680 \\ (0.012554) \end{array}$ | $\begin{array}{r} 2.38376 \\ (0.013218) \end{array}$ | $\begin{array}{r} 2.63859 \\ (0.013956) \end{array}$ |
| 100 | $\begin{array}{r} 2.26264 \\ (0.010035) \end{array}$ | $\begin{array}{r} 2.48866 \\ (0.010565) \end{array}$ | $\begin{array}{r} 2.77011 \\ (0.011154) \end{array}$ |
| n |  | Levy Prior |  |
| 20 | $\begin{array}{r} 1.27054 \\ (0.049989) \end{array}$ | $\begin{array}{r} 1.34243 \\ (0.052619) \end{array}$ | $\begin{array}{r} 1.42286 \\ (0.055541) \end{array}$ |
| 40 | $\begin{array}{r} 1.69351 \\ (0.024999) \end{array}$ | $\begin{array}{r} 1.86554 \\ (0.026314) \end{array}$ | $\begin{array}{r} 2.01136 \\ (0.027776) \end{array}$ |
| 60 | $\begin{array}{r} 1.90254 \\ (0.016663) \end{array}$ | $\begin{array}{r} 2.19742 \\ (0.017856) \end{array}$ | $\begin{array}{r} 2.32432 \\ (0.018518) \end{array}$ |
| 80 | $\begin{array}{r} 2.01472 \\ (0.012499) \end{array}$ | $\begin{array}{r} 2.29894 \\ (0.013158) \end{array}$ | $\begin{array}{r} 2.52262 \\ (0.013889) \end{array}$ |
| 100 | $\begin{array}{r} 2.20627 \\ (0.009999) \end{array}$ | $\begin{array}{r} 2.40058 \\ (0.010526) \end{array}$ | $\begin{array}{r} 2.64965 \\ (0.011111) \end{array}$ |

## LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION

Table 5. Bayes estimates and the posterior risks under ELF for $\tau \in 2.5$.

| n | Uniform Prior |  |  |
| :---: | :---: | :---: | :---: |
|  | No Censoring | 5\% Censoring | 10\% Censoring |
| 20 | $\begin{array}{r} 2.63866 \\ (0.024792) \end{array}$ | $\begin{array}{r} 3.10757 \\ (0.025787) \end{array}$ | $\begin{array}{r} 3.56083 \\ (0.026520) \end{array}$ |
| 40 | $\begin{array}{r} 2.56586 \\ (0.012448) \end{array}$ | $\begin{array}{r} 3.06196 \\ (0.012508) \end{array}$ | $\begin{array}{r} 3.46458 \\ (0.012576) \end{array}$ |
| 60 | $\begin{array}{r} 2.53490 \\ (0.008310) \end{array}$ | $\begin{array}{r} 3.03388 \\ (0.008570) \end{array}$ | $\begin{array}{r} 3.42366 \\ (0.008987) \end{array}$ |
| 80 | $\begin{array}{r} 2.52287 \\ (0.006237) \end{array}$ | $\begin{array}{r} 3.00312 \\ (0.006286) \end{array}$ | $\begin{array}{r} 3.15751 \\ (0.006721) \end{array}$ |
| 100 | $\begin{array}{r} 2.51440 \\ (0.004992) \end{array}$ | $\begin{array}{r} 2.901795 \\ (0.005235) \end{array}$ | $\begin{array}{r} 3.003575 \\ (0.005982) \end{array}$ |
| n | Exponential Prior |  |  |
| 20 | $\begin{array}{r} 2.56510 \\ (0.024792) \end{array}$ | $\begin{array}{r} 2.69689 \\ (0.025787) \end{array}$ | $\begin{array}{r} 3.05465 \\ (0.026520) \end{array}$ |
| 40 | $\begin{array}{r} 2.52434 \\ (0.012448) \end{array}$ | $\begin{array}{r} 2.58528 \\ (0.012508) \end{array}$ | $\begin{array}{r} 3.02735 \\ (0.012576) \end{array}$ |
| 60 | $\begin{array}{r} 2.50708 \\ (0.008310) \end{array}$ | $\begin{array}{r} 2.561238 \\ (0.008570) \end{array}$ | $\begin{array}{r} 3.017921 \\ (0.008987) \end{array}$ |
| 80 | $\begin{array}{r} 2.48248 \\ (0.006237) \end{array}$ | $\begin{array}{r} 2.52515 \\ (0.006286) \end{array}$ | $\begin{array}{r} 3.00984 \\ (0.006721) \end{array}$ |
| 100 | $\begin{array}{r} 2.46838 \\ (0.004992) \end{array}$ | $\begin{array}{r} 2.49894 \\ (0.005235) \end{array}$ | $\begin{array}{r} 2.91496 \\ (0.005982) \end{array}$ |
| n | Gamma Prior |  |  |
| 20 | $\begin{array}{r} 1.33972 \\ (0.025879) \end{array}$ | $\begin{array}{r} 1.44818 \\ (0.024988) \end{array}$ | $\begin{array}{r} 1.52916 \\ (0.025776) \end{array}$ |
| 40 | $\begin{array}{r} 1.76606 \\ (0.012763) \end{array}$ | $\begin{array}{r} 1.96735 \\ (0.012456) \end{array}$ | $\begin{array}{r} 2.12581 \\ (0.011955) \end{array}$ |
| 60 | $\begin{array}{r} 1.94527 \\ (0.008429) \end{array}$ | $\begin{array}{r} 2.21469 \\ (0.008322) \end{array}$ | $\begin{array}{r} 2.44627 \\ (0.008047) \end{array}$ |
| 80 | $\begin{array}{r} 2.07237 \\ (0.006304) \end{array}$ | $\begin{array}{r} 2.36455 \\ (0.006255) \end{array}$ | $\begin{array}{r} 2.62396 \\ (0.006071) \end{array}$ |
| 100 | $\begin{array}{r} 2.15873 \\ (0.005034) \end{array}$ | $\begin{array}{r} 2.47250 \\ (0.005010) \end{array}$ | $\begin{array}{r} 2.75845 \\ (0.004880) \end{array}$ |
| n |  | Levy Prior |  |
| 20 | $\begin{array}{r} 1.23549 \\ (0.025422) \end{array}$ | $\begin{array}{r} 1.31738 \\ (0.024519) \end{array}$ | $\begin{array}{r} 1.39072 \\ (0.023289) \end{array}$ |
| 40 | $\begin{array}{r} 1.66838 \\ (0.012605) \end{array}$ | $\begin{array}{r} 1.84774 \\ (0.012314) \end{array}$ | $\begin{array}{r} 1.97503 \\ (0.0117967) \end{array}$ |
| 60 | $\begin{array}{r} 1.87576 \\ (0.008380) \end{array}$ | $\begin{array}{r} 2.10021 \\ (0.008254) \end{array}$ | $\begin{array}{r} 2.30080 \\ (0.007957) \end{array}$ |
| 80 | $\begin{array}{r} 2.011420 \\ (0.006276) \end{array}$ | $\begin{array}{r} 2.26947 \\ (0.006214) \end{array}$ | $\begin{array}{r} 2.49758 \\ (0.006016) \end{array}$ |
| 100 | $\begin{array}{r} 2.30955 \\ (0.005017) \\ \hline \end{array}$ | $\begin{array}{r} 2.39526 \\ (0.004983) \\ \hline \end{array}$ | $\begin{array}{r} 2.65130 \\ (0.004843) \\ \hline \end{array}$ |

## SINDHU ET AL.

Table 6. The $95 \%$ credible intervals for $\tau \in 2.5$.

| n | Lower Limit | Uniform Prior <br> Upper Limit | Difference |
| ---: | ---: | ---: | ---: |
| 20 | 2.10503 | 5.23490 | 3.12987 |
| 40 | 2.44587 | 4.67921 | 2.23334 |
| 60 | 2.58722 | 4.39961 | 1.81239 |
| 80 | 2.71041 | 4.29493 | 1.58452 |
| 100 | 2.77531 | 4.19040 | 1.41509 |
| n |  | Exponential Prior |  |
| 20 | 1.84980 | 4.60018 | 2.75038 |
| 40 | 2.28485 | 4.37117 | 2.08632 |
| 60 | 2.47071 | 4.20149 | 1.73078 |
| 80 | 2.61670 | 4.14644 | 1.52974 |
| 100 | 2.69796 | 4.07361 | 1.37565 |
| n |  | Gamma Prior |  |
| 20 | 1.06688 | 2.58544 | 1.51856 |
| 40 | 1.60787 | 3.04682 | 1.43895 |
| 60 | 1.91272 | 3.23551 | 1.32279 |
| 80 | 2.13391 | 3.36978 | 1.23587 |
| 100 | 2.27978 | 3.43369 | 1.15391 |
| n |  | Inverse Levy Prior |  |
| 20 | 0.86467 | 2.17747 | 1.31280 |
| 40 | 1.41811 | 2.72520 | 1.30709 |
| 60 | 1.74630 | 2.97690 | 1.23060 |
| 80 | 1.98529 | 3.15093 | 1.16564 |
| 100 | 2.14761 | 3.24636 | 1.09875 |

## Graphical Representation of Posterior Risks under Different Priors

The graphs reveal that posterior risks under different informative and non informative priors. It is observed that both the priors (uniform and exponential) yield the approximately the identical posterior inferences under ELF and SLELF.


Figure 1. Effect of posterior risk under PLF with no censoring


Figure 2. Effect of posterior risk under PLF with 10\% censoring

## SINDHU ET AL.



Figure 3. Effect of posterior risk under WSELF with no censoring


Figure 4. Effect of posterior risk under WSELF with $10 \%$ censoring

LEFT CENSORED DATA FROM THE GUMBEL TYPE II DISTRIBUTION


Figure 5. Effect of posterior risk under SLELF with no censoring


Figure 6. Effect of posterior risk under SLELF with 10\% censoring

## SINDHU ET AL.



Figure 7. Effect of posterior risk under ELF with no censoring


Figure 8. Effect of posterior risk under ELF with 10\% censoring

## Conclusion

The simulation study displayed some interesting properties of the Bayes estimates. The risks under said loss functions are reduced as the sample size increases. The effect of censoring on estimation of $\tau$ is in the form of overestimation under uniform and exponential priors and underestimation assuming gamma and inverse Levy priors. Larger degrees of censoring results in bigger sizes of over or underestimation.

However, the parameter $\tau$ is either underestimated or overestimated depending upon the prior distribution to be used when censoring is not done. Then extent of this over or under estimation is directly proportional to amount of censoring rates and inversely proportional to the sample size. Further, the increase in sample size reduces the posterior risks of $\tau$.

Another interesting remark concerning the risks of the estimates is that increasing (decreasing) the censoring rate increasing (reduces) the risks of the estimates under said loss functions. The performance of squared-log error loss function and entropy loss function is independent of choice of parametric value. In comparison of informative priors and the uniform prior, the inverse Levy prior provides the better estimates as the corresponding risks are least under said loss functions except ELF and SLELF. Although the uniform and the exponential priors are equally efficient under ELF and SLELF, therefore they produce more efficient estimates as compared to the other informative priors.

The credible intervals are in accordance with the point estimates, that is, the width of credible interval is inversely proportional to sample size. From the Table 6, appended above, it can be revealed that the effect of the prior information is in the form of narrower width of interval. The credible interval assuming inverse Levy prior is much narrower than the credible intervals assuming informative and non-informative priors.

It is the use of prior information that makes a difference in terms of gain in precision. To see the effects of the posterior risks assuming different priors Figures 1-8 are prepared. It is observed from all the figures that posterior risk decreases with the increase in sample size under all loss functions. It is evident from Figures 5-8 that behavior of posterior risks is similar in all aspects. The study can further be extended by considering generalized versions of the distribution under variety of circumstances.

## SINDHU ET AL.

## References

Aslam, M. (2003). An application of prior predictive distribution to elicit the prior density. Journal of Statistical Theory and Applications, 2(1), 70-83.

Bagger, J. (2005). Wage growth and turnover in Denmark. Denmark: University of Aarhus.

Bansal, A. K. (2007). Bayesian Parametric Inference. New Delhi, India: Narosa Publishing House.

Coburn, A. F., McBride, T. D., \& Ziller, E. C. (2002). Patterns of Health Insurance Coverage among Rural and Urban Children. Medical Care Research and Review, 59(3), 272-292. doi: 10.1177/1077558702059003003

Corsini, G., Gini, F., Greco, M. V., \& Verrazzani, L. (2002). Cramer-Rao bounds and estimation of the parameters of the Gumbel distribution. IEEE Transactions on Aerospace and Electronic Systems, 31(3), 1202-1204. doi: 10.1109/7.395217

Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. Technometrics, 35(2), 185-191. doi: 10.1080/00401706.1993.10485040

Grimshaw, S. D., Collings, B. J., Larsen, W. A., \& Hurt, C. R. (2001). Eliciting Factor Importance in a Designed Experiment. Technometrics, 43(2), 133-146. doi: 10.1198/004017001750386251

Jenkinson, D. (2005). The elicitation of probabilities: A review of the statistical literature. BEEP Working Paper. Department of Probability and Statistics, Sheffield: University of Sheffield.

Kadane, J. B. (1980). Predictive and Structural Methods for Eliciting Prior Distributions. In H. Jeffreys \& A. Zellner (Eds.), Bayesian analysis in econometrics and statistics: Essays in honor of Harold Jeffreys. Amsterdam: North-Holland.

Kotz, S., \& Nadarajah, S. (2000). Extreme value distributions. Theory and applications. London: Imperial College Press.

León, C. J., Vázquez-Polo, F. J., \& González, R. L. (2003). Elicitation of Expert Opinion in Benefit Transfer of Environmental Goods. Environmental and Resource Economics, 26(2), 199-210. doi: 10.1023/A:1026307420804

Mousa, M. A M. A., Jaheen, Z. F., \& Ahmad, A. A. (2002). Bayesian estimation, prediction and characterization for the Gumbel model based on
records. Statistics: A Journal of Theoretical and Applied Statistics, 36(1), 65-74. doi: 10.1080/02331880210929

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). Uncertain Judgements: Eliciting experts' probabilities. Chichester: John Wiley \& Sons.

Saleem, M. \& Aslam, M. (2009). On Bayesian analysis of the Rayleigh survival time assuming the random censor time. Pakistan Journal of Statistics, 25(2), 71-82.

# Preliminary Tests of Normality When Comparing Three Independent Samples 

Björn Lantz

Chalmers University of Technology Göteborg, Sweden

Roy Andersson
Jönköping University
Jönköping, Sweden

## Peter Manfredsson

University of Borås
Borås, Sweden

This paper uses simulation to explore the performance of a two-stage procedure where a preliminary Shapiro-Wilk test is used to choose between the ANOVA and Kruskal-Wallis tests as a three-sample location test. The results suggest that the two-stage procedure actually seems to be preferable when conducting such location tests.

Keywords: Normality, assumptions, preliminary tests, ANOVA, Kruskal-Wallis, Shapiro-Wilk

## Introduction

It is common among applied researchers in psychology to conduct data analyses as two-stage procedures where one or more preliminary tests precede the test of interest (Keselman, Othman, \& Wilcox, 2013). For example, when a researcher plans to compare two population means with Student's $t$-test, the underlying normality assumption is often checked with a preliminary goodness-of-fit test. If the null hypothesis of normality is rejected, the Mann-Whitney test (or some other non-parametric test) is used to analyze the data. If the null hypothesis of normality is not rejected, the underlying homoscedasticity assumption may be checked in a similar manner. If the null hypothesis of homoscedasticity is rejected, Welch's $t$ test (or some other robust test) is used. If data were neither significantly non-normal nor significantly heteroscedastic, Student's $t$-test is used to compare the two means.

The normality assumption issue is highly relevant for data analyses in psychological research. For example, in the empirical study of achievement and psychometric measures conducted by Micceri (1989), significant non-normality contaminations were found in all 440 measures, including tail weights from the uniform to the double exponential, exponential level asymmetry, and bimodality. Furthermore, recent research has shown that most real data samples are at least

[^7]
## PRELIMINARY TESTS OF NORMALITY

slightly non-normal in terms of skewness and kurtosis (Blanca, Arnau, LópezMontiel, Bono, \& Bendayan, 2013) and that the variance heterogeneity assumption is violated in a nontrivial number of published studies (Ruscio \& Roche, 2012). However, there are several conceptual reasons why the use of a two-stage test procedure with a preliminary test of normality and/or the homoscedasticity assumption may be problematic in practice (Wells \& Hinze, 2006):

- The probability of a type I error as well as a type II error in the procedure may be heavily distorted. This is because the distribution of the location test statistic is not only related to the parental distribution(s), but also conditional on the preliminary test since both type I errors and type II errors are possible in the first stage. For example, even if a parental population is significantly contaminated from the exponential distribution, many samples will not look non-normal enough to fail the normality test. This is because of the random component of the sampling procedure. However, the samples that pass the normality test will often be significantly different from the other samples, not only in terms of shape but also in terms of mean and/or standard deviation.
- A preliminary test in which the null hypothesis of normality is not rejected does not constitute proof that the normality assumption holds. In fact, no null hypothesis is strictly ever true when empirical data are considered (Cohen, 1994). From this perspective, normality assumptions are always violated.
- Preliminary test procedures rely on assumptions themselves. This means that, strictly speaking, those assumptions also need to be tested. This would however also require new assumptions, and so on, and so forth.
- Even though a preliminary test correctly indicates that a normality assumption does not hold, a parametric test with higher power than the corresponding non-parametric test might still be valid because of high robustness against the current type of non-normality.

Recently, the performance of different two-stage procedures, where samples are checked with preliminary tests of normality before univariate or bivariate location tests, have been studied. For example, Rochon and Kieser (2011) examined the type I error rate of the one-sample Student's $t$-test with a preliminary normality test. They found an increase in the type I error rate for conditional samples compared to unconditional ones, especially when parental distributions were

## LANTZ ET AL

skewed. Schucany and Ng (2006) found similar results for the one-sample Student's $t$-test with a preliminary normality test, concluding that graphical diagnostics are probably better in practice than formal pretests. Rochon, Gondan, and Kieser (2012) examined a two-stage procedure where a preliminary normality test was used to decide between the two-sample Student's $t$-test and the MannWhitney test in the second stage. They concluded that even though the two-stage procedure might be considered incorrect from a formal perspective, the procedure seemed to satisfactorily maintain the nominal significance level and had acceptable power properties in the investigated examples. Rasch, Kubinger, and Moder (2011), on the other hand, found that it is preferable to use Welch's $t$-test without pre-testing for normality rather than the two-stage procedure including Student's $t$-test as a standard test, and that the corresponding non-parametric test should not be used in the given context. Preliminary tests have also recently been discussed in related contexts by, for example, Lantz (2013), Zimmerman (2004, 2011, 2014), Shuster (2005, 2009), and Schoder, Himmelmann, and Wilhelm (2006).

Overall, there seems to be a general consensus in the literature that two-stage procedures including preliminary tests are unnecessary at best, or harmful at worst, in a one-sample or a two-sample location test context. However, there does not seem to exist any similar literature based on simulated two-stage location tests for three (or more) groups, even though both Othman, Keselman, and Wilcox (2015) and Keselman, Othman, and Wilcox (2014) analyze the two-stage procedure problem itself in a multi-group context based on simulations. The focus in both papers is on the normality screening rather on the two-stage procedure as a whole, though.

We thus seek to answer the following question in this paper: what are the properties of a two-stage procedure where a normality test at the first stage is used to decide between the omnibus one-way ANOVA and the Kruskal-Wallis test in the second stage? The purpose of this paper is to present the results from a simulation study designed to shed light on this issue. In the next section the methodology of the study is described. The results of the simulations are then presented and discussed in relation to previous research. Finally, the paper concludes with the implications of these results for use in statistical analysis in practice.

## Methodology

In the simulations, random samples from three independent groups were drawn from four different distributions, in line with the typical contaminations found by

## PRELIMINARY TESTS OF NORMALITY

Micceri (1989) in his empirical study of achievement and psychometric measures. The distributions used are also the ones typically used in this type of study (e.g., Rochon et al., 2012), that is, the normal, the uniform, the exponential, and the Laplace distributions. The uniform distribution represents a decent approximation of the normal distribution, while the exponential and the Laplace distributions represent two different types of distinct non-normality in terms of skewness and kurtosis. The normal distribution is included for the purpose of comparability.

The probability density function of the normal distribution is given by

$$
\begin{equation*}
\mathrm{f}(x)=\frac{\mathrm{e}^{-(x-\mu)^{2} / 2 \sigma^{2}}}{\sqrt{2 \pi \sigma^{2}}} \tag{1}
\end{equation*}
$$

with mean $\mu$ and variance $\sigma^{2}$. It has no skewness and by definition no excess kurtosis.
The probability density function of the uniform distribution is given by

$$
\mathrm{f}(x)= \begin{cases}\frac{1}{b-a}, & a \leq x \leq b  \tag{2}\\ 0, & \text { otherwise }\end{cases}
$$

The uniform distribution is symmetric, like the normal distribution, and slightly platykurtic.

The probability density function of the exponential distribution is given by

$$
\mathrm{f}(x)= \begin{cases}\lambda \mathrm{e}^{-\lambda x}, & x \geq 0  \tag{3}\\ 0, & x<0\end{cases}
$$

where $\lambda$ is the rate parameter. It represents a distinct form of non-normality due to its heavy skewness to the right and its strong leptokurtic form. In reality, it can often approximate, e.g., the time between events or the time of events.

The probability density function of the Laplace distribution, finally, is given by

$$
\begin{equation*}
\mathrm{f}(x)=\frac{\mathrm{e}^{-|x-\mu| / b}}{2 b} \tag{4}
\end{equation*}
$$

## LANTZ ET AL

which means that it is symmetric and significantly leptokurtic with an excess kurtosis of 3. At first glance, one might think that the Laplace distribution resembles the normal distribution. The huge difference, however, is that outliers are much more common due to the fatter tails. Hence, it represents an important form of nonnormality where wild randomness exists (for realistic examples of such cases, see e.g., Mandelbrot \& Taleb, 2006).

In the simulations, the standard deviation was kept constant at 1 for all distributions in all cases while the mean values were varied to accomplish five different effect sizes in order to evaluate actual significance as well as actual power. Table 1 shows the manner in which the true mean values of the distributions were shifted to achieve a suitable range of effect sizes (see Cohen, 1992), ranging from no effect ( $f=0.00$ ) to a very large effect ( $f=0.65$ ).

The simulated data sets for the three groups were subject to individual normality screening at various significance levels based on the Shapiro-Wilk test with the Royston algorithm (Royston, 1992), that is, the default algorithm in SPSS and other statistical software. The Shapiro-Wilk test has recently been found to have the best power among the tests commonly used for normality screening (Marmolejo-Ramos \& González-Burgos, 2013; Razali \& Wah, 2011), even though other researchers recommend other tests, such as the Anderson-Darling test (see Keselman et al., 2013), for normality screening.

If the normality hypothesis for at least one group was rejected, a location test was performed with the Kruskal-Wallis test (Kruskal \& Wallis, 1952) at the 0.05 significance level. If not, a location test was performed with the omnibus one-way ANOVA at the 0.05 significance level. This two-stage procedure was repeated 100,000 times for each combination of effect size and distribution, and for three different sample sizes ( $n=15, n=30$, and $n=60$ in each group). All 100,000 data sets were also analyzed with the ANOVA without a preliminary test, as well as with the Kruskal-Wallis test without a preliminary test. This was done in both cases for each combination of effect size, distribution, and sample size. All simulation procedures were conducted using Microsoft Excel 2010.

Table 1. The different combinations of mean values

| Effect size $\boldsymbol{f}$ | $\boldsymbol{\mu}_{\mathbf{1}}$ | $\boldsymbol{\mu}_{\mathbf{2}}$ | $\boldsymbol{\mu}_{3}$ |
| ---: | ---: | ---: | ---: |
| 0.00 | 0.000 | 0.000 | 0.000 |
| 0.10 | 0.000 | 0.123 | 0.246 |
| 0.25 | 0.000 | 0.307 | 0.614 |
| 0.40 | 0.000 | 0.490 | 0.980 |
| 0.65 | 0.000 | 0.796 | 1.592 |

## PRELIMINARY TESTS OF NORMALITY

## Results

## Estimated Type I Error Probabilities

This section presents the results when all samples were drawn from distributions with the same mean value. Table 2 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure (the ANOVA without the preliminary test, the Kruskal-Wallis test without the preliminary test, or the two-stage procedure), sample size ( $n=15, n=30$, or $n=60$ in each group) and distribution (normal, uniform, exponential, or Laplace). For example, the two-stage procedure (TSP) where the preliminary Shapiro-Wilk test for normality was conducted on a significance level of 0.1 yielded 4,970 significant tests when $n=30$. Hence, the estimated type I error probability for this specific combination of distribution, test procedure, and sample size was $4.97 \%$ when samples were taken from exponential distributions.

Table 2. Estimated type I error probabilities

| Distribution | Method | $\boldsymbol{n}=\mathbf{1 5}$ | $\boldsymbol{n}=\mathbf{3 0}$ | $\boldsymbol{n}=\mathbf{6 0}$ |
| ---: | ---: | ---: | ---: | ---: |
| Normal | ANOVA | $4.99 \%$ | $4.99 \%$ | $4.93 \%$ |
|  | Kruskal-Wallis | $4.79 \%$ | $4.86 \%$ | $4.86 \%$ |
|  | TSP 0.1 | $5.13 \%$ | $5.17 \%$ | $5.06 \%$ |
|  | TSP 0.05 | $5.14 \%$ | $5.15 \%$ | $5.05 \%$ |
|  | TSP 0.01 | $5.06 \%$ | $5.06 \%$ | $4.97 \%$ |
| Uniform | TSP 0.005 | $5.03 \%$ | $5.03 \%$ | $4.96 \%$ |
|  | ANOVA | $5.12 \%$ | $5.10 \%$ | $5.08 \%$ |
|  | Kruskal-Wallis | $4.72 \%$ | $4.90 \%$ | $4.99 \%$ |
|  | TSP 0.1 | $4.88 \%$ | $4.93 \%$ | $4.99 \%$ |
|  | TSP 0.05 | $5.00 \%$ | $4.96 \%$ | $4.99 \%$ |
|  | TSP 0.01 | $5.09 \%$ | $5.08 \%$ | $5.01 \%$ |
|  | TSP 0.005 | $5.11 \%$ | $5.11 \%$ | $5.03 \%$ |
|  | ANOVA | $4.46 \%$ | $4.59 \%$ | $4.74 \%$ |
|  | Kruskal-Wallis | $4.75 \%$ | $4.83 \%$ | $4.85 \%$ |
|  | TSP 0.1 | $4.82 \%$ | $4.83 \%$ | $4.85 \%$ |
|  | TSP 0.05 | $4.95 \%$ | $4.83 \%$ | $4.85 \%$ |
|  | TSP 0.01 | $5.49 \%$ | $4.84 \%$ | $4.85 \%$ |
|  | TSP 0.005 | $5.69 \%$ | $4.87 \%$ | $4.85 \%$ |
|  | ANOVA | $4.77 \%$ | $4.93 \%$ | $4.90 \%$ |
|  | Lruskal-Wallis | $4.78 \%$ | $4.93 \%$ | $4.92 \%$ |
|  | TSP 0.1 | $4.92 \%$ | $4.97 \%$ | $4.92 \%$ |
|  | TSP 0.05 | $4.97 \%$ | $5.00 \%$ | $4.92 \%$ |
|  | TSP 0.01 | $5.00 \%$ | $5.04 \%$ | $4.97 \%$ |
|  | TSP 0.005 | $4.96 \%$ | $5.06 \%$ | $4.98 \%$ |

The overall picture seems to be that the two pure tests both perform in a similar way as the two stage process. The only, but rather minor, exception seems to be that that the two stage process generates slightly more type I errors when samples of a small size are drawn from an exponential distribution. This tendency is amplified when the preliminary test is conducted at a smaller significance level, but diminishes when the sample size becomes larger. The reason is probably that the normality screening of samples taken from exponential distributions favors samples with smaller standard deviations (Rochon \& Kieser, 2011).

## Estimated Power under Exponential Distribution

This section presents the results when all samples were drawn from exponential distributions with different mean values. Table 3 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure (the ANOVA without the preliminary test, the Kruskal-Wallis test without the preliminary test, or the two-stage procedure), sample size ( $n=15$, $n=30$, or $n=60$ in each group), and effect size ( $f=0.10, f=0.25, f=0.40$, or $f=0.65$ ). For example, the ANOVA without the preliminary test yielded 12,740 significant tests when the effect size was $f=0.10$ and when $n=30$. Hence, the proportion of significant tests for this specific combination test procedure, sample

Table 3. Estimated power under the exponential distribution

| Method | Sample size $\boldsymbol{n}$ | $\boldsymbol{f}=\mathbf{0 . 1 0}$ | $\boldsymbol{f}=\mathbf{0 . 2 5}$ | $\boldsymbol{f}=\mathbf{0 . 4 0}$ | $\boldsymbol{f}=\mathbf{0 . 6 5}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| ANOVA | 15 | $8.45 \%$ | $32.05 \%$ | $66.47 \%$ | $95.54 \%$ |
|  | 30 | $12.74 \%$ | $56.37 \%$ | $91.57 \%$ | $99.65 \%$ |
| Kruskal-Wallis | 60 | $21.05 \%$ | $85.11 \%$ | $99.20 \%$ | $99.42 \%$ |
|  | 15 | $13.34 \%$ | $52.24 \%$ | $85.71 \%$ | $99.18 \%$ |
| TSP 0.1 | 30 | $24.15 \%$ | $84.81 \%$ | $99.05 \%$ | $99.69 \%$ |
|  | 60 | $45.35 \%$ | $98.63 \%$ | $99.42 \%$ | $99.42 \%$ |
|  | 15 | $13.40 \%$ | $52.27 \%$ | $85.70 \%$ | $99.18 \%$ |
| TSP 0.05 | 30 | $24.15 \%$ | $84.81 \%$ | $99.05 \%$ | $99.69 \%$ |
|  | 60 | $45.35 \%$ | $98.63 \%$ | $99.42 \%$ | $99.42 \%$ |
|  | 15 | $13.53 \%$ | $52.29 \%$ | $85.64 \%$ | $99.17 \%$ |
| TSP 0.01 | 30 | $24.15 \%$ | $84.81 \%$ | $99.05 \%$ | $99.69 \%$ |
|  | 60 | $45.35 \%$ | $98.63 \%$ | $99.42 \%$ | $99.42 \%$ |
|  | 15 | $13.82 \%$ | $51.53 \%$ | $84.66 \%$ | $98.97 \%$ |
| TSP 0.005 | 30 | $24.16 \%$ | $84.80 \%$ | $99.05 \%$ | $99.69 \%$ |
|  | 60 | $45.35 \%$ | $98.63 \%$ | $99.42 \%$ | $99.42 \%$ |
|  | 15 | $13.71 \%$ | $50.44 \%$ | $83.43 \%$ | $98.76 \%$ |
|  | 30 | $24.18 \%$ | $84.78 \%$ | $99.04 \%$ | $99.69 \%$ |
|  | 60 | $45.35 \%$ | $98.63 \%$ | $99.42 \%$ | $99.42 \%$ |

## PRELIMINARY TESTS OF NORMALITY

size, and effect size was $12.74 \%$ when samples were taken from exponential distributions.

The two-stage procedure (regardless of the significance level of the ShapiroWilk test) and the Kruskal-Wallis test perform similarly at all combinations of effect size and sample size. However, the ANOVA has substantially less power than both other procedures. Furthermore, this pattern remains the same regardless of the sample size. The main reason is of course that the preliminary normality screening in the two-stage procedure in most cases favors the Kruskal-Wallis test at the second stage.

## Estimated Power under Laplace Distribution

This section presents the results when all samples were drawn from Laplace distributions with different mean values. Table 4 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

As when samples were drawn from exponential distributions, the ANOVA has a lower power than both the two-stage procedure (regardless of the significance level of the Shapiro-Wilk test) and the Kruskal-Wallis test when the samples come from Laplace distributions. The effect is somewhat smaller, however.

Table 4. Estimated power under the Laplace distribution

| Method | Sample size $\boldsymbol{n}$ | $\boldsymbol{f}=\mathbf{0 . 1 0}$ | $\boldsymbol{f}=\mathbf{0 . 2 5}$ | $\boldsymbol{f}=\mathbf{0 . 4 0}$ | $\boldsymbol{f}=\mathbf{0 . 6 5}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| ANOVA | 15 | $8.32 \%$ | $30.06 \%$ | $64.72 \%$ | $96.08 \%$ |
|  | 30 | $12.34 \%$ | $54.84 \%$ | $91.80 \%$ | $99.41 \%$ |
| Kruskal-Wallis | 60 | $20.70 \%$ | $84.70 \%$ | $98.73 \%$ | $98.90 \%$ |
|  | 15 | $9.56 \%$ | $37.45 \%$ | $74.34 \%$ | $98.09 \%$ |
|  | 30 | $15.70 \%$ | $68.64 \%$ | $96.86 \%$ | $99.44 \%$ |
| TSP 0.1 | 60 | $28.16 \%$ | $93.86 \%$ | $98.89 \%$ | $98.90 \%$ |
|  | 15 | $9.54 \%$ | $36.47 \%$ | $72.93 \%$ | $97.90 \%$ |
|  | 30 | $15.50 \%$ | $67.61 \%$ | $96.51 \%$ | $99.44 \%$ |
| TSP 0.05 | 60 | $28.08 \%$ | $93.72 \%$ | $98.89 \%$ | $98.90 \%$ |
|  | 15 | $9.49 \%$ | $35.60 \%$ | $71.92 \%$ | $97.72 \%$ |
|  | 30 | $15.29 \%$ | $66.64 \%$ | $96.18 \%$ | $99.44 \%$ |
| TSP 0.01 | 60 | $27.94 \%$ | $93.55 \%$ | $98.89 \%$ | $98.90 \%$ |
|  | 15 | $9.19 \%$ | $33.53 \%$ | $69.24 \%$ | $97.20 \%$ |
|  | 30 | $14.65 \%$ | $63.70 \%$ | $95.20 \%$ | $99.43 \%$ |
| TSP 0.005 | 60 | $27.15 \%$ | $92.65 \%$ | $98.88 \%$ | $98.90 \%$ |
|  | 15 | $9.01 \%$ | $32.74 \%$ | $68.23 \%$ | $96.97 \%$ |
|  | 30 | $14.27 \%$ | $62.36 \%$ | $94.72 \%$ | $99.43 \%$ |
|  | 60 | $26.70 \%$ | $92.09 \%$ | $98.86 \%$ | $98.90 \%$ |

## LANTZ ET AL

The Kruskal-Wallis test has slightly higher power than the two-stage procedure, and this effect is amplified when the preliminary test is conducted at a smaller significance level irrespective of the sample size. The reason is probably that normality screening at a lower significance level favors the ANOVA at the second stage because the Laplace distribution, despite its leptokurtic shape, resembles the normal distribution more than the exponential distribution does due to its symmetry and unimodality.

## Estimated Power under Uniform Distribution

This section presents the results when all samples were drawn from uniform distributions with different mean values. Table 5 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

In line with previous research (see Schmider, Ziegler, Danay, Beyer, \& Buhner, 2010, for a review), the ANOVA shows slightly higher power than the Kruskal-Wallis test when samples are drawn from uniform distributions. The main reason is of course that the uniform distribution, in terms of skewness and/or kurtosis, does not impose an equally serious violation of normality as the Laplace and exponential distributions do.

Table 5. Estimated power under the uniform distribution

| Method | Sample size $\boldsymbol{n}$ | $\boldsymbol{f}=\mathbf{0 . 1 0}$ | $\boldsymbol{f}=\mathbf{0 . 2 5}$ | $\boldsymbol{f}=\mathbf{0 . 4 0}$ | $\boldsymbol{f}=\mathbf{0 . 6 5}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| ANOVA | 15 | $8.25 \%$ | $27.83 \%$ | $62.92 \%$ | $97.66 \%$ |
|  | 30 | $12.13 \%$ | $53.63 \%$ | $92.81 \%$ | $100.00 \%$ |
| Kruskal-Wallis | 60 | $20.35 \%$ | $85.70 \%$ | $99.91 \%$ | $100.00 \%$ |
|  | 15 | $7.75 \%$ | $25.05 \%$ | $56.24 \%$ | $94.61 \%$ |
|  | 30 | $11.53 \%$ | $49.05 \%$ | $88.43 \%$ | $99.96 \%$ |
| TSP 0.1 | 60 | $19.43 \%$ | $81.37 \%$ | $99.64 \%$ | $100.00 \%$ |
|  | 15 | $8.08 \%$ | $26.69 \%$ | $59.52 \%$ | $95.37 \%$ |
|  | 30 | $11.62 \%$ | $49.56 \%$ | $88.70 \%$ | $99.96 \%$ |
| TSP 0.05 | 60 | $19.43 \%$ | $81.37 \%$ | $99.64 \%$ | $100.00 \%$ |
|  | 15 | $8.21 \%$ | $27.41 \%$ | $61.28 \%$ | $96.16 \%$ |
|  | 30 | $11.77 \%$ | $50.59 \%$ | $89.37 \%$ | $99.96 \%$ |
| TSP 0.01 | 60 | $19.43 \%$ | $81.38 \%$ | $99.64 \%$ | $100.00 \%$ |
|  | 15 | $8.28 \%$ | $27.84 \%$ | $62.83 \%$ | $97.40 \%$ |
|  | 30 | $12.13 \%$ | $53.05 \%$ | $91.67 \%$ | $99.97 \%$ |
| TSP 0.005 | 60 | $19.64 \%$ | $81.88 \%$ | $99.65 \%$ | $100.00 \%$ |
|  | 15 | $8.28 \%$ | $27.85 \%$ | $62.91 \%$ | $97.57 \%$ |
|  | 30 | $12.16 \%$ | $53.49 \%$ | $92.30 \%$ | $99.98 \%$ |
|  | 60 | $19.89 \%$ | $82.69 \%$ | $99.67 \%$ | $100.00 \%$ |

## PRELIMINARY TESTS OF NORMALITY

In general, the ANOVA performs somewhat better than the two-stage procedure while the Kruskal-Wallis test performs somewhat worse. As one might expect, the performance of the two-stage procedure approaches the performance of the ANOVA when the normality tests are conducted at a lower significance level as that favors the ANOVA at the second stage. However, the difference in performance between the Kruskal-Wallis test and the two-stage procedure also diminishes when the sample size is larger.

## Estimated Power under Normal Distribution

This section presents the results when all samples were drawn from normal distributions with different mean values. Table 6 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

As one would expect, the Kruskal-Wallis test performs somewhat worse than the ANOVA. The two-stage procedure on the other hand has a performance very similar to the ANOVA, which is easy to understand as the Shapiro-Wilk test only favors the Kruskal-Wallis test at the second stage in a few cases.

Table 6. Estimated power under the normal distribution

| Method | Sample size $\boldsymbol{n}$ | $\boldsymbol{f}=\mathbf{0 . 1 0}$ | $\boldsymbol{f}=\mathbf{0 . 2 5}$ | $\boldsymbol{f}=\mathbf{0 . 4 0}$ | $\boldsymbol{f}=\mathbf{0 . 6 5}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
| ANOVA | 15 | $8.29 \%$ | $28.55 \%$ | $63.53 \%$ | $97.05 \%$ |
|  | 30 | $12.23 \%$ | $54.16 \%$ | $92.60 \%$ | $99.99 \%$ |
| Kruskal-Wallis | 60 | $20.50 \%$ | $85.40 \%$ | $99.87 \%$ | $100.00 \%$ |
|  | 15 | $7.80 \%$ | $26.63 \%$ | $60.49 \%$ | $96.13 \%$ |
|  | 30 | $11.65 \%$ | $51.75 \%$ | $91.21 \%$ | $99.98 \%$ |
| TSP 0.1 | 60 | $19.71 \%$ | $83.61 \%$ | $99.82 \%$ | $100.00 \%$ |
|  | 15 | $8.43 \%$ | $28.48 \%$ | $63.15 \%$ | $96.90 \%$ |
|  | 30 | $12.38 \%$ | $53.85 \%$ | $92.30 \%$ | $99.99 \%$ |
| TSP 0.05 | 60 | $20.65 \%$ | $85.00 \%$ | $99.86 \%$ | $100.00 \%$ |
|  | 15 | $8.45 \%$ | $28.69 \%$ | $63.47 \%$ | $97.02 \%$ |
|  | 30 | $12.39 \%$ | $54.15 \%$ | $92.51 \%$ | $99.99 \%$ |
| TSP 0.01 | 60 | $20.68 \%$ | $85.26 \%$ | $99.87 \%$ | $100.00 \%$ |
|  | 15 | $8.37 \%$ | $28.67 \%$ | $63.63 \%$ | $97.10 \%$ |
|  | 30 | $12.31 \%$ | $54.28 \%$ | $92.65 \%$ | $99.99 \%$ |
| TSP 0.005 | 60 | $20.60 \%$ | $85.45 \%$ | $99.88 \%$ | $100.00 \%$ |
|  | 15 | $8.36 \%$ | $28.64 \%$ | $63.63 \%$ | $97.10 \%$ |
|  | 30 | $12.29 \%$ | $54.25 \%$ | $92.66 \%$ | $99.99 \%$ |
|  | 60 | $20.58 \%$ | $85.46 \%$ | $99.88 \%$ | $100.00 \%$ |

## LANTZ ET AL

## Conclusion

A preliminary test of normality before conducting a location test will yield one of four possible outcomes:

- Incorrectly rejecting $\mathrm{H}_{0}$ (i.e. a type I error), resulting in the use of a location test with less power than necessary at the second stage.
- Correctly rejecting $\mathrm{H}_{0}$, resulting in the (correct) use of a non-parametric location test at the second stage.
- Incorrectly 'accepting' $\mathrm{H}_{0}$ (i.e. a type II error), resulting in the use of an invalid location test (i.e. with uncertain actual power and significance) at the second stage.
- Correctly 'accepting' $\mathrm{H}_{0}$, resulting in the (correct) use of a parametric location test at the second stage.

Therefore, the probability of a type I error as well as of a type II error of the entire two-stage procedure may be heavily distorted, if it is at all possible to determine. In this study, we have used simulations in order to shed some light on this problem. While we have been unable to see any specific disturbance in the type I error probability of the two-stage procedure, the effect on power exhibits some interesting patterns in comparison to the 'pure' methods. The overall impression is that the two-stage procedure performs similarly to the ANOVA, but slightly better than the Kruskal-Wallis test when the parent distributions are 'relatively normally' distributed. On the other hand, the two-stage procedure performs similarly to the Kruskal-Wallis test, but substantially better than the ANOVA, when the parent distributions are characterized by a more distinct violation of normality. These observed patterns are also relatively insensitive to the sample sizes.

The choice of level of significance for the preliminary tests also requires some thought. If we, for example, want to compare six groups and choose to use $\alpha=0.10$ during the normality screening, the overall probability of a type I error, leading us to use a less powerful non-parametric test to compare the means in the second stage, would be around $50 \%$. On the other hand, since the ANOVA typically perform a lot worse than the Kruskal-Wallis when there is a more distinct violation of normality while the Kruskal-Wallis only perform slightly worse when normality actually holds, type II errors are potentially a lot more harmful than type I errors in the first stage of the two-stage procedure.

## PRELIMINARY TESTS OF NORMALITY

Hence, in contrast to previous similar research on bivariate situations, the two-stage procedure seems in general to be the preferable choice when conducting location tests for three samples as neither the ANOVA nor the Kruskal-Wallis test as one-stage procedures perform noticeably better than the two-stage procedure, while the two-stage procedure is substantially better than the ANOVA when data are distinctly non-normally distributed. This is especially so when the normality screening is conducted at a relatively high significance level. Hence, the two-stage procedure seems to have no practical shortcoming but an apparent practical advantage. The theoretical weakness, of course, is that the true probability of type I and type II errors may be unknown, which, in addition to the fact that the ANOVA is known to be relatively robust to non-normally distributed data when groups sizes are roughly equal, albeit more sensitive to non-normality when group sizes are unequal, should be borne in mind (Schmider et al., 2010; Wilcox, 2012; Field, 2013).

Future research should extend the design in this study, for example, by using different sample sizes in the groups, and/or by including other statistical distributions in order to evaluate other types of non-normality than those related to skewness and kurtosis. Further research in this field should also aim at comparing other types of parametric methods with their non-parametric counterparts as twostage procedures, as well as comparing two-stage procedures with robust procedures in general such as bootstrapping. Screening for other types of violations, for example, heteroscedasticity, in the first stage would also be interesting to consider.

## References

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., \& Bendayan, R. (2013). Skewness and kurtosis in real data samples. Methodology, 9(2), 78-84. doi: 10.1027/1614-2241/a000057

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ( $p<.05$ ). American Psychologist, 49(12), 997-1003. doi: 10.1037/0003-066X.49.12.997

Field, A. (2013). Discovering statistics using IBM SPSS Statistics. London: SAGE Publications.

Keselman, H. J., Othman, A. R., \& Wilcox, R. R. (2013). Preliminary testing for normality: Is this a good practice? Journal of Modern Applied

## LANTZ ET AL

Statistical Methods, 12(2), 2-19. Retrieved from
http://digitalcommons.wayne.edu/jmasm/vol12/iss2/2/
Keselman, H. J., Othman, A. R., \& Wilcox, R. R. (2014). Testing for normality in the multi-group problem: Is this a good practice? Clinical Dermatology, 2(1), 29-43. doi: 10.11138/cderm/2014.2.1.029

Kruskal, W. H., \& Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583621. doi: 10.2307/2280779

Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods. British Journal of Mathematical and Statistical Psychology, 66(2), 224-244. doi: 10.1111/j.2044-8317.2012.02047.x

Mandelbrot, B., \& Taleb, N. (2006, March 24). A focus on the exceptions that prove the rule. Financial Times.

Marmolejo-Ramos, F., \& González-Burgos, J. (2013). A power comparison of various tests of univariate normality on ex-Gaussian distributions. Methodology, 9(4), 137-149. doi: 10.1027/1614-2241/a000059

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156-166. doi: 10.1037/00332909.105.1.156

Othman, A. R., Keselman, H. J., \& Wilcox, R. R. (2015). Assessing normality: Applications in multi-group designs. Malaysian Journal of Mathematical Sciences, 9(1), 53-65. Retrieved from http://einspem.upm.edu.my/journal/volume9.1.php

Rasch, D., Kubinger, K. D., \& Moder, K. (2011). The two-sample $t$ test: Pre-testing its assumptions does not pay off. Statistical Papers, 52(1), 219-231. doi: 10.1007/s00362-009-0224-x

Razali, N., \& Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2(1), 21-33.

Rochon, J., Gondan, M., \& Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. BMC Medical Research Methodology, 12(81). doi: 10.1186/1471-2288-12-81

Rochon, J., \& Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample $t$-test. British Journal of Mathematical and Statistical Psychology, 64(3), 410-426. doi: 10.1348/20448317.002003

## PRELIMINARY TESTS OF NORMALITY

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for nonnormality. Statistics and Computing, 2(3), 117-119. doi: 10.1007/BF01891203

Ruscio, J., \& Roche, B (2012). Variance heterogeneity in published psychological research. Methodology, 8(1), 1-11. doi: 10.1027/16142241/a000034

Schmider, E., Ziegler, M., Danay, E., Beyer, L., \& Buhner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology, 6(4), 147-151. doi: 10.1027/16142241/a000016

Schoder, V., Himmelmann, A., \& Wilhelm, K. P. (2006). Preliminary testing for normality: Some statistical aspects of a common concept. Clinical and Experimental Dermatology, 31(6), 757-761. doi: 10.1111/j.1365-
2230.2006.02206.x

Schucany, W. R., \& Ng, T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student $t$. Communications in Statistics - Theory and Methods, 35(12), 2275-2286. doi: 10.1080/03610920600853308

Shuster, J. J. (2005). Diagnostics for assumptions in moderate to large simple clinical trials: Do they really help? Statistics in Medicine, 24(16), 24312438. doi: 10.1002/sim. 2175

Shuster, J. J. (2009). Student $t$-tests for potentially abnormal data. Statistics in Medicine, 28(16), 2170-2184. doi: 10.1002/sim. 3581

Wells, C. S., \& Hinze, J. M. (2006). Dealing with assumptions underlying statistical tests. Psychology in the Schools, 44(5), 495-502. doi:
10.1002/pits. 20241

Wilcox, R. R. (2012). Introduction to robust estimation and hypothesis testing. San Diego, CA: Academic Press.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. British Journal of Mathematical and Statistical Psychology, 57(1), 173-181. doi: 10.1348/000711004849222

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. British Journal of Mathematical and Statistical Psychology, 64(3), 388-409. doi: 10.1348/000711010X524739

Zimmerman, D. W. (2014). Consequences of choosing samples in hypothesis testing to ensure homogeneity of variance. British Journal of Mathematical and Statistical Psychology, 67(1), 1-29. doi: 10.1111/bmsp. 12001

# Estimation of Population Mean on Recent Occasion under Non-Response in h-Occasion Successive Sampling 

Anup Kumar Sharma<br>Indian School of Mines<br>Dhanbad, India

Garib Nath Singh<br>Indian School of Mines<br>Dhanbad, India


#### Abstract

In this article, an attempt has been made to study on general estimation procedures of population mean on recent occasion when non-response occurs in $h$-occasion successive sampling. Suggested estimators have advantageously influenced the estimation procedures in the presence of non-response. Detailed properties of the suggested estimation procedures have been examined and compared with the estimation process of the same circumstances but in the absence of non-response. Empirical studies have been carried out to demonstrate the performances of the estimates and suitable recommendations have been made.


Keywords: Non-response, successive sampling, study variable, variance

## Introduction

Successive sampling was developed for estimation of population parameters on recent point of time (occasion), when the population parameters changes over successive points of time (occasion). It is a sampling method to provide reliable and fruitful estimates of population parameters over different desire points of time (occasion). Jessen (1942) initiated a technique with the help of past information to provide the effective estimates on current occasion in two-occasion successive sampling. Later, this technique was extended by Yates (1949), Patterson (1950), Tikkiwal (1951), Eckler (1955), Rao and Graham (1964), Gupta (1979), Binder and Hidiroglou (1988), Kish (1998), McLaren and Steel (2000), Singh, Kennedy and Wu (2001), Steel and McLaren (2002) among others. Sen (1971, 1973) applied this theory in designing the estimators of population mean using information on two or more auxiliary variables which was readily available on

[^8]
## ESTIMATION OF POPULATION MEAN UNDER NON-RESPONSE

previous occasion in two-occasion successive sampling. Singh, Singh and Shukla (1991), Singh and Singh (2001) made an efficient use of auxiliary variable on current occasion and subsequently Singh (2003) uses this methodology for $h$ occasion successive sampling in estimation of current population mean.

In many situations, information on an auxiliary variable may be readily available on the first as well as on the second occasion. Utilizing the auxiliary information on both occasions, Feng and Zou (1997), Biradar and Singh (2001), Singh (2005), Singh and Karna (2009), Singh and Prasad (2010), Singh, Prasad, and Karna (2011), Singh, Majhi, Maurya, and Sarma (2015) and Singh and Sharma $(2014,2015)$ have proposed several estimators of population mean on current (second) occasion in two-occasion successive sampling.

Non-response is a common problem almost encountered in all sample surveys and successive sampling is more prone to this problem because of its repetitive nature. For example, in agriculture yield surveys, it might be possible that crop on certain plots are destroyed due to some natural calamities or disease so that yield on these plots are impossible to be measured. Hansen and Hurwitz (1946) suggested a method of sub sampling of non-respondents to address the problems of non-response in mail surveys. Later on Cochran (1977) and Okafor and Lee (2000) extended this technique for the case when besides the information on character under study, information is also available on one auxiliary character. More recently, Choudhary, Bathla, and Sud (2004), Singh and Priyanka (2007), and Singh and Kumar (2008) used the Hansen and Hurwitz (1946) technique for the estimation of population mean on current occasion in context of sampling on two occasions.

Motivated with the above arguments and using Hansen and Hurwitz (1946) method, the aim of the present work is to suggest the estimation procedure for population mean at $h^{\text {th }}$ (recent) occasion when the non-response occurs on $h^{\text {th }}$ occasion, $(h-1)^{\text {th }}$ (previous) occasion and simultaneously on both $h^{\text {th }}$ and $(h-1)^{\text {th }}$ occasions in $h$-occasion successive (rotation) sampling. The properties of the proposed estimation procedure have been examined and compared with the similar estimation but under complete response. Empirical studies are carried out and suitable recommendations have been made.

## Notations

Let $U=\left(U_{1}, U_{2},--, U_{\mathrm{N}}\right)$ be the finite population of $N$ units, which has been sampled over $h$ occasions. The character under studies are denoted by $y_{h}$ and $y_{h-1}$ on the $h^{\text {th }}$ and $(h-1)^{\text {th }}$ occasions respectively. Assume that the non-response occur

## SHARMA \& SINGH

on $h^{\text {th }}$ occasion, $(h-1)^{\text {th }}$ occasion and simultaneously on both $h^{\text {th }}$ and $(h-1)^{\text {th }}$ occasions, so that the population can be divided into two classes, those who will respond at the first attempt and those who will not. Let the sizes of these two classes be $N_{h}$ and $N^{*}{ }_{h}$ on the $h^{\text {th }}$ occasion and the corresponding sizes on $(h-1)^{\text {th }}$ occasion be $N_{h-1}$ and $N^{*}{ }_{h-1}$. Let a simple random sample (without replacement) of size $n$ be selected on the $h^{\text {th }}$ occasion which consist of $n_{h}^{\prime}=n \lambda_{h}$ units common to the units observed on the $(h-1)^{\text {th }}$ occasion and $n_{h}^{\prime \prime}=n \mu_{h}$ units drawn afresh on the $h^{\text {th }}$ occasion i.e. $n=n_{h}^{\prime}+n_{h}^{\prime \prime}$. Here $\lambda_{h}$ and $\mu_{h}\left(\lambda_{h}+\mu_{h}=1\right)$ are the fractions of matched and unmatched samples, respectively, on the $h^{\text {th }}$ occasion. The values of $\lambda_{h}$ and $\mu_{h}$ should be chosen optimally. Assume that in the unmatched portion of the sample on the $h^{\text {th }}$ occasion $n_{h}^{\prime \prime}$ units respond and $n_{h 2}^{\prime \prime}$ units do not respond. Let $n_{h 2 s}^{\prime \prime}$ denote the size of sub sample drawn from the non-response class in the unmatched portion of the sample on the $h^{\text {th }}$ occasion and their response collected by direct contact or interview. Similarly, $n_{h 1}^{\prime}$ units respond and $n_{h 2}^{\prime}$ units do not respond in the sample of matched units and let $n_{h 2 s}^{\prime}$ denote the size of sub sample drawn from the non-response class in the matched portion of the sample on the $h^{\text {th }}$ occasion and their response collected by direct contact or interview. Following are the list of notations, which are considered for their further use:
$\bar{Y}_{h}: \quad$ The population mean of the study variable $y_{h}$ on the $h^{\text {th }}$ occasion.
The sample mean of the study variable based on $n_{h}^{\prime}$ units common to the units observed on the $(h-1)^{\mathrm{th}}$ occasion.

The sample mean of the study variable based on $n_{h}^{\prime \prime}$ units drawn afresh on the $h^{\text {th }}$ occasion.
$\rho_{h, h-1}: \quad$ The correlation between the study variables $y_{h}$ and $y_{h-1}$.
$S_{h y}^{2}:$
$W_{h-1}^{*}=\frac{N_{h-1}^{*}}{N}: \quad$ The proportion of non-responding units in the population on the $(h-1)^{\text {th }}$ occasion.
$W_{h}^{*}=\frac{N_{h}^{*}}{N}: \quad$ The proportion of non-responding units in the population on the $h^{\text {th }}$ occasion.
$f\left(\frac{n}{N}\right): \quad$ The sampling fraction.
$f_{1}=\frac{n_{h 2}^{\prime}}{n_{h 2 s}^{\prime}}$.
$f_{2}=\frac{n_{h 2}^{\prime \prime}}{n_{h 2 s}^{\prime \prime}}$.

## Formulation of Estimator

For estimating the population mean $\bar{Y}$ on the $h^{\text {th }}$ occasion, a sample mean and a regression type estimator are suggested. First is the Hansen and Hurwitz (1946) type estimator, say $\omega_{h}^{\prime \prime}$, which is based on $n_{h}^{\prime \prime}$ sample units drawn afresh on $h^{\text {th }}$ occasion such that out of these $n_{h}^{\prime \prime}$ units, $n_{h 1}^{\prime \prime}$ units respond and remaining $n_{h 2}^{\prime \prime}\left(=n_{h}^{\prime \prime}-n_{h 1}^{\prime \prime}\right)$ units do not respond. Hence, $\omega_{h}^{\prime \prime}$ is defined as

$$
\begin{equation*}
\omega_{h}^{\prime \prime}=\bar{y}_{h}^{\prime \prime *} \tag{1}
\end{equation*}
$$

where

$$
\bar{y}_{h}^{\prime^{\prime *}}=\frac{n_{h 1}^{\prime \prime} \bar{y}_{1}^{\prime \prime}+n_{h 2}^{\prime \prime} \bar{y}_{h 2 s}^{\prime \prime}}{n_{h}^{\prime \prime}}
$$

The second estimator is based on the sample of size $n_{h}^{\prime}$, which is common to the units observed on the $(h-1)^{\text {th }}$ occasion. Because non-response is occurred on the previous occasion, therefore, again Hansen and Hurwitz (1946) type estimator are considered. The second estimator, $\omega_{h}^{\prime}$, for estimating the population mean on $h^{\text {th }}$ occasion is a regression type estimator, and is defined as

$$
\begin{equation*}
\omega_{h}^{\prime}=\bar{y}_{h}^{\prime *}+\beta_{h, h-1}\left(\omega_{h-1}-\bar{y}_{h-1}^{\prime *}\right) \tag{2}
\end{equation*}
$$

where

## SHARMA \& SINGH

$$
\bar{y}_{h}^{\prime *}=\frac{n_{h 1}^{\prime} \bar{y}_{h 1}^{\prime}+n_{n 2}^{\prime} \bar{y}_{h 2}^{\prime}}{n_{h}^{\prime}}, \bar{y}_{h-1}^{\prime *}=\frac{n_{h 1}^{\prime} \bar{y}_{(h-1) 1}^{\prime}+n_{h 2}^{\prime} \bar{y}_{(h-1) 2}^{\prime}}{n_{h}^{\prime}}
$$

and $\beta_{h, h-1}$ is population regression coefficient between the study variable $y_{h}$ and $y_{h-1}$.

The resulting estimator $\omega_{h}$ is a convex linear combination of the estimators $\omega_{h}^{\prime \prime}$ and $\omega_{h}^{\prime}$. The estimator $\omega_{h}$ is defined as

$$
\begin{equation*}
\omega_{h}=\varphi_{h} \omega_{h}^{\prime \prime}+\left(1-\varphi_{h}\right) \omega_{h}^{\prime} \tag{3}
\end{equation*}
$$

where $\varphi_{h}\left(0 \leq \varphi_{h} \leq 1\right)$ is the unknown constant to be determined under certain criterion.

Remark 1: For estimating the mean on $h^{\text {th }}$ occasion the estimator $\omega_{h}^{\prime \prime}$ is suitable, which implies that more belief on $\omega_{h}^{\prime \prime}$ could be shown by choosing $\varphi_{h}$ as 1 (or close to 1 ), while for estimating the change from one occasion to the next, the estimator $\omega_{h}^{\prime}$ could be more useful so $\varphi_{h}$ might be chosen as 0 (or close to 0 ). For asserting both the problems simultaneously, the suitable (optimum) choice of $\varphi_{h}$ is required.

Remark 2: (i) Assume that the correlation between variables observed on two occasions, more than one occasion apart is zero. (ii) For practical application the population regression coefficient will be estimated by their respective sample estimates.

## Properties of the Estimator $\omega_{h}$

Because $\omega_{h}^{\prime \prime}$ and $\omega_{h}^{\prime}$ are sample mean and difference type estimators respectively, they are unbiased for population mean $\bar{Y}_{h}$. Therefore, the resulting estimator $\omega_{h}$ defined in equation (3) is also an unbiased estimator of $\bar{Y}_{h}$. The variance of the estimator $\omega_{h}$ is shown in following theorem.

Theorem 1: Variance of the estimator $\omega_{h}$ to the first order of approximations is obtained as

$$
\begin{equation*}
V\left(\omega_{h}\right)=\varphi_{h}^{2} V\left(\omega_{h}^{\prime \prime}\right)+\left(1-\varphi_{h}\right)^{2} V\left(\omega_{h}^{\prime}\right)+2 \varphi_{h}\left(1-\varphi_{h}\right) C\left(\omega_{h}^{\prime \prime}, \omega_{h}^{\prime}\right) \tag{4}
\end{equation*}
$$

where

$$
\begin{gather*}
V\left(\omega_{h}^{\prime \prime}\right)=\left[\frac{1}{n_{h}^{\prime \prime}}\left\{1+W_{h}^{*}\left(f_{2}-1\right)\right\}-\frac{1}{N}\right] S_{h y}^{2}  \tag{5}\\
V\left(\omega_{h}^{\prime}\right)=\left[\frac{1}{n_{h}^{\prime}}\left\{1+W_{h-1}^{*}\left(f_{1}-1\right)\right\}\left(1-\rho_{h, h-1}^{2}\right)+\frac{\varphi_{h-1}}{n_{h-1}^{\prime \prime}} \rho_{h, h-1}^{2}-\frac{1}{N}\right] S_{h y}^{2} \tag{6}
\end{gather*}
$$

and

$$
\begin{equation*}
C\left(\omega_{h}^{\prime \prime}, \omega_{h}^{\prime}\right)=-\frac{1}{N} S_{h y}^{2} \tag{7}
\end{equation*}
$$

Remark 3: Following Hansen and Hurwitz (1946) technique, some variances which are used in Theorem 1, are evaluated as given below:

$$
\begin{aligned}
V\left(\bar{y}_{h}^{\prime \prime *}\right) & =V\left[E\left(\bar{y}_{h}^{\prime \prime *} \mid n_{h 1}^{\prime \prime}, n_{h 2}^{\prime \prime}\right)\right]+E\left[V\left(\bar{y}_{h}^{\prime \prime *} \mid n_{h 1}^{\prime \prime}, n_{h 2}^{\prime \prime}\right)\right] \\
& =V\left(\bar{y}_{h}^{\prime \prime}\right)+E\left[\frac{n_{h 2}^{\prime \prime}}{n_{h}^{\prime \prime 2}}\left(f_{2}-1\right) S_{h y}^{2}\left(n_{h 2}^{\prime \prime}\right)\right] \\
& =\left(\frac{1}{n_{h}^{\prime \prime}}-\frac{1}{N}\right) S_{h y}^{2}+\frac{\left(f_{2}-1\right)}{n^{\prime \prime}} S_{h y}^{2}\left(N_{h}^{*}\right) \frac{N_{h}^{*}}{N}
\end{aligned}
$$

where $S_{h y}^{2}\left(N_{h}^{*}\right)$ is the population variance of non response class on $h^{\text {th }}$ occasion. Further we assume that $S_{h y}^{2}\left(N_{h}^{*}\right)=S_{h y}^{2}$, and hence

$$
\begin{equation*}
V\left(\bar{y}_{h}^{\prime \prime *}\right)=\left[\left(\frac{1}{n_{h}^{\prime \prime}}-\frac{1}{N}\right)+\frac{W_{h}^{*}\left(f_{2}-1\right)}{n_{h}^{\prime \prime}}\right] S_{h y}^{2} \tag{8}
\end{equation*}
$$

Similarly

## SHARMA \& SINGH

$$
\begin{gather*}
V\left(\bar{y}_{h}^{\prime *}\right)=\left[\left(\frac{1}{n_{h}^{\prime}}-\frac{1}{N}\right)+\frac{W_{h-1}^{*}\left(f_{1}-1\right)}{n_{h}^{\prime}}\right] S_{h y}^{2}  \tag{9}\\
V\left(\bar{y}_{h-1}^{\prime *}\right)=\left[\left(\frac{1}{n_{h}^{\prime}}-\frac{1}{N}\right)+\frac{W_{h-1}^{*}\left(f_{1}-1\right)}{n_{h}^{\prime}}\right] S_{(h-1) y}^{2} \tag{10}
\end{gather*}
$$

where

$$
f_{1}=\frac{n_{h 2}^{\prime}}{n_{h 2 s}^{\prime}} ; f_{2}=\frac{n_{h 2}^{\prime \prime}}{n_{h 2 s}^{\prime \prime}}
$$

## Minimum Variance of the Estimator $\omega_{h}$

Substituting the values of variances and covariance from equations (5), (6) and (7) in equation (4) we have the expression of the exact variance of the proposed estimator $\omega_{h}$. Now, minimize the variance of $\omega_{h}$, which is shown in equation (4). Define a function $f(x, y)$, where the variables $x$ and $y$ are interpreted as $\varphi_{h}$ and $\mu_{h}$ respectively, which represents the expression of the variance of $\omega_{h}$ given in equation (4). Thus, variance of $\omega_{h}$ is reduce to following equation

$$
\begin{equation*}
f(x, y)=\frac{S}{n}\left[\frac{x^{2}}{y} \Delta_{2}+(1-x)^{2}\left(\frac{\Delta_{1}}{1-y}+\gamma\right)-f\right] \tag{11}
\end{equation*}
$$

where

$$
\begin{aligned}
S & =S_{h y}^{2}, \quad \alpha=1-\rho_{h, h-1}^{2}, \quad \gamma=t_{h-1} \rho_{h, h-1}^{2}, \quad \Delta_{1}=\alpha+W_{h-1}^{*}\left(f_{1}-1\right) \alpha, \\
\Delta_{2} & =1+W_{h}^{*}\left(f_{2}-1\right), t_{h-1}=\frac{\varphi_{h-1}}{\mu_{h-1}}, \quad \mu_{h}=1-\lambda_{h}, \quad \text { and } f=\frac{n}{N} .
\end{aligned}
$$

To find the minimum variance, we differentiate the equation (11) with respect to $x$ and $y$ respectively and then equate to zero,

$$
\begin{equation*}
\frac{x}{y} \Delta_{2}=\frac{1-x}{1-y}\left[\Delta_{1}+\gamma(1-y)\right] \tag{12}
\end{equation*}
$$

## ESTIMATION OF POPULATION MEAN UNDER NON-RESPONSE

and

$$
\begin{equation*}
\frac{x}{y} \sqrt{\Delta_{2}}=\frac{1-x}{1-y} \sqrt{\Delta_{1}} \tag{13}
\end{equation*}
$$

From equations (12) and (13),

$$
\begin{equation*}
y=1-\sqrt{\Delta_{1}}\left(\sqrt{\Delta_{2}}-\sqrt{\Delta_{1}}\right) \gamma^{-1} \tag{14}
\end{equation*}
$$

Again, from equations (13) and (14), if

$$
\begin{equation*}
\frac{y}{x}=1+\left(\sqrt{\Delta_{2}}-\sqrt{\Delta_{1}}\right)^{2} \gamma^{-1} \tag{15}
\end{equation*}
$$

then

$$
\begin{equation*}
t_{h}=\left[1+r_{h} t_{h}^{-1}\right]^{-1}=\frac{x}{y} \tag{16}
\end{equation*}
$$

where

$$
\begin{equation*}
r_{h}=\left(\sqrt{\Delta_{2}}-\sqrt{\Delta_{1}}\right)^{2}(1-\alpha)^{-1} \tag{17}
\end{equation*}
$$

Because the values of $\alpha$ depend on the values of correlation. Therefore, $\alpha \geq 0$ and consequently $r_{h}$ is real. After iteration,

$$
\begin{equation*}
t_{h}=\left[1-\sum_{j=1}^{h} \prod_{k=j}^{h} r_{k}\right]^{-1} \tag{18}
\end{equation*}
$$

Hence, minimum variance of $\omega_{h}$ is obtained from equations (11) and (12) which is as follows

$$
\begin{equation*}
V\left(\omega_{h}\right)_{\mathrm{opt}}=f(x, y)_{\mathrm{opt}}=\frac{S}{n}\left[t_{h} \Delta_{2}-f\right] \tag{19}
\end{equation*}
$$

## SHARMA \& SINGH

## Special Cases

Case 1: When non-response occurs only on (h-1) ${ }^{\text {th }}$ (previous) occasion.

For the case when non-response occurs only on $(h-1)^{\text {th }}$ occasion, the estimator for population mean $\bar{Y}_{h}$ on recent occasion may be structured as

$$
\begin{equation*}
\omega_{h}^{*}=\varphi_{h}^{*} \tau_{h}^{\prime \prime}+\left(1-\varphi_{h}^{*}\right) \omega_{h}^{\prime} \tag{20}
\end{equation*}
$$

where $\tau_{h}^{\prime \prime}=\bar{y}_{h}^{\prime \prime}$ and $\omega_{h}^{\prime}=\bar{y}_{h}^{\prime *}+\beta_{h, h-1}\left(\omega_{h-1}^{*}-\bar{y}_{h-1}^{\prime *}\right) . \varphi_{h}^{*}$ is unknown constant to be determined so as to minimize the variance of the estimator $\omega_{h}^{*}$.

## Properties of the estimator $\omega_{h}^{*}$

Because $\tau_{h}^{\prime \prime}$ and $\omega_{h}^{\prime}$ are sample mean and difference type estimators respectively, they are unbiased for population mean $\bar{Y}_{h}$. Therefore, the resulting estimator $\omega_{h}^{*}$ is defined in equation (20) is also unbiased estimator of $\bar{Y}_{h}$.

Theorem 2: variance of the estimator $\omega_{h}^{*}$ is obtained as

$$
\begin{equation*}
V\left(\omega_{h}^{*}\right)=\varphi_{h}^{* 2} V\left(\tau_{h}^{\prime \prime}\right)+\left(1-\varphi_{h}^{*}\right)^{2} V\left(\omega_{h}^{\prime}\right)+2 \varphi_{h}^{*}\left(1-\varphi_{h}^{*}\right) C\left(\tau_{h}^{\prime \prime}, \omega_{h}^{\prime}\right) \tag{21}
\end{equation*}
$$

where

$$
\begin{align*}
& V\left(\tau_{h}^{\prime \prime}\right)=\left(\frac{1}{n_{h}^{\prime \prime}}-\frac{1}{N}\right) S_{h y}^{2}  \tag{22}\\
& V\left(\omega_{h}^{\prime}\right)=\left[\frac{1}{n_{h}^{\prime}}\left\{1+W_{h-1}^{*}\left(f_{1}-1\right)\right\}\left(1-\rho_{h, h-1}^{2}\right)+\frac{\varphi_{h-1}^{*}}{n_{h-1}^{\prime \prime}} \rho_{h, h-1}^{2}-\frac{1}{N}\right] S_{h y}^{2} \tag{23}
\end{align*}
$$

and

$$
\begin{equation*}
C\left(\tau_{h}^{\prime \prime}, \omega_{h}^{\prime}\right)=-\frac{1}{N} S_{h y}^{2} \tag{24}
\end{equation*}
$$

## Minimum Variance of the estimator $\omega_{h}^{*}$

Similarly, represent the expression of the variance of $\omega_{h}^{*}$ in equation (21) as

$$
\begin{equation*}
f^{*}\left(x^{*}, y^{*}\right)=\frac{S}{n}\left[\frac{x^{* 2}}{y^{*}}+\left(1-x^{*}\right)^{2}\left(\frac{\Delta_{1}}{1-y^{*}}+\gamma\right)-f\right] \tag{25}
\end{equation*}
$$

To find the minimum variance,

$$
\begin{gather*}
\frac{x^{*}}{y^{*}}=\frac{1-x^{*}}{1-y^{*}}\left[\Delta_{1}+\gamma\left(1-y^{*}\right)\right]  \tag{26}\\
\frac{x^{*}}{y^{*}}=\frac{1-x^{*}}{1-y^{*}} \sqrt{\Delta_{1}} \tag{27}
\end{gather*}
$$

From equations (26) and (27),

$$
\begin{equation*}
y^{*}=1-\sqrt{\Delta_{1}}\left(1-\sqrt{\Delta_{1}}\right) \gamma^{-1} \tag{28}
\end{equation*}
$$

Further,

$$
\begin{align*}
& \frac{y^{*}}{x^{*}}=1+\left(1-\sqrt{\Delta_{1}}\right)^{2} \gamma^{-1}  \tag{29}\\
& t_{h}^{*}=\left[1+r_{h}^{* *} t_{h-1}^{*-1}\right]^{-1}=\frac{x^{*}}{y^{*}} \tag{30}
\end{align*}
$$

where

$$
\begin{equation*}
r_{h}^{*}=\left(1-\sqrt{\Delta_{1}}\right)^{2}(1-\alpha)^{-1} \tag{31}
\end{equation*}
$$

## SHARMA \& SINGH

$$
\begin{equation*}
t_{h}^{*}=\left[1+\sum_{j=1}^{h} \prod_{k=j}^{h} r_{k}^{*}\right]^{-1} \tag{32}
\end{equation*}
$$

From (25) and (26) minimum variance of $\omega_{h}^{*}$ is expressed as

$$
\begin{equation*}
V\left(\omega_{h}^{*}\right)_{\mathrm{opt}}=f^{*}\left(x^{*}, y^{*}\right)_{\mathrm{opt}}=\frac{S}{n}\left[t_{h}^{*}-f\right] \tag{33}
\end{equation*}
$$

## Case 2: When non-response occurs only on $h^{\text {th }}$ (recent) occasion

The estimator for the population mean $\bar{Y}_{h}$ on recent occasion for this case may be given as

$$
\begin{equation*}
\omega_{h}^{* *}=\varphi_{h}^{* *} \omega_{h}^{\prime \prime}+\left(1-\varphi_{h}^{* *}\right) \tau_{h}^{\prime} \tag{34}
\end{equation*}
$$

where $\omega_{h}^{\prime \prime}$ is defined in equation (1) and $\tau_{h}^{\prime}=\bar{y}_{h}^{\prime}+\beta_{h, h-1}\left(\omega_{h-1}^{* *}-\bar{y}_{h-1}^{\prime}\right)$ and $\varphi_{h}^{* * *}$ is unknown constant to be determined so as to minimize the variance of the estimator $\omega_{h}^{* *}$.

## Properties of the estimators $\omega_{n}^{* \prime}$

Because $\omega_{h}^{\prime \prime}$ and $\tau_{h}^{\prime}$ are sample mean and difference type estimators respectively, they are unbiased for population mean $\bar{Y}_{h}$. Therefore, the resulting estimator $\omega_{h}^{* *}$ defined in equation (34) is also unbiased estimator of $\bar{Y}_{h}$.

Theorem 3: Variance of estimators $\omega_{h}^{* *}$ is obtained as

$$
\begin{equation*}
V\left(\omega_{h}^{* *}\right)=\varphi_{h}^{* * 2} V\left(\omega_{h}^{\prime \prime}\right)+\left(1-\varphi_{h}^{* * *}\right)^{2} V\left(\tau_{h}^{\prime}\right)+2 \varphi_{h}^{* *}\left(1-\varphi_{h}^{* *}\right) C\left(\omega_{h}^{\prime \prime}, \tau_{h}^{\prime}\right) \tag{35}
\end{equation*}
$$

where $V\left(\omega_{h}^{\prime \prime}\right)$ is shown in equation (5),

$$
\begin{equation*}
V\left(\tau_{h}^{\prime}\right)=\left[\frac{1}{n_{h}^{\prime}}\left(1-\rho_{h, h-1}^{2}\right)+\frac{\varphi_{h-1}^{* *}}{n_{h-1}^{\prime \prime}} \rho_{h, h-1}^{2}-\frac{1}{N}\right] S_{h y}^{2} \tag{36}
\end{equation*}
$$

and

$$
\begin{equation*}
C\left(\Delta_{h}^{\prime \prime}, \tau_{h}^{\prime}\right)=-\frac{1}{N} S_{h y}^{2} \tag{37}
\end{equation*}
$$

## Minimum Variance of the estimator $\omega_{h}^{* *}$

The expression of the variance of $\omega_{h}^{* *}$ shown in equation (35) is reduced to the following form

$$
\begin{equation*}
f^{* * *}\left(x^{* * *}, y^{* *}\right)=\frac{S}{n}\left[\frac{x^{* * 2}}{y^{* *}} \Delta_{2}+\left(1-x^{* *}\right)^{2}\left(\frac{\alpha}{1-y^{* * *}}+\gamma\right)-f\right] \tag{38}
\end{equation*}
$$

To find the minimum variance,

$$
\begin{gather*}
\frac{x^{* *}}{y^{* *}}=\frac{1-x^{* *}}{1-y^{* *}}\left[\alpha+\gamma\left(1-y^{* *}\right)\right]  \tag{39}\\
\frac{x^{* *}}{y^{* *}} \sqrt{\Delta_{2}}=\frac{1-x^{* *}}{1-y^{* *}} \sqrt{\alpha} \tag{40}
\end{gather*}
$$

From equations (39) and (40)

$$
\begin{equation*}
y^{* *}=1-\sqrt{\alpha}\left(\sqrt{\Delta_{2}}-\sqrt{\alpha}\right) \gamma^{-1} \tag{41}
\end{equation*}
$$

then

$$
\begin{align*}
& \frac{y^{* *}}{x^{* *}}=1+\left(\sqrt{\omega_{2}}-\sqrt{\beta}\right)^{2} \gamma^{-1}  \tag{42}\\
& t_{h}^{* *}=\left[1+r_{h}^{* * * *-1} t_{h-1}^{*-1}\right]^{\frac{x^{* *}}{y^{* * *}}} \tag{43}
\end{align*}
$$

where

## SHARMA \& SINGH

$$
\begin{gather*}
r_{h}^{* *}=\left(\sqrt{\Delta_{2}}-\sqrt{\alpha}\right)^{2}(1-\alpha)^{-1}  \tag{44}\\
t_{h}^{* *}=\left[1+\sum_{j=1}^{h} \prod_{k=j}^{h} r_{k}^{* *}\right]^{-1} \tag{45}
\end{gather*}
$$

Thus, from (38) and (39) minimum variance of $\Delta_{h}^{* *}$ is obtained as

$$
\begin{equation*}
V\left(\omega_{h}^{* *}\right)_{\mathrm{opt}}=f^{* *}\left(x^{* *}, y^{* *}\right)_{\mathrm{opt}}=\frac{S}{n}\left[t_{h}^{* *} \Delta_{2}-f\right] \tag{46}
\end{equation*}
$$

## Efficiency Comparison

To examine the loss in precision of the estimators $\omega_{h}, \omega_{h}^{*}$ and $\omega_{h}^{* *}$ due to nonresponse, the percent relative loss in precision of estimator $\omega_{h}, \omega_{h}^{*}$ and $\omega_{h}^{* *}$ with respect to the estimator $\tau_{h}$, have been computed for different choices of $\rho_{h, h-1}$. The estimator $\tau_{h}$ is defined under the similar circumstances as the estimator $\omega_{h}$ but in the absence of non-response. Hence the estimator $\tau_{h}$ is given as

$$
\begin{equation*}
\tau_{h}=\varphi_{h} \tau_{h}^{\prime \prime}+\left(1-\varphi_{h}\right) \tau_{h}^{\prime} \tag{47}
\end{equation*}
$$

where $\tau_{h}^{\prime \prime}=\bar{y}_{h}^{\prime \prime}, \quad \tau_{h}^{\prime}=\bar{y}_{h}^{\prime}+\beta_{h, h-1}\left(\tau_{h-1}-\bar{y}_{h-1}^{\prime}\right)$ and $\psi_{h}$ is unknown constant to be determined by the minimization of the variance of $\tau_{h}$.

Following Sukhatme, Sukhatme, Sukhatme, and Asok (1984) the optimum variance of $\tau_{h}$ is given by

$$
\begin{equation*}
V\left(\tau_{h}\right)_{\mathrm{opt}}=\frac{S}{n}\left[\hat{t}_{h}-f\right] \tag{48}
\end{equation*}
$$

where $\hat{t}_{h}=\left[1+\sum_{j=1}^{h} \prod_{k=j}^{h} \hat{r}_{k}\right]^{-1}$ and $\hat{r}_{k}=(1-\sqrt{\alpha})(1+\sqrt{\alpha})^{-1}$.

## ESTIMATION OF POPULATION MEAN UNDER NON-RESPONSE

Remark 4: To compare the performance of the estimators $\omega_{h}, \omega_{h}^{*}$ and $\omega_{h}^{* *}$ with respect to $\tau_{h}$, the assumptions $W_{h}^{*}=W_{h-1}^{*}\left(\right.$ say $\left.W^{*}\right)$ are introduced. The percent relative losses in precision of the estimators $\omega_{h}, \omega_{h}^{*}$ and $\omega_{h}^{* *}$ with respect to $\tau_{h}$ under their respective optimality conditions are given by

$$
L=\frac{V\left(\omega_{h}\right)_{\mathrm{opt}}-V\left(\tau_{h}\right)_{\mathrm{opt}}}{V\left(\omega_{h}\right)_{\mathrm{opt}}} \times 100 \quad L^{*}=\frac{V\left(\omega_{h}^{*}\right)_{\mathrm{opt}}-V\left(\tau_{h}\right)_{\mathrm{opt}}}{V\left(\omega_{h}^{*}\right)_{\mathrm{opt}}} \times 100
$$

and

$$
L^{* *}=\frac{V\left(\omega_{h}^{* *}\right)_{\mathrm{opt}}-V\left(\tau_{h}\right)_{\mathrm{opt}}}{V\left(\omega_{h}^{* *}\right)_{\mathrm{opt}}} \times 100
$$

The expressions of $\mu_{h(\mathrm{opt})}, \mu_{h(\mathrm{opt})}^{*}, \mu_{h(\mathrm{opt})}^{* *}$ and the percent relative losses are given in terms of the population correlation coefficients. Therefore, they have been computed for different choices of correlation $\rho_{h, h-1}$. Percent relative losses in precision of the estimators $\omega_{h}, \omega_{h}^{*}$ and $\omega_{h}^{* *}$ have been computed for different choices of $f, f_{1}, f_{2}, W_{h}^{*}, W_{h-1}^{*}$ and $\rho_{h, h-1}$.

Presented in Tables 1-3 are the optimum values of $\mu_{h(\mathrm{opt})}, \mu_{h(\mathrm{opt})}^{*}, \mu_{h(\mathrm{opt})}^{* *}$ and the percent relative losses with respect to $\tau_{h}$.

## SHARMA \& SINGH

Table 1. Percent relative loss $L$ in precision of $\omega_{h}$ with respect to $\tau_{h}$ for $f=0.1$.

| Occasions ( $h$ ) | $\rho_{h, h-1} \rightarrow$ |  | 0.5 |  | 0.7 |  | 0.9 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $f \downarrow$ | $W^{*} \downarrow$ | $f_{2} \downarrow$ | $\mu_{h(\text { (pt) }}$ | $L$ | $\mu_{h(\text { (opt })}$ | $L$ | $\mu_{h(\text { (pt) }}$ | $L$ |
|  | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.3668 \\ & 0.2053 \end{aligned}$ | $\begin{aligned} & 4.2873 \\ & 6.5394 \end{aligned}$ | $\begin{aligned} & 0.5122 \\ & 0.4443 \end{aligned}$ | $\begin{aligned} & 4.8124 \\ & 8.2798 \end{aligned}$ | $\begin{aligned} & 0.6702 \\ & 0.6451 \end{aligned}$ | $\begin{aligned} & 5.1467 \\ & 9.2212 \end{aligned}$ |
| $2^{1.0}$ | 0.4 | 1.5 2.0 | 0.2053 | ${ }_{\text {* }}^{6.5394}$ | $\begin{aligned} & 0.4443 \\ & 0.3164 \end{aligned}$ | $\begin{array}{r} 8.2798 \\ 12.5942 \end{array}$ | $\begin{aligned} & 0.6451 \\ & 0.5978 \end{aligned}$ | $\begin{array}{r} 9.2212 \\ 15.1928 \end{array}$ |
|  | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.6201 \\ & 0.4431 \end{aligned}$ | $\begin{aligned} & 13.0095 \\ & 17.0978 \end{aligned}$ | $\begin{aligned} & 0.5745 \\ & 0.5001 \end{aligned}$ | $\begin{aligned} & 12.0443 \\ & 15.8719 \end{aligned}$ | $\begin{aligned} & 0.6632 \\ & 0.6357 \end{aligned}$ | $\begin{array}{r} 9.6643 \\ 13.6380 \end{array}$ |
| 2.0 | 0.4 | $\begin{aligned} & 1.5 \\ & 2.0 \end{aligned}$ | $\begin{aligned} & 0.7100 \\ & 0.3503 \end{aligned}$ | $\begin{aligned} & 22.6145 \\ & 28.9318 \end{aligned}$ | $\begin{aligned} & 0.5681 \\ & 0.4167 \end{aligned}$ | $\begin{aligned} & 21.4502 \\ & 26.8537 \end{aligned}$ | $\begin{aligned} & 0.6309 \\ & 0.5750 \end{aligned}$ | $\begin{aligned} & 17.4019 \\ & 23.0665 \end{aligned}$ |
|  | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.2822 \\ & 0.0379 \end{aligned}$ | $\begin{aligned} & 3.1253 \\ & 3.2392 \end{aligned}$ | $\begin{aligned} & 0.4008 \\ & 0.2794 \end{aligned}$ | $\begin{aligned} & 2.8605 \\ & 3.8429 \end{aligned}$ | $\begin{aligned} & 0.5172 \\ & 0.4545 \end{aligned}$ | $\begin{aligned} & 1.5974 \\ & 1.9800 \end{aligned}$ |
| 1.0 | 0.4 | 1.5 2.0 | 0.0379 | $3.2392$ | $\begin{aligned} & 0.2794 \\ & 0.0094 \end{aligned}$ | $\begin{aligned} & 3.8429 \\ & 1.4728 \end{aligned}$ | $\begin{aligned} & 0.4545 \\ & 0.3205 \end{aligned}$ | $\begin{aligned} & 1.9800 \\ & 0.0642 \end{aligned}$ |
| 3.0 | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.6048 \\ & 0.3951 \end{aligned}$ | $\begin{aligned} & 13.3281 \\ & 16.9222 \end{aligned}$ | $\begin{aligned} & 0.5128 \\ & 0.4000 \end{aligned}$ | $\begin{aligned} & 12.6047 \\ & 15.0408 \end{aligned}$ | $\begin{aligned} & 0.5275 \\ & 0.4640 \end{aligned}$ | $\begin{array}{r} 9.3567 \\ 10.2194 \end{array}$ |
| 2.0 | 0.4 | $\begin{aligned} & 1.5 \\ & 2.0 \end{aligned}$ | $\begin{aligned} & 0.7042 \\ & 0.2849 \end{aligned}$ | $\begin{aligned} & 22.9963 \\ & 28.6079 \end{aligned}$ | $\begin{aligned} & 0.5128 \\ & 0.2806 \end{aligned}$ | $\begin{aligned} & 22.2920 \\ & 25.3154 \end{aligned}$ | $\begin{aligned} & 0.4778 \\ & 0.3413 \end{aligned}$ | $\begin{aligned} & 16.7377 \\ & 16.6839 \end{aligned}$ |
|  | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.2709 \\ & 0.0027 \end{aligned}$ | $\begin{aligned} & 3.1253 \\ & 3.2392 \end{aligned}$ | $\begin{aligned} & 0.3753 \\ & 0.2305 \end{aligned}$ | $\begin{aligned} & 2.8605 \\ & 3.8429 \end{aligned}$ | $\begin{aligned} & 0.4462 \\ & 0.3521 \end{aligned}$ | $\begin{aligned} & 1.5974 \\ & 1.9800 \end{aligned}$ |
| 1.0 | 0.4 | 1.5 2.0 | 0.0027 | 3.2392 | 0.2305 | $\underset{* *}{ }$ | $\begin{aligned} & 0.3521 \\ & 0.1294 \end{aligned}$ | $\begin{aligned} & 1.9800 \\ & 0.0642 \end{aligned}$ |
| 4 | 0.2 | 1.5 2.0 | $\begin{aligned} & 0.6042 \\ & 0.3910 \end{aligned}$ | $\begin{aligned} & 13.3281 \\ & 16.9222 \end{aligned}$ | $\begin{aligned} & 0.5039 \\ & 0.3800 \end{aligned}$ | $\begin{aligned} & 12.6047 \\ & 15.0408 \end{aligned}$ | $\begin{aligned} & 0.4728 \\ & 0.3830 \end{aligned}$ | $\begin{array}{r} 9.3567 \\ 10.2194 \end{array}$ |
| 2.0 | 0.4 | 1.5 2.0 | $\begin{aligned} & 0.7041 \\ & 0.2784 \end{aligned}$ | $\begin{array}{r} 22.9963 \\ 28.6079 \end{array}$ | $\begin{aligned} & 0.5057 \\ & 0.2488 \\ & \hline \end{aligned}$ | $\begin{array}{r} 22.2920 \\ 25.3154 \\ \hline \end{array}$ | $\begin{aligned} & 0.4142 \\ & 0.2127 \\ & \hline \end{aligned}$ | $\begin{array}{r} 16.7377 \\ 16.6839 \\ \hline \end{array}$ |

[^9]
## ESTIMATION OF POPULATION MEAN UNDER NON-RESPONSE

Table 2. Percent relative loss $L^{*}$ in precision of $\omega_{h}^{*}$ with respect to $\tau_{h}$ for $f=0.1$.

| Occasions (h) |  | $\rho_{h, h-1} \rightarrow$ | 0.5 |  | 0.6 |  | 0.8 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $f \downarrow$ | $W^{*} \downarrow$ | $\mu_{h(\mathrm{ppt})}^{*}$ | $L^{*}$ | $\mu_{h(\mathrm{opt})}^{*}$ | $L^{*}$ | $\mu_{h(\mathrm{opt})}^{*}$ | $L^{*}$ |
| 2 | 1.5 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | $\begin{aligned} & 0.6668 \\ & 0.8053 \end{aligned}$ | $\begin{aligned} & 3.9706 \\ & 6.3585 \end{aligned}$ | $\begin{aligned} & 0.6163 \\ & 0.6524 \end{aligned}$ | $\begin{aligned} & 3.6895 \\ & 6.7430 \end{aligned}$ | $\begin{aligned} & 0.6936 \\ & 0.6920 \end{aligned}$ | $\begin{aligned} & 2.4170 \\ & 4.6254 \end{aligned}$ |
|  | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.8053 | $\underset{* *}{6.3585}$ | $\begin{aligned} & 0.6524 \\ & 0.7327 \end{aligned}$ | $\begin{array}{r} 6.7430 \\ 11.2723 \end{array}$ | $\begin{aligned} & 0.6920 \\ & 0.6917 \end{aligned}$ | $\begin{aligned} & 4.6254 \\ & 8.5195 \end{aligned}$ |
|  | 1.5 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | $\begin{aligned} & 0.6556 \\ & 0.8032 \end{aligned}$ | $\begin{aligned} & 4.3688 \\ & 6.8491 \end{aligned}$ | $\begin{aligned} & 0.5670 \\ & 0.6188 \end{aligned}$ | $\begin{aligned} & 4.6766 \\ & 8.3301 \end{aligned}$ | $\begin{aligned} & 0.5822 \\ & 0.5882 \end{aligned}$ | $\begin{aligned} & 3.8818 \\ & 7.2990 \end{aligned}$ |
| 3 | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.8032 | ${ }_{\text {c }}^{6} 8.8491$ | $\begin{aligned} & 0.6188 \\ & 0.7196 \end{aligned}$ | $\begin{array}{r} 8.3301 \\ 13.3894 \end{array}$ | $\begin{aligned} & 0.5882 \\ & 0.6024 \end{aligned}$ | $\begin{array}{r} 7.2990 \\ 13.0285 \end{array}$ |
| 4 | 1.5 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | $\begin{aligned} & 0.6552 \\ & 0.8032 \end{aligned}$ | $\begin{aligned} & 4.4016 \\ & 6.8849 \end{aligned}$ | $\begin{aligned} & 0.5607 \\ & 0.6156 \end{aligned}$ | $\begin{aligned} & 4.8963 \\ & 8.6488 \end{aligned}$ | $\begin{aligned} & 0.5416 \\ & 0.5532 \end{aligned}$ | $\begin{aligned} & 4.6755 \\ & 8.6882 \end{aligned}$ |
| 4 | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.8032 | $6.8849$ | $\begin{aligned} & 0.6156 \\ & 0.7189 \end{aligned}$ | $\begin{array}{r} 8.6488 \\ 13.7590 \\ \hline \end{array}$ | $\begin{aligned} & 0.5532 \\ & 0.5766 \end{aligned}$ | $\begin{array}{r} 8.6882 \\ 15.2022 \\ \hline \end{array}$ |

*Note "*" indicate $\mu_{h(\mathrm{opt})}^{*}$ does not exist.

Table 3. Percent relative loss $L^{* *}$ in precision of $\omega_{h}^{* *}$ with respect to $\tau_{h}$ for $f=0.1$.

| Occasions ( $h$ ) |  | $\rho_{h, h-1} \rightarrow$ | 0.5 |  | 0.6 |  | 0.7 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $f_{2} \downarrow$ | $W^{*} \downarrow$ | $\mu_{h(\mathrm{opt})}^{* *}$ | $L^{* *}$ | $\mu_{h(\mathrm{ptt})}^{* *}$ | $L^{* *}$ | $\mu_{h(\mathrm{opt})}^{* *}$ | $L^{* *}$ |
| 2 | 1.5 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | $\begin{aligned} & 0.3668 \\ & 0.2053 \end{aligned}$ | $\begin{aligned} & 4.2873 \\ & 6.5394 \end{aligned}$ | $\begin{aligned} & 0.5122 \\ & 0.4443 \end{aligned}$ | $\begin{aligned} & 4.8124 \\ & 8.2798 \end{aligned}$ | $\begin{aligned} & 0.6702 \\ & 0.6451 \end{aligned}$ | $\begin{aligned} & 5.1467 \\ & 9.2212 \end{aligned}$ |
|  | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.2053 | ${ }_{6}^{6.5394}$ | $\begin{aligned} & 0.4443 \\ & 0.3164 \end{aligned}$ | $\begin{array}{r} 8.2798 \\ 12.5942 \end{array}$ | $\begin{aligned} & 0.6451 \\ & 0.5978 \end{aligned}$ | $\begin{array}{r} 9.2212 \\ 15.1928 \end{array}$ |
| 3 | 1.5 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | $\begin{aligned} & 0.2822 \\ & 0.0379 \end{aligned}$ | $\begin{aligned} & 3.1253 \\ & 3.2392 \end{aligned}$ | $\begin{aligned} & 0.4008 \\ & 0.2794 \end{aligned}$ | $\begin{aligned} & 2.8605 \\ & 3.8429 \end{aligned}$ | $\begin{aligned} & 0.5172 \\ & 0.4545 \end{aligned}$ | $\begin{aligned} & 1.5974 \\ & 1.9800 \end{aligned}$ |
| 3 | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.0379 | 3.2392 | $\begin{aligned} & 0.2794 \\ & 0.0094 \end{aligned}$ | $\begin{aligned} & 3.8429 \\ & 1.4728 \end{aligned}$ | $\begin{aligned} & 0.4545 \\ & 0.3205 \end{aligned}$ | $\begin{aligned} & 1.9800 \\ & 0.0642 \end{aligned}$ |
|  | 1.5 | 0.2 | 0.2709 | 2.9384 | 0.3753 | 2.2653 | 0.4462 | -0.6762 |
| 4 |  | 0.4 | 0.0027 | 2.4725 | 0.2305 | 2.2186 | 0.3521 | -3.1638 |
|  | 2.0 | $\begin{aligned} & 0.2 \\ & 0.4 \end{aligned}$ | 0.0027 | 2.4725 | 0.2305 | ${ }_{\text {** }}^{2.2186}$ | $\begin{aligned} & 0.3521 \\ & 0.1294 \\ & \hline \end{aligned}$ | $\begin{array}{r} -3.1638 \\ -13.0652 \\ \hline \end{array}$ |

*Note: "*" indicate $\mu_{h(\mathrm{opt})}^{* *}$ does not exist.

## SHARMA \& SINGH

## Results

## Behavior of Estimator $\omega_{h}$,

## From Table 1,

(a) For the fixed values of $h, f_{1}, f_{2}$ and $W^{*}$, the value of $\mu_{h(\mathrm{opt})}$ are mostly increased while the values of $L$ are almost decreased when the values of $\rho_{h, h-1}$ are increased.
(b) For the fixed values of $h, f_{1}, W^{*}$ and $\rho_{h, h-1}$, the values of $\mu_{h(\text { opt })}$ decrease while $L$ increases with the increasing value of $f_{2}$. This trend shows the larger fresh sample is required to be replaced on the recent occasion.
(c) For the fixed values of $h, f_{2}, W^{*}$, and $\rho_{h, h-1}$, the values of $\mu_{h(\mathrm{opt})}$ and $L$ are increasing with the increasing values of $f_{1}$.
(d) For the fixed values of $h, f_{1}, f_{2}$ and $\rho_{h, h-1}$, the values of $\mu_{h(\text { opt })}$ almost decrease while $L$ increases with the increasing value of $W^{*}$. This behavior shows that the higher the non-response rate, the larger fresh sample is required to be replaced on the recent occasion.
(e) For the fixed values of $h, f_{1}, W^{*}$ and $\rho_{h, h-1}$, the values of $\mu_{h(\mathrm{opt})}$ and $L$ are almost decreasing with the increasing values of number of occasions ( $h$ ). This phenomenon suggests that smaller fresh sample is required on the recent occasion which leads in reducing the cost of the survey.

## Behavior of Estimator $\omega_{h}^{*}$

From Table 2,
(a) For the fixed values of $h, f_{1}$, and $W^{*}$, no patterns are visible in the values of $\mu_{h(\text { opt })}^{*}$ and $L^{*}$ with the increasing value of $\rho_{h, h-1}$.
(b) For the fixed values of $h, W^{*}$, and $\rho_{h, h-1}$, the values of $\mu_{h(\text { opt })}^{*}$ and $L^{*}$ are increasing with the increasing values of $f_{1}$.
(c) For the fixed values of $h, f_{1}$, and $\rho_{h, h-1}$, the values of $\mu_{h(\mathrm{opt})}^{*}$ and $L^{*}$ increase with the increasing values of $W^{*}$.
(d) For the fixed values of $f_{1}, W^{*}$ and $\rho_{h, h-1}$, the values of $\mu_{h(\text { opt })}^{*}$ are decreasing while the values of $L^{*}$ are increasing with the increasing values of number of occasions (h). This event suggests that smaller fresh sample is required on the recent occasion so that cost of the survey is reduced.

## Behavior of Estimator $\omega_{h}^{*}$

From Table 3,
(a) For the fixed values of $h, f_{2}$, and $W^{*}$ the values of $\mu_{h(\mathrm{opt})}^{* *}$ and $L^{* *}$ are almost increased when the value of $\rho_{h, h-1}$ is increased.
(b) For the fixed values of $h, \rho_{h, h-1}$, and $W^{*}$ the values of $\mu_{h(\mathrm{opt})}^{* *}$ decrease while $L^{* *}$ increases with the increasing values of $f_{2}$. This phenomenon indicates that if a highly correlated auxiliary variable is available it pays in terms of reducing the cost of the survey and smaller fresh sample is required at the recent occasion.
(c) For the fixed values of $h, f_{2}$, and $\rho_{h, h-1}$, the values of $\mu_{h(\mathrm{opt})}^{* *}$ decreases while the values of $L^{* *}$ does not follow any certain pattern with the increasing value of $W^{*}$.
(d) For the fixed values of $f_{2}, W^{*}$ and $\rho_{h, h-1}$, the values of $\mu_{h(\text { opt })}^{* *}$ and $L^{* *}$ are decreasing with the increasing values of number of occasions (h). This behavior suggests that lower the non-response is useful and smaller fresh sample is required at the recent occasion which leads in the minimizing the survey cost.

## SHARMA \& SINGH

## Conclusion

On the basis of preceding interpretations, it may be concluded that the proposed estimation procedure is more useful and fruitful in the estimate of population mean when non-response occur on $h^{\text {th }}$ occasion, $(h-1)^{\text {th }}$ occasion and simultaneously on both $h^{\text {th }}$ and $(h-1)^{\text {th }}$ occasions in the $h$-occasion successive sampling. It is also visible from the empirical studies that the percent relative loss in precision is not so high. Hence, the proposed estimators $\omega_{h}, \omega_{h}^{*}$, and $\omega_{h}^{* *}$ are performing well in terms of precision even in the presence of non-responses. Thus they are reliable and may be recommended to the survey statisticians and practitioners for its practical applications.

## Acknowledgement

Authors are thankful to the University Grants Commission, New Delhi and Indian School of Mines, Dhanbad for providing the financial assistance and necessary infrastructure to carry out the present work.

## References

Binder, D. A., \& Hidiroglou. (1988). Sampling in Time. In P. R. Krishaiah and C. R. Rao (Eds.), Sampling (pp. 187-211). Amsterdam: North Holland.

Biradar, R. S., \& Singh, H. P. (2001). Successive sampling using auxiliary information on both occasions. Bulletin of the Calcutta Statistical Association, 51(203-204), 243-251.

Choudhary, R. K, Bathla, H. V. L., \& Sud, U. C., (2004). On non-response in sampling on two occasions. Journal of the Indian Society of Agricultural Statistics 58(3), 331-343.

Cochran, W. G. (1977). Sampling techniques (3 ${ }^{\text {rd }}$ ed.). New York, NY: John Wiley \& Sons.

Eckler, A. R. (1955). Rotation sampling. The Annals of Mathematical Statistics, 26(4), 664-685. doi:10.1214/aoms/1177728427

Feng, S., \& Zou, G. (1997). Sample rotation method with auxiliary variable. Communications in Statistics - Theory and Methods, 26I6), 1497-1509. doi:10.1080/03610929708831996

Gupta, P. C. (1979). Sampling on two successive occasions. Journal of Statistical Research, 13, 7-16.

## ESTIMATION OF POPULATION MEAN UNDER NON-RESPONSE

Hansen, M. H., \& Hurwitz, W. N. (1946). The problem of the non-response in sample surveys. Journal of the American Statistical Association, 41(236), 517529. doi:10.1080/01621459.1946.10501894

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts (No. 304). Agricultural Experiment Station, Agricultural Research Bulletins (No. 304). Ames, IA: Iowa State College.

Kish, L. (1998). Space/time variations and rolling samples. Journal of Official Statistics, 14(1), 31-46.

McLaren, C. H., \& Steel, D. G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. Survey Methodology, 26(2), 163-172.

Okafor, F. C., \& Lee, H. (2000). Double Sampling for ratio and regression estimation with sub-sampling the non-respondents. Survey Methodology, 26(2), 183-188.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. Series B (Methodological), 12(2), 241-255.

Rao, J. N. K., \& Graham, J. E. (1964). Rotation design for sampling on repeated occasions. Journal of the American Statistical Association, 59(306), 492-509. doi:10.1080/01621459.1964.10482175

Sen, A. R. (1971). Successive sampling with two auxiliary variables. Sankhyā: The Indian Journal of Statistics, Series B (1960-2002), 33(3/4), 371-378.

Sen, A. R. (1973). 346. Note: Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics, 29(2), 381-385. doi:10.2307/2529401

Singh, A. C., Kennedy, B. \& Wu, S. (2001). Regression composite estimation for the Canadian labour force survey with a rotating panel design. Survey Methodology, 27(1), 33-44.

Singh, G. N. (2003). Estimation of population mean using auxiliary information on recent occasion in h occasions successive sampling. Statistics in Transition, 6(4), 523-532.

Singh, G. N. (2005). On the use of chain-type ratio estimator in successive sampling. Statistics in Transition, 7(1), 21-26.

Singh, G. N., \& Karna, J. P. (2009). Estimation of population mean on current occasion in two-occasion successive sampling. Metron, 67(1), 87-103.

## SHARMA \& SINGH

Singh, G. N., Majhi, D., Maurya, S., \& Sarma, A. K. (2015). Some effective rotation patterns in estimation of population mean in two-occasion successive sampling. Communications in Statistics - Theory \& Methods, 44(12), 2571-2585. doi:10.1080/03610926.2013.785000.

Singh, G. N., \& Prasad, S. (2010). Some estimators of population mean in two occasions rotation patterns. Advances in Modelling. Series A. General Mathematics, 47(2), 1-18.

Singh, G. N., Prasad, S., \& Karna, J. P. (2011). Some classes of estimators for population mean at current occasion in two-occasion successive sampling, Journal of Statistical Research, 45(1), 21-36.

Singh, G. N., \& Priyanka, K. (2007). Effect of non-response on current occasion in search of good rotation patterns on successive occasions. Statistics in Transition-New Series, 8(2), 273-292.

Singh, G. N., \& Sharma, A. K. (2014). An alternative rotation patterns in two occasion successive sampling. International Journal of Mathematics and Statistics, 15(3), 9-22.

Singh, G. N., \& Sharma, A. K. (2015). Search of some effective rotation patterns in estimation of population mean on current occasion in two-occasion successive sampling. Journal of Statistics Application and Probability, 4(1), 127138.

Singh, G. N., \& Singh, V. K. (2001). On the use of auxiliary information in successive sampling. Journal of the Indian Society of Agricultural Statistics, 54(1), 1-12.

Singh, H. P., \& Kumar, S. (2008). A regression approach to the estimation of finite population mean in the presence of non-response. Australian \& New Zealand Journal of Statistics, 50(4), 395-408.
doi:10.1111/j.1467-842X.2008.00525.x
Singh, V. K., Singh, G. N., \& Shukla, D. (1991). An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions. Journal of Scientific Research, 41 C, 149-159.

Steel D. G., \& McLaren, C. H. (2002). In search of good rotation pattern. Submitted to world Scientific.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. \& Asok, C. (1984).
Sampling theory of surveys with applications ( $3^{\text {rd }}$ revised ed.). Ames, IA: Iowa State University Press.

Tikkiwal, B. D. (1951). Theory of successive sampling (Unpublished dissertation). Institute of Agricultural Research Statistics, New Delhi, India.

Yates, F. (1949). Sampling methods for census and surveys. London: Charles Griffin and Company Limited.

# Efficient and Unbiased Estimation Procedure of Population Mean in Two-Phase Sampling 

Reba Maji<br>Indian School of Mines<br>Dhanbad, India

Arnab Bandyopadhyay<br>Asansol Engineering College<br>Asansol, Inida

G. N. Singh<br>Indian School of Mines<br>Dhanbad, India

In this paper, an unbiased regression-ratio type estimator has been developed for estimating the population mean using two auxiliary variables in double sampling. Its properties are studied under two different cases. Empirical studies and graphical simulation have been done to demonstrate the efficiency of the proposed estimator over other estimators.

Keywords: Double sampling, study variable, auxiliary variable, chain-type, regression, bias, variance, efficiency

## Introduction

The use of supplementary information on auxiliary variable for estimating the finite population mean of the variable under study has played an eminent role in sampling theory and practices. Auxiliary information may be truthfully utilized at the planning, design, and estimation stages to develop improved estimation procedures in sample surveys. Ratio, product, and regression methods of estimation are good examples in this context. Use of auxiliary information at the estimation stage was introduced during the 1930's with a comprehensive theory provided by Cochran (1940). Sometimes, information on auxiliary variable may be readily available for all the units of a population; for example, tonnage (or seat capacity) of each vehicle or ship is known in survey sampling of transportation, and number of beds available in different hospitals may be known well in advance in health care surveys. If such information is lacking, it is sometimes relatively cheap to take a large preliminary sample where an auxiliary variable alone is measured. Such practice is applicable in two-phase (or double) sampling. Two-phase sampling happens to be a powerful and cost-effective (economical) technique to generate reliable estimates of the unknown population parameters of the auxiliary variables in a first phase sample.

[^10]In order to construct an efficient estimator of the population mean of the auxiliary variable in a first-phase (preliminary) sample, Chand (1975) introduced the technique of chaining another auxiliary variable with the first auxiliary variable by using the ratio estimator in the first phase sample. This estimator is known as the chain-type ratio estimator. This work was further extended by Kiregyera (1980; 1984), Sahoo and Sahoo (1993), Tracy, Singh, and Singh (1996), Singh and Espejo (2007), Gupta and Shabbir (2007), Dash and Mishra (2011), Shukla, Pathak, and Thakur (2012), and Choudhury and Singh (2012), among others, who proposed various chain-type ratio and regression estimators. It may be noted that the most of these estimation procedures of the population mean in two-phase sampling are biased which becomes a serious drawback for their practical applications.

Encouraged and fascinated with the work discussed earlier, we have proposed an unbiased regression-ratio type estimator of the population mean and studied its properties under two different structures of two-phase sampling. Performances of the proposed estimator have been examined through empirical and graphical means of comparisons. Suitable recommendations to the survey statistician are made.

## Methodology

## Sample Structure and Some Existing Estimation Procedures

Let $y_{k}, x_{k}$, and $z_{k}$ be the values of the study variable $y$, first auxiliary variable $x$, and second auxiliary variable $z$, respectively, associated with the $k^{\text {th }}$ unit of the finite population $\mathrm{U}=\left(U_{1}, U_{2}, U_{3}, \ldots, U_{N}\right)$. The intent is to estimate the population mean $\bar{Y}$ of the study variable $y$ in the presence of auxiliary variables $x$ and $z$ when the population mean $\bar{X}$ of $x$ is unknown but information on $z$ is readily available for all the units of population.

To estimate $\bar{Y}$, a first-phase sample $\mathrm{S}^{\prime}\left(\mathrm{S}^{\prime} \subset \mathrm{U}\right)$ of size $n$ is drawn via a simple random sampling without replacement (SRSWOR) scheme from the entire population U and observed for the auxiliary variable $x$ to furnish the estimate of $\bar{X}$. Next, a second-phase sample $S$ of size $m(m \leq n)$ is drawn by SRSWOR according to the following rules to observe the study variable $y$ :

Case I: Second-phase sample is drawn as a subsample of the first-phase sample
Case II: Second-phase sample is drawn independently of the first-phase sample

## MAJI ET AL.

The case where the second sample is drawn independent of the first was considered by Bose (1943).

In the sections below, we use the following notations:
$\bar{x}_{m}, \bar{x}_{n}, \bar{y}_{m}, \bar{z}_{m}, \bar{z}_{n}$ : Sample mean of the respective variables of the sample sizes shown in subscripts.
$\bar{X}, \bar{Y}, \bar{Z}$ : Population mean of $x, y$, and $z$, respectively.
$\rho_{y x}, \rho_{y z}, \rho_{x z}$ : Correlation coefficient between the variables shown in subscripts.
$C_{x}, C_{y}, C_{z}$ : Coefficient of variance of $x, y$, and $z$ respectively.
$S_{y z}$ : Population covariance between $y$ and $z$.
$S_{z}^{2}$ : Population mean square of $z$.
$s_{y z}(m)$ : Sample covariance between $y$ and $z$ based on the sample of size $m$.
$s_{z}^{2}(m)$ : Sample mean square of $z$ based on the sample of size $m$.
$\beta_{y z}$ : Population regression coefficient between the variables $y$ and $z$.
$b_{x z}(n), b_{y z}(m), b_{y x}(m)$ : Sample regression coefficient between the variables shown in subscripts and based on samples of the size indicated in braces.

To estimate the population mean $\bar{Y}$, the classical ratio estimator is presented as

$$
\begin{equation*}
\bar{y}_{r}=\frac{\bar{y}_{m}}{\bar{x}_{m}} \bar{X} \tag{1}
\end{equation*}
$$

If $\bar{X}$ is unknown, we estimate $\bar{Y}$ under the two-phase sampling set up as

$$
\begin{equation*}
t_{1}=\frac{\bar{y}_{m}}{\bar{x}_{m}} \bar{x}_{n} \tag{2}
\end{equation*}
$$

S. K. Srivastava (1971) generalized the ratio method of estimation, and its structure in two-phase sampling is given as

$$
\begin{equation*}
t_{2}=\bar{y}_{m}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{\alpha} \tag{3}
\end{equation*}
$$

where $\alpha$ is a real scalar which can be suitably determined by minimizing the mean square error (MSE) of the estimator.

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

The way in which the estimate of $\bar{Y}$ is improved using the auxiliary information on $x$ can also be extended to improve the estimate of $\bar{X}$ in the firstphase sample if another auxiliary variable, $z$, closely related to $x$ but remotely related to $y$ is used. Thus, assuming that the population mean of the auxiliary variable $z$ is known, Chand (1975) proposed a chain-type ratio estimator as

$$
\begin{equation*}
t_{3}=\frac{\bar{y}_{m}}{\bar{x}_{m}} \frac{\bar{x}_{n}}{\bar{z}_{n}} \bar{Z} \tag{4}
\end{equation*}
$$

Similarly, for negative correlation between the variables $y$ and $x$, the chain-type product estimator is defined as

$$
\begin{equation*}
t_{4}=\bar{y}_{m} \frac{\bar{x}_{m}}{\bar{x}_{n}} \frac{\bar{z}_{n}}{\bar{Z}} \tag{5}
\end{equation*}
$$

Kiregyera (1984) suggested the chain linear regression estimator in double sampling as

$$
\begin{equation*}
t_{5}=\bar{y}_{m}+b_{y x}(m)\left[\bar{x}_{n}+b_{x z}(n)\left(\bar{Z}-\bar{z}_{n}\right)-\bar{x}_{m}\right] \tag{6}
\end{equation*}
$$

Singh and Espejo (2007) considered a ratio-product type estimator in double sampling as

$$
\begin{equation*}
t_{6}=\bar{y}_{m}\left[k \frac{\bar{x}_{n}}{\bar{x}_{m}}+(1-k) \frac{\bar{x}_{m}}{\bar{x}_{n}}\right] \tag{7}
\end{equation*}
$$

## Proposed Estimator

The suggested unbiased regression-ratio type estimator for estimating the population mean $\bar{Y}$ is

$$
\begin{equation*}
T_{R}=\sum_{i=1}^{3} d_{i} \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{i} \tag{8}
\end{equation*}
$$

where $\bar{y}_{m}^{*}=\bar{y}_{m}+b_{y z}(m)\left(\bar{Z}-\bar{z}_{m}\right)$ and the $d_{i}(i=1,2,3)$ are real scalars suitably chosen so that

## MAJI ET AL.

$$
\begin{equation*}
\sum_{i=1}^{3} d_{i}=1 \tag{9}
\end{equation*}
$$

Remark 1: The estimator $T_{R}$ is proposed under the following conditions:

1. The sum of the weights is one.
2. The weights of the linear form are chosen such that the approximate bias is zero.
3. The approximate variance attains minimum.

## Properties of the Estimator $\boldsymbol{T}_{\boldsymbol{R}}$

Note from (8) that the proposed estimator $T_{R}$ is biased for $\bar{Y}$. Following Remark 1, it may be made unbiased for $\bar{Y}$. The variance $\mathrm{V}($.$) up to the first order of$ approximations are derived under large sample approximations using the following transformations:

$$
\begin{array}{ll}
\bar{y}_{m}=\bar{Y}\left(1+e_{1}\right), & \bar{x}_{m}=\bar{X}\left(1+e_{2}\right), \quad \bar{x}_{n}=\bar{X}\left(1+e_{3}\right) \\
\bar{z}_{m}=\bar{Z}\left(1+e_{4}\right), & s_{y z}(m)=S_{y z}\left(1+e_{5}\right), \quad s_{z}^{2}(m)=S_{z}^{2}\left(1+e_{6}\right)
\end{array}
$$

where $\mathrm{E}\left(e_{i}\right)=0$ and $\left|e_{i}\right|<1$ for all $i=1, \ldots, 6$.
Under the above transformations the estimator $T_{R}$ takes the following form:

$$
\begin{equation*}
T_{R}=\sum_{i=1}^{3} d_{i}\left\{\bar{Y}\left(1+e_{1}\right)-\beta_{y z} \bar{Z} e_{4}\left(1+e_{5}\right)\left(1+e_{6}\right)^{-1}\right\} \times\left\{\left(1+e_{3}\right)\left(1+e_{2}\right)^{-1}\right\} \tag{10}
\end{equation*}
$$

The bias and mean square error of the estimator was derived separately for the Cases I and II of the two-phase sampling structure.

Case I: $\quad$ When the second phase sample is drawn as a subsample of the first phase sample.

In this case we have the following expected values of the sample statistics:

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

$$
\left.\begin{array}{lll}
\mathrm{E}\left(e_{1}^{2}\right)=f_{m} C_{y}^{2}, & \mathrm{E}\left(e_{s}^{2}\right)=f_{m} C_{x}^{2}, & \mathrm{E}\left(e_{3}^{2}\right)=f_{1} C_{x}^{2}, \\
\mathrm{E}\left(e_{4}^{2}\right)=f_{m} C_{z}^{2}, & \mathrm{E}\left(e_{1} e_{2}\right)=f_{m} \rho_{y x} C_{y} C_{x}, & \mathrm{E}\left(e_{1} e_{3}\right)=f_{1} \rho_{y x} C_{y} C_{x}, \\
\mathrm{E}\left(e_{1} e_{4}\right)=f_{m} \rho_{y z} C_{y} C_{z}, & \mathrm{E}\left(e_{2} e_{3}\right)=f_{1} C_{x}^{2}, & \mathrm{E}\left(e_{2} e_{4}\right)=f_{m} \rho_{x z} C_{x} C_{z},  \tag{11}\\
\mathrm{E}\left(e_{3} e_{4}\right)=f_{1} \rho_{x z} C_{x} C_{z}, & \mathrm{E}\left(e_{4} e_{5}\right)=f_{m} \frac{\mu_{012}}{\bar{Z} S_{y z}}, & \mathrm{E}\left(e_{4} e_{6}\right)=f_{m} \frac{\mu_{003}}{\bar{Z} S_{z}^{2}}
\end{array}\right\}
$$

where

$$
\begin{aligned}
& f_{m}=\frac{1}{m}-\frac{1}{N}, \quad f_{1}=\frac{1}{n}-\frac{1}{N}, \\
& \mu_{p q r}=\frac{1}{N} \sum_{i=1}^{N}\left(x_{i}-\bar{X}\right)^{p}\left(y_{i}-\bar{Y}\right)^{q}\left(z_{i}-\bar{Z}\right)^{r}
\end{aligned}
$$

and $p, q, r \geq 0$ are integers.
Expanding the terms of (10) binomially and using the results from (11), we have derived the expression of bias and mean square error of the estimator $T_{R}$ up to the first order of approximations as

$$
\begin{align*}
\mathrm{B}\left(T_{R}\right)= & \mathrm{E}\left(T_{R}-\bar{Y}\right) \\
=f_{m} & \beta_{y z}\left(\frac{\mu_{003}}{S_{z}^{2}}-\frac{\mu_{012}}{S_{y z}}\right)+f_{2} P\left(\bar{Y} C_{x}^{2}-\bar{Y} \rho_{y x} C_{y} C_{x}+\bar{Z} \beta_{y z} \rho_{x z} C_{x} C_{z}\right)  \tag{12}\\
& \quad-d_{2} f_{2} \bar{Y} C_{x}^{2} \\
\mathrm{M}\left(T_{R}\right)= & \mathrm{E}\left(T_{R}-\bar{Y}\right) \\
= & \bar{Y}^{2} f_{m} C_{y}^{2}+\bar{Z}^{2} \beta_{y z}^{2} f_{m} C_{z}^{2}-2 \overline{Y Z} \beta_{y z} f_{m} \rho_{y z} C_{y} C_{z}+P^{2} \bar{Y}^{2} f_{2} C_{x}^{2}  \tag{13}\\
& \quad+2 P \bar{Y} f_{2}\left(\bar{Z} \beta_{y z} \rho_{x z} C_{x} C_{z}-\bar{Y} \rho_{y x} C_{y} C_{x}\right)
\end{align*}
$$

where

$$
\begin{equation*}
P=\sum_{i=1}^{3} i d_{i}, \quad f_{2}=\frac{1}{m}-\frac{1}{n} \tag{14}
\end{equation*}
$$

## MAJI ET AL.

Minimization of mean square error in (13) with respect to $P$ yields its optimum value as

$$
\begin{equation*}
P=\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right) \frac{C_{y}}{C_{x}} \tag{15}
\end{equation*}
$$

Substituting the optimum value of $P$ in (13), we obtain the minimum mean square error of $T_{R}$ as

$$
\begin{equation*}
\operatorname{Min} \cdot \mathrm{M}\left(T_{R}\right)=f_{m} S_{y}^{2}\left(1-\rho_{y z}^{2}\right)-f_{2} S_{y}^{2}\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right)^{2} \tag{16}
\end{equation*}
$$

Further, from (14) and (15),

$$
\begin{equation*}
P=\sum_{i=1}^{3} i d_{i}=\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right) \frac{C_{y}}{C_{x}} \tag{17}
\end{equation*}
$$

which we will denote with $R$.
From (9) and (17), it may be noted that the two equations in three unknowns are not sufficient to find the unique values of the $d_{i}(i=1,2,3)$. In order to get unique values of the $d_{i}$, impose the linear constraint

$$
\begin{equation*}
\mathrm{B}\left(T_{R}\right)=0 \tag{18}
\end{equation*}
$$

Thus, from (12),

$$
\begin{equation*}
K d_{1}+\left(2 K-\bar{Y} f_{2} C_{x}^{2}\right) d_{2}+3 K d_{3}=M \tag{19}
\end{equation*}
$$

where

$$
K=f_{2}\left(\bar{Y} C_{x}^{2}-\bar{Y} \rho_{y x} C_{y} C_{x}+\bar{Z} \beta_{y z} \rho_{x z} C_{x} C_{z}\right), \quad M=f_{m} \beta_{y z}\left(\frac{\mu_{012}}{S_{y z}}-\frac{\mu_{003}}{S_{z}^{2}}\right)
$$

Equations (9), (17), and (19) can be written in matrix form as

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

$$
\left(\begin{array}{ccc}
1 & 1 & 1  \tag{20}\\
1 & 2 & 3 \\
K & 2 K-\bar{Y}_{2} C_{x}^{2} & 3 K
\end{array}\right) \times\left(\begin{array}{l}
d_{1} \\
d_{2} \\
d_{3}
\end{array}\right)=\left(\begin{array}{c}
1 \\
R \\
M
\end{array}\right)
$$

Solving (20), the unique values of the $d_{i}$ are

$$
\left.\begin{array}{l}
d_{1}=\frac{3}{2}+\frac{1}{2 f_{2} \bar{Y} C_{x}^{2}}\left[M-R\left(K+f_{2} \bar{Y} C_{x}^{2}\right)\right] \\
d_{2}=\frac{1}{f_{2} \bar{Y} C_{x}^{2}}(R K-M)  \tag{21}\\
d_{3}=\frac{1}{2 f_{2} \bar{Y} C_{x}^{2}\left[(M-R K)+(R-1) f_{2} \bar{Y} C_{x}^{2}\right]}
\end{array}\right\}
$$

From (21), substituting the values of $d_{1}, d_{2}$, and $d_{3}$ into (8) yields the unbiased optimum regression-ratio type estimator as

$$
\begin{align*}
T_{R}=\left[\frac{3}{2}\right. & \left.+\frac{1}{2 f_{2} \bar{Y} C_{x}^{2}}\left\{M-R\left(K+f_{2} \bar{Y} C_{x}^{2}\right)\right\}\right] \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right) \\
& +\frac{1}{f_{2} \bar{Y} C_{x}^{2}}(R K-M) \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{2}  \tag{22}\\
& +\left[\frac{1}{2 f_{2} \bar{Y} C_{x}^{2}}\left\{(M-R K)+(R-1) f_{2} \bar{Y} C_{x}^{2}\right\}\right] \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{3}
\end{align*}
$$

whose variance up to the first degree of approximations is given by

$$
\begin{equation*}
\mathrm{V}\left(T_{R}\right)=f_{m} S_{y}^{2}\left(1-\rho_{y z}^{2}\right)-f_{2} S_{y}^{2}\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right)^{2} \tag{23}
\end{equation*}
$$

Case II: When the second-phase sample is drawn independently of the firstphase sample.

In this case, the expected values of the sample statistics are:

## MAJI ET AL.

$$
\left.\begin{array}{lll}
\mathrm{E}\left(e_{1}^{2}\right)=f_{m} C_{y}^{2}, & \mathrm{E}\left(e_{1}^{2}\right)=f_{m} C_{x}^{2}, & \mathrm{E}\left(e_{3}^{2}\right)=f_{1} C_{x}^{2}, \\
\mathrm{E}\left(e_{4}^{2}\right)=f_{m} C_{z}^{2}, & \mathrm{E}\left(e_{1} e_{2}\right)=f_{m} \rho_{y x} C_{y} C_{x}, & \mathrm{E}\left(e_{1} e_{4}\right)=f_{m} \rho_{y z} C_{y} C_{z}, \\
\mathrm{E}\left(e_{2} e_{4}\right)=f_{m} \rho_{x z} C_{x} C_{z}, & \mathrm{E}\left(e_{4} e_{5}\right)=f_{m} \frac{\mu_{012}}{\overline{\mathrm{Z}} S_{y z}}, & \mathrm{E}\left(e_{4} e_{6}\right)=f_{m} \frac{\mu_{003}}{\overline{\mathrm{Z}} S_{z}^{2}},  \tag{24}\\
\mathrm{E}\left(e_{1} e_{3}\right)=\mathrm{E}\left(e_{2} e_{3}\right)=\mathrm{E}\left(e_{3} e_{4}\right)=0 &
\end{array}\right\}
$$

Proceeding as in Case I, the unbiased optimum regression-ratio type estimator is obtained as

$$
\begin{align*}
& T_{R}=\frac{1}{f_{1}+2 f_{m}}\left[\left(f_{1}+f_{m}\right)(3-G)-\frac{1}{\bar{Y} C_{x}^{2}}(W G-A)\right] \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right) \\
&+ \frac{1}{f_{1}+2 f_{m}}\left[\frac{2}{\bar{Y} C_{x}^{2}}(W G-A)-f_{1}(3-G)\right] \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{2}  \tag{25}\\
&+ {\left[1-\frac{1}{f_{1}+2 f_{m}}\left\{f_{m}(3-G)+\frac{1}{\bar{Y} C_{x}^{2}}(W G-A)\right\}\right] \bar{y}_{m}^{*}\left(\frac{\bar{x}_{n}}{\bar{x}_{m}}\right)^{3} }
\end{align*}
$$

with variance up to the first order of approximations as

$$
\begin{equation*}
\mathrm{V}\left(T_{R}\right)=f_{m} S_{y}^{2}\left(1-\rho_{y z}^{2}\right)-\frac{f_{m}^{2}}{f_{1}+f_{m}} S_{y}^{2}\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right)^{2} \tag{26}
\end{equation*}
$$

where

$$
\begin{aligned}
W & =f_{1} \bar{Y} C_{x}^{2}+f_{m} \bar{Y} C_{x}^{2}-f_{m} \bar{Y} \rho_{y x} C_{y} C_{x}+f_{m} \bar{Z} \beta_{y z} \rho_{x z} C_{x} C_{z} \\
G & =\frac{f_{m}}{f_{1}+f_{m}}\left(\rho_{y x}-\rho_{y z} \rho_{x z}\right) \frac{C_{y}}{C_{x}} \\
A & =f_{m} \beta_{y z}\left(\frac{\mu_{012}}{S_{y z}}-\frac{\mu_{003}}{S_{z}^{2}}\right)
\end{aligned}
$$

Remark 2: The unique value of the scalars $d_{i}$ depend on unknown population parameters such as $\beta_{x z}, \beta_{y z}, \mu_{012}, \mu_{003}, C_{x}, C_{y}, C_{z}, \bar{X}, \bar{Y}, \rho_{y x}$, and $\rho_{x z}$. Thus, to make the estimator practicable, these unknown population parameters may be estimated with

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

their respective sample estimates or from past data or guessed from experience gathered over time. Such problems are also considered by Reddy (1978), Tracy et al. (1996), and Singh and Espejo (2007).

## Results

## Efficiency Comparison

To examine the performance of our proposed estimator, we have considered some contemporary estimators of population mean which are discussed in a previous section. The mean square errors/minimum mean square errors of the estimators $t_{i}$ $(i=1,2, \ldots, 6)$ are given below for both cases of two-phase sampling structure considered in this paper:

## Case I:

$$
\begin{aligned}
& \mathrm{M}\left(t_{1}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{2} C_{x}^{2}-2 f_{2} \rho_{y x} C_{y} C_{x}\right] \\
& \operatorname{Min} . \mathrm{M}\left(t_{2}\right)=S_{y}^{2}\left[f_{m}-f_{2} \rho_{y x}^{2}\right] \\
& \mathrm{M}\left(t_{3}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{2} C_{x}^{2}+f_{1} C_{z}^{2}-2 f_{2} \rho_{y x} C_{y} C_{x}-2 f_{1} \rho_{y z} C_{y} C_{z}\right] \\
& \operatorname{M}\left(t_{4}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{2} C_{x}^{2}+f_{1} C_{z}^{2}+2 f_{2} \rho_{y x} C_{y} C_{x}+2 f_{1} \rho_{y z} C_{y} C_{z}\right] \\
& \operatorname{M}\left(t_{5}\right)=S_{y}^{2}\left[f_{m}\left(1-\rho_{y x}^{2}\right)+f_{1}\left\{\rho_{y x}^{2}\left(1+\rho_{x z}^{2}\right)-2 \rho_{y x} \rho_{x z} \rho_{y z}\right\}\right] \\
& \operatorname{M}\left(t_{6}\right)=S_{y}^{2}\left[f_{m}-f_{2} \rho_{y x}^{2}\right]
\end{aligned}
$$

## Case II:

$$
\begin{aligned}
& \mathrm{M}\left(t_{1}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{m} C_{x}^{2}+f_{1} C_{x}^{2}-2 f_{m} \rho_{y x} C_{y} C_{x}\right] \\
& \operatorname{Min} . \mathrm{M}\left(t_{2}\right)=f_{m} S_{y}^{2}\left[1-\theta \rho_{y x}^{2}\right] \\
& \mathrm{M}\left(t_{3}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{m} C_{x}^{2}+f_{1} C_{z}^{2}-2 f_{m} \rho_{y x} C_{y} C_{x}-2 f_{1} \rho_{x z} C_{x} C_{z}\right] \\
& \mathrm{M}\left(t_{4}\right)=\bar{Y}^{2}\left[f_{m} C_{y}^{2}+f_{m} C_{x}^{2}+f_{1} C_{x}^{2}+f_{1} C_{z}^{2}+2 f_{m} \rho_{y x} C_{y} C_{x}-2 f_{1} \rho_{x z} C_{x} C_{z}\right] \\
& \mathrm{M}\left(t_{5}\right)=S_{y}^{2}\left[f_{m}-f_{2} \rho_{y x}^{2}-f_{1} \rho_{y x}^{2} \rho_{x z}^{2}\right] \\
& \mathrm{M}\left(t_{6}\right)=f_{m} S_{y}^{2}\left[1-\theta \rho_{y x}^{2}\right]
\end{aligned}
$$

## MAJI ET AL.

where $\theta=\frac{f_{m}}{f_{1}+f_{m}}$.
The superiority of the suggested estimator has been demonstrated over the estimators $t_{i}(i=1,2, \ldots, 6)$ through numerical illustrations and graphical interpretation.

## Numerical Illustrations

Five natural population data sets were selected to illustrate the efficiency of the proposed estimator. The source of the populations, the nature of the variables $y, x$, $z$ and the values of the various parameters are as follows:

Population I: (Murthy, 1967)
y: Area under wheat in 1964.
$x$ : Area under wheat in 1963.
$z$ : Cultivated area in 1961.

Population II: (Sukhatme \& Sukhatme, 1970)
y: Area (acres) under wheat in 1937.
$x$ : Area (acres) under wheat in 1936.
$z$ : Total cultivated area (acres) in 1931.

Population III: (S. K. Srivastava, 1971)
$y$ : yield per plant.
$x$ : Height of the plant.
$z$ : Base diameter.

Population IV: (Anderson, 1958)
$y$ : Head length of second son.
$x$ : Head length of first son.
$z$ : Head breadth of second son.

Population V: (R. S. Srivastava, Srivastava, \& Khare, 1989)
$y$ : measurement of weight of children.
$x$ : Mid-arm circumference of children.
$z$ : Skull circumference of children.

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

Table 1. Parametric values of different populations

| Population | $\boldsymbol{N}$ | $\boldsymbol{n}$ | $\boldsymbol{m}$ | $\overline{\boldsymbol{Y}} \boldsymbol{r}$ | $\boldsymbol{\rho}_{\boldsymbol{y x}}$ | $\boldsymbol{\rho}_{\mathbf{y z}}$ | $\boldsymbol{\rho}_{\boldsymbol{x z}}$ | $\boldsymbol{C}_{\boldsymbol{y}}$ | $\boldsymbol{C}_{\boldsymbol{x}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | $\boldsymbol{C}_{\boldsymbol{z}}$

Table 2. PREs of different estimators (Case I)

| Population | $\overline{\boldsymbol{y}}_{\boldsymbol{m}}$ | $\boldsymbol{t}_{\boldsymbol{1}}$ | $\boldsymbol{t}_{\boldsymbol{2}}$ | $\boldsymbol{t}_{\boldsymbol{3}}$ | $\boldsymbol{t}_{\boldsymbol{4}}$ | $\boldsymbol{t}_{\boldsymbol{5}}$ | $\boldsymbol{t}_{\boldsymbol{6}}$ | $\boldsymbol{T}_{\boldsymbol{R}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| I | 100 | $*$ | 380.6032 | $*$ | $*$ | 602.4841 | 380.6032 | 859.4502 |
| II | 100 | 147.7505 | 148.5310 | 566.9582 | ${ }^{*}$ | 557.1949 | 148.5310 | 591.1479 |
| III | 100 | 128.6936 | 140.7686 | 159.0693 | ${ }^{*}$ | 156.4518 | 140.7686 | 202.8928 |
| IV | 100 | 122.5372 | 126.6649 | 178.8188 | $*$ | 190.0258 | 126.6649 | 201.7059 |
| V | 100 | 120.9633 | 120.9751 | 131.9087 | $*$ | 118.3101 | 120.9751 | 165.7976 |

Table 3. PREs of different estimators (Case II)

| Population | $\overline{\boldsymbol{y}}_{\boldsymbol{m}}$ | $\boldsymbol{t}_{\boldsymbol{1}}$ | $\boldsymbol{t}_{\boldsymbol{2}}$ | $\boldsymbol{t}_{\boldsymbol{3}}$ | $\boldsymbol{t}_{\boldsymbol{4}}$ | $\boldsymbol{t}_{\boldsymbol{5}}$ | $\boldsymbol{t}_{\boldsymbol{6}}$ | $\boldsymbol{T}_{\boldsymbol{R}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| I | 100 | $*$ | 426.8195 | $*$ | $*$ | 577.2247 | 426.8195 | 862.0124 |
| II | 100 | $*$ | 214.2050 | 286.3820 | $*$ | 330.5549 | 214.2050 | 590.4718 |
| III | 100 | $*$ | 159.5903 | 139.4500 | $*$ | 143.0017 | 159.5903 | 242.1787 |
| IV | 100 | $*$ | 146.8638 | 120.7684 | $*$ | 158.6188 | 146.8638 | 202.5066 |
| V | 100 | $*$ | 126.1805 | 116.6850 | $*$ | 121.0860 | 126.1805 | 200.0489 |

The values of various parameters obtained from the above populations are presented in Table 1.

To have a tangible idea about the performance of the proposed estimator $T_{R}$, the percent relative efficiencies (PREs) of $T_{R}$ and other estimators were computed with respect to the sample mean estimator $\bar{y}_{m}$, and the results are demonstrated in Tables 2-3. The PRE of an estimator $T$ with respect to a sample mean estimator $\bar{y}$ is defined as

$$
\begin{equation*}
\operatorname{PRE}=\frac{\mathrm{V}(\bar{y})}{\mathrm{M}(T)} \times 100 \tag{27}
\end{equation*}
$$

where $\mathrm{M}(T)$ denotes the MSE/Minimum MSE of an estimator $T$.

## MAJI ET AL.

## Graphical Interpretation

The performance of the proposed estimator is illustrated by means of pictorial representation for different choices of correlations. This could not only improve the readability of the results but also allows the comparison of a much denser grid of different correlation values. For $N=100, n=50, m=20$, and different values of $\rho_{y x}$, $\rho_{y z}, \rho_{x z}$, the PREs of the proposed estimator $T_{R}$ with respect to $\bar{y}_{m}$ are computed and presented in Figures 1-2. Note that the $X$-axis, $Y$-axis, and $Z$-axis are denoting $\rho_{y x}$, $\rho_{y z}$, and PRE, respectively, and that $\rho_{x z}$ is assumed to be 0.5 .


Figure 1. PRE of $T_{R}$ (Case I)

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN



Figure 2. PRE of $T_{R}$ (Case II)

## Conclusion

From Table 2 and Table 3, it may be observed that, under different structures of two-phase sampling set up, the suggested estimator $T_{R}$ is superior to the existing ones. It can also be noted that, for high positive values of correlation coefficients, the estimator $T_{R}$ yields impressive gains in efficiencies over the conventional estimators of population mean.

From Figures 1 and 2, it is observed that, for fixed values of $\rho_{x z}$, the PRE of the proposed estimator is increasing with increasing values of $\rho_{y x}$ and $\rho_{y z}$. This phenomenon indicates that suggested estimator could perform satisfactorily if highly positive correlated auxiliary variables are available.

Therefore, the proposed estimator $T_{R}$ is more justified in comparison with the previous work of similar nature. Hence, it may be recommended to the survey practitioners for their use in real life problems.

## MAJI ET AL.

## References

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. New York: John Wiley \& Sons.

Bose, C. (1943). Note on the sampling error in the method of double sampling. Sankhyā: The Indian Journal of Statistics, 6(3), 329-330.

Chand, L. (1975). Some ratio type estimators based on two or more auxiliary variables (Unpublished doctoral thesis). Iowa State University, Ames, IA.

Choudhury, S., \& Singh, B. K. (2012). A class of chain ratio-product type estimators with two auxiliary variables under double sampling scheme. Journal of the Korean Statistical Society, 41(2), 247-256. doi: 10.1016/j.jkss.2011.09.002

Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for ratio gain to total produce. The Journal of Agricultural Science, 30(02), 262-275. doi: 10.1017/S0021859600048012

Dash, P. R., \& Misra, G. (2011). An improved class of estimators in twophase sampling using two auxiliary variables. Communication in Statistics Theory and Methods, 40(24), 4347-4352. doi: 10.1080/03610920903094683

Gupta, S., \& Shabbir, J. (2007). On the use of transformed auxiliary variables in estimating population mean by using two auxiliary variables. Journal of Statistical Planning and Inference, 137(5), 1606-1611. doi:
10.1016/j.jspi.2006.09.008

Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. Metrika, 27(1), 217-223. doi: 10.1007/BF01893599

Kiregyera, B. (1984). Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika, 31(1), 215-226. doi: 10.1007/BF01915203

Murthy, M. N. (1967). Sampling theory and methods. Calcutta, India: Statistical Publishing Society.

Reddy, V. N. (1978). A study on the use of prior knowledge on certain population parameters in estimation. Sankhyā: The Indian Journal of Statistics, 40(1), 29-37.

Sahoo, J., \& Sahoo, L. N. (1993). A class of estimators in two-phase sampling using two auxiliary variables. Journal of the Indian Statistical Association, 37, 107-114.

## EFFICIENT AND UNBIASED ESTIMATION OF POPULATION MEAN

Shukla, D., Pathak, S., \& Thakur, N. S. (2012). Estimation of population mean using two auxiliary sources in sample surveys. Statistics in Transition - new series, 13(1), 21-36.

Singh, H. P., \& Espejo, M. R. (2007). Double sampling ratio-product estimator of a finite population mean in sampling surveys. Journal of Applied Statistics, 34(1), 74-85. doi: 10.1080/02664760600994562

Srivastava, R. S., Srivastava, S. P., \& Khare, B. B. (1989). Chain ratio type estimator for ratio of two population means using auxiliary character. Communications in Statistics - Theory and Methods, 18(10), 3917-3926. doi: 10.1080/03610928908830131

Srivastava, S. K. (1971). Generalized estimator for the mean of a finite population using multi auxiliary information. Journal of the American Statistical Association, 66(334), 404-407. doi: 10.1080/01621459.1971.10482277

Sukhatme, P. V., \& Sukhatme, B. V. (1970). Sampling theory of surveys with application. Ames, IA: Iowa State University Press.

Tracy, D. S., Singh, H. P., \& Singh, R. (1996). An alternative to the ratio-cum-product estimator in sample surveys. Journal of Statistical Planning and Inference, 53(3), 375-387. doi: 10.1016/0378-3758(95)00136-0

# An Improved Generalized Estimation Procedure of Current Population Mean in Two-Occasion Successive Sampling 

G. N. Singh<br>Indian School of Mines<br>Dhanbad, India

Alok Kumar Singh<br>Indian School of Mines<br>Dhanbad, India

Anup Kumar Sharma<br>Indian School of Mines<br>Dhanbad, India

The present work is an attempt to make use of several auxiliary variables on both occasions for improving the precision of estimates for the current population mean in two-occasion successive sampling. A generalized exponential-cum-regression type estimator of the current population mean is proposed and its optimum replacement strategy has been discussed. Empirical studies are carried out to show the dominance of the proposed estimation procedure over the sample mean estimator and natural successive sampling estimator. Empirical results have been interpreted and suitable recommendations are put forward to survey practitioners.

Keywords: Successive sampling, auxiliary information, bias, mean square error, optimum replacement strategy

## Introduction

There are many problems of practical interest in different fields of the applied and environmental sciences where the various characters of interest have tendencies to change over time. It is often required to monitor the behaviors of such characters at different points of time (occasions) and the patterns of variations occurring over the period of time. For example, an investigator or owner involved in the cold drinks industry may be interested (a) to know the average or total sale of cold drinks in the different seasons, (b) to know the pattern of change in average or total sale of cold drinks in two different seasons, or (c) they may be simultaneously interested to know both (a) and (b). These kinds of problems are well answered by the tools of successive (rotation) sampling.

[^11]
## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

The theory of successive (rotation) sampling was initiated by Jessen (1942), where the idea of using the available information gathered on previous occasions during the past surveys was suggested. Jessen (1942) used past information in order to make current estimates more precise in agronomical surveys. This idea was further explored by Patterson (1950), Rao and Graham (1964), Gupta (1979), Das (1982), and Chaturvedi and Tripathi (1983), among others. Sen (1971) extended this theory by utilizing the information on two auxiliary variables, which was available on previous occasions, and suggested estimators of current the population mean in two-occasion successive sampling. Sen (1972; 1973) generalized his idea for several auxiliary variables. V. K. Singh, Singh, and Shukla (1991) and G. N. Singh and Singh (2001) used the auxiliary information from the current occasion for estimating the current population mean in two-occasion successive sampling. G. N. Singh (2003) extended this work for h-occasion successive sampling.

In many situations, information on an auxiliary variable may be readily available on the first as well as on the second occasion. For instance, to study the problems related to the public health and welfare of a state or a country, several factors that can be treated as auxiliary variables, such as the number of beds, doctors, and supporting staff in different hospitals, the amount of funds available for medicine, etc. may be known well in advance. Likewise, in other cases, there may be information available on several auxiliary variables and, if efficiently utilized, the estimates could be made more precise.

Utilizing the auxiliary information on both occasions, Feng and Zou (1997), Biradar and Singh (2001), G. N. Singh (2005), G. N. Singh and Priyanka (2006; 2007; 2008; 2010), G. N. Singh and Karna (2009), H. P. Singh and Vishwakarma (2009), G. N. Singh and Prasad (2010), G. N. Singh, Karna, and Prasad (2011), H. P. Singh, Tailor, Singh, and Kim (2011), G. N. Singh and Prasad (2013), and G. N. Singh and Homa (2013) proposed varieties of estimators of the population mean on the current (second) occasions in two-occasion successive sampling.

Motivated with these arguments, the objective of the present work is to propose a more precise estimator of the population mean on the current occasion using the information on $p(p \geq 2)$ stable auxiliary variables which are readily available on both occasions. Utilizing the information on $p$ auxiliary variables, a generalized exponential-cum-regression type estimator of the current population mean in two-occasion successive sampling has been proposed. The dominance of the proposed estimator has been shown over the sample mean and natural successive sampling estimators. Empirical studies have been carried out to justify the proposition of estimator. Results are interpreted, and suitable recommendations have been made.

## SINGH ET AL.

## Formulation of Estimator

Let $\mathrm{U}=\left(U_{1}, U_{2}, \ldots, U_{N}\right)$ be a finite population of $N$ units which has been sampled over two occasions, and let the character under study be denoted by $x(y)$ on the first (second) occasion. It is assumed that the information on $p$ stable (non-negative integer constant) auxiliary variables $z_{j}(j=1,2, \ldots, p)$, whose population means are known and closely related to $x$ and $y$, are available on the first (second) occasion. Let a simple random sample (without replacement) of size $n$ be drawn on the first occasion. A random subsample of size $m=n \lambda$ is retained (matched) for its use on the second occasion, while a fresh simple random sample (without replacement) of size $u=(n-m)=n \mu$ is drawn on the second occasion from the entire population so that the sample size on the second occasion is also $n$. Here $\lambda$ and $\mu(\lambda+\mu=1)$ are the fractions of the matched and fresh samples, respectively, on the current (second) occasion. The values of $\lambda$ or $\mu$ would be chosen optimally.

The following notations have been considered for use below:
$\bar{X}(\bar{Y})$ : The population mean of the study variable $x(y)$ on the first (second) occasion.
$\bar{Z}_{j}$ : Population mean of the $j^{\text {th }}(j=1,2, \ldots, p)$ auxiliary variable.
$\bar{x}_{n}, \bar{x}_{m}, \bar{y}_{u}, \bar{y}_{m}, \bar{z}_{j n}, \bar{z}_{j u}, \bar{z}_{j m},(j=1,2, \ldots, p)$ : The sample means of the respective variables based on the sample sizes shown in the subscript.
$\rho_{y x}, \rho_{y z_{j}}, \rho_{x z_{j}}, \rho_{z_{j} z_{k}}$ : Population correlation coefficients between the variables shown in the subscript.
$S_{x}^{2}=(N-1)^{-1} \sum_{i=1}^{N}\left(x_{i}-\bar{X}\right)^{2}$ : Population variance of the variable $x$.
$S_{y}^{2}, S_{z_{j}}^{2}$ : Population variances of the variables $y$ and $z_{j}(j=1,2, \ldots, p)$, respectively.

To estimate the population mean $\bar{Y}$ on the current (second) occasion, two independent estimators are suggested. One is a generalized exponential type estimator based on a sample of size $u(=n \mu)$ drawn afresh on the second occasion and given by

$$
\begin{equation*}
T_{u}=\bar{y}_{u} \exp \sum_{j=1}^{p}\left\{\frac{\bar{Z}_{j}-\bar{z}_{j u}}{\bar{Z}_{j}+\bar{z}_{j u}}\right\} \tag{1}
\end{equation*}
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

The second estimator is a generalized exponential-cum-regression type estimator based on the sample of size $m(=n \lambda)$ common to both the occasions and is defined as

$$
\begin{equation*}
T_{m}=\bar{y}_{m}^{*}+\sum_{j=1}^{p} b_{y z_{j}}^{(m)}\left(\bar{Z}_{j}-\bar{z}_{j m}\right) \tag{2}
\end{equation*}
$$

where

$$
\begin{aligned}
& \bar{y}_{m}^{*}=\bar{y}_{m}+b_{y x}^{(m)}\left(\bar{x}_{n}^{*}-\bar{x}_{m}^{*}\right), \quad \bar{x}_{n}^{*}=\bar{x}_{n} \exp \sum_{j=1}^{p}\left\{\frac{\bar{Z}_{j}-\bar{z}_{j n}}{\bar{Z}_{j}+\bar{z}_{j n}}\right\} \\
& \bar{x}_{m}^{*}=\bar{x}_{m} \exp \sum_{j=1}^{p}\left\{\frac{\bar{Z}_{j}-\bar{z}_{j m}}{\bar{Z}_{j}+\bar{z}_{j m}}\right\}
\end{aligned}
$$

Combining the estimators $T_{u}$ and $T_{m}$, we have the final estimator $T$ of $\bar{Y}$ given as

$$
\begin{equation*}
T=\varphi T_{u}+(1-\varphi) T_{m} \tag{3}
\end{equation*}
$$

where $\varphi(0 \leq \varphi \leq 1)$ is an unknown constant (scalar) to be determined under certain criterion.

Remark 1: The estimator $T_{u}$ is suitable for estimating the population mean on the current occasion, while the estimator $T_{m}$ is more appropriate for estimating change over two occasions. These two estimators may be derived from the estimator $T$ by choosing $\varphi$ as 1 or 0 , respectively. To handle both problems simultaneously, an optimum choice of $\varphi$ is required.

## Properties of the Proposed Estimator

## Bias and Mean Square Error

Because the estimators $T_{u}$ and $T_{m}$ are generalized exponential and generalized exponential-cum-regression type estimators, they are biased estimators of the population mean $\bar{Y}$. Therefore, the resulting estimator $T$ is also a biased estimator of $\bar{Y}$. The bias $\mathrm{B}($.$) and mean square error \mathrm{M}($.$) of the estimator T$ is derived under

## SINGH ET AL.

large sample assumption and up to the first order of approximations using the following transformations:

$$
\begin{aligned}
& \bar{y}_{u}=\left(1+e_{1}\right) \bar{Y}, \quad \bar{y}_{m}=\left(1+e_{2}\right) \bar{Y}, \quad \bar{x}_{m}=\left(1+e_{3}\right) \bar{X}, \quad \bar{x}_{n}=\left(1+e_{4}\right) \bar{X}, \\
& \bar{z}_{j u}=\left(1+e_{5 j}\right) \bar{Z}_{j}, \quad \bar{z}_{j m}=\left(1+e_{6 j}\right) \bar{Z}_{j}, \quad \bar{z}_{j n}=\left(1+e_{7 j}\right) \bar{Z}_{j}, \quad s_{y z_{j}}=\left(1+e_{8 j}\right) S_{y z_{j}}, \\
& s_{z_{j}}^{2}(m)=\left(1+e_{9 j}\right) S_{z_{j}}^{2}, \quad s_{x z_{j}}(m)=\left(1+e_{10 j}\right) S_{x z_{j}}, \quad s_{y x}(m)=\left(1+e_{11}\right) S_{y x}, \\
& s_{x}^{2}(m)=\left(1+e_{12}\right) S_{x}^{2}
\end{aligned}
$$

such that $\mathrm{E}\left(e_{i}\right)=0$ and $\mathrm{E}\left(e_{h j}\right)=0,\left|e_{i}\right| \leq 1$ for $i=1,2,3,4,11,12$ and $\left|e_{h j}\right| \leq 1$ for $h=5,6, \ldots, 10, j=1,2,3, \ldots, p$.

Under the above transformations, the estimators $T_{u}$ and $T_{m}$ take the following forms:

$$
\begin{gather*}
T_{u}=\bar{Y}\left(1+e_{1}\right) \exp \sum_{j=1}^{p}\left[-\frac{e_{5 j}}{2}\left(1+\frac{e_{5 j}}{2}\right)^{-1}\right]  \tag{4}\\
T_{m}=\bar{Y}\left(1+e_{2}\right)-\frac{\left(1+e_{11}\right)}{\left(1+e_{12}\right)} \beta_{y x} \bar{X}\binom{\left(1+e_{4}\right) \exp \sum_{j=1}^{p} \frac{e_{7 j}}{2}\left(1+\frac{e_{7 j}}{2}\right)^{-1}}{-\left(1+e_{3}\right) \exp \sum_{j=1}^{p} \frac{e_{6 j}}{2}\left(1+\frac{e_{6 j}}{2}\right)^{-1}} \tag{5}
\end{gather*}
$$

Thus, there are the following theorems:
Theorem 1: The bias of the estimator $T$ to the first order of approximations is obtained as

$$
\begin{equation*}
\mathrm{B}(T)=\varphi \mathrm{B}\left(T_{u}\right)+(1-\varphi) \mathrm{B}\left(T_{m}\right) \tag{6}
\end{equation*}
$$

where

$$
\begin{equation*}
\mathrm{B}\left(T_{u}\right)=\left(\frac{1}{u}-\frac{1}{N}\right) \bar{Y}\left\{\frac{3}{8} \sum_{j=1}^{p} \frac{\alpha_{002}}{\bar{z}_{j}^{2}}+\frac{1}{8} \sum_{j=1}^{p} \frac{\alpha_{002}}{\bar{z}_{j} \bar{z}_{k}}-\frac{1}{2} \sum_{j=1}^{p} \frac{a_{010} \alpha_{001}}{\bar{Y}_{j}}\right\} \tag{7}
\end{equation*}
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

and

$$
\begin{equation*}
\mathrm{B}\left(T_{m}\right)=\left(\frac{1}{m}-\frac{1}{n}\right)\binom{\sum_{j=1}^{p}\left(\frac{1}{2} \frac{\alpha_{100} \alpha_{001}}{\bar{z}_{j}}-\frac{3}{8} \bar{X} \frac{\alpha_{002}}{\bar{z}_{j}^{2}}-\frac{1}{8} \bar{X} \frac{\alpha_{002}}{\bar{z}_{j} \bar{z}_{k}}+\frac{1}{2} \bar{X} \frac{\alpha_{111}}{\bar{z}_{j} S_{y x}}\right)}{-\sum_{j=1}^{p}\left(\frac{1}{2} \frac{\alpha_{201}}{\bar{z}_{j} S_{x}^{2}}-\frac{\alpha_{012}}{S_{y z_{j}}}+\frac{\alpha_{003}}{S_{z_{j}}^{2}}\right)-\frac{\alpha_{210}}{S_{y x}}+\frac{\alpha_{200}}{S_{x}^{2}}} \tag{8}
\end{equation*}
$$

where $\alpha_{p q r}=\mathrm{E}\left[\left(x_{i}-\bar{X}\right)^{p}\left(y_{i}-\bar{Y}\right)^{q}\left(z_{j}-\bar{Z}_{j}\right)^{r}\right]$ for integers $p, q, r \geq 0$ and $j=1,2, \ldots, p$.

Proof: The bias of the estimator $T$ is given by

$$
\begin{align*}
\mathrm{B}(T) & =\mathrm{E}[T-\bar{Y}]=\varphi \mathrm{E}\left(T_{u}-\bar{Y}\right)+(1-\varphi) \mathrm{E}\left(T_{m}-\bar{Y}\right)  \tag{9}\\
& =\varphi \mathrm{B}\left(T_{u}\right)+(1-\varphi) \mathrm{B}\left(T_{m}\right)
\end{align*}
$$

where $\mathrm{B}\left(T_{u}\right)=\mathrm{E}\left(T_{u}-\bar{Y}\right)$ and $\mathrm{B}\left(T_{m}\right)=\mathrm{E}\left(T_{m}-\bar{Y}\right)$.
To derive the $\mathrm{B}\left(T_{u}\right)$, proceed as follows:

$$
\begin{equation*}
\mathrm{E}\left(T_{u}-\bar{Y}\right)=\mathrm{E}\left(\bar{Y}\left(1+e_{1}\right) \exp \sum_{j=1}^{p}\left[-\frac{e_{5 j}}{2}\left(1+\frac{e_{5 j}}{2}\right)^{-1}\right]-\bar{Y}\right) \tag{10}
\end{equation*}
$$

Now, expanding the right hand side of (10) binomially and exponentially and taking expectations and retaining the terms up to the first order of approximations, we have the expression of the bias of the estimator $T_{u}$ as given in (7).

Similarly, the bias of the estimator $T_{m}$ is written as

$$
\begin{align*}
& \mathrm{E}\left(T_{m}-\bar{Y}\right)= \\
& \mathrm{E}\left[\bar{Y} e_{2}-\frac{\left(1+e_{11}\right)}{\left(1+e_{12}\right)} \beta_{y x} \bar{X}\left\{\begin{array}{l}
\left(1+e_{4}\right) \exp \left[\sum_{j=1}^{p} \frac{e_{7 j}}{2}\left(1+\frac{e_{7 j}}{2}\right)^{-1}\right] \\
\left.-\left(1+e_{3}\right) \exp \left[\sum_{j=1}^{p} \frac{e_{6 j}}{2}\left(1+\frac{e_{6 j}}{2}\right)^{-1}\right]\right]
\end{array}\right]\right. \tag{11}
\end{align*}
$$

## SINGH ET AL.

Expanding (11) binomially and exponentially, taking expectations both sides, and retaining the terms up to the first order of approximations yields the expression of the bias of the estimator $T_{m}$ as shown in (8).

Theorem 2: The mean square error of the estimator $T$ to the first degree of approximation is obtained as

$$
\begin{equation*}
\mathrm{M}(T)=\varphi^{2} \mathrm{M}\left(T_{u}\right)+(1-\varphi)^{2} \mathrm{M}\left(T_{m}\right)+2 \varphi(1-\varphi) \mathrm{C}\left(T_{u}, T_{m}\right) \tag{12}
\end{equation*}
$$

where

$$
\begin{gather*}
\mathrm{M}\left(T_{u}\right)=\left(\frac{1}{u}-\frac{1}{N}\right) S_{y}^{2}\left[1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}+\sum_{j \neq k=1}^{p} \rho_{z_{j} z_{k}}\right]  \tag{13}\\
\mathrm{M}\left(T_{m}\right)=S_{y}^{2}\left[\left(\frac{1}{m}-\frac{1}{N}\right)\left[\begin{array}{c}
1-\sum_{j=1}^{p}\left(\rho_{y z_{j}}^{2}+\rho_{y z_{j}} \rho_{y z_{k}} \rho_{z_{j} z_{k}}-\rho_{y x} \rho_{y z_{j}}^{2}\right)-2 \rho_{y x}^{2} \\
+\rho_{y x}^{2}\left(1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}+\sum_{j \neq k=1}^{p} \rho_{z_{j} z_{k}}\right)
\end{array}\right]\right]  \tag{14}\\
\mathrm{C}\left(T_{u}, T_{m}\right)=-\frac{1}{N} S_{y}^{2}\left[1-\sum_{j=1}^{p} \rho_{y z_{j}}^{2}\right] \tag{15}
\end{gather*}
$$

Proof: The mean square error of the estimator $T$ is given by

$$
\begin{align*}
\mathrm{M}(T) & =\mathrm{E}[T-\bar{Y}]^{2}=\mathrm{E}\left[\varphi\left(T_{u}-\bar{Y}\right)+(1-\varphi)\left(T_{m}-\bar{Y}\right)\right]^{2}  \tag{16}\\
& =\varphi^{2} \mathrm{M}\left(T_{u}\right)+(1-\varphi)^{2} \mathrm{M}\left(T_{m}\right)+2 \varphi(1-\varphi) \mathrm{C}\left(T_{u}, T_{m}\right)
\end{align*}
$$

where $\mathrm{C}\left(T_{u}, T_{m}\right)=\mathrm{E}\left[\left(T_{u}-\bar{Y}\right)\left(T_{m}-\bar{Y}\right)\right], \mathrm{M}\left(T_{u}\right)=\mathrm{E}\left(T_{u}-\bar{Y}\right)^{2}, \mathrm{M}\left(T_{m}\right)=\mathrm{E}\left(T_{m}-\bar{Y}\right)^{2}$. To derive the $\mathrm{M}\left(T_{u}\right)$, proceed as follows:

$$
\begin{equation*}
\mathrm{E}\left(T_{u}-\bar{Y}\right)^{2}=\mathrm{E}\left(\bar{Y}\left(1+e_{1}\right) \exp \sum_{j=1}^{p}\left[-\frac{e_{5 j}}{2}\left(1+\frac{e_{5 j}}{2}\right)^{-1}\right]-\bar{Y}\right)^{2} \tag{17}
\end{equation*}
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

Now, expanding the right hand side of (17) binomially and exponentially and taking expectations and retaining the terms up to the first order of approximations, we have the expression of the mean square error of the estimator $T_{u}$ as given in (13).

The mean square error of the estimator $T_{m}$ is written as

$$
\mathrm{E}\left(T_{m}-\bar{Y}\right)^{2}=\mathrm{E}\left[\bar{Y} e_{2}-\frac{\left(1+e_{11}\right)}{\left(1+e_{12}\right)} \beta_{y x} \bar{X}\left\{\begin{array}{l}
\left(1+e_{4}\right) \exp \left[\sum_{j=1}^{p} \frac{e_{7 j}}{2+e_{7 j}}\right]  \tag{18}\\
-\left(1+e_{3}\right) \exp \left[\sum_{j=1}^{p} \frac{e_{6 j}}{2+e_{6 j}}\right]
\end{array}\right\}\right]^{2}
$$

Expanding (18) binomially and exponential, taking expectations both sides, and retaining the terms up to the first order of approximations, the expression is derived for the mean square error of the estimator $T_{m}$ as shown in (14). Similarly, the expectation of $\mathrm{C}\left(T_{u}, T_{m}\right)$ may be derived in the form shown in (15).

Remark 2: The above results are derived under the assumption that the coefficients of variation of variables $x, y, z_{j}$, and $z_{k}$ are approximately equal. We have also considered the intuitive assumptions $\rho_{x z_{j}}=\rho_{y z_{j}}(j=1,2,3, \ldots, p)$, as suggested by Cochran (1984) and Feng and Zou (1997). In the light of these assumptions, the expression of $\mathrm{M}\left(T_{m}\right)$ takes the form as shown in (14).

## Minimum Mean Square Errors of the Estimator $\boldsymbol{T}$

Because the mean square error of the estimator $T$ in (12) is a function of the unknown constant (scalar) $\varphi$, it can be minimized with respect to $\varphi$ and, subsequently, the optimum value of $\varphi$ is obtained as

$$
\begin{equation*}
\varphi_{\text {opt }}=\frac{\mathrm{M}\left(T_{m}\right)-\mathrm{C}\left(T_{u}, T_{m}\right)}{\mathrm{M}\left(T_{u}\right)+\mathrm{M}\left(T_{m}\right)-2 \mathrm{C}\left(T_{u}, T_{m}\right)} \tag{19}
\end{equation*}
$$

From (19), substituting the value of $\varphi_{\text {opt }}$ in (12), we get the optimum mean square error of the estimator $T$ as

$$
\begin{equation*}
\mathrm{M}(T)_{\mathrm{opt}}=\frac{\mathrm{M}\left(T_{u}\right) \mathrm{M}\left(T_{m}\right)-\left\{\mathrm{C}\left(T_{u}, T_{m}\right)\right\}^{2}}{\mathrm{M}\left(T_{u}\right)+\mathrm{M}\left(T_{m}\right)-2 \mathrm{C}\left(T_{u}, T_{m}\right)} \tag{20}
\end{equation*}
$$

## SINGH ET AL.

Further substituting the values from (13)-(15) into (19) and (20), the simplified values of $\varphi_{\text {opt }}$ and $\mathrm{M}(T)_{\text {opt }}$ are obtained as

$$
\begin{align*}
\varphi_{\text {opt }} & =\frac{\mu\left(A_{9}+\mu A_{8}\right)}{A_{9}-\mu A_{12}+\mu^{2} A_{13}}  \tag{21}\\
\mathrm{M}(T)_{\text {opt }} & =\left[\frac{A_{18}-\mu^{2} A_{20}-\mu A_{21}}{A_{9}+\mu^{2} A_{13}-\mu A_{12}}\right] \frac{S_{y}^{2}}{n} \tag{22}
\end{align*}
$$

where

$$
\begin{aligned}
& A_{1}=1-\sum_{j=1}^{p} \rho_{y z_{j}}^{2}+\rho_{y z_{j}} \rho_{y z_{k}} \rho_{z_{j} z_{k}}, \quad A_{2}=\rho_{y x}^{2}\left(1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}+\frac{1}{4} \sum_{j \neq k=1}^{p} \rho_{z_{j} z_{k}}\right), \\
& A_{3}=\rho_{y x}\left(2 \rho_{y x}-\sum_{j=1}^{p} \rho_{y z_{j}} \rho_{y z_{j}}\right), \quad A_{4}=1-\sum_{j=1}^{p} \rho_{y z_{j}}^{2}, \quad A_{5}=A_{1}+A_{2}-A_{3}, \quad A_{6}=A_{2}-A_{3}, \\
& A_{7}=A_{1}-A_{4}, \quad A_{8}=A_{6}+f A_{7}, \quad A_{9}=A_{5}-A_{8}, \quad A_{10}=1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}+\frac{1}{4} \sum_{j \neq k=1}^{p} \rho_{z_{j} z_{k}}, \\
& A_{11}=A_{10}-A_{4}, \quad A_{12}=A_{10}-A_{9}+f A_{11}, \quad A_{13}=A_{8}+f A_{11}, \quad A_{14}=A_{10} A_{1}+A_{6} A_{9}, \\
& A_{15}=A_{10} A_{1}, \quad A_{16}=A_{6} A_{10}, \quad A_{17}=A_{4}^{2}-A_{15}, \quad A_{18}=A_{14}-A_{16}-f A_{15}, \\
& A_{19}=A_{14}-A_{15}-A_{16}+f A_{17}, \quad A_{20}=f A_{16}-f^{2} A_{17}, \quad A_{21}=f A_{19}-A_{16}
\end{aligned}
$$

where $f=n / N$.

## Optimum Replacement Strategy of the Estimator T

The optimum mean square error $\mathrm{M}(T)_{\text {opt }}$ in (22) is a function on $\mu$, the fraction of the sample to be drawn afresh at the second occasion. It is an important factor in reducing the cost of the survey, therefore, to determine the optimum value of $\mu$ so that $\bar{Y}$ may be estimated with maximum precision and minimum cost. We thus minimize $\mathrm{M}(T)_{\text {opt }}$ with respect to $\mu$ which results in a quadratic equation in $\mu$, which is shown as

$$
\begin{equation*}
\mu^{2} D_{1}-2 \mu D_{2}-D_{3}=0 \tag{23}
\end{equation*}
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

where $D_{1}=A_{12} A_{20}+A_{13} A_{21}, D_{2}=A_{10} A_{20}+A_{13} A_{18}, D_{3}=A_{10} A_{21}-A_{12} A_{18}$.
Solving (23) for $\mu$, the solutions of $\mu$ (say $\hat{\mu}$ ) are given as

$$
\begin{equation*}
\hat{\mu}=\frac{D_{2} \pm \sqrt{D_{2}^{2}+D_{1} D_{3}}}{D_{1}} \tag{24}
\end{equation*}
$$

From (24), it is clear that the real values of $\hat{\mu}$ exist IFF the quantity under the square root is greater than or equal to zero. For any combinations of correlations which satisfy this condition for real solutions, two real values of $\hat{\mu}$ are possible. Hence, while choosing the values of $\hat{\mu}$, it should be remembered that $0 \leq \hat{\mu} \leq 1$, and that all other values of $\hat{\mu}$ are said to be inadmissible. If both the values of $\hat{\mu}$ are admissible, the lowest one is the best choice as it reduces the cost of the survey. From (24), substituting the admissible value of $\hat{\mu}$ (say $\mu_{0}$ ) in (22), we have the optimum value of mean square error of the estimator $T$, which is shown below:

$$
\begin{equation*}
\mathrm{M}(T)_{\mathrm{opt}}^{*}=\left[\frac{A_{18}-\mu_{0}^{2} A_{20}-\mu_{0} A_{21}}{A_{10}+\mu_{0}^{2} A_{13}-\mu_{0} A_{12}}\right] \frac{S_{y}^{2}}{n} \tag{25}
\end{equation*}
$$

## Special Case

When the $p$ auxiliary variates are mutually uncorrelated, i.e., $\rho_{z_{j} z_{k}}=0$ for $\mathrm{j} \neq k=1,2,3, \ldots, p$, then the expression of the optimum values of $\mu$ and $\mathrm{M}(T)_{\mathrm{opt}}$ reduce to

$$
\begin{equation*}
\hat{\mu}=\frac{D_{2}^{*} \pm \sqrt{D_{2}^{* 2}+D_{1}^{*} D_{3}^{*}}}{D_{1}^{*}} \tag{26}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{M}(T)_{o p t}^{*}=\left[\frac{A_{18}^{*}-\mu_{0}^{2} A_{20}^{*}-\mu_{0} A_{21}^{*}}{A_{10}^{*}+\mu_{0}^{2} A_{13}^{*}-\mu_{0} A_{12}^{*}}\right] \frac{S_{y}^{2}}{n} \tag{27}
\end{equation*}
$$

where

## SINGH ET AL.

$$
\begin{aligned}
& D_{1}^{*}=A_{12}^{*} A_{20}^{*}+A_{13}^{*} A_{21}^{*}, \quad D_{2}^{*}=A_{10}^{*} A_{20}^{*}+A_{13}^{*} A_{18}^{*}, \quad D_{3}^{*}=A_{10}^{*} A_{21}^{*}-A_{12}^{*} A_{18}^{*}, \\
& A_{1}^{*}=1-\sum_{j=1}^{p} \rho_{y z_{j}}^{2}, \quad A_{2}^{*}=\rho_{y x}^{2}\left(1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}\right), \quad A_{5}^{*}=A_{1}^{*}+A_{2}^{*}-A_{3}, \quad A_{6}^{*}=A_{2}^{*}-A_{3}, \\
& A_{7}^{*}=A_{1}^{*}-A_{4}, \quad A_{8}^{*}=A_{6}^{*}+f A_{7}^{*}, \quad A_{9}^{*}=A_{5}^{*}-A_{8}^{*}, \quad A_{10}^{*}=1+\frac{p}{4}-\sum_{j=1}^{p} \rho_{y z_{j}}, \\
& A_{11}^{*}=A_{10}^{*}-A_{4}, \quad A_{12}^{*}=A_{10}^{*}-A_{9}^{*}+f A_{11}^{*}, \quad A_{13}^{*}=A_{8}^{*}+f A_{11}^{*}, \quad A_{14}^{*}=A_{10}^{*} A_{1}^{*}+A_{6}^{*} A_{9}^{*}, \\
& A_{15}^{*}=A_{10}^{*} A_{1}^{*}, \quad A_{16}^{*}=A_{6}^{*} A_{10}^{*}, \quad A_{17}^{*}=A_{4}^{2}-A_{15}^{*}, \quad A_{18}^{*}=A_{14}^{*}-A_{16}^{*}-f A_{15}^{*}, \\
& A_{19}^{*}=A_{14}^{*}-A_{15}^{*}-A_{16}^{*}+f A_{17}^{*}, \quad A_{20}^{*}=f A_{16}^{*}-f^{2} A_{17}^{*}, \quad A_{21}^{*}=f A_{19}^{*}-A_{16}^{*}
\end{aligned}
$$

## Efficiency Comparison

The percent relative efficiencies of the estimator $T$ with respect to (i) the sample mean estimator $\bar{y}_{n}$ when there is no matching and (ii) $\hat{\bar{Y}}=\varphi_{u}{ }^{*} \bar{y}_{u}+\left(1-\varphi^{*}\right) \bar{y}_{m}^{\prime}$ when no additional auxiliary information is used at any occasion, where $\bar{y}_{m}^{\prime}=\bar{y}_{m}+\beta_{y x}\left(\bar{x}_{n}-\bar{x}_{m}\right)$, have been obtained for different choices of the correlations involved. Since $\bar{y}_{n}$ and $\hat{\bar{Y}}$ are unbiased estimators of $\bar{Y}$ following Sukhatme, Sukhatme, Sukhatme, and Asok (1984), the variance of $\bar{y}_{n}$ and optimum variance of $\hat{\bar{Y}}$ are given by

$$
\begin{gather*}
\mathrm{V}\left(\bar{y}_{n}\right)=\left(\frac{1}{n}-\frac{1}{N}\right) S_{y}^{2}  \tag{28}\\
\mathrm{~V}(\hat{\bar{Y}})_{\mathrm{opt}}=\left[1+\sqrt{1-\rho_{y x}^{2}}\right] \frac{S_{y}^{2}}{2 n}-\frac{S_{y}^{2}}{N} \tag{29}
\end{gather*}
$$

The percent relative efficiencies $E_{1}$ and $E_{2}$ of $T$ (under optimal condition) with respect to $\bar{y}_{n}$ and $\hat{\bar{Y}}$, respectively, are given by

$$
E_{1}=\frac{\mathrm{V}\left(\bar{y}_{n}\right)}{\mathrm{M}(T)_{\mathrm{opt}}^{*}} \times 100, \quad E_{2}=\frac{\mathrm{V}(\hat{\bar{Y}})_{\mathrm{opt}}}{\mathrm{M}(T)_{\mathrm{opt}}^{*}} \times 100
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

## Empirical Study

The expression of the optimum $\mu$ (i.e., $\mu_{0}$ ) and the percent relative efficiencies $E_{1}$ and $E_{2}$ are in terms of population correlation coefficients. Therefore, the values of $\mu_{0}, E_{1}$, and $E_{2}$ have been computed for different choices of positive correlations, while the value of $f(=n / N)$ (sampling fraction) is chosen to be 0.1 . For empirical studies, cases of $p=2$ and 3 have been considered.

## Case 1

For $p=2$ and assuming that the two auxiliary variables are correlated, i.e., $\rho_{z_{1} z_{2}} \neq 0$, the values of $A_{1}, A_{2}, A_{3}, A_{4}, A_{9}$, and $A_{10}$ take the form

$$
\begin{aligned}
& A_{1}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}\right)+2 \rho_{y z_{1}} \rho_{y z_{2}} \rho_{z_{1} z_{2}}, \quad A_{2}=\rho_{y x}^{2}\left(\frac{3}{2}+2 \rho_{z_{1} z_{2}}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}\right)\right), \\
& A_{3}=\rho_{y x}\left(2 \rho_{y x}-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}\right)\right), \quad A_{4}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}\right), \\
& A_{10}=\frac{3}{2}+\frac{1}{2} \rho_{z z_{2}}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}\right)
\end{aligned}
$$

Substituting these values in (24) and (25) yields the values of optimum $\mathrm{M}(T)_{\mathrm{opt}}^{*}$, $E_{1}$, and $E_{2}$. For different choices of correlations, Tables 1-2 show the optimum values of $\mu$ (i.e., $\mu_{0}$ ) and percent relative efficiencies $E_{1}$ and $E_{2}$ of the estimator $T$ (under optimal condition) with respect to $\bar{y}_{n}$ and $\hat{\bar{Y}}$, respectively.

## Case 2

For $p=2$ and assuming that the two auxiliary variables are uncorrelated, i.e., $\rho_{z_{1} z_{2}}=0$, the values of $A_{1}^{*}, A_{2}^{*}$, and $A_{10}^{*}$ take the form

$$
A_{1}^{*}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}\right), \quad A_{2}^{*}=\rho_{y x}^{2}\left(\frac{3}{2}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}\right)\right), \quad A_{10}^{*}=\frac{3}{2}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}\right)
$$

Using these values in (26) and (27), the optimum values of $\mu, E_{1}$, and $E_{2}$ are shown in Table 3.

## SINGH ET AL.

Table 1. Optimum values of $\mu$ and percent relative efficiencies of $T$ with respect to $\bar{y}_{n}$ and $\hat{\bar{Y}}$ for $\rho_{y x}=0.3$

| $\rho_{y z_{2}}$ | $\rho_{y z_{1}}$ | 0.5 |  |  | 0.6 |  |  | 0.7 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\rho_{z_{1} z_{2}}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ |
| 0.4 | 0.3 | 0.7265 | 132.18 | 128.80 | 0.4034 | 168.19 | 163.88 | 0.3533 | 247.90 | 241.56 |
|  | 0.4 | 0.7101 | 123.49 | 120.33 | 0.3839 | 152.06 | 148.16 | 0.3872 | 210.05 | 204.67 |
|  | 0.5 | 0.7001 | 115.87 | 112.91 | 0.3700 | 138.74 | 135.19 | 0.3360 | 181.76 | 177.10 |
|  | 0.6 | 0.6932 | 109.13 | 106.34 | 0.3595 | 127.57 | 124.31 | 0.2952 | 159.92 | 155.82 |
| 0.6 | 0.3 | 0.6022 | 149.75 | 145.92 | 0.3207 | 193.77 | 188.81 | 0.3411 | 306.04 | 298.21 |
|  | 0.4 | 0.5842 | 138.08 | 134.54 | 0.2827 | 170.36 | 166.00 | 0.2607 | 243.37 | 237.15 |
|  | 0.5 | 0.5719 | 128.09 | 124.81 | 0.2548 | 151.91 | 148.03 | 0.1979 | 200.80 | 195.67 |
|  | 0.6 | 0.5630 | 119.44 | 116.39 | 0.2335 | 137.04 | 133.53 | 0.1475 | 170.24 | 165.88 |
| 0.8 | 0.3 | 0.5882 | 174.09 | 169.63 | 0.3120 | 234.31 | 228.31 | 0.2669 | 418.57 | 407.85 |
|  | 0.4 | 0.5517 | 157.66 | 153.63 | 0.2512 | 197.91 | 192.84 | 0.1760 | 298.60 | 290.96 |
|  | 0.5 | 0.5252 | 144.02 | 140.34 | 0.2051 | 170.92 | 166.55 | 0.1034 | 228.96 | 223.10 |
|  | 0.6 | 0.5050 | 132.53 | 129.14 | 0.1690 | 150.21 | 146.36 | 0.0441 | 184.02 | 179.31 |

Table 2. Optimum values of Optimum values of $\mu$ and percent relative efficiencies of $T$ with respect to $\bar{y}_{n}$ and $\hat{\bar{Y}}$ for $\rho_{y x}=0.5$

| $\rho_{y z_{2}}$ | $\rho_{y z}$ |  | 0.5 |  | 0.6 |  |  | 0.7 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\rho_{z_{1} z_{2}}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ |
| 0.4 | 0.3 | * | -- | -- | 0.3809 | 170.91 | 158.19 | 0.3568 | 249.91 | 231.31 |
|  | 0.4 | 0.7440 | 123.96 | 114.73 | 0.3815 | 152.39 | 141.05 | 0.3875 | 208.31 | 192.80 |
|  | 0.5 | 0.6739 | 115.31 | 106.73 | 0.3779 | 137.57 | 127.33 | 0.3431 | 178.37 | 165.10 |
|  | 0.6 | 0.6411 | 107.74 | 100.72 | 0.3738 | 125.42 | 116.08 | 0.3120 | 155.88 | 144.27 |
| 0.6 | 0.3 | 0.6748 | 152.04 | 140.72 | 0.2932 | 196.4 | 181.78 | 0.3407 | 307.53 | 284.64 |
|  | 0.4 | 0.5913 | 138.47 | 128.16 | 0.2827 | 170.36 | 157.68 | 0.2673 | 240.90 | 222.97 |
|  | 0.5 | 0.5579 | 127.09 | 117.63 | 0.2740 | 150.50 | 139.29 | 0.2192 | 197.38 | 182.69 |
|  | 0.6 | 0.5386 | 117.45 | 108.71 | 0.2669 | 134.83 | 124.80 | 0.1851 | 166.90 | 154.48 |
| 0.8 | 0.3 | 0.6209 | 176.39 | 163.26 | 0.3015 | 236.81 | 219.19 | 0.2670 | 418.02 | 386.91 |
|  | 0.4 | 0.5506 | 157.54 | 145.81 | 0.2567 | 197.19 | 182.51 | 0.1865 | 294.69 | 272.76 |
|  | 0.5 | 0.5134 | 142.28 | 131.69 | 0.2285 | 168.91 | 156.34 | 0.1319 | 225.58 | 208.79 |
|  | 0.6 | 0.4896 | 129.72 | 120.07 | 0.2089 | 147.72 | 136.73 | 0.0924 | 181.87 | 168.33 |

[^12]
## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

Table 3. Optimum values of Optimum values of $\mu$ and percent relative efficiencies of $T$ with respect to $\overline{y_{n}}$ and $\hat{\bar{Y}}$ for $\rho_{z_{1} z_{2}}=0.0$

| $\boldsymbol{\rho}_{\boldsymbol{y} x}$ | $\boldsymbol{\rho}_{y z_{1}}$ | 0.5 |  |  | 0.6 |  |  | 0.7 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\boldsymbol{\rho}_{y z_{2}}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ | $\mu_{0}$ | $E_{1}$ | $E_{2}$ |
| 0.5 | 0.3 | 0.4203 | 138.95 | 128.61 | 0.4317 | 158.35 | 146.57 | 0.3727 | 187.50 | 173.54 |
|  | 0.4 | 0.4874 | 156.18 | 144.55 | 0.5084 | 181.27 | 167.78 | 0.4661 | 219.80 | 203.44 |
|  | 0.5 | 0.5084 | 181.27 | 167.78 | 0.5359 | 216.08 | 200.00 | 0.4938 | 272.88 | 252.57 |
|  | 0.6 | 0.4661 | 219.80 | 203.44 | 0.4938 | 272.88 | 252.57 | 0.4112 | 371.39 | 343.75 |
| 0.7 | 0.3 | 0.5632 | 157.93 | 132.85 | 0.5865 | 185.15 | 155.75 | 0.6036 | 228.38 | 192.11 |
|  | 0.4 | 0.6015 | 182.74 | 153.71 | 0.6312 | 220.87 | 185.79 | 0.6662 | 287.47 | 241.82 |
|  | 0.5 | 0.6312 | 220.87 | 185.79 | 0.6759 | 281.34 | 236.66 | 0.7712 | 417.32 | 351.04 |
|  | 0.6 | 0.6662 | 287.47 | 241.82 | 0.7712 | 417.32 | 351.04 | * | -- | -- |
| 0.9 | 0.3 | 0.9388 | 268.34 | 184.24 | * | -- | -- | * | -- | -- |
|  | 0.4 | * | -- | -- | * | -- | -- | * | -- | -- |
|  | 0.5 | * | -- | -- | * | -- | -- | * | -- | -- |
|  | 0.6 | * | -- | -- | * | -- | -- | * | -- | -- |

Note: "*" indicates $\mu_{0}$ does not exists and "--" indicates no gain.

## Case 3

For $p=3$ and assuming that the two auxiliary variables are correlated, i.e., $\rho_{z_{j} z_{k}} \neq 0$ for $j \neq k=1,2,3$, the values of $A_{1}, A_{2}, A_{3}, A_{4}$, and $A_{10}$ take the form

$$
\begin{aligned}
& A_{1}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}+\rho_{y z_{3}}^{2}\right)+2\left(\rho_{y z_{1}} \rho_{y z_{2}} \rho_{z_{1} z_{2}}+\rho_{y z_{1}} \rho_{y z_{3}} \rho_{z z_{1}}+\rho_{y z_{2}} \rho_{y z_{3}} \rho_{z_{2} z_{3}}\right), \\
& A_{2}=\rho_{y x}^{2}\left(\frac{7}{4}+2\left(\rho_{z z_{2}}+\rho_{z_{1} z_{3}}+\rho_{z 2 z_{3}}\right)-\left(\rho_{y z_{1}}+\rho_{y z_{2}}+\rho_{y z_{3}}\right)\right), \\
& A_{3}=\rho_{y x}\left(2 \rho_{y x}-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}+\rho_{y z_{3}}^{2}\right)\right), \quad A_{4}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}+\rho_{y z_{3}}^{2}\right), \\
& A_{10}=\frac{7}{4}+\frac{1}{2}\left(\rho_{z z_{2}}+\rho_{z_{1} z_{3}}+\rho_{z_{2} z_{3}}\right)-\left(\rho_{y z_{1}}+\rho_{y z_{2}}+\rho_{y z_{3}}\right)
\end{aligned}
$$

In this case there are seven different correlations. For a few sets of these seven correlations, the optimum value of $\mu$ (i.e., $\mu_{0}$ ) and percent relative efficiencies $E_{1}$ and $E_{2}$ of the estimator $T$ (under optimal condition) with respect to $\bar{y}_{n}$ and $\hat{\bar{Y}}$ have been computed and shown below:

## SINGH ET AL.

Set 1:

$$
\begin{aligned}
& \rho_{y x}=0.3, \quad \rho_{y z_{1}}=0.5, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.5, \quad \rho_{z_{1} z_{2}}=0.3, \quad \rho_{z_{1} z_{3}}=0.4, \\
& \rho_{z_{2} z_{3}}=0.6, \quad \mu_{0}=0.3664, \quad E_{1}=104.37, \quad E_{2}=101.37
\end{aligned}
$$

Set 2:

$$
\begin{aligned}
& \rho_{y x}=0.3, \quad \rho_{y z_{1}}=0.6, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.5, \quad \rho_{z_{1} z_{2}}=0.3, \quad \rho_{z_{1} z_{3}}=0.4, \\
& \rho_{z_{2} z_{3}}=0.6, \quad \mu_{0}=0.2900, \quad E_{1}=110.60, \quad E_{2}=107.77
\end{aligned}
$$

Set 3:

$$
\begin{aligned}
& \rho_{y x}=0.3, \quad \rho_{y z_{1}}=0.7, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.5, \quad \rho_{z_{1} z_{2}}=0.3, \quad \rho_{z_{1} z_{3}}=0.4, \\
& \rho_{z_{2} z_{3}}=0.6, \quad \mu_{0}=0.2393, \quad E_{1}=119.33, \quad E_{2}=116.27
\end{aligned}
$$

## Set 4:

$$
\begin{aligned}
& \rho_{y x}=0.3, \quad \rho_{y z_{1}}=0.8, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.5, \quad \rho_{z_{1} z_{2}}=0.3, \quad \rho_{z_{1} z_{3}}=0.4, \\
& \rho_{z_{2} z_{3}}=0.6, \quad \mu_{0}=0.2105, \quad E_{1}=131.59, \quad E_{2}=128.22
\end{aligned}
$$

## Case 4

For $p=3$ and assuming that the two auxiliary variables are uncorrelated, i.e., $\rho_{z_{1} z_{2}}=0$ for $j \neq k=1,2,3$, the values of $A_{1}, A_{2}$, and $A_{10}$ take the form

$$
\begin{aligned}
& A_{1}=1-\left(\rho_{y z_{1}}^{2}+\rho_{y z_{2}}^{2}+\rho_{y z_{3}}^{2}\right), \quad A_{2}=\rho_{y x}^{2}\left(\frac{7}{4}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}+\rho_{y z_{3}}\right)\right), \\
& A_{10}=\frac{7}{4}-\left(\rho_{y z_{1}}+\rho_{y z_{2}}+\rho_{y z_{3}}\right)
\end{aligned}
$$

For a few sets of the above four correlations, the values of the optimum value of $\mu$ (i.e., $\mu_{0}$ ) and percent relative efficiencies $E_{1}$ and $E_{2}$ are shown below:

## Set 1:

$$
\begin{aligned}
& \rho_{y x}=0.3, \quad \rho_{y z_{1}}=0.5, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.4, \quad \mu_{0}=0.6004, \quad E_{1}=382.42, \\
& E_{2}=372.64
\end{aligned}
$$

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

Set 2:
$\rho_{y x}=0.4, \quad \rho_{y z_{1}}=0.5, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.4, \quad \mu_{0}=0.7981, \quad E_{1}=397.89$, $E_{2}=379.44$

Set 3:
$\rho_{y x}=0.5, \quad \rho_{y z_{1}}=0.5, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.4, \quad \mu_{0}=0.3807, \quad E_{1}=449.12$,
$E_{2}=415.67$

## Set 4:

$\rho_{y x}=0.6, \quad \rho_{y z_{1}}=0.5, \quad \rho_{y z_{2}}=0.6, \quad \rho_{y z_{3}}=0.4, \quad \mu_{0}=0.6317, \quad E_{1}=568.19$,
$E_{2}=505.06$

## Conclusion

1. From Tables 1-2 it is vindicated that:
a. For the fixed values of $\rho_{y x}, \rho_{z_{1} z_{2}}$, and $\rho_{y z_{1}}$, the values of $\mu_{0}$ decrease and $E_{1}$ and $E_{2}$ increase with the increasing values of $\rho_{y z_{2}}$. Similarly, for fixed values of $\rho_{y x}, \rho_{z_{1} z_{2}}$, and $\rho_{y z_{2}}$, the optimum value of $\mu_{0}$ decrease and $E_{1}$ and $E_{2}$ increase with the increasing values of $\rho_{y z_{1}}$. These patterns indicate that a smaller fresh sample on the current occasion is required if highly correlated auxiliary variables are available.
b. For the fixed values of $\rho_{y x}, \rho_{y z_{1}}$, and $\rho_{y z_{2}}$, the values of $\mu_{0}, E_{1}$, and $E_{2}$ decrease with the increasing values of $\rho_{z_{1} z_{2}}$; this means that the auxiliary variables are quite sensitive with respect to the relation between them.
2. From Table 3, i.e., when the auxiliary variables are uncorrelated, it has been observed that
a. For fixed values of $\rho_{y z_{1}}$ and $\rho_{y z_{2}}$, the values of $E_{1}$ and $E_{2}$ increase with increasing value of $\rho_{y x}$, while no definite patterns are observed in $\mu_{0}$.

## SINGH ET AL.

b. For fixed values of $\rho_{y x}$ and $\rho_{y z_{1}}$, the values of $E_{1}$ and $E_{2}$ increase with increasing value of $\rho_{y z_{2}}$, while no definite patterns are observed in $\mu_{0}$. Similar patterns are visible for the case when the values of $\rho_{y x}$ and $\rho_{y z_{2}}$ are fixed and increasing values of $\rho_{y z_{1}}$ are observed.
3. For $p=3$ and when the three auxiliary variables are uncorrelated, for fixed values of $\rho_{y x}, \rho_{z_{1} z_{2}}, \rho_{z_{2} z_{3}}, \rho_{z_{1} z_{3}}, \rho_{y z_{2}}$, and $\rho_{y z_{3}}$, the values of $\mu_{0}$ decrease while $E_{1}$ and $E_{2}$ increase with the increasing values of $\rho_{y z_{1}}$. Similar patterns are observed if the case for the increasing values of $\rho_{y z_{2}}$ or $\rho_{y z_{3}}$ is taken into account.
4. For $p=3$ and when the three auxiliary variables are mutually correlated, we observed that no specific pattern is seen as for so many combinations of correlations the optimum values of $\mu_{0}$ do not exist. This behavior suggests that the correlation between the auxiliary variable do not play a significant role in terms of the proposed estimator.
5. It could be seen that the results are more appreciable for one and two auxiliary variables, while when the number of auxiliary variables increases, the expressions become complex due to the increase in the number of correlations. Hence, practically, it is more realistic to use two auxiliary variables out of several available auxiliary variables.

Thus, it is clear that the use of the auxiliary variables is highly rewarding in terms of the proposed estimator. It is also clear that, if the information on highly correlated auxiliary variables is used, only a relatively small fraction of the sample on the current (second) occasion is desired to be replaced by a fresh sample, which reduces the cost of the survey. Hence, it can be recommended for future use.

## References

Biradar, R. S., \& Singh, H. P. (2001). Successive sampling using auxiliary information on both occasions. Bulletin of the Calcutta Statistical Association, 51(203-204), 243-251.

Chaturvedi, D. K., \& Tripathi, T. P. (1983). Estimation of population ratio on two occasions using multivariate auxiliary information. Journal of the Indian Statistical Association, 21, 113-120.

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

Cochran, W. G. (1984). Sampling techniques (3rd ed.). New Delhi, India: Wiley Eastern.

Das, A. K. (1982). Estimation of population ratio on two occasions. Journal of the Indian Society of Agricultural Statistics, 34(2), 1-9.

Feng, S., \& Zou, G. (1997). Sample rotation method with auxiliary variable. Communications in Statistics - Theory and Methods, 26(6), 1497-1509. doi: 10.1080/03610929708831996

Gupta, P. C. (1979). Sampling on two successive occasions. Journal of Statistical Research, 13(13), 7-16.

Jessen, R. J. (1942). Statistical Investigation of a Sample Survey for obtaining farm facts. Agricultural Research Bulletins, 304, 1-104. Retrieved from http://lib.dr.iastate.edu/ag_researchbulletins/27/

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society. Series B (Methodological), 12(2), 241-255. Available from http://www.jstor.org/stable/2983982

Rao, J. N. K., \& Graham, J. E. (1964). Rotation design for sampling on repeated occasions. Journal of the American Statistical Association, 59(306), 492509. doi: 10.1080/01621459.1964.10482175

Sen, A. R. (1971). Successive sampling with two auxiliary variables. Sankhyā: The Indian Journal of Statistics, Series B, 33(3/4), 371-378. Available from http://www.jstor.org/stable/25051764

Sen, A. R. (1972). Successive sampling with $p(p \geq 1)$ auxiliary variables. The Annals of Mathematical Statistics, 43(6), 2031-2034. Available from http://www.jstor.org/stable/2240218

Sen, A. R. (1973). Note: Theory and application of sampling on repeated occasions with several auxiliary variables. Biometrics, 29(2), 381-385. doi: 10.2307/2529401

Singh, G. N. (2003). Estimation of population mean using auxiliary information on recent occasion in h-occasion successive sampling. Statistics in Transition, 6(4), 523-532.

Singh, G. N. (2005). On the use of chain-type ratio estimator in successive sampling. Statistics in Transition, 7(1), 21-26.

Singh, G. N., \& Homa, F. (2013). Effective rotation patterns in successive sampling over two occasions. Journal of Statistical Theory and Practice, 7(1), 146-155. doi: 10.1080/15598608.2012.755484

## SINGH ET AL.

Singh, G. N., \& Karna, J. P. (2009). Estimation of population mean on current occasions in two-occasion successive sampling. Metron, 67(1), 69-85.

Singh, G. N., Karna, J. P., \& Prasad, S. (2011). On the use of multiple auxiliary variables in estimation of current population mean in two-occasion successive sampling. Sri Lankan Journal of Applied Statistics, 12, 101-116. doi: 10.4038/sljastats.v12i0.4970

Singh, G. N., \& Prasad, S. (2010). Some estimator of population mean in two-occasion rotation patterns. Association for the Advancement of Modelling \& Simulation Techniques in Enterprises, 47(2), 1-18.

Singh, G. N., \& Prasad, S. (2013). Best linear unbiased estimators of population mean on current occasion in two-occasion successive sampling. Statistics in Transition - New Series, 14(1), 57-74.

Singh, G. N., \& Priyanka, K. (2006). On the use of chain-type ratio to difference estimator in successive sampling. International Journal of Applied Mathematics and Statistics, 5(S06), 41-49.

Singh, G. N., \& Priyanka, K. (2007). On the use of auxiliary information in search of good rotation patterns on successive occasions. International Journal of Statistics and Economics, 1(A07), 42-60.

Singh, G. N., \& Priyanka, K. (2008). On the use of several auxiliary variates to improve the precision of estimates at current occasion. Journal of the Indian Society of Agricultural Statistics, 62(3), 253-265. Retrieved from http://isas.org.in/jisas/jsp/volume/vol62/7-G.N.Singh.pdf

Singh, G. N., \& Priyanka, K. (2010). Estimation of population mean at current occasion in presence of several varying auxiliary variates in two-occasion successive sampling. Statistics in Transition - New Series, 11(1), 105-126.

Singh, G. N., \& Singh, V. K. (2001). On the use of auxiliary information in successive sampling. Journal of the Indian Society of Agricultural Statistics, 54(1), 1-12. Retrieved from
http://www.isas.org.in/jsp/volume/vol54/G.N.\ Singh\ .pdf
Singh, H. P., Tailor, R., Singh, S., \& Kim, J.-M., (2011). Estimation of population variance in successive sampling. Quality \& Quantity, 45(3), 477-494. doi: 10.1007/s11135-009-9309-9

Singh, H. P., \& Vishwakarma, G. K. (2009). A general procedure for estimating population mean in successive sampling. Communications in Statistics - Theory and Methods, 38(2), 293-308. doi: 10.1080/03610920802169594

## AN IMPROVED GENERALIZED ESTIMATION PROCEDURE

Singh, V. K., Singh, G. N., \& Shukla, D. (1991). An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions. Journal of Scientific Research, 41(C), 149-159.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., \& Asok, C. (1984). Sampling theory of surveys with applications. Ames, IA: Iowa State University Press.

# A New Exponential Type Estimator for the Population Mean in Simple Random Sampling 

Gamze Özel Kadilar

Hacettepe University
Ankara, Turkey

This paper provides a new exponential type estimator in simple random sampling for population mean. It is shown that proposed exponential type estimator is always more efficient than estimators considered by Bahl and Tuteja (1991) and Singh, Chauhan, Sawan, and Smarandache (2009). From numerical examples it is also observed that proposed modified ratio estimator performs better than existing estimators.

Keywords: Simple random sampling, Ratio and regression-type estimator, Auxiliary information, Mean squared error, Efficiency

## Introduction

The auxiliary information in sampling theory is used for improved estimation of parameters enhancing the efficiencies of the estimators. The problem of estimating the population mean in the presence of an auxiliary variable has been widely discussed in finite population sampling literature. The use of auxiliary information is well-known to improve the precision of the estimate of the population mean for the study variable. In survey sampling, ratio, product and difference methods of estimation are good examples in this context. Ratio method of estimation is quite effective when there is high positive correlation between study and auxiliary variables. However, if correlation is negative (high), the product method of estimation can be employed efficiently. In recent years, a number of research papers on ratio-type, exponential ratio-type and regressiontype estimators have appeared, based on different types of transformations. The main aspect of the paper is to obtain an estimator to predict the population mean

[^13]
## NEW EXPONENTIAL TYPE ESTIMATOR FOR THE POPULATION MEAN

which is more efficient than the ratio, product, exponential estimators of Bahl and Tuteja (1991) and Singh et al. (2009).

Consider a finite population $U=U_{1}, U_{2}, \ldots, U_{N}$ of $N$ units. Let $Y$ and $X$ stand for the variable under study and auxiliary variables, respectively. Let ( $y_{i}, x_{i}$ ), $i=1,2, \ldots, n$, denote the $n$ pair of sample observations for the study and auxiliary variables, respectively, drawn from the population size $N$ using simple random sampling without replacement (SRSWOR). Let $\bar{X}$ and $\bar{Y}$ be the population means of auxiliary and study variables, respectively, and let $\bar{x}$ and $\bar{y}$ be the respective sample means. It is well known that the sample mean $\bar{y}$ is an unbiased estimator of $\bar{Y}$ and under SRSWOR its variance is given by

$$
\begin{equation*}
V(\bar{y})=\gamma \bar{Y}^{2} C_{y}^{2} \tag{1}
\end{equation*}
$$

where $\gamma=n^{-1}(1-f), f=n / N, C_{y}^{2}=S_{y}^{2} / \bar{Y}^{2}$, and $S_{y}^{2}$ is the variance of the study variable.

Ratio and product type estimators in the simple random sampling (SRS) were considered by Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999), Singh and Tailor (2003), Singh, Tailor, Tailor, and Kakran (2004), Singh and Tailor (2005), Yadav and Kadilar (2013a), Singh et al. (2009), Singh, Chauhan, Sawan, and Smarandache (2011), Yadav and Kadilar (2013b), etc. When information is available on $X$ that is positively correlated with $Y$, the ratio estimator is suitable for estimating the population mean and it is given by

$$
\bar{y}_{R}=\bar{y} \bar{X} / \bar{x} .
$$

The mean squared error (MSE) of this estimator is

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{y}_{R}\right)=\gamma \bar{Y}^{2}\left[C_{y}^{2}-2 \rho C_{x} C_{y}+C_{x}^{2}\right], \tag{2}
\end{equation*}
$$

where $C_{x}^{2}=S_{x}^{2} / \bar{X}^{2}$, and $S_{x}^{2}$ is the variance of the auxiliary variable.
When there is a negative high correlation between $Y$ and $X$, the product estimator for $\bar{Y}$ was defined by Robson (1957) as

$$
\bar{y}_{p}=\bar{y} \bar{x} / \bar{X}
$$

and the MSE of the product estimator is given by

## GAMZE ÖZEL KADILAR

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{y}_{p}\right)=\gamma \bar{Y}^{2}\left[C_{y}^{2}+2 \rho C_{x} C_{y}+C_{x}^{2}\right] . \tag{3}
\end{equation*}
$$

Bahl and Tuteja (1991) suggested an exponential ratio type estimator for the population mean as

$$
\bar{y}_{B T}=\bar{y} \exp \left[\frac{(\bar{X}-\bar{x})}{(\bar{X}+\bar{x})}\right]
$$

and the MSE of the estimator is given by

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{y}_{B T}\right)=\gamma \bar{Y}^{2}\left[C_{y}^{2}-2 \rho C_{x} C_{y}+C_{x}^{2} / 4\right] . \tag{4}
\end{equation*}
$$

The auxiliary information associated with $X$ such as mean, median, coefficient of variation, skewness, kurtosis or correlation coefficient can be used to improve the efficiency of the estimators. Singh et al. (2009) defined a modified exponential ratio estimator using auxiliary variable information for estimating $\bar{Y}$ as

$$
\bar{y}=\bar{y} \exp \left[\frac{(a \bar{X}+b)-(a \bar{x}+b)}{(a \bar{X}+b)+(a \bar{x}+b)}\right],
$$

where ( $a \neq 0$ ), $b$ are either real numbers or the functions of the known parameters of the auxiliary variable such as coefficient of variation $\left(C_{x}\right)$, coefficient of kurtosis ( $\beta_{2}(x)$ ), and correlation coefficient.

The MSE of the modified exponential estimator is given by

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{y}_{S}\right)=\gamma \bar{Y}^{2}\left(C_{y}^{2}+\theta^{2} C_{x}^{2}-2 \theta \rho C_{x} C_{y}\right), \tag{5}
\end{equation*}
$$

where $\theta=a \bar{X} / 2(a \bar{X}+b)$.

## Suggested Estimator

Following Bahl and Tuteja (1991) and Singh et al. (2009), a modified exponential type estimator is defined for estimating $\bar{Y}$ as

## NEW EXPONENTIAL TYPE ESTIMATOR FOR THE POPULATION MEAN

$$
\begin{equation*}
\bar{y}_{P R}=\bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)^{\alpha} \exp \left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right) . \tag{6}
\end{equation*}
$$

To obtain the MSE of $\bar{y}_{P R 1}$, write $\bar{y}=\bar{Y}\left(1+e_{0}\right)$ and $\bar{x}=\bar{X}\left(1+e_{1}\right)$ such that $E\left(e_{0}\right)=E\left(e_{1}\right)=0 \quad$ and $\quad E\left(e_{0}^{2}\right)=\gamma C_{Y}^{2} \quad, \quad E\left(e_{1}^{2}\right)=\gamma C_{X}^{2} \quad, \quad E\left(e_{0} e_{1}\right)=\gamma \rho C_{y} C_{x}$. Expressing (6), in terms of $e$ 's,

$$
\begin{align*}
\bar{y}_{P R} & =\bar{Y}\left(1+e_{0}\right)\left(1+e_{1}\right)^{\alpha} \exp \left[\frac{\bar{X}-\bar{X}\left(1+e_{1}\right)}{\bar{X}+\bar{X}\left(1+e_{1}\right)}\right] \\
& =\bar{Y}\left(1+e_{0}\right)\left(1+e_{1}\right)^{\alpha} \exp \left[\frac{e_{1}}{2}\left(1+\frac{e_{1}}{2}\right)^{-1}\right] . \tag{7}
\end{align*}
$$

Expanding the right hand side of (7) and retaining terms up to the second power of $e$ 's,

$$
\begin{align*}
\bar{y}_{P R} & =\bar{Y}\left(1+e_{0}\right)\left(1+e_{1}\right)^{\alpha} \exp \left[\frac{e_{1}}{2}\left(1+\frac{e_{1}}{2}+\frac{e_{1}^{2}}{4}+\ldots\right)\right] \\
& =\bar{Y}\left(1+e_{0}\right)\left[1+\alpha e_{1}+\frac{\alpha(\alpha-1)}{2} e_{1}^{2}+\ldots\right]\left(1+\frac{e_{1}}{2}+\frac{2 e_{1}^{2}}{8}\right) \tag{8}
\end{align*}
$$

From (8),

$$
\begin{equation*}
\bar{y}_{P R}-\bar{Y} \cong \bar{Y}\left[\alpha e_{1}+\frac{\alpha(\alpha-1)}{2} e_{1}^{2}+\frac{e_{1}}{2}+\frac{\alpha e_{1}^{2}}{2}+\frac{3 e_{1}^{2}}{8}+e_{0}+\alpha e_{1} e_{0}+\frac{e_{0} e_{1}}{2}\right] . \tag{9}
\end{equation*}
$$

Squaring (9) and then taking expectation of both sides, the MSE of the estimator $\bar{y}_{P R}$ is

$$
\begin{equation*}
\operatorname{MSE}\left(\bar{y}_{P R}\right)=\gamma \bar{Y}^{2}\left(C_{y}^{2}+C_{x}^{2} / 4+2 \alpha \rho C_{x} C_{y}+\rho C_{x} C_{y}+\alpha^{2} C_{x}^{2}+\alpha C_{x}^{2}\right) \tag{10}
\end{equation*}
$$

## GAMZE ÖZEL KADILAR

Obtain the optimum $\alpha$ to minimize $\operatorname{MSE}\left(\bar{y}_{P R}\right)$. Differentiating $\operatorname{MSE}\left(\bar{y}_{P R}\right)$ with respect to $\alpha$ and equating the derivative to zero, optimum value of $\alpha$ is given by

$$
\alpha_{o p t}=\left(C_{x}-2 \rho C_{y}\right) / 2 C_{x} .
$$

Substituting the value of $\alpha_{o p t}$ in (10), we get the minimum value of $\operatorname{MSE}\left(\bar{y}_{P R}\right)$ as

$$
\begin{equation*}
\operatorname{MSE}_{\min }\left(\bar{y}_{P R}\right)=\gamma \bar{Y}^{2} C_{y}^{2}\left(1-\rho^{2}\right) \tag{11}
\end{equation*}
$$

It follows from (11) that the proposed estimator $\bar{y}_{P R}$ at its optimum condition is equal efficient as that of the usual linear regression estimator.

## Efficiency Comparisons

In this section, the MSE of traditional estimators $\bar{y}, \bar{y}_{R}, \bar{y}_{P}, \bar{y}_{S}$, and $\bar{y}_{B T}$ are compared with the MSE of the proposed estimator $\bar{y}_{P R}$. From (1)-(5) and (11),

$$
\begin{gather*}
{\left[\operatorname{Var}(\bar{y})-\operatorname{MSE}_{\text {min }}\left(\bar{y}_{p r l}\right)\right]=\lambda \bar{Y}^{2} \rho^{2}>0,}  \tag{12}\\
{\left[\operatorname{MSE}\left(\bar{y}_{R}\right)-\operatorname{MSE}_{\text {min }}\left(\bar{y}_{p r l}\right)\right]=\lambda \bar{Y}^{2}\left(C_{x}-\rho C_{y}\right)^{2}>0,}  \tag{13}\\
{\left[\operatorname{MSE}\left(\bar{y}_{P}\right)-\operatorname{MSE}_{\text {min }}\left(\bar{y}_{p r l}\right)\right]=\lambda \bar{Y}^{2}\left(C_{x}+\rho C_{y}\right)^{2}>0,}  \tag{14}\\
{\left[\operatorname{MSE}\left(\bar{y}_{B T}\right)-\operatorname{MSE}_{\text {min }}\left(\bar{y}_{p r l}\right)\right]=\lambda \bar{Y}^{2}\left(C_{x} / 2-\rho C_{y}\right)^{2}>0,}  \tag{15}\\
{\left[\operatorname{MSE}\left(\bar{y}_{S}\right)-\operatorname{MSE}_{\text {min }}\left(\bar{y}_{p r l}\right)\right]=\lambda \bar{Y}^{2}\left(C_{x} \theta-\rho C_{y}\right)^{2}>0,} \tag{16}
\end{gather*}
$$

It is observed that $\bar{y}_{P R}$ is always more efficient than the traditional estimators $\bar{y}, \bar{y}_{R}, \bar{y}_{P}, \bar{y}_{S}$, and $\bar{y}_{B T}$, because the conditions from (12) to (16) are always satisfied.

## NEW EXPONENTIAL TYPE ESTIMATOR FOR THE POPULATION MEAN

## Numerical Illustrations

The appropriateness of the proposed estimator has been verified with the help of the following data sets given in Table 1.

Table 1. Statistics of populations

| Parameters | Population 1 | Population 2 | Population 3 | Population 4 |
| ---: | ---: | ---: | ---: | ---: |
| $N$ | 80 | 104 | 80 | 10 |
| $n$ | 20 | 20 | 20 | 4 |
| $\bar{Y}$ | 11.264 | 6.254 | 51.826 | 5.920 |
| $\bar{X}$ | 51.826 | 13931.680 | 2.851 | 3.590 |
| $\rho$ | 0.941 | 0.860 | 0.915 | 1.680 |
| $C_{y}$ | 0.750 | 1.860 | 0.354 | 0.144 |
| $C_{x}$ | 0.354 | 1.650 | 0.948 | 0.128 |
| $\beta_{2}(x)$ | 0.063 | 17.516 | 1.300 | 0.381 |

The explanation of the data sets in Table 1 from various sources is given as follows:

Population 1. Source Murthy (1967): Y is the fixed capital and X is the output of the 80 factories.
Population 2. Source Shabbir, Haq, and Gupta (2014): The study variable Y is the level of apple production (in 1000 tons) and the auxiliary variable X is the number of apple trees in 104 villages in 1999.
Population 3. Source Murthy (1967): The auxiliary variable X is the number of workers and the study variable is the output for 80 factories in a region.
Population 4. Source Cochran (1977): The auxiliary variable X is the number of rooms and the study variable is the number of persons.

The Percent Relative Efficiencies (PREs) of different estimators of the population mean with respect to the sample mean based on Populations 1-4 are given in Table 2.

## GAMZE ÖZEL KADILAR

Table 2. PREs of different estimators of population mean with respect to sample mean $\bar{y}$.

| Estimators | 1 | Population |  |  |
| ---: | ---: | ---: | ---: | ---: |
|  | 2 | 3 | 4 |  |
| $\bar{y}$ | 100.000 | 100.000 | 100.000 | 100.000 |
| $\bar{y}_{R}$ | 298.972 | 382.945 | 30.586 | 158.823 |
| $\bar{y}_{P}$ | 47.369 | 30.186 | 7.651 | 34.089 |
| $\bar{y}_{B T}$ | 163.521 | 230.504 | 292.078 | 161.440 |
| $\bar{y}_{S}$ | 104.054 | 323.244 | 319.840 | 8.197 |
| $\bar{y}_{P R 1}$ | 873.218 | 384.025 | 614.345 | 173.748 |

From the values of Table 2, it is observed that the MSE of the proposed estimator is less than the mean squared errors of all the existing estimators. Note that $\bar{y}_{S}$ requires the auxiliary variable information, on the other hand, one can reach the minimum MSE value using the proposed estimator without auxiliary variable information.

## Conclusion

As an improved exponential ratio estimator for estimating the population mean was proposed. The proposed estimator is better than the mentioned existing estimators in literature, in the sense of having lesser mean squared error. Hence, the proposed estimator is recommended for its practical use for estimating the population mean when the auxiliary information is available.

## References

Bahl, S., \& Tuteja, R. K. (1991). Ratio and product type exponential estimator. Journal of Information and Optimization Sciences, 12(1), 159-163. doi:10.1080/02522667.1991.10699058

Cochran, W. G. (1977). Sampling techniques (3 ${ }^{\text {rd }}$ ed.). Wiley Eastern Limited.

Murthy, M. N. (1967). Sampling theory and methods. Calcutta, India: Statistical Publishing Society.

## NEW EXPONENTIAL TYPE ESTIMATOR FOR THE POPULATION MEAN

Robson, D. S. (1957). Applications of multivariate polykays to the theory of unbiased ratio-type estimation. Journal of American Statistical Association, 52(280), 511-522. doi:10.1080/01621459.1957.10501407

Shabbir, J., Haq, A., \& Gupta, S. (2014). A new difference-cum-exponential type estimator of finite population mean in simple random sampling. Revista Colombiana de Estadística, 37(1), 197-209. doi:10.15446/rce.v37n1.44366

Singh, H. P. \& Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean. Statistics in Transition, 6(4), 555-560.

Singh, H. P., \& Tailor, R. (2005). Estimation of finite population mean with coefficient of variation of an auxiliary character. Statistica, 65(3), 407-413. doi:10.6092/issn.1973-2201/1299

Singh, H. P., Tailor, R., Tailor, R., \& Kakran, M. S. (2004). An improved estimator of population mean using power transformation. Journal of the Indian Society Agricultural Statistics, 58(2), 223-230.

Singh, R., Chauhan, P., Sawan, N., \& Smarandache, F. (2009). Improvement in estimating the population mean using exponential estimator in simple random sampling. International Journal of Statistics and Economics, 3(A09), 13-18.

Singh, R., Chauhan, P., Sawan, N., \& Smarandache, F. (2011). Improved exponential estimator for population variance using two auxiliary variables. Italian Journal of Pure and Applied Mathematics, 28, 101-109.

Sisodia, B. V. S., \& Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. Journal of the Indian Society Agricultural Statistics, 33(2), 13-18.

Upadhyaya, L. N., \& Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. Biometrical Journal, 41(5), 627-636. doi:10.1002/(SICI)1521-4036(199909)41:5<627::AID-BIMJ627>3.0.CO;2-W

Yadav, S. K., \& Kadilar, C. (2013a). Improved Class of Ratio and Product Estimators. Applied Mathematics and Computation, 219 (22), 10726-10731

Yadav, S. K., \& Kadilar, C. (2013b). Improved exponential type ratio estimator of population variance. Revista Colombiana de Estadística, 36(1), 145-152.

# A Comprehensive Review of the Two-Sample Independent or Paired Binary Data, with or without Stratum Effects 

Dewi Rahardja
U.S. Department of Defense

Fort Meade, Maryland

Ying Yang<br>U. S. Food and Drug Administration<br>Silver Spring, Maryland

Zhiwei Zhang<br>U. S. Food and Drug Administration<br>Silver Spring, Maryland


#### Abstract

Various statistical hypotheses testing for discrete or categorical or binary data have been extensively discussed in the literature. A comprehensive review is given for the twosample binary or categorical data testing methods on data with or without Stratum Effects. The review includes traditional methods such as Fisher's Exact, Pearson's Chi-Square, McNemar, Bowker, Stuart-Maxwell, Breslow-Day and, Cochran-Mantel-Haenszel, as well as newly developed ones. We also provide the roadmap, in a figure or diagram format to which methods are available in the literature. In addition, the implementation of these methods in popular statistical software packages such as SAS and/or R is also presented. This will be helpful for researchers to determine which (categorical-data) testing method is available to use in various fields of study such as clinical trials, epidemiology, etc., both for the design phase of a study in prospective study, crosssectional or retrospective study analysis.


Keywords: Cochran-Mantel-Haenszel (CMH) test, common odds ratio (OR), common risk difference (CRD), homogeneous stratum effect (HSE), McNemar's test, paired binary data, stratified data, Bowker's test, marginal homogeneity, stratified test, Stuart-Maxwell's test, symmetry, Fisher's exact test, chi-squared test

## Introduction

Many real-world data, such as data in clinical trials, financial data, epidemiology, sociology, etc. often use outcome variables that are categorical or binary in nature, that is, for example, in binary case, there are two possible outcomes for each subject. Without loss of generality (WLOG), these two outcomes are mutually exclusive and are categorized as success or failure. A frequent task in many fields of study, such as medical statistics (or any other field) is to compare two (independent or paired) binomial proportions. It can occur both in randomized

[^14]controlled trials and in observational studies. The outcomes of two groups can be summarized in a single $2 \times 2$ contingency table. The number of subjects in each group ( $n_{1+}$ and $n_{2+}$ ) is assumed to be fixed by the design. Assume that the subjects in group 1 have probability of success equal to $p_{1}$, and that the subjects in group 2 have probability of success equal to $p_{2}$. The following Table 1 illustrates a single $2 \times 2$ contingency table.

Table 1. Comparing 2 groups of binomial data in a single $2 \times 2$ contingency-table format.

|  | Success | Failure | $n_{12}$ |
| :---: | ---: | ---: | ---: |
| Group 1 | $n_{11}$ | $n_{22}$ | $n_{2+}$ |
| Group 2 | $n_{21}$ |  |  |

Let $n=\left\{n_{11}, n_{12}, n_{21}, n_{22}\right\}$ be the observed values as in Table 1 . The number of successes in group 1 is binomially distributed with parameters $n_{1+}$ and $p_{1}$. In a similar manner, the number of successes in group 2 is binomially distributed with parameters $n_{2+}$ and $p_{2}$. The parameters $p_{1}$ and $p_{2}$ are estimated by the sample proportions

$$
\hat{p}_{1}=\frac{n_{11}}{n_{1+}} \quad \text { and } \quad \hat{p}_{2}=\frac{n_{21}}{n_{2+}}
$$

which are the maximum likelihood estimates.
The followings are the three most common measures to compare between two groups in a study. They are, the proportion difference, proportion (risk) ratio, and the odds ratio:

| Parameter: | Notation |
| :---: | :---: |
| Difference: | $\hat{p}_{1}-\hat{p}_{2}$ |
| Risk ratio: | $O R=\hat{p}_{1} / \hat{p}_{2}$ |
|  |  |
| Odds Ratio: | $\frac{\hat{p}_{1} /\left(1-\hat{p}_{1}\right)}{\hat{p}_{2} /\left(1-\hat{p}_{2}\right)}$ |

The two groups being considered can be classified either as independent or matched pairs. Independent groups mean that the two samples taken are

## RAHARDJA ET AL.

independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent or paired outcomes. The two variables may be two responses on a single individual or two responses from a matched pair (as in matched case-control studies). Table 2 summarizes the outcomes of matched pair two groups in a $2 \times 2$ contingency-table format.

Table 2. Matched pair two groups in a $2 \times 2$ contingency-table format.

| Conse |  | Failure | Total |
| ---: | ---: | ---: | ---: |
| Control | Success | (pait | $n_{12}\left(p_{12}\right)$ |
| $n_{1+}\left(p_{1+}\right)$ |  |  |  |
| Success | $n_{21}\left(p_{21}\right)$ | $n_{22}\left(p_{22}\right)$ | $n_{2+}\left(p_{2+}\right)$ |
| Failure | $n_{+1}\left(p_{+1}\right)$ | $n_{+2}\left(p_{+2}\right)$ |  |
| Total |  |  |  |

where $p_{+1}$ and $p_{1+}$ are the marginal probabilities of a success response for the case and control subjects, respectively.

In a stratified design (or multiple $2 \times 2$ contingency tables), the subjects are selected from two or more strata which are formed from important covariates such as gender, income level, marital status, etc. The number of subjects in each of the two groups in each stratum is set (fixed) by the design. A separate $2 \times 2$ table is formed for each stratum. Hence, there are multiple $2 \times 2$ contingency tables. The data can be represented as a set of $K 2 \times 2$ tables as the following Table 3 .

Table 3. Comparing 2 groups of binomial data in a multiple $2 \times 2$ contingency-table format.

|  | Success | Failure |  |
| :---: | ---: | ---: | ---: |
| Group 1 | $n_{11 k}$ | $n_{12 k}$ | $n_{1+k}$ |
| Group 2 | $n_{21 k}$ | $n_{22 k}$ | $n_{2+k}$ |

where $k=1, \ldots, K$ stratum.
Thus, the purpose of this review of is to consider the existing testing methods in the literatures on the two independent or matched pair samples with binary data with or without stratum effects.

## Hypothesis Testing

Consider two independent groups without stratum effect (i.e., a single $2 \times 2$ contingency table). The hypotheses for two independent proportions can be written as $\mathrm{H}_{0}: p_{1}=p_{2}$ and $H_{1}: p_{1} \neq p_{2}$. A Chi-square test is often used test the hypotheses.

In SAS PROC FREQ, the CHISQ option is used in the TABLES statement to obtain the test statistic and its associated $p$-value. By the famous rule of thumb, the Cochran's rule, the Chi-square test assumes that the expected value for each cell is five or higher. However, if this assumption is not met, the Fisher's exact test can be used regardless of how small the expected frequency is. The Fisher's exact test can be used with the FISHER option on the TABLES statement. However, Fisher's exact test is computationally explosive for large sample size and hence the Chi-square test is needed for large sample size approximation.

When subjects from two groups are independently sampled from two or more strata (i.e., with stratum effect; or a multiple $2 \times 2$ contingency table), the null hypothesis of the interest can be to test whether odds ratios are the same across strata, that is, $H_{0}: O R_{1}=O R_{2}=\ldots=O R_{k}$ (or, homogeneity across strata). The Breslow-Day (BD) test (1980) for homogeneous odds ratios across strata can be used to test for the stratum effect. If the BD test is rejected, then the treatment comparison should be performed by strata; otherwise, the Cochran Mantel Haenzel (CMH) test (Cochran, 1954; Mantel \& Haenszel, 1959) can be used to test whether the common odds ratios across strata is equal to 1 , i.e., if all the $O R_{i}=1$, for $i=1, \ldots, k$. In SAS PROC FREQ, the CMH option can be used for testing whether the common odds ratios are equal to one. The CMH option also provides logit estimates of the common odds ratio and the common relative risks.

Next, consider binary data collected on matched pairs. The sampling unit is not one individual but a pair of related individuals, which could be two parts of or two occasions for the same individual. For example, the binary response is a voter's choice from two presidential candidates and the two occasions could be two different time points before the presidential election.

For unstratified paired binary data, McNemar's test (1947) is commonly used to test whether the risk difference is zero. Such a null hypothesis is more commonly known as marginal homogeneity or symmetry of the $2 \times 2$ contingency table. This null hypothesis of homogeneity can be written as $H_{0}: p_{1 .}=p_{.1}$, where $p_{1 .}=p_{11}+p_{12}$ and $p .1=p_{11}+p_{21}$, or equivalently, the null hypothesis of symmetry, $H_{0}: p_{12}=p_{21}$. McNemar's test can be calculated using AGREE option in PROC FREQ. Developed from asymptotic theory, McNemar's test requires a large

## RAHARDJA ET AL.

number of observations (say 5, by Cochran's rule) in each cell of discordance. For small samples, an exact binomial test can be used to test the null hypothesis of symmetry.

When the paired categorical random variables take $K(K>2)$ values, Bowker's test (1948) can be used to test the symmetry $H_{0}: p_{i j}=p_{j i}$, for all $i \neq j$, where $i, j \in\{1,2, \ldots, K\}$.

If the test of marginal homogeneity is of the interest, the generalization of the McNemar's test, commonly referred to as generalized McNemar's or StuartMaxwell test (1955) can be used, $H_{0}: p_{i}=p . i$, where $i=1,2, \ldots, K$. Note that for $K>2$, the null hypothesis of symmetry is not equivalent to the null hypothesis of homogeneity. In fact, rejection of marginal homogeneity implies rejection of symmetry, but not vice versa. Therefore, practitioners need to decide which hypothesis to test for a particular application. A SAS macro by Sun and Yang (2008) has been developed for Stuart-Maxwell statistic.

Zhao, Rahardja, Wang and Shen (2014) considered a series of independent paired binary data in which the series is defined by a stratification factor, the null hypothesis of interest is to test the homogeneous stratum effects. In analogy, this is similar to the Breslow-Day (BD) test (1980) for homogeneous odds ratios across a series of stratified $2 \times 2$ contingency tables in which the binary data are unpaired. The null hypothesis can be written as $H_{0}: p_{1.1}-p_{.11}=\ldots=p_{1 . K}-p_{.1 K}$, or equivalently, $H_{0}: p_{121}-p_{211}=\ldots=p_{12 K}-p_{21 K}$. The R-code of testing HSE in stratified paired binary data is available in the Zhao et al (2014) manuscript. If the homogeneous stratum effect (HSE) test is rejected, then the data should be analyzed by strata; otherwise the common risk difference (CRD) test for paired binary data in Zhao-Rahardja (2013) manuscript can be used to estimate the CRD. The test for CRD is analogous to the CMH test when the binary data are unpaired. Table 4 summarizes the above discussion.

## Roadmap

WLOG, the figure/diagram below (see Figure 1) provides the roadmap for practitioners to choose a suitable testing method for their categorical data analysis. In the Figure 1, the roadmap is provided by whether or not stratification table or multiple contingency tables is necessary.

Table 4. Listing of Sample Type with the appropriate testing, test statistics, and SAS command or R code.


## RAHARDJA ET AL.



Figure 1. Categorical-Data Roadmap by Stratify or Not

## COMPREHENSIVE REVIEW OF CATEGORICAL-DATA TESTING METHODS

## Summary

Categorical Data or most commonly binary or dichotomous outcome (i.e., success vs. failure, dead vs. alive, 1 vs. 0 ) is very common in real-data applications such as clinical trials, financial data, epidemiology, sociology, etc. The analysis of such categorical outcomes has a long history, beginning with the single $2 \times 2$ table, multiple/stratified $2 \times 2$ tables, matched/paired $2 \times 2$ tables, to big table such as $K \times K$ tables. In this paper, we provide a comprehensive review of the hypothesis testing procedures that are available in the literature for various types of categorical data. In summary, this review will be helpful for the practitioners in various fields of study (such as clinical trials, financial data, epidemiology, sociology, etc.) to determine the appropriate method according to the provided roadmap in Figure 1.

## Disclaimer

This research represents the authors own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other U.S. Federal agency.

## References

Bowker, A. H. (1948). A test for symmetry in contingency tables. Journal of the American Statistical Association, 43(244), 572-574. doi: 10.2307/2280710

Breslow, N. E., \& Day, N. E. (1980). Statistical methods in cancer research: The analysis of case-control studies. Lyon, France: International Agency for Research on Cancer.

Cochran, W. G. (1954). Some methods for strengthening the common $\chi^{2}$ tests. Biometrics, 10(4), 417-451. doi: 10.2307/3001616

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), 153-157. doi: 10.1007/BF02295996

Mantel, N., \& Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719-748.

## RAHARDJA ET AL.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. British Journal of Psychiatry, 116(535), 651-655. doi: 10.1192/bjp.116.535.651

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42(3/4), 412-416. doi: 10.2307/2333387

Sun, X., \& Yang, Z. (2008). Generalized McNemar's Test for Homogeneity of the Marginal Distributions (Paper 382-2008). SAS Global Forum 2008, Statistics and Data Analysis, San Antonio, Texas.

Zhao, Y. D., \& Rahardja, D. (2013). Estimation of the common risk difference in stratified paired binary data with homogeneous stratum effect. Journal of Biopharmaceutical Statistics, 23(4), 848-855.
doi: 10.1080/10543406.2013.789888
Zhao, Y. D., Rahardja, D., Wang, D.-H., \& Shen, H. (2014). Testing homogeneity of stratum effects in stratified paired binary data. Journal of Biopharmaceutical Statistics, 24(3), 600-607.
doi: 10.1080/10543406.2014.888440

# Evaluation of the Addition of Firth's Penalty Term to the Bradley-Terry Likelihood 

Paul Meyvisch<br>Janssen Pharmaceutica NV<br>Beerse, Belgium


#### Abstract

A major shortcoming of the Bradley-Terry model is that the maximum likelihood estimates are infinite-valued in the presence of separation, and may be unreliable when data are nearly separated. A well-known solution consists of the addition of Firth's penalty term to the log-likelihood function, and solving this penalized likelihood through logistic regression. The maximum likelihood estimates with and without Firth's penalty are compared in a large and heterogeneous population of table tennis players, showing that exact penalized maximum likelihood estimates can be reasonably approximated using a well-chosen Minorization-Maximization (MM) algorithm.


Keywords: Bradley-Terry, Firth, MM algorithm, table tennis

## Introduction

Consider the evaluation of the addition of Firth's penalty term to the BradleyTerry likelihood function, with an application to a large dataset of table tennis players. The problem of rating table tennis players falls into the topic of binary paired comparison modeling, provided the victory margin is ignored. A binary paired-comparison experiment is used to assess the relative worth of $t$ objects even though they can only be compared two at a time, and when the result of such a comparison can only be that one of the objects is preferred to the other. Zermelo (1929) is generally credited with being the first to address the problem of estimating the strengths of players. The model and various parts of the theory have been rediscovered over the intervening years and were first described in detail by Bradley \& Terry (1952).

Suppose there are $m$ players and define $\pi=\left(\pi_{1}, \ldots, \pi_{m}\right)^{\prime}$ to be the vector of the player's strengths. The Bradley-Terry model assumes that the probability $p_{i j}$ of player $i$ defeating player $j$ is:

[^15]
## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

$$
\begin{equation*}
p_{i j}=\frac{\pi_{i}}{\pi_{i}+\pi_{j}} \tag{1}
\end{equation*}
$$

Any constant multiple of the strengths $\pi_{i}$ estimates also satisfy (1), so they can be scaled to satisfy an additional constraint such as $\sum_{i} \pi_{I}=1$ or $\pi_{I}=1$ for sake of identifiability.

If each pair of players $i$ and $j$ plays $n_{i j}$ games against each other, with player $i$ winning $v_{i j}$ times and losing $d_{i j}$ times, and all games assumed independent, the likelihood takes the form:

$$
\begin{equation*}
L(\pi)=\left[\prod_{i=1}^{m} \prod_{j: j \neq i} \frac{\pi_{i}^{v_{i j}} \pi_{i j}^{d_{i j}}}{\left(\pi_{i}+\pi_{j}\right)^{n_{i j}}}\right]^{\frac{1}{2}}, \tag{2}
\end{equation*}
$$

where $v_{i j}=d_{j i}$ and $n_{i j}=n_{j i .}$
As noted by Ford (1957), if it is possible to partition the set of players into two groups $A$ and $B$ such that there are never any intergroup comparisons, then there is no basis for rating any player in A with respect to any player in B (Hunter, 2004). It is therefore assumed that the tournament is completely connected, i.e., there is a chain of matches which links any given pair of players. In order for the maximum likelihood estimates of the strengths to exist, a second condition is required which will be further denoted as Ford's Assumption: In every possible partition of the players into two nonempty subsets, some player in the second set beats some player in the first set at least once (Ford, 1957). As a special case, Ford's Assumption is not satisfied if group A consists of only one player who has lost or won all games. The maximum likelihood estimate for this player will be infinite-valued.

The likelihood can alternatively be expressed as a function of $\theta=\left(\theta_{1}, \ldots, \theta_{m}\right)^{\prime}$ with $\theta_{i}=\log \left(\pi_{i}\right), \forall i: 1, \ldots, m$. Using (1), the probability $p_{i j}$ then becomes:

$$
\begin{equation*}
p_{i j}=\frac{\exp \left(\theta_{i}-\theta_{j}\right)}{1+\exp \left(\theta_{i}-\theta_{j}\right)} \tag{3}
\end{equation*}
$$

The Bradley-Terry model can hence be solved using logistic regression (Agresti, 2002). Details as to how this model can practically be fit are provided by

## PAUL MEYVISCH

So (1995). The non-existence of maximum likelihood estimates is a well-known and understood problem in logistic regression models and has been denoted by Albert \& Anderson (1984) as separation.

The log-likelihood takes the form:

$$
\begin{equation*}
l(\theta)=\frac{1}{2} \sum_{i=1}^{m} \sum_{j: j \neq i}\left[v_{i j} \theta_{i}+d_{i j} \theta_{j}-n_{i j} \log \left(\exp \theta_{i}+\exp \theta_{j}\right)\right], \tag{4}
\end{equation*}
$$

Extensions to the Bradley-Terry model have been proposed in the literature but are not considered here. Hunter (2004) provides an interesting review.

## Firth's penalty term

The phenomenon of separation or monotone likelihood is observed in the fitting process of a logistic model if at least one parameter estimate diverges to $\pm \infty$. It is believed that separation is unpredictable because it is primarily caused by random variation as it may depend on the outcome of a few matches. Furthermore, it is demonstrated by Heinze (2006) that maximum likelihood estimation by logistic regression may give questionable results in the presence of so-called nearly separated data. This situation occurs when the existence of the maximum likelihood estimates depends on the presence of a few particular observations. A solution proposed by Heinze \& Schemper (2002) and Heinze (2006) to separation and near-separation is to penalize the log-likelihood, as described by Firth (1993). The basic idea is to introduce a bias term into the standard likelihood function which itself goes to zero as $n \rightarrow \infty$, but for small $n$ operates to counteract the $O\left(n^{-1}\right)$ bias present here. The penalty function used is Jeffreys invariant prior (Jeffreys, 1961). One of the advantages of the addition of Firth's penalization term is that no arbitrary data manipulation is involved. It is also justified from the use of Jeffreys prior, in the sense that it is non-informative, thereby implying that maximal weight is given to the data. It should also be noted that the interpretation of the model is not changed in any way. Firth (1993) demonstrated that, for a broad class of generalized linear models, this penalized likelihood is asymptotically consistent and eliminates the usual small-sample bias found in maximum likelihood estimates.

The suggested penalized log-likelihood function takes the following form:

$$
\begin{equation*}
l^{*}(\theta)=l(\theta)+\frac{1}{2} \log |I(\theta)|, \tag{5}
\end{equation*}
$$

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

where $I(\theta)$ is the Fisher Information Matrix of $\theta$.

## Case Study

The impact of the addition of Firth's penalty using a motivational (simplified) example will be demonstrated. The evaluation will be done on a large data set of table-tennis players. The data that are used for analysis consist of all recorded match results during the sports season of 2006-2007 of a population of 770 players from the province of Vlaams-Brabant (Belgium). It is shown in Figure 1(a) that the population is highly heterogeneous, both in terms of strengths as number of matches played. It is noted that, in line with existing rating systems, the estimates for $\theta$ were linearly transformed to fall roughly between 0 and 3,000 (Glickman, 1995 and 1999) and (Marcus, 2001).

The transformation used was such that a difference of 100 points between 2 players corresponds with odds of 2 for the highest rated player to win. The median (Q1-Q3) number of matches per player equals 61 (35-78). The primary objective is to rate each player in this pool using the penalized and unpenalized maximum likelihood estimates, and to provide Wald-based and profile likelihood $95 \%$ confidence intervals. The differences between penalized and unpenalized maximum likelihood estimates will be investigated. Additionally, the differences between both types of confidence intervals will be discussed.

Consistent with local regulation, a simplified log-likelihood was used to allow the new rating of the $i^{\text {th }}$ player, $\forall i: 1, \ldots, m$ to depend only on the ratings of each of his/her opponents, which are by way of simplification (naively) considered constant during the season. Therefore, $\forall i: 1, \ldots, m$ :

$$
\begin{equation*}
l\left(\theta_{i}\right)=\sum_{j: j \neq i}\left[v_{i j} \theta_{i}+d_{i j} \theta_{j}^{c}-n_{i j} \log \left(\exp \theta_{i}+\exp \theta_{j}^{c}\right)\right], \tag{6}
\end{equation*}
$$

where $\theta_{j}^{c}$ indicates the (scalar-valued) rating of the $j^{\text {th }}$ player.
This log-likelihood (6) can, contrary to (4), not be considered a logistic regression model but has to be optimized using Newton's Method or through an appropriate Minorization Maximization (MM) algorithm. Maximum likelihood estimation using (6) will better allow an evaluation of the impact of separation as it will, unlike model (4), not depend on a linear combination of regressors. It can indeed be verified that monotonicity of the log-likelihood (6) is only to occur when a player loses or wins all matches. It is therefore expected that the phenomenon of near-separation is more simply expressed as a function of the

## PAUL MEYVISCH

victory rate. Application of (4) to the same data set will be presented in before the conclusion of this article. It can easily be shown that the score function of the penalized log-likelihood can be expressed as:

$$
\begin{equation*}
S\left(\theta_{i}\right)=\sum_{j: j \neq i}\left[\left(v_{i j}-n_{i j} p_{i j}\right)\right]+\frac{1}{2}\left[1-2 \frac{\sum_{j: j \neq i} I_{j}\left(\theta_{i}\right) p_{i j}}{I\left(\theta_{i}\right)}\right], \tag{7}
\end{equation*}
$$

where the Fisher Information $I\left(\theta_{i}\right)=\Sigma_{j: j \neq i} n_{i j} p_{i j}\left(1-p_{i j}\right)$ is alternatively expressed as $\Sigma_{j: j \neq i} I_{j}\left(\theta_{i}\right)$. It should also be noted that $p_{i j}$ is equal to the expression in (3) with $\theta_{j}$ replaced by $\theta_{j}{ }^{c}$.

Rearranging some of the terms and denoting the total number of wins for the $i^{\text {th }}$ player as $V_{i}$ results in

$$
\begin{equation*}
V_{i}+\frac{1}{2}=\sum_{j: j \neq i}\left[n_{i j}+\frac{I_{j}\left(\theta_{i}\right)}{I\left(\theta_{i}\right)}\right] p_{i j}, \tag{8}
\end{equation*}
$$

This expression has a simple interpretation in terms of data adjustments: add $1 / 2$ match to the players total number of wins and add a fraction of a match to the total number of matches played against the $j^{\text {th }}$ player. The fractions to be added depend on the unknown $\theta_{i}$.

Prior to fitting the data, note Ford's Assumption is not satisfied for about $5 \%$ of the players, and hence, the maximum likelihood estimates of these players will be infinite-valued. Removing these players from the data by no means guarantees the maximum likelihood estimates of the remaining players to exist, as some of the latter may have only won matches against those that are removed. To solve this problem, two virtual games for every single player are added, i.e., one win and one loss against a (virtual) player of equal strength. These virtual players are added with their given strengths at the right-hand side of (6). The introduction of virtual matches may dilute the difference between penalized and unpenalized maximum likelihood estimates for every single player; however, given the size and the heterogeneity of the data, the overall relationship between both estimates can still reliably be expressed.

As observed from Figure 1(b), the penalized maximum likelihood (PML) estimates are slightly more conservative, i.e., the estimate is pulled towards the center. Players with a low victory rate, i.e. $\leq 20 \%$, have a PML estimate which is slightly higher than the ML estimate. The reverse phenomenon is observed for

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

players with a high, perhaps $\geq 80 \%$, victory rate. The small-sample bias reduction is also evident in the subset of players who have played fewer than 30 matches. The shrinkage towards the mean is more pronounced compared to players on whom more information is available.


Figure 1. Supporting figures of Case Study

Although the symmetry of the profile likelihood may be enhanced by the addition of a penalization term, it is important to bear in mind that the resulting profile likelihood may still be asymmetric, in particular in the presence of nearseparation. Heinze \& Schemper (2002) therefore advise against the use of Waldbased confidence intervals and propose the profile penalized likelihood confidence interval as a more suitable solution. The discrepancy between Wald and profile likelihood $95 \%$ confidence intervals is graphically presented in Figure 1(c). For this purpose, the percent overlap of both confidence intervals is defined as the length of the intersecting interval, divided by the length of its union. It shows that both confidence intervals match very well when victory rates are close

## PAUL MEYVISCH

to $50 \%$. However, as the victory rate is an indicator of the likelihood's asymmetry, it is not surprising that the discrepancy is increased with increasing victory or defeat rate. It is also shown that the discrepancy is more pronounced for players on whom less data is available. Compared to Wald-based confidence limits, profile likelihood confidence limits are slightly shifted towards higher values for players with a high victory rate. The reverse phenomenon is observed for players with a low victory rate. Finally, it is seen from Figure 1(d) that the length of the profile likelihood confidence interval is not only dependent on the number of matches played but also on the victory/defeat rate. It may not come as a surprise that the precision of the estimates is lowest for extreme victory rates.

## Optimizing the penalized Bradley-Terry log-likelihood

It was shown by Firth (1993) and Heinze \& Schemper (2002) that maximum penalized likelihood estimates in logistic regression models are obtained by splitting each original observation $i$ into two new observations having response values $y_{i}$ and $1-y_{i}$ with iteratively updated weights $1+h_{i} / 2$ and $h_{i} / 2$ respectively (using their notation). It is also argued that the splitting of each original observation into a response and non-response guarantees finite estimates. It is further shown that the $h_{i}$ 's are obtained from the $i^{\text {th }}$ diagonal elements of the hat matrix whose elements are refreshed at every iteration. Mathematical details are provided by Firth (1993) and Heinze \& Schemper (2002).

This led to the development of software to allow calculation of Firth-type estimates. Direct implementation of the methodology in a SAS macro, S-plus library and R package owes to Heinze \& Ploner (2004). An additional R package to fit the Bradley-Terry logistic model was developed by Firth (2005). Implementation in $\log$ Xact version 8 by Cytel (Cytel, n.d.) has become available in 2005. As of 2008, users of SAS version 9.2 can apply Firth's correction as an option to the LOGISTIC procedure.

Because of the recent advancements in software development for logistic regression, maximum likelihood estimation using a Minorization-Maximization (MM) algorithm seems to be of lesser use from a practical point of view. In addition, an MM algorithm method to obtain the maximum penalized likelihood estimates has so far not been developed. However, it is important to note that some of the extensions to the Bradley-Terry model cannot be fitted using logistic regression (Hunter, 2004) and the MM algorithm may need to be used here as an alternative. In the next sections, the approximate score equations and an MM

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

algorithm for approximate maximum penalized likelihood estimation will be presented.

## Approximating the penalized score equation

From (4) it follows that

$$
\begin{equation*}
\frac{\partial l(\theta)}{\partial \theta_{i}}=\sum_{j: j \neq i}\left(v_{i j}-n_{i j} p_{i j}\right) \tag{9}
\end{equation*}
$$

because $\frac{\partial p_{i j}}{\partial \theta_{i}}=p_{i j}\left(1-p_{i j}\right)$ and $\frac{\partial p_{i j}}{\partial \theta_{j}}=-p_{i j}\left(1-p_{i j}\right)$, the information matrix $I(\theta)$ has diagonal elements

$$
\begin{equation*}
I(\theta)_{i i}=-\frac{\partial^{2} l(\theta)}{\partial^{2} \theta_{i}}=\sum_{j: j \neq i}\left[n_{i j} p_{i j}\left(1-p_{i j}\right)\right] \tag{10}
\end{equation*}
$$

and off-diagonal elements

$$
\begin{equation*}
I(\theta)_{i j}=-\frac{\partial^{2} l(\theta)}{\partial \theta_{i} \partial \theta_{j}}=-n_{i j} p_{i j}\left(1-p_{i j}\right) \tag{11}
\end{equation*}
$$

Differentiation of $\log |I(\theta)|$ in (5) requires derivatives of a log determinant with respect to the vector $\boldsymbol{\theta}$. To avoid that optimization of the penalized score equation would require major matrix operations at every iteration, lengthening the computational process and likely making it less stable, suggesting an approximate rather that an exact approach. The approximation consists of imposing the score function to be of a similar structure as (7) to obtain:

$$
\begin{equation*}
S_{\text {approx }}(\theta)=\frac{\partial l(\theta)}{\partial \theta_{i}}=\sum_{j: j \neq i}\left[\left(v_{i j}-n_{i j} p_{i j}\right)\right]+\frac{1}{2}\left[1-2 \frac{\sum_{j: j \neq i} I_{j}(\theta)_{i i} p_{i j}}{I(\theta)_{i i}}\right] \tag{12}
\end{equation*}
$$

The term $I_{j}(\theta)_{i i}$ in the numerator is the $j^{\text {th }}$ contribution to $I(\theta)_{i i}$ and is equal to $n_{i j} p_{i j}\left(1-p_{i j}\right)$. Setting this expression (12) to zero and rearranging some of the terms results in:

## PAUL MEYVISCH

$$
\begin{equation*}
V_{i}+\frac{1}{2}=\sum_{j: j \neq i}\left[n_{i j}+\frac{I_{j}(\theta)_{i i}}{I(\theta)_{i i}}\right] p_{i j} \tag{13}
\end{equation*}
$$

The same reasoning as Firth (1993) is applied, i.e., that each original observation $x_{i j}$ (i.e., a win or a loss of the $i^{\text {th }}$ player against the $j^{\text {th }}$ player) can be split into 2 new observations having response values $x_{i j}$ and $1-x_{i j}$ with iteratively updated weights $1+g_{i j} / 2$ and $g_{i j} / 2$ respectively. Note that the weights $g_{i j}$ are an approximation to the diagonal elements of the hat matrix introduced earlier if we were to express (5) as a logistic regression model. The weights are updated at each iteration and depend on the unknown $\theta$. It can then be verified that the approximation to the likelihood function $l^{*}(\theta)$ can alternatively be expressed as:

$$
l_{\text {approx }}^{*}(\theta)=\frac{1}{2} \sum_{i=1}^{m} \sum_{j: j \neq i}\left[\begin{array}{r}
{\left[v_{i j}+n_{i j} \frac{g_{i j}}{2}\right] \theta_{i}+\left[d_{i j}+n_{i j} \frac{g_{i j}}{2}\right] \theta_{j}}  \tag{14}\\
-\left[n_{i j}+n_{i j} g_{i j}\right] \log \left(\exp \theta_{i}+\exp \theta_{j}\right)
\end{array}\right]
$$

Optimizing (14) for $\theta_{i}$, it is easily verified from (8) that

$$
\begin{equation*}
\sum_{j: j \neq i} n_{i j} g_{i j}=\sum_{j: j \neq i} \frac{I_{j}(\theta)_{i i}}{I(\theta)_{i i}}=1 \tag{15}
\end{equation*}
$$

Expressions (14) and (15) will allow construction of a MM-algorithm.

## Minorization-Maximization algorithms

Hunter (2004) demonstrated optimization of the unpenalized log-likelihood function is obtained using a specific case of a general class of algorithms referred to here as Minorization-Maximization (MM) algorithms and shows that convergence is reached provided Ford's Assumption holds.

An MM algorithm operates by creating at each iteration a surrogate function $Q(\theta)$ that minorizes the log-likelihood function $l(\theta)$. This is to say $Q(\theta) \leq l(\theta)$ with equality if and only if $\theta=\theta^{(k)}$. When now the surrogate function is maximized, the log-likelihood is driven uphill. This combination of a minorization and a maximization step is repeated until convergence.

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

The strict concavity of the logarithm function implies for positive $x$ and $y$ that $-\log (x) \geq 1-\log (y)-x / y$ with equality if $x=y$. As shown in Hunter (2004), fixing $\theta^{(k)}$ and defining the function

$$
Q_{k}^{*}(\theta)=\frac{1}{2} \sum_{i=1}^{m} \sum_{j: j \neq i}\left[\begin{array}{l}
{\left[v_{i j}+n_{i j} \frac{g_{i j}}{2}\right] \theta_{i}+\left[d_{i j}+n_{i j} \frac{g_{i j}}{2}\right] \theta_{j}} \\
+\left[n_{i j}+n_{i j} g_{i j}\right]\left[\begin{array}{l}
1-\log \left(\exp \left(\theta_{i}^{(k)}\right)+\exp \left(\theta_{j}^{(k)}\right)\right) \\
-\frac{\exp \left(\theta_{i}\right)+\exp \left(\theta_{j}\right)}{\exp \left(\theta_{i}^{(k)}\right)+\exp \left(\theta_{j}^{(k)}\right)}
\end{array}\right]
\end{array}\right]
$$

it can be seen that $Q_{k}^{*}(\theta)$ minorizes $l^{*}{ }_{\text {approx }}(\theta)$ as

$$
\begin{equation*}
Q_{k}^{*}(\theta) \leq l_{\text {approx }}^{*}(\theta) \tag{16}
\end{equation*}
$$

with equality if $\theta=\theta^{(k)}$.
Using (15), optimization of $Q_{k}^{*}(\theta)$ for $\theta_{i}$ is now straightforward with solution

$$
\begin{equation*}
\exp \left(\theta_{i}^{k+1}\right)=\left(V_{i}+\frac{1}{2}\right)\left[\sum_{j: j \neq i} \frac{n_{i j}+\frac{I_{j}(\theta)_{i i}}{I(\theta)_{i i}}}{\exp \left(\theta_{i}^{(k)}\right)+\exp \left(\theta_{j}^{(k)}\right)}\right]^{-1} \tag{17}
\end{equation*}
$$

Similarly, minorization and maximization of the unpenalized log-likelihood function $l(\theta)$ is achieved with

$$
\begin{equation*}
\exp \left(\theta_{i}^{k+1}\right)=\left(V_{i}\right)\left[\sum_{j: j \neq i} \frac{n_{i j}}{\exp \left(\theta_{i}^{(k)}\right)+\exp \left(\theta_{j}^{(k)}\right)}\right]^{-1} \tag{18}
\end{equation*}
$$

## PAUL MEYVISCH

## Application

The same data will be used. Approximate maximum penalized likelihood estimates will be produced using (17). In addition equation (17) will be generalized such that the $1 / 2$ match to the player's total number of wins can be modified at both sides of the equation:

$$
\begin{equation*}
V_{i}+a=\sum_{j: j \neq i}\left[n_{i j}+2 a \frac{I_{j}(\theta)_{i i}}{I(\theta)_{i i}}\right] p_{i j} \tag{19}
\end{equation*}
$$

Comparisons with the exact penalized maximum likelihood estimates obtained using logistic regression are compared with the approximate penalized likelihood estimates for $a=0.3,0.5,0.7$ and 1 . A comparison between exact penalized likelihood estimates and unpenalized likelihood estimates is presented in Figure 2(a).

Unlike the results shown in Figure 1(b), the differences between both estimates are not only a function of the percentage of wins and of the sample size. This is because separation can occur as a result of a non-trivial linear combination of regressors, which can potentially occur at any sample size or victory rate. Also note the far larger presence of players with low rather than high victory rates in the data. It is further shown in Figure 2(b) that unpenalized estimates obtained using either logistic regression or by the MM algorithm (18) effectively give the same results. An investigation of the effect of the value $a$ for the added match in (19) is presented in Figures 2(c) to 2(f). It is shown in Figure 2(c) that the approximate penalized ML estimates (for $a=1$ ) strongly differ from the exact penalized ML estimates.

It is also clear from Figures 2(c) and 2(d) that approximations implied by values of a larger than 0.5 result in a too strong correction of the unpenalized ML estimates, when compared to the exact penalized ML estimates. The reverse phenomenon is observed for $a=0.3$ (see Figure 2(f)) and for any value of a lower than 0.3 (results not shown). For these small values of $a$, the comparison with the exact penalized ML estimates will become more and more similar to the pattern observed in Figure 2(a), for $a \rightarrow 0$. It is clear from Figure 2(e) that choosing $a=0.5$ resulted in the best fit. Similar results were obtained through simulations (data not shown). A value of $a=0.5$ always yielded results that are sufficiently close to the exact values. It was observed that the correction implied by the exact results, both on the real data as on the simulation, was always slightly larger

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

compared to the approximate results. However, differences between the exact and the approximate estimates were always negligible.


Figure 2. Supporting figures of Case Study

## Conclusion

The objective of this work was to evaluate the effect of the addition of Firth's penalty term to the Bradley-Terry log-likelihood. One of the fundamental differences between the current work and earlier applications of strength estimation in the literature, such as in Agresti (2002) and Firth (2005), is due to the size and degree of imbalance of the data. Application of the implied models to a sufficiently large and heterogeneous pool of players allows better characterization of the impact of the penalty term. The differences between penalized and unpenalized ML estimates were generally more pronounced when the number of matches were relatively low or when victory or defeat rates were in

## PAUL MEYVISCH

the high range. Findings due to Heinze \& Schemper (2002) such as the recommended use of profile likelihood confidence intervals over Wald-based confidence limits, in presence of asymmetric likelihood functions, also carry over to the Bradley-Terry model.

A secondary objective consisted of the development of a MM algorithm for optimization of the penalized log-likelihood. Direct application of the MM algorithm to this type of data may seem inefficient due to the availability of logistic regression software that can easily produce Firth-type maximum likelihood estimates. However, some of the extensions of the Bradley-Terry model cannot be expressed as a logistic regression model and MM algorithms can be used as an alternative as they tend to give fast, simple-to-code iterations, where each iteration moves in the right direction. When applied to the full size of the data, the MM algorithm converged within an acceptable time frame and behaved stably for any set of starting values. Although exact results were not obtained with the proposed MM algorithm, the approximate values were shown to be sufficiently close to the exact values when applied to the data at hand. The applicability of these results may need to be confirmed on other data sets. A favorable feature of the proposed MM algorithm is that it is constructed in such a way that major matrix operations at every single iteration are avoided. As convergence of the algorithm is only obtained after several hundreds of iterations, the gain in processing time is expected to be considerable. In a next step, approximate MM algorithms will need to be constructed on some of the wellknown extensions of the Bradley-Terry model. This will be a subject for further research.

## References

Agresti A. (2002). Categorical data analysis (2nd edition). New York: John Wiley \& Sons.

Albert, A. \& Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1), 1-10. doi: 10.1093/biomet/71.1.1

Bradley, R. A. \& Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. Biometrika, 39(3-4), 324-345. doi: 10.1093/biomet/39.3-4.324

Cytel. (n.d.). LogXact® 11: Exact Inference for Logistic Regression. Retrieved October 5, 2014, from http://www.cytel.com/software/logxact

## ADDING FIRTH'S PENALTY TO THE BRADLEY-TERRY LIKELIHOOD

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27-38. doi: 10.1093/biomet/80.1.27

Firth, D. (2005). Bradley-Terry models in R. Journal of Statistical Software, 12(1), 1-12. doi: 10.18637/jss.v012.i01

Ford, L. R. (1957). Solution of a ranking problem from binary comparisons. American Mathematical Monthly, 64(8.2), 28-33. doi: 10.2307/2308513

Glickman M. E. (1995). Chess rating systems. American Chess Journal, 3, 59-102.

Glickman M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. Journal of the Royal Statistical Society C: Applied Statistics, 48(3), 377-394. doi: 10.1111/1467-9876.00159

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine, 25(24), 4216-4226. doi: 10.1002/sim. 2687

Heinze, G. \& Ploner, M. (2004). A SAS macro, S-plus library and R package to perform logistic regression without convergence problems (Technical Report 2/2004). Vienna, Austria: Department of Medical Computer Sciences, Medical University of Vienna.

Heinze, G. \& Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine, 21(16), 2409-2419. doi: 10.1002/sim. 1047

Hunter, D. R. (2004). MM Algorithms for generalized Bradley-Terry models. The Annals of Statistics, 32(1), 386-408. doi: 10.1214/aos/1079120141

Jeffreys, A. (1961). The theory of probability. Cambridge, UK: Cambridge University Press.

Marcus, D. J. (2001). New table-tennis rating system. Journal of the Royal Statistical Society, Series D (The Statistician), 50(2), 191-208. doi: 10.1111/14679884.00271

SAS Institute, Inc. (2009). SAS/STAT users guide, version 9.2.: Cary, NC: SAS Institute Inc.

So, Y. (1995). A tutorial on logistic regression. [PDF]. Retrieved from http://www.ats.ucla.edu/stat/SAS/library/logistic.pdf

Zermelo, E. (1929). Die berechnung der Turnier-Ergebnisse als ein maximumproblem der Wahrscheinlichkeitsrechnung. Mathemathische Zeitschrift, 29(1), 436-460. doi: 10.1007/BF01180541

# A New Test for Correlation on Bivariate Nonnormal Distributions 

Ping Wang<br>Great Basin College<br>Elko, NV

Ping Sa<br>University of North Florida<br>Jacksonville, FL

A new method to conduct a right-tailed test for the correlation on bivariate non-normal distribution is proposed. The comparative simulation study shows that the new test controls the type I error rates well for all the distributions considered. An investigation of the power performance is also provided.

Keywords: Correlation, Edgeworth expansion, Cornish-Fisher inverse expansion, type I error rate, power performance

## Introduction

Bivariate data is data that has two variables. In the bivariate case, the study of the relationship between the two variables is at least as important as analyzing each variable individually. The most popular measure of the strength of the linear relation between two variables is the correlation coefficient, denoted by $\rho$. The Pearson product-moment correlation, $r$, is the most frequently-used estimator for $\rho$. Another widely-used estimator is the Spearman's rank correlation, denoted by $r_{s}$.

## Tests Based on Pearson Product-Moment Correlation

Pearson (1896) developed the initial mathematical formulas for the sample correlation coefficient. Let $\left(X_{i}, Y_{i}\right), i=1, \ldots, n$ be a random sample, the statistic $r$ is given by:

$$
r=\frac{s_{X Y}}{s_{X} s_{Y}}=\frac{\sum\left(X_{i}-\bar{X}\right)\left(Y_{i}-\bar{Y}\right)}{\left[\sum\left(X_{i}-\bar{X}\right)^{2} \sum\left(Y_{i}-\bar{Y}\right)^{2}\right]^{1 / 2}}
$$

Ms. Wang is the Director of the Academic Success and Testing Center. Email her at: ping.wang @gbcnv.edu. Dr. Sa is a Professor in the Department of Mathematics and Statistics. Email her at: psa@unf.edu.

## WANG \& SA

where $s_{X Y}$ is the sample covariance of $X$ and $Y, n$ is the sample size, $s_{X}$ and $s_{Y}$ are the sample standard deviations, and $\bar{X}$ and $\bar{Y}$ are the sample means for the variables $X$ and $Y$, respectively.

The Pearson product-moment correlation $r$ is the maximum likelihood estimator of the parameter $\rho$ when the population has a bivariate normal distribution. Although $r$ is a biased estimator, the bias is negligible when the sample size is large. Researchers have done intensive work on the distribution of $r$ when the population is bivariate normal (Fisher, 1915; Stuart \& Ord, 1994).
$r$ can be used to test $\mathrm{H}_{0}: \rho=0$ when the population is a bivariate normal distribution. The test statistic

$$
t_{r}^{*}=r \sqrt{\frac{n-2}{1-r^{2}}}
$$

follows the Student's $t$-distribution with $n-2$ degrees of freedom under $\mathrm{H}_{0}$.
$r$ can also be used to test $\mathrm{H}_{0}: \rho=\rho_{0}$, for $-1 \leq \rho_{0} \leq 1$. The sampling distribution of $r$ is complicated and unstable even when the population is bivariate normal. Fisher (1921) introduced a remarkable transformation of $r$, which tends to normality much faster. When the sample size $n$ is moderately large, given

$$
r^{*}=\frac{1}{2} \ln \frac{1+r}{1-r}, \quad \rho^{*}=\frac{1}{2} \ln \frac{1+\rho}{1-\rho}
$$

the distribution of $r^{*}-\rho^{*}$ approaches to normal with an approximate mean $\rho / 2(n-1)$ and variance $1 / n-3$. Note that $n>50$ is an adequate sample size for the above approximation (see David, 1938).

To test $\mathrm{H}_{0}: \rho=\rho_{0}$, the test statistic is:

$$
\begin{equation*}
z_{F}=\frac{\frac{1}{2} \ln \frac{1+r}{1-r}-\frac{1}{2} \ln \frac{1+\rho_{0}}{1-\rho_{0}}-\frac{\rho_{0}}{2(n-1)}}{\sqrt{\frac{1}{n-3}}} \tag{1}
\end{equation*}
$$

$z_{F}$ has approximately a standard normal distribution under $\mathrm{H}_{0}$.

## A NEW TEST FOR CORRELATION

## Test Based on Spearman Rank Correlation

Spearman (1904) proposed a rank correlation which can be used to measure the relationship between two variables when the distribution is neither bivariate normal nor transformed to a bivariate normal. The Spearman rank correlation, $r_{s}$, is a nonparametric version of the Pearson product-moment correlation. Let ( $R_{1 i}, R_{2 i}$ ), $i=1, \ldots, n$ be the paired rank data of two variables, $r_{s}$ is given by:

$$
r_{s}=\frac{s_{R_{1} R_{2}}}{s_{R_{1}} s_{R_{2}}}=\frac{\sum\left(R_{1 i}-\bar{R}_{1}\right)\left(R_{2 i}-\bar{R}_{2}\right)}{\left[\sum\left(R_{1 i}-\bar{R}_{1}\right)^{2} \sum\left(R_{2 i}-\bar{R}_{2}\right)^{2}\right]^{1 / 2}}
$$

where $s_{R_{1} R_{2}}$ is the sample covariance of the paired ranks and $s_{R_{1}}$ and $\bar{R}_{1}, s_{R_{2}}$ and $\bar{R}_{2}$ are the sample standard deviation and the sample mean of the ranks of the two variables, respectively.

The Spearman rank correlation $r_{s}$ can be used to test:

$$
\mathrm{H}_{0} \text { : there is no association between the rank pairs }
$$

The test statistic is

$$
t_{s}^{*}=r_{s} \sqrt{\frac{n-2}{1-r_{s}^{2}}}
$$

which follows the Student's $t$-distribution with $n-2$ degrees of freedom under $\mathrm{H}_{0}$.

## Other Tests on Correlation

The test based on $r$ can only be used when the population is bivariate normal or the sample size is relatively large. Although the test based on $r_{s}$ is applicable to the distribution-free case, it is less powerful and limited to test for zero correlation. However, in real world situations, most distributions are not bivariate normal and the sample sizes may not be large. Furthermore, a test of non-zero correlation is often required. It is desired to develop methods to meet these needs.

Beasley et al. (2007) proposed two new approaches to test a non-zero by using the bootstrapping method. Their methods do not require any knowledge of the population. One is the hypothesis-imposed univarite sampling bootstrap (HI) and the other one is the observed-imposed univariate sampling bootstrap (OI). Two tests
are conducted on populations with various combinations of normal and skewed variates with $\rho \geq 0.4$ and the sample size $n \geq 10$. Their study demonstrated that although OI is preferable to HI under the significance level of 0.05 , the type I error rates are still slightly inflated. Also, the simulated populations are limited to the combinations between normal and skewed populations. The methods are not evaluated under the situations that both variables are non-normal. Another drawback of these two methods is that they are computer-intensive methods. Unfortunately, most practitioners do not have the computer programming skills to implement these methods.

Beversdorf and Sa (2011) proposed tests of correlation for bivariate nonnormal data with small sample sizes. The tests investigated are Fisher's Z transformation $z_{F}$ and the saddlepoint approximation $r_{L}$. They found that $z_{F}$ and $r_{L}$ have extremely similar performance which could control the type I error rates well when a left-tailed test was performed under all the bivariate non-normal distributions considered. Both methods essentially failed to control the type I error rates when a right-tailed test is desired.

The purpose of this study is to develop a new right-tailed test on bivariate non-normal distributions with non-zero correlation. The new test statistic is derived using the Edgeworth expansion and the Cornish-Fisher inverse expansion.

## Methodology

## Edgeworth Expansion

The Edgeworth expansion was derived by Edgeworth (1905), and uses a series to approximate a probability distribution in terms of its cumulants. Let $\hat{\theta}$ be an estimator of an unknown parameter $\theta$, and $\sqrt{n}(\hat{\theta}-\theta)$ be asymptotically normally distributed with mean zero and variance $\sigma^{2}$. Hall (1983) developed the Edgeworth expansion of the distribution function of $\sqrt{n}(\hat{\theta}-\theta)$ as a power series in $\sqrt{n}$.

$$
\begin{equation*}
\mathrm{P}\left\{\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma} \leq u\right\}=\Phi(u)+n^{-1 / 2} \mathrm{p}_{1}(u) \phi(u)+\ldots+n^{-j / 2} \mathrm{p}_{j}(u) \phi(u)+\ldots \tag{2}
\end{equation*}
$$

## A NEW TEST FOR CORRELATION

where $\Phi(u), \phi(u)$, and $\mathrm{p}_{j}(u)$ denote the standard normal distribution function, its density function, and a polynomial function with coefficients depending on cumulants of $\hat{\theta}-\theta$, respectively.

The inverse of the Edgeworth expansion, obtained by inverting the formula (2), is known as the Cornish-Fisher expansion:

$$
\mathrm{P}\left\{\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma} \leq z+n^{-1 / 2} \mathrm{p}_{11}(z)+n^{-1} \mathrm{p}_{21}(z)+\ldots+n^{-j / 2} \mathrm{p}_{j 1}(z)+\ldots\right\}=\Phi(z)
$$

where $z$ is the percentile of the standard normal distribution and the $\mathrm{p}_{j 1}$ are polynomials defined in terms of $\mathrm{p}_{j} s$ (Hall, 1992).

## Proposed Test Procedure

Assume that a bivariate population has finite cumulants and a correlation coefficient $\rho$. Let $\kappa_{01}, \kappa_{10}, \kappa_{02}, \kappa_{20}, \kappa_{11}, \ldots$ up to order six be the product cumulants for the bivariate population. Then $r_{*}=\sqrt{n}(r-\rho)$ has a limiting normal distribution with mean zero and constant variance $\sigma^{2}$, where $\sigma^{2}$ is of the form (Nakagawa \& Niki, 1992):

$$
\begin{gathered}
\sigma^{2}=\frac{1}{\kappa_{20} \kappa_{02}}\left(\frac{1}{4} \kappa_{40} \kappa_{20}^{-2} \kappa_{11}^{2}-\kappa_{31} \kappa_{20}^{-1} \kappa_{11}+\frac{1}{2} \kappa_{22} \kappa_{20}^{-1} \kappa_{11}^{2} \kappa_{02}^{-1}+\kappa_{22}+\kappa_{20} \kappa_{02}+\kappa_{20}^{-1} \kappa_{11}^{4} \kappa_{02}^{-1}\right. \\
\left.-\kappa_{13} \kappa_{11} \kappa_{02}^{-1}+\frac{1}{4} \kappa_{11}^{2} \kappa_{04} \kappa_{02}^{-2}-2 \kappa_{11}^{2}\right)
\end{gathered}
$$

Bhattacharya and Ghosh (1978) provided the Edgeworth expansion of $R$, where

$$
R=\frac{r_{*}}{\sigma}=\frac{\sqrt{n}(r-\rho)}{\sigma}
$$

as

## WANG \& SA

$$
\begin{align*}
\mathrm{P}(R<u)=\Phi(u)-\phi(u) & {\left[\frac{1}{\sqrt{n}}\left\{\frac{1}{6} \frac{v_{3}}{\sigma^{3}} \mathrm{H}_{2}(u)+\frac{v_{1}}{\sigma}\right\}+\frac{1}{n}\left\{\frac{1}{72} \frac{v_{3}^{2}}{\sigma^{6}} \mathrm{H}_{5}(u)\right.\right.} \\
& +\left(\frac{1}{6} \frac{v_{1} v_{3}}{\sigma^{4}}+\frac{1}{24} \frac{v_{4}}{\sigma^{4}} \mathrm{H}_{3}(u)\right)  \tag{3}\\
& \left.\left.+\left(\frac{1}{2} \frac{v_{2}}{\sigma^{2}}+\frac{1}{2} \frac{v_{2}}{\sigma^{2}}\right) \mathrm{H}_{1}(u)\right\}\right]+\mathrm{O}\left(\frac{1}{n \sqrt{n}}\right)
\end{align*}
$$

where $\Phi(u)$ and $\phi(u)$ denote the standard normal distribution function and its density function, $\mathrm{O}(1 / n \sqrt{n})$ is the big-oh function of order $1 / n \sqrt{n}, \mathrm{H}_{1}(u), \mathrm{H}_{2}(u)$, $\mathrm{H}_{3}(u)$, and $\mathrm{H}_{5}(u)$ are Hermite polynomials with

$$
\mathrm{H}_{1}(u)=u ; \quad \mathrm{H}_{2}(u)=u^{2}-1 ; \quad \mathrm{H}_{3}(u)=u^{3}-3 u ; \quad \mathrm{H}_{5}(u)=u^{5}-10 u^{3}+15 u,
$$

and $v_{1}, v_{2}, v_{3}$, and $v_{4}$ are parameters such that

$$
\frac{1}{\sqrt{n}} \frac{v_{1}}{\sigma}, \quad 1+\frac{1}{n} \frac{v_{2}}{\sigma^{2}}, \quad \frac{1}{\sqrt{n}} \frac{v_{3}}{\sigma^{3}}, \quad \frac{1}{n} \frac{v_{4}}{\sigma^{4}}
$$

are the approximate cumulants of $R$. The explicit forms of $v_{1}, v_{2}, v_{3}$, and $v_{4}$ were provided by Nakagawa and Niki (1992). Formulas for calculating $v_{1}$ and $v_{3}$ are listed in Appendix A. Formulas for calculating $v_{2}$ and $v_{4}$ are not needed in this study.

Nakagawa and Niki (1992) applied the inverted Edgeworth expansion to the distribution of $R$ of order $1 / n$ :

$$
\begin{align*}
\mathrm{P}(R \leq z & +\frac{1}{\sqrt{n}}\left\{\frac{1}{6} \frac{v_{3}}{\sigma^{3}}\left(z^{2}-1\right)+\frac{v_{1}}{\sigma}\right\}+\frac{1}{n}\left\{z^{3}\left(-\frac{1}{18} \frac{v_{3}^{2}}{\sigma^{6}}+\frac{1}{24} \frac{v_{4}}{\sigma^{4}}\right)\right. \\
& \left.\left.+z\left(\frac{1}{2} \frac{v_{2}}{\sigma^{2}}+\frac{5}{36} \frac{v_{3}^{2}}{\sigma^{6}}-\frac{1}{8} \frac{v_{4}}{\sigma^{4}}\right)\right\}\right)  \tag{4}\\
& =\Phi(z)+\mathrm{O}\left(\frac{1}{n \sqrt{n}}\right)
\end{align*}
$$

If only order $1 / \sqrt{n}$ is required, then (4) can be reduced to a simpler form:

## A NEW TEST FOR CORRELATION

$$
\begin{equation*}
\mathrm{P}\left(R \leq z+B_{1}\right)=\Phi(z)+\mathrm{O}\left(\frac{1}{n}\right) \tag{5}
\end{equation*}
$$

where $B_{1}=\frac{1}{\sqrt{n}}\left\{\frac{1}{6} \frac{v_{3}}{\sigma^{3}}\left(z^{2}-1\right)+\frac{v_{1}}{\sigma}\right\}$.
To test $\mathrm{H}_{0}: \rho=\rho_{0}$ versus $\mathrm{H}_{\mathrm{a}}: \rho>\rho_{0}$, the intuitive decision rule is:

$$
\text { reject } \mathrm{H}_{0}: \rho=\rho_{0} \text { when } R>z+B_{1} \text {, i.e. }
$$

reject $\mathrm{H}_{0}: \rho=\rho_{0}$ under the significance level $\alpha$, if

$$
\begin{equation*}
\frac{\sqrt{n}\left(r-\rho_{0}\right)}{\sigma}>z_{\alpha}+\frac{1}{\sqrt{n}}\left\{\frac{1}{6} \frac{v_{3}}{\sigma^{3}}\left(z_{\alpha}^{2}-1\right)+\frac{v_{1}}{\sigma}\right\} \tag{6}
\end{equation*}
$$

Since negative values of $B_{1}$ might increase type I errors, the following adjustment is proposed: Define

$$
\begin{equation*}
B_{2}=\max \left(z_{\alpha}, z_{\alpha}+\frac{1}{\sqrt{n}}\left\{\frac{1}{6} \frac{v_{3}}{\sigma^{3}}\left(z_{\alpha}^{2}-1\right)+\frac{v_{1}}{\sigma}\right\}\right) \tag{7}
\end{equation*}
$$

The decision rule is adjusted to:

$$
\begin{equation*}
\text { reject } \mathrm{H}_{0} \text { if } \frac{\sqrt{n}\left(r-\rho_{0}\right)}{\sigma}>B_{2} \tag{8}
\end{equation*}
$$

All the parameters in (6) and (7) can be written in terms of the product cumulants. These product cumulants $\kappa_{i j}$ are estimated by their corresponding unbiased estimators $k_{i j}$. Detailed formulas are provided in Appendix B.

For the special case of $\rho_{0}=0, \kappa_{01}=\kappa_{10}=0, \kappa_{02}=\kappa_{20}=1$, and $\kappa_{p q}=0$ for $p+q \geq 5$, Nakagawa and Niki (1992) gave the simplified forms for parameters $\sigma^{2}$, $v_{1}$, and $v_{3}$ as follows:

$$
\begin{align*}
& \sigma^{2}=1+\kappa_{22} \\
& v_{1}=-\frac{1}{2}\left(\kappa_{13}+\kappa_{31}\right)  \tag{9}\\
& v_{3}=3 \kappa_{12} \kappa_{21}+\kappa_{03} \kappa_{30}-3 \kappa_{22}\left(\kappa_{13}+\kappa_{31}\right)
\end{align*}
$$

## WANG \& SA

To test $\mathrm{H}_{0}: \rho=0$, (6) and (7) are evaluated with the parameters given in (9). Again, all the parameters are estimated by their corresponding unbiased estimators.

## Simulation Study

The simulation study was implemented to evaluate type I error rates, to investigate the power performance, and to compare with the existing Fisher's Z transformation method on the type I error rates.

## Simulation Description

Fleishman (1978) proposed a method to generate univariate non-normal random variables with desired coefficients of skewness $\beta$ and kurtosis $\gamma$. Vale and Maurelli (1983) extended Fleishman's method to the bivariate non-normal case with a specified correlation coefficient. Five parameters, including two sets of skewness and kurtosis and one correlation coefficient, are required to generate the bivariate non-normal data using Vale and Maurelli method.

Seven levels of the skewness, $-3.0,-1.2,-0.5,0.0,0.5,1.2$, and 3.0 , and five levels of the kurtosis, $0.0,4.0,10.0,14.0$, and 25.0 , were considered, and 24 combinations were selected. Moreover, five correlation coefficients, $0.0,0.5,0.6$, 0.75 , and 0.9 , three significance levels, $0.10,0.05$, and 0.01 , and two sample sizes, 15 and 30 , were used in the simulation study.

Two new methods and the Fisher's Z transformation method were evaluated. The method using (6) was denoted by Zb , and the one using (8) was denoted by Zc . The Fisher's Z transformation method (1) was shortened as Zf. Both Zb and Zc methods were evaluated with two critical values, $z_{\alpha}$ and $t_{(\alpha, n-2)}$.

## The Algorithm of the Test on Correlation:

1. Input the desired $\rho_{X, Y}$ and two sets of skewness and kurtosis, $\left(\beta_{1}, \gamma_{1}\right)$ and $\left(\beta_{2}, \gamma_{2}\right)$.
2. Generate $n$ bivariate non-normal random variates $(X, Y)$ based on the given parameters.
3. Calculate $z_{F}$ in (1), Zb in (6), and Zc in (8).
4. Compare the tests with their critical values; count one if the test is rejected.
5. Repeat (2) - (4) 99,999 times.

## A NEW TEST FOR CORRELATION

6. Calculate the type I error rate, the proportion of the false rejection (out of 100,000 ) for each test.

In the power study, an extra parameter $\rho_{a}$ is input in step (1) and used to generate the data as the true population correlation. However, all of the test statistics in step (3) are evaluated under $\rho_{0}$. All fo the simulations were run with Fortran 77 for Windows on an IBM T61 Laptop Computer.

## Simulation Results

## Type I Error Rate Comparisons

Tables 1-3 provide the comparative study of the type I error rates on various bivariate non-normal distributions with significance levels $0.10,0.05$, and 0.01 and sample size 30 . Comparisons were made among the tests $\mathrm{Zf}, \mathrm{Zb}$, and Zc with two critical values, $z_{\alpha}$ and $t_{(\alpha, n-2)}$, while Zf only used the critical value, $z_{\alpha}$. The correlation coefficients $0.00,0.50,0.60,0.75$, and 0.90 were targeted during the simulation study. A total of 24 bivariate non-normal distributions with various population conditions were examined.

Table 1 shows the results on testing a zero correlation. It can be observed that the Zc method controls the type I error rates well. On the contrary, the Zb method do not control type I error rates at all. Almost all of the type I error rates obtained by the Zb method are slightly inflated except for a few cases. The Zf method can control the type I error rates as long as the skewness and kurtosis are small. Once theses parameters increase, Zf becomes unstable and fails to control the type I error rates in many cases.

More specifically, in testing $\rho_{0}=0$ on a distribution which is bivariate normal or very close to bivariate normal, Zf controls type I error rates a bit better than Zc . However, for the non-normal distributions, Zc is better than Zf in controlling type I error rates.

Tables 2 and 3 give the results for right-tailed tests on the non-zero correlation. It is quite interesting to see that the hypothesized value $\rho_{0}$ actually affects the type I error rate performance. When $n=30$ and $\rho_{0}=0.5$, both the Zf and Zb methods basically fail to control the type I error rates with very few exceptions. The type I error rates obtained by the Zc method have better performance. However, the cases with controlled type I error rates are restricted to the distributions with small to moderate skewness and kurtosis. When $\rho_{0}$ increases to 0.6 , the Zc method successfully controls the type I error rates for nearly all the distributions considered
with the $t$ critical point. As $\rho_{0}$ increases, the type I error rates get more conservative. This tendency can be observed on both Zb and Zc methods.

The Zf method fails completely in the right-tailed test on non-zero correlation with only a few exceptions. This result confirms with the study by Beversdorf and Sa (2011). Their study shows that Zf can properly control the type I error rates on the left-tailed test but not on the right-tailed test. Therefore, it is fair to conclude that, for the right-tailed test, the only method that can properly control the type I error rates is the Zc method with the $t$ critical point.

Due to the similar results in the study, only the moderate sample size 30 and significance levels of 0.05 and 0.01 are reported in the tables.

## Power Results

The power performance of the proposed test is also evaluated. Tables 4 and 5 provide the power performance to test $\rho_{0}=0$ when $\rho_{a}=0.0,0.2,0.4,0.6$, and 0.8 with significance levels 0.05 and 0.01 . Table 6 provides a small-scale investigation on the power performance to test $\rho_{0}=0.55$ and $\rho_{a}=0.6$ and 0.7 .

Both the Zf and Zc methods perform well in testing $\rho_{0}=0$. In testing on an exactly- or nearly-normal distribution, the power from Zf and Zc converges to 1 quickly. When $\rho_{a}=0.6$, both achieve a power of 0.99 ; when $\rho_{a}=0.8$, the power rates are essentially 1 . For the distributions with large skewness and kurtosis, the Zc method, which is the only one with controlled type I error rates, converges to 1 more slowly but still reasonably well. A small-scale power study to test non-zero correlation is presented in Table 6. At significance level 0.10, sample size 30, $\rho_{0}=0.55$ versus $\rho_{a}=0.6$ and 0.7 , it is observed that the power of Zc steadily increases when $\rho_{a}$ moves away from $\rho_{0}$.

## Conclusions

This study proposed a new right-tailed test for the correlation of bivariate nonnormal distributions. This new test adapts the inverse Edgeworth expansion for the standardized correlation $R=\frac{\sqrt{n}(r-\rho)}{\sigma}$ by Nakagawa and Niki (1992).

This newly proposed test can be conducted without any knowledge of the populations. The simulation study shows that this new right-tailed test has the best performance in controlling the type I error rates. The proposed method, along with the $t$ critical point, can be used to test both $\rho_{0}=0$ and any value of $\rho_{0}$ when $\rho_{0}>0.5$.

The power performance of the new test was also evaluated. Zc is as powerful as Zf when testing $\rho_{0}=0$. To test non-zero correlations, it is meaningless to

## A NEW TEST FOR CORRELATION

compare the two tests since Zf fails to control type I error rates. The power examination of the Zc method shows that the power steadily goes up when $\rho_{a}$ moves away from $\rho_{0}$.

The new test does have its own limitations. It cannot control the type I error rates well when the population has a small correlation and it is a right-tailed test. In order to better control the type I error rates, a higher-order Edgeworth expansion may be considered. Unfortunately, this might lead to tedious computations when higher-order terms are introduced in the test.

## References

Beasley, W. H., DeShea, L., Toothaker, L. E., Mendoza, J. L., Bard, D. E., \& Rodgers, J. L. (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. Psychological Methods, 12(4), 414-433. doi: 10.1037/1082-989X.12.4.414

Beversdorf, L. M., \& Sa, P. (2011). Tests for correlation on bivariate nonnormal distributions. Journal of Modern Applied Statistical Methods, 10(2), 699709. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol10/iss2/29/

Bhattacharya, R. N., \& Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. The Annals of Statistics, 6(2), 434-451. Available from http://www.jstor.org/stable/2958887

David, F. N. (1938). Tables of the correlation coefficient. London: Cambridge University Press.

Edgeworth, F. Y. (1905). The law of error. Transactions of the Cambridge Philosophical Society, 20, 36-65, 113-141.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10(4), 507-521. doi: 10.2307/2331838

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. Metron, 1, 3-32.

Fisher, R. A. (1930). Moments and product moments of sampling distributions. Proceedings of the London Mathematical Society, s2-30(1), 199238. doi: $10.1112 / \mathrm{plms} / \mathrm{s} 2-30.1 .199$

Fleishman, A. I. (1978). A method of simulating non-normal distributions. Psychometrika, 43(4), 521-531. doi: 10.1007/BF02293811

Hall, P. (1983). Inverting an Edgeworth expansion. The Annals of Statistics, 11(2), 569-576. doi: 10.1214/aos/1176346162

## WANG \& SA

Hall, P. (1992). The bootstrap and Edgeworth expansion. New York, NY: Springer-Verlag.

Nakagawa, S., \& Niki, N. (1992). Distribution of sample correlation coefficient for nonnormal populations. Journal of the Japanese Society of Computational Statistics, 5(1), 1-19. doi: 10.5183/jjscs1988.5.1

Pearson, K. (1896). Mathematical contributions to the theory of evolution.
III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, 187, 253-318. Available from http://www.jstor.org/stable/90707

Spearman, C. (1904). "General intelligence," Objectively determined and measured. The American Journal of Psychology, 15(2), 201-293. doi: 10.2307/1412107

Stuart, A., \& Ord, J. K. (1994). Kendall's advanced theory of statistics (6th ed.) (Vol. 1). New York, NY: Halsted Press.

Vale, D. C., \& Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. Psychometrika, 48(3), 465-471. doi: 10.1007/BF02293687

## A NEW TEST FOR CORRELATION

## Appendix A: Formulae Used in Edgeworth Expansion of $\boldsymbol{R}$

Let $\eta=\frac{1}{\sqrt{\kappa_{20} \kappa_{02}}}$. Then

$$
\begin{aligned}
& \nu_{1}=\eta\left(\frac{3}{8} \kappa_{40} \kappa_{20}^{-2} \kappa_{11}-\frac{1}{2} \kappa_{31} \kappa_{20}^{-1}+\frac{1}{4} \kappa_{22} \kappa_{20}^{-1} \kappa_{11} \kappa_{02}^{-1}+\frac{1}{2} \kappa_{20}^{-1} \kappa_{11}^{3} \kappa_{02}^{-1}-\frac{1}{2} \kappa_{13} \kappa_{02}^{-1}\right. \\
& \left.+\frac{3}{8} \kappa_{11} \kappa_{04} \kappa_{02}^{-2}-\frac{1}{2} \kappa_{11}\right) \\
& \nu_{3}=\eta^{3}\left(-\frac{1}{8} \kappa_{60} \kappa_{20}^{-3} \kappa_{11}^{3}+\frac{3}{4} \kappa_{51} \kappa_{20}^{-2} \kappa_{11}^{2}-\frac{3}{2} \kappa_{42} \kappa_{20}^{-1} \kappa_{11}-\frac{3}{8} \kappa_{42} \kappa_{20}^{-2} \kappa_{11}^{3} \kappa_{02}^{-1}+\frac{9}{16} \kappa_{40}^{2} \kappa_{20}^{-4} \kappa_{11}^{3}\right. \\
& -3 \kappa_{40} \kappa_{31} \kappa_{20}^{-3} \kappa_{11}^{2}+\frac{3}{2} \kappa_{40} \kappa_{22} \kappa_{20}^{-2} \kappa_{11}+\frac{3}{2} \kappa_{40} \kappa_{22} \kappa_{20}^{-3} \kappa_{11}^{3} \kappa_{02}^{-1}-\frac{3}{2} \kappa_{40} \kappa_{20}^{-2} \kappa_{13} \kappa_{11}^{2} \kappa_{02}^{-1} \\
& +\frac{3}{8} \kappa_{40} \kappa_{20}^{-2} \kappa_{11}^{3} \kappa_{04} \kappa_{02}^{-2}-3 \kappa_{40} \kappa_{20}^{-2} \kappa_{11}^{3}+3 \kappa_{40} \kappa_{20}^{-3} \kappa_{11}^{5} \kappa_{02}^{-1}+\frac{3}{2} \kappa_{33} \kappa_{20}^{-1} \kappa_{11}^{2} \kappa_{02}^{-1}+\kappa_{33} \\
& +\frac{15}{4} \kappa_{31}^{2} \kappa_{20}^{-2} \kappa_{11}-3 \kappa_{31} \kappa_{22} \kappa_{20}^{-1}-\frac{9}{2} \kappa_{31} \kappa_{22} \kappa_{20}^{-2} \kappa_{11}^{2} \kappa_{02}^{-1}+\frac{9}{2} \kappa_{31} \kappa_{20}^{-1} \kappa_{13} \kappa_{11} \kappa_{02}^{-1} \\
& -\frac{3}{2} \kappa_{31} \kappa_{20}^{-1} \kappa_{11}^{2} \kappa_{04} \kappa_{02}^{-2}+12 \kappa_{31} \kappa_{20}^{-1} \kappa_{11}^{2}-12 \kappa_{31} \kappa_{20}^{-2} \kappa_{11}^{4} \kappa_{02}^{-1}-\frac{1}{2} \kappa_{30}^{2} \kappa_{20}^{-3} \kappa_{11}^{3} \\
& +3 \kappa_{30} \kappa_{21} \kappa_{20}^{-2} \kappa_{11}^{2}-3 \kappa_{30} \kappa_{20}^{-1} \kappa_{12} \kappa_{11}+\kappa_{30} \kappa_{03}-\frac{3}{8} \kappa_{24} \kappa_{20}^{-1} \kappa_{11}^{3} \kappa_{02}^{-2}-\frac{3}{2} \kappa_{24} \kappa_{11} \kappa_{02}^{-1} \\
& +3 \kappa_{22}^{2} \kappa_{20}^{-1} \kappa_{11} \kappa_{02}^{-1}+\frac{3}{2} \kappa_{22}^{2} \kappa_{20}^{-2} \kappa_{11}^{3} \kappa_{02}^{-2}-\frac{9}{2} \kappa_{22} \kappa_{20}^{-1} \kappa_{13} \kappa_{11}^{2} \kappa_{02}^{-2}+\frac{3}{2} \kappa_{22} \kappa_{20}^{-1} \kappa_{11}^{3} \kappa_{04} \kappa_{02}^{-3} \\
& +6 \kappa_{22} \kappa_{20}^{-1} \kappa_{11}^{3} \kappa_{02}^{-1}+6 \kappa_{22} \kappa_{20}^{-2} \kappa_{11}^{5} \kappa_{02}^{-2}-3 \kappa_{22} \kappa_{13} \kappa_{02}^{-1}+\frac{3}{2} \kappa_{22} \kappa_{11} \kappa_{04} \kappa_{02}^{-2}-12 \kappa_{22} \kappa_{11} \\
& -3 \kappa_{21}^{2} \kappa_{20}^{-1} \kappa_{11}-\frac{3}{2} \kappa_{21}^{2} \kappa_{20}^{-2} \kappa_{11}^{3} \kappa_{02}^{-1}+6 \kappa_{21} \kappa_{20}^{-1} \kappa_{12} \kappa_{11}^{2} \kappa_{02}^{-1}+3 \kappa_{21} \kappa_{12}-3 \kappa_{21} \kappa_{11} \kappa_{03} \kappa_{02}^{-1} \\
& -6 \kappa_{20} \kappa_{11} \kappa_{02}-12 \kappa_{20}^{-1} \kappa_{13} \kappa_{11}^{4} \kappa_{02}^{-2}-\frac{3}{2} \kappa_{20}^{-1} \kappa_{12}^{2} \kappa_{11}^{3} \kappa_{02}^{-2}+3 \kappa_{20}^{-1} \kappa_{11}^{5} \kappa_{04} \kappa_{02}^{-3}-18 \kappa_{20}^{-1} \kappa_{11}^{5} \kappa_{02}^{-1} \\
& +6 \kappa_{20}^{-2} \kappa_{11}^{7} \kappa_{02}^{-2}+\frac{3}{4} \kappa_{15} \kappa_{11}^{2} \kappa_{02}^{-2}+\frac{15}{4} \kappa_{13}^{2} \kappa_{11} \kappa_{02}^{-2}-3 \kappa_{13} \kappa_{11}^{2} \kappa_{04} \kappa_{02}^{-3}+12 \kappa_{13} \kappa_{11}^{2} \kappa_{02}^{-1} \\
& -3 \kappa_{12}^{2} \kappa_{11} \kappa_{02}^{-1}+3 \kappa_{12} \kappa_{11}^{2} \kappa_{03} \kappa_{02}^{-2}-\frac{1}{8} \kappa_{11}^{3} \kappa_{06} \kappa_{02}^{-3}+\frac{9}{16} \kappa_{11}^{3} \kappa_{04}^{2} \kappa_{02}^{-4}-3 \kappa_{11}^{3} \kappa_{04} \kappa_{02}^{-2} \\
& \left.-\frac{1}{2} \kappa_{11}^{3} \kappa_{03}^{2} \kappa_{02}^{-3}+18 \kappa_{11}^{3}\right)
\end{aligned}
$$

## WANG \& SA

## Appendix B: $\boldsymbol{k}$-Statistics

Fisher (1930) introduced $k$-statistics as the unbiased estimator of the $m^{\text {th }}$ cumulant $\kappa_{m}$, i.e. $\mathrm{E}\left(k_{m}\right)=\kappa_{m}$. Define the power sum of a univariate data as: $s_{m}=\sum_{i=1}^{n} x_{i}^{m}$, the first six $k$-statistics in terms of the corresponding $\kappa_{m}$ are (See Stuart \& Ord, 1994):

$$
\begin{aligned}
& \begin{array}{l}
k_{1}=\frac{1}{n^{[1]}} s_{1}, \quad k_{2}=\frac{1}{n^{[2]}}\left(n s_{2}-s_{1}^{2}\right), \\
k_{3}= \\
\frac{1}{n^{[3]}}\left(n^{2} s_{3}-3 n s_{2} s_{1}+2 s_{1}^{3}\right),
\end{array} \\
& k_{4}=\frac{1}{n^{[4]}}\left\{\left(n^{3}+n^{2}\right) s_{4}-4\left(n^{2}+n\right) s_{3} s_{1}-3\left(n^{2}-n\right) s_{2}^{2}+12 n s_{2} s_{1}^{2}-6 s_{1}^{4}\right\}, \\
& k_{5}=\frac{1}{n^{[5]}}\left\{\left(n^{4}+5 n^{3}\right) s_{5}-5\left(n^{3}+5 n^{2}\right) s_{4} s_{1}-10\left(n^{3}-n^{2}\right) s_{3} s_{2}+20\left(n^{2}+2 n\right) s_{3} s_{1}^{2}\right. \\
&\left.+30\left(n^{2}-n\right) s_{2}^{2} s_{1}-60 n s_{2} s_{1}^{3}+24 s_{1}^{5}\right\}, \\
& k_{6}=\frac{1}{n^{[6]}}\left\{\left(n^{5}+16 n^{4}+11 n^{3}-4 n^{2}\right) s_{6}-6\left(n^{4}+16 n^{3}+11 n^{2}-4 n\right) s_{5} s_{1}\right. \\
& \quad-15 n(n-1)^{2}(n+4) s_{4} s_{2}-10\left(n^{4}-2 n^{3}+5 n^{2}-4 n\right) s_{3}^{2} \\
&+30\left(n^{3}+9 n^{2}+2 n\right) s_{4} s_{1}^{2}+120\left(n^{3}-n\right) s_{3} s_{2} s_{1}+30\left(n^{3}-3 n^{2}+2 n\right) s_{2}^{3} \\
&\left.-120\left(n^{2}+3 n\right) s_{3} s_{1}^{3}-270\left(n^{2}-n\right) s_{2}^{2} s_{1}^{2}+360 n s_{2} s_{1}^{4}-120 s_{1}^{6}\right\} \\
&=\frac{1}{n^{[6]}}\left\{\ell_{1} s_{6}-6 \ell_{2} s_{5} s_{1}-15 \ell_{3} s_{4} s_{2}-10 \ell_{4} s_{3}^{2}+30 \ell_{5} s_{4} s_{1}^{2}+120 \ell_{6} s_{3} s_{2} s_{1}\right. \\
&\left.+30 \ell_{7} s_{2}^{3}-120 \ell_{8} s_{3} s_{1}^{2}-270 \ell_{9} s_{2}^{2} s_{1}^{2}+360 \ell_{10} s_{2} s_{1}^{4}-120 s_{1}^{6}\right\}
\end{aligned}
$$

Stuart and Ord (1994) also provided an approach to derive the multivariate $k$ statistics. Define $s_{r t}=\sum_{i=1}^{n} x_{i}^{r} y_{i}^{t}$, where $\left(x_{i}, y_{i}\right), i=1,2, \ldots, n$ are the bivariate random observations. The following multivariate $k$-statistics can be derived:

$$
\begin{aligned}
& k_{11}=\frac{1}{n^{[2]}}\left(n s_{11}-s_{10} s_{01}\right) \\
& k_{21}=\frac{1}{n^{[3]}}\left(n^{2} s_{21}-2 n s_{11} s_{10}-n s_{20} s_{01}+2 s_{10}^{2} s_{01}\right)
\end{aligned}
$$

## A NEW TEST FOR CORRELATION

$$
\begin{aligned}
& k_{12}=\frac{1}{n^{[3]}}\left(n^{2} s_{12}-2 n s_{11} s_{01}-n s_{02} s_{10}+2 s_{01}^{2} s_{10}\right) \\
& k_{31}=\left\{\frac{1}{n^{[4]}}\left(n^{3}+n^{2}\right) s_{31}-3\left(n^{2}+n\right) s_{21} s_{10}-\left(n^{2}+n\right) s_{30} s_{01}-3\left(n^{2}-n\right) s_{20} s_{11}\right. \\
& \left.+6 n s_{11} s_{10}^{2}+6 n s_{20} s_{10} s_{01}-6 s_{10}^{3} s_{01}\right\} \\
& k_{13}=\left\{\frac{1}{n^{[4]}}\left(n^{3}+n^{2}\right) s_{13}-3\left(n^{2}+n\right) s_{12} s_{01}-\left(n^{2}+n\right) s_{03} s_{10}-3\left(n^{2}-n\right) s_{02} s_{11}\right. \\
& \left.+6 n s_{11} s_{01}^{2}+6 n s_{02} s_{01} s_{10}-6 s_{01}^{3} s_{10}\right\} \\
& k_{22}=\left\{\frac{1}{n^{[4]}}\left(n^{3}+n^{2}\right) s_{22}-2\left(n^{2}+n\right) s_{21} s_{01}-2\left(n^{2}+n\right) s_{12} s_{10}-2\left(n^{2}-n\right) s_{11}^{2}\right. \\
& \left.-\left(n^{2}-n\right) s_{20} s_{02}+2 n s_{02} s_{10}^{2}+8 n s_{11} s_{10} s_{01}+2 n s_{20} s_{01}^{2}-6 s_{10}^{2} s_{01}^{2}\right\} \\
& k_{51}=\frac{1}{n^{[6]}}\left\{\ell_{1} s_{51}-5 \ell_{2} s_{41} s_{10}-\ell_{2} s_{50} s_{01}-10 \ell_{3} s_{31} s_{20}-5 \ell_{3} s_{40} s_{11}-10 \ell_{4} s_{30} s_{21}\right. \\
& +20 \ell_{5} s_{31} s_{10}^{2}+10 \ell_{5} s_{40} s_{10} s_{01}+60 \ell_{6} s_{21} s_{20} s_{10}+40 \ell_{6} s_{30} s_{11} s_{10} \\
& +20 \ell_{6} s_{30} s_{20} s_{01}+30 \ell_{7} s_{20}^{2} s_{11}-60 \ell_{8} s_{21} s_{10}^{3}-60 \ell_{8} s_{30} s_{10}^{2} s_{01} \\
& -180 \ell_{9} s_{20} s_{11} s_{10}^{2}-90 \ell_{9} s_{20}^{2} s_{10} s_{01}+120 \ell_{10} s_{11} s_{10}^{4} \\
& \left.+240 \ell_{10} s_{20} s_{10}^{3} s_{01}-120 s_{10}^{5} s_{01}\right\} \\
& k_{42}=\frac{1}{n^{[6]}}\left\{\ell_{1} s_{42}-4 \ell_{2} s_{32} s_{10}-2 \ell_{2} s_{41} s_{01}-6 \ell_{3} s_{22} s_{20}-8 \ell_{3} s_{31} s_{11}-\ell_{3} s_{40} s_{02}\right. \\
& -6 \ell_{4} s_{21}^{2}-4 \ell_{4} s_{30} s_{12}+12 \ell_{5} s_{22} s_{10}^{2}+16 \ell_{5} s_{31} s_{10} s_{01}+2 \ell_{5} s_{40} s_{01}^{2} \\
& +24 \ell_{6} s_{12} s_{20} s_{10}+12 \ell_{6} s_{21} s_{20} s_{01}+48 \ell_{6} s_{21} s_{11} s_{10}+8 \ell_{6} s_{30} s_{02} s_{10} \\
& +12 \ell_{6} s_{21} s_{20} s_{01}+16 \ell_{6} s_{30} s_{11} s_{01}+12 \ell_{7} s_{20} s_{11}^{2}+6 \ell_{7} s_{20}^{2} s_{02} \\
& -24 \ell_{8} s_{12} s_{10}^{3}-72 \ell_{8} s_{21} s_{10}^{2} s_{01}-24 \ell_{8} s_{30} s_{10} s_{01}^{2}-72 \ell_{9} s_{11}^{2} s_{10}^{2} \\
& -36 \ell_{9} s_{20} s_{02} s_{10}^{2}-144 \ell_{9} s_{20} s_{11} s_{10} s_{01}-18 \ell_{9} s_{20}^{2} s_{01}^{2}+24 \ell_{10} s_{02} s_{10}^{4} \\
& \left.+192 \ell_{10} s_{11} s_{10}^{3} s_{01}+144 \ell_{10} s_{20} s_{10}^{2} s_{01}^{2}-120 s_{10}^{4} s_{01}^{2}\right\}
\end{aligned}
$$

## WANG \& SA

$$
\begin{aligned}
k_{24}=\frac{1}{n^{[6]}}\{ & \ell_{1} s_{24}-4 \ell_{2} s_{23} s_{01}-2 \ell_{2} s_{14} s_{10}-6 \ell_{3} s_{22} s_{02}-8 \ell_{3} s_{13} s_{11}-\ell_{3} s_{04} s_{20} \\
& -6 \ell_{4} s_{12}^{2}-4 \ell_{4} s_{03} s_{21}+12 \ell_{5} s_{22} s_{01}^{2}+16 \ell_{5} s_{13} s_{01} s_{10}+2 \ell_{5} s_{04} s_{10}^{2} \\
& +24 \ell_{6} s_{21} s_{02} s_{01}+12 \ell_{6} s_{12} s_{02} s_{10}+48 \ell_{6} s_{12} s_{11} s_{01}+8 \ell_{6} s_{03} s_{20} s_{01} \\
& +12 \ell_{6} s_{12} s_{02} s_{10}+16 \ell_{6} s_{03} s_{11} s_{10}+12 \ell_{7} s_{02} s_{11}^{2}+6 \ell_{7} s_{02}^{2} s_{20} \\
& -24 \ell_{8} s_{21} s_{01}^{3}-72 \ell_{8} s_{12} s_{01}^{2} s_{10}-24 \ell_{8} s_{03} s_{01} s_{10}^{2}-72 \ell_{9} s_{11}^{2} s_{01}^{2} \\
& -36 \ell_{9} s_{02} s_{20} s_{01}^{2}-144 \ell_{9} s_{02} s_{11} s_{01} s_{10}-18 \ell_{9} s_{02}^{2} s_{10}^{2}+24 \ell_{10} s_{20} s_{01}^{4} \\
& \left.+192 \ell_{10} s_{11} s_{01}^{3} s_{10}+144 \ell_{10} s_{02} s_{01}^{2} s_{10}^{2}-120 s_{01}^{4} s_{10}^{2}\right\} \\
k_{33}=\frac{1}{n^{[6]}}\{ & \ell_{1} s_{33}-3 \ell_{2} s_{23} s_{10}-3 \ell_{2} s_{32} s_{01}-3 \ell_{3} s_{13} s_{20}-3 \ell_{3} s_{31} s_{02}-9 \ell_{3} s_{22} s_{11} \\
& -9 \ell_{4} s_{21} s_{12}-\ell_{4} s_{30} s_{03}+6 \ell_{5} s_{13} s_{10}^{2}+18 \ell_{5} s_{22} s_{10} s_{01}+6 \ell_{5} s_{31} s_{01}^{2} \\
& +6 \ell_{6} s_{20} s_{10} s_{03}+36 \ell_{6} s_{12} s_{11} s_{10}+18 \ell_{6} s_{12} s_{20} s_{01}+36 \ell_{6} s_{21} s_{11} s_{01} \\
& +18 \ell_{6} s_{21} s_{10} s_{02}+6 \ell_{6} s_{30} s_{01} s_{02}+6 \ell_{7} s_{11}^{3}+12 \ell_{7} s_{20} s_{11} s_{02} \\
& -6 \ell_{8} s_{10}^{3} s_{03}-54 \ell_{8} s_{12} s_{10}^{2} s_{01}-54 \ell_{8} s_{21} s_{10} s_{01}^{2}-6 \ell_{8} s_{30} s_{01}^{3} \\
& -54 \ell_{9} s_{11} s_{10}^{2} s_{02}-108 \ell_{9} s_{11}^{2} s_{10} s_{01}-54 \ell_{9} s_{20} s_{10} s_{01} s_{02} \\
& -54 \ell_{9} s_{20} s_{11} s_{01}^{2}+72 \ell_{10} s_{10}^{3} s_{02} s_{01}+72 \ell_{10} s_{20} s_{10} s_{01}^{3} \\
& \left.+216 \ell_{10} s_{10}^{2} s_{01}^{2} s_{11}-120 s_{10}^{3} s_{01}^{3}\right\}
\end{aligned}
$$

where

$$
\begin{array}{ll}
n^{[1]}=n & n^{[2]}=n(n-1) \\
n^{[3]}=n(n-1)(n-2) & n^{[4]}=n(n-1)(n-2)(n-3) \\
n^{[5]}=n(n-1)(n-2)(n-3)(n-4) & \\
n^{[6]}=n(n-1)(n-2)(n-3)(n-4)(n-5) &
\end{array}
$$

and

## A NEW TEST FOR CORRELATION

$$
\begin{array}{ll}
\ell_{1}=n^{5}+16 n^{4}+11 n^{3}-4 n^{2} & \ell_{2}=n^{4}+16 n^{3}+11 n^{2}-4 n \\
\ell_{3}=n^{4}+2 n^{3}-7 n^{2}+4 n & \ell_{4}=n^{4}-2 n^{3}+5 n^{2}-4 n \\
\ell_{5}=n^{3}+9 n^{2}+2 n & \ell_{6}=n^{3}-n \\
\ell_{7}=n^{3}-3 n^{2}+2 n & \ell_{8}=n^{2}+3 n \\
\ell_{9}=n^{2}-n & \ell_{10}=n
\end{array}
$$

## WANG \& SA

## Appendix C: Tables

Table 1. Comparison of type I error rates ( $\rho_{0}=0, n=30$ )

|  |  |  | $\alpha=0.10$ |  |  | $\alpha=0.05$ |  |  | $\alpha=0.01$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\boldsymbol{r}$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | $z_{\alpha}$ | 0.0985 | 0.1171 | 0.1124 | 0.0491 | 0.0679 | 0.0632 | 0.0102 | 0.0262 | 0.0191 |
| 0.0 | 0 | $t_{a}$ |  | 0.1124 | 0.1079 |  | 0.0628 | 0.0580 |  | 0.0227 | 0.0153 |
| 0.0 | 1 | $z_{\alpha}$ | 0.0987 | 0.1268 | 0.1153 | 0.0510 | 0.0755 | 0.0640 | 0.0101 | 0.0306 | 0.0173 |
| 0.0 | 1 | $t_{\alpha}$ |  | 0.1213 | 0.1095 |  | 0.0697 | 0.0579 |  | 0.0273 | 0.0137 |
| 0.0 | 10 | $z_{\alpha}$ | 0.1034 | 0.1512 | 0.1061 | 0.0549 | 0.0958 | 0.0476 | 0.0144 | 0.0588 | 0.0083 |
| 0.0 | 10 | $t_{\alpha}$ |  | 0.1453 | 0.0998 |  | 0.0900 | 0.0413 |  | 0.0570 | 0.0057 |
| 0.5 | 0 | $z a$ | 0.0988 | 0.1118 | 0.1051 | 0.0508 | 0.0637 | 0.0569 | 0.0110 | 0.0232 | 0.0163 |
| 0.5 | 0 | $t_{a}$ |  | 0.1064 | 0.1021 |  | 0.0578 | 0.0541 |  | 0.0195 | 0.0140 |
| 1.0 | 0 | $z_{\alpha}$ | 0.1015 | 0.1014 | 0.0954 | 0.0559 | 0.0541 | 0.0499 | 0.0131 | 0.0164 | 0.0126 |
| 1.0 | 0 | $t_{a}$ |  | 0.0959 | 0.0901 |  | 0.0488 | 0.0446 |  | 0.0138 | 0.0098 |
| 0.5 | 1 | $z a$ | 0.0999 | 0.1225 | 0.1117 | 0.0516 | 0.0709 | 0.0604 | 0.0111 | 0.0275 | 0.0162 |
| 0.5 | 1 | $t_{a}$ |  | 0.1168 | 0.1063 |  | 0.0649 | 0.0548 |  | 0.0242 | 0.0128 |
| 1.2 | 4 | $z_{\alpha}$ | 0.1055 | 0.1265 | 0.1015 | 0.0562 | 0.0734 | 0.0495 | 0.0143 | 0.0330 | 0.0104 |
| 1.2 | 4 | $t_{\alpha}$ |  | 0.1209 | 0.0958 |  | 0.0674 | 0.0435 |  | 0.0304 | 0.0075 |
| 1.2 | 10 | $z a$ | 0.1048 | 0.1433 | 0.1022 | 0.0580 | 0.0927 | 0.0469 | 0.0156 | 0.0539 | 0.0078 |
| 1.2 | 10 | $t_{\alpha}$ |  | 0.1380 | 0.0963 |  | 0.0871 | 0.0412 |  | 0.0522 | 0.0056 |

4
5
Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

## A NEW TEST FOR CORRELATION

1
2

|  |  |  | $\alpha=0.10$ |  |  | $\alpha=0.05$ |  |  | $\alpha=0.01$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\boldsymbol{r}$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 1.2 | 10 | $z_{\alpha}$ | 0.1048 | 0.1433 | 0.1022 | 0.0580 | 0.0927 | 0.0469 | 0.0156 | 0.0539 | 0.0078 |
| 1.2 | 10 | $t_{a}$ |  | 0.1380 | 0.0963 |  | 0.0871 | 0.0412 |  | 0.0522 | 0.0056 |
| 1.2 | 25 | $z_{\alpha}$ | 0.1026 | 0.1508 | 0.0913 | 0.0593 | 0.1022 | 0.0361 | 0.0214 | 0.0745 | 0.0046 |
| 1.2 | 25 | $t_{\alpha}$ |  | 0.1456 | 0.0852 |  | 0.0975 | 0.0304 |  | 0.0734 | 0.0031 |
| -1.2 | 4 | $z_{\alpha}$ | 0.1050 | 0.1262 | 0.1029 | 0.0561 | 0.0720 | 0.0482 | 0.0146 | 0.0322 | 0.0103 |
| -1.2 | 4 | $t_{a}$ |  | 0.1208 | 0.0972 |  | 0.0663 | 0.0429 |  | 0.0297 | 0.0076 |
| -1.2 | 10 | $z a$ | 0.1051 | 0.1446 | 0.1021 | 0.0563 | 0.0915 | 0.0456 | 0.0159 | 0.0543 | 0.0075 |
| -1.2 | 10 | $t_{a}$ |  | 0.1386 | 0.0960 |  | 0.0853 | 0.0393 |  | 0.0525 | 0.0054 |
| -1.2 | 25 | $z a$ | 0.1024 | 0.1507 | 0.0920 | 0.0589 | 0.1030 | 0.0358 | 0.0205 | 0.0739 | 0.0047 |
| -1.2 | 25 | $t_{a}$ |  | 0.1450 | 0.0855 |  | 0.0978 | 0.0303 |  | 0.0730 | 0.0033 |
| 1.2 | 4 | $z a$ | 0.1057 | 0.1375 | 0.1043 | 0.0579 | 0.0860 | 0.0490 | 0.0145 | 0.0436 | 0.0090 |
| 1.2 | 10 | $t a$ |  | 0.1315 | 0.0984 |  | 0.0797 | 0.0429 |  | 0.0416 | 0.0067 |
| 1.2 | 4 | $z a$ | 0.1064 | 0.1498 | 0.1012 | 0.0562 | 0.0947 | 0.0429 | 0.0155 | 0.0599 | 0.0065 |
| 1.2 | 25 | $t a$ |  | 0.1437 | 0.0950 |  | 0.0888 | 0.0373 |  | 0.0583 | 0.0045 |
| -1.2 | 4 | $z a$ | 0.1047 | 0.1374 | 0.1032 | 0.0563 | 0.0841 | 0.0475 | 0.0143 | 0.0446 | 0.0087 |
| -1.2 | 10 | $t a$ |  | 0.1313 | 0.0970 |  | 0.0781 | 0.0418 |  | 0.0425 | 0.0063 |
| -1.2 | 4 | $z \alpha$ | 0.1052 | 0.1477 | 0.1004 | 0.0575 | 0.0968 | 0.0437 | 0.0162 | 0.0590 | 0.0066 |
| -1.2 | 25 | $t_{\alpha}$ |  | 0.1421 | 0.0940 |  | 0.0908 | 0.0377 |  | 0.0576 | 0.0050 |

3
4
5
Table 1 (continued). Comparison of type I error rates ( $\rho_{0}=0, n=30$ )

Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

## WANG \& SA

1 Table 1 (continued). Comparison of type I error rates ( $\rho_{0}=0, n=30$ )


## A NEW TEST FOR CORRELATION

1
2

| $\beta$ | $r$ | CP | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | $z_{a}$ | 0.0488 | 0.0335 | 0.0155 | 0.0497 | 0.0183 | 0.0053 | 0.0499 | 0.0081 | 0.0007 | 0.0496 | 0.0009 | 0.0000 |
| 0.0 | 0 | $t_{a}$ |  | 0.0320 | 0.0130 |  | 0.0189 | 0.0047 |  | 0.0085 | 0.0006 |  | 0.0009 | 0.0000 |
| 0.0 | 1 | $z_{\alpha}$ | 0.0517 | 0.0537 | 0.0221 | 0.0525 | 0.0362 | 0.0092 | 0.0529 | 0.0218 | 0.0025 | 0.0543 | 0.0044 | 0.0002 |
| 0.0 | 1 | $t_{a}$ |  | 0.0522 | 0.0186 |  | 0.0373 | 0.0083 |  | 0.0226 | 0.0023 |  | 0.0044 | 0.0001 |
| 0.0 | 10 | $z_{\alpha}$ | 0.0917 | 0.1477 | 0.0502 | 0.1025 | 0.1265 | 0.0302 | 0.1180 | 0.1037 | 0.0153 | 0.1346 | 0.0474 | 0.0045 |
| 0.0 | 10 | $t_{a}$ |  | 0.1476 | 0.0441 |  | 0.1293 | 0.0277 |  | 0.1063 | 0.0143 |  | 0.0478 | 0.0043 |
| 0.5 | 0 | $z_{\alpha}$ | 0.0555 | 0.0344 | 0.0158 | 0.0550 | 0.0202 | 0.0058 | 0.0567 | 0.0106 | 0.0011 | 0.0590 | 0.0014 | 0.0001 |
| 0.5 | 0 | $t_{a}$ |  | 0.0330 | 0.0133 |  | 0.0205 | 0.0052 |  | 0.0111 | 0.0010 |  | 0.0014 | 0.0001 |
| 1.0 | 0 | $z_{\alpha}$ | 0.0489 | 0.0269 | 0.0121 | 0.0449 | 0.0156 | 0.0037 | 0.0308 | 0.0128 | 0.0010 | 0.0078 | 0.0072 | 0.0000 |
| 1.0 | 0 | $t_{a}$ |  | 0.0256 | 0.0096 |  | 0.0162 | 0.0033 |  | 0.0135 | 0.0009 |  | 0.0073 | 0.0000 |
| 0.5 | 1 | $z_{\alpha}$ | 0.0558 | 0.0548 | 0.0225 | 0.0558 | 0.0366 | 0.0087 | 0.0571 | 0.0230 | 0.0028 | 0.0580 | 0.0049 | 0.0002 |
| 0.5 | 1 | $t_{a}$ |  | 0.0534 | 0.0194 |  | 0.0375 | 0.0079 |  | 0.0239 | 0.0026 |  | 0.0050 | 0.0002 |
| 1.2 | 4 | $z_{a}$ | 0.0760 | 0.0998 | 0.0386 | 0.0809 | 0.0776 | 0.0198 | 0.0875 | 0.0597 | 0.0083 | 0.0931 | 0.0223 | 0.0016 |
| 1.2 | 4 | $t_{a}$ |  | 0.0982 | 0.0337 |  | 0.0790 | 0.0179 |  | 0.0612 | 0.0076 |  | 0.0224 | 0.0015 |
| 1.2 | 10 | $z_{\alpha}$ | 0.0973 | 0.1505 | 0.0529 | 0.1053 | 0.1264 | 0.0309 | 0.1218 | 0.1036 | 0.0162 | 0.1370 | 0.0490 | 0.0044 |
| 1.2 | 10 | $t_{a}$ |  | 0.1501 | 0.0470 |  | 0.1288 | 0.0282 |  | 0.1057 | 0.0153 |  | 0.0494 | 0.0043 |
| 1.2 | 25 | $z_{\alpha}$ | 0.1551 | 0.2064 | 0.0770 | 0.1735 | 0.1803 | 0.0533 | 0.2002 | 0.1566 | 0.0337 | 0.2268 | 0.0936 | 0.0137 |
| 1.2 | 25 | $t_{\alpha}$ |  | 0.2065 | 0.0697 |  | 0.1829 | 0.0492 |  | 0.1595 | 0.0321 |  | 0.0943 | 0.0132 |

4 Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

## WANG \& SA

1 2

|  |  |  | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\boldsymbol{r}$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| -1.2 | 4 | $z_{a}$ | 0.0762 | 0.0984 | 0.0376 | 0.0803 | 0.0777 | 0.0199 | 0.0864 | 0.0609 | 0.0087 | 0.0926 | 0.0228 | 0.0018 |
| -1.2 | 4 | $t_{\alpha}$ |  | 0.0968 | 0.0330 |  | 0.0790 | 0.0182 |  | 0.0623 | 0.0082 |  | 0.0229 | 0.0017 |
| -1.2 | 10 | $z_{\alpha}$ | 0.0955 | 0.1500 | 0.0515 | 0.1060 | 0.1265 | 0.0320 | 0.1224 | 0.1073 | 0.0170 | 0.1363 | 0.0485 | 0.0050 |
| -1.2 | 10 | $t_{\alpha}$ |  | 0.1501 | 0.0458 |  | 0.1286 | 0.0290 |  | 0.1097 | 0.0160 |  | 0.0488 | 0.0049 |
| -1.2 | 25 | $z_{a}$ | 0.1580 | 0.2076 | 0.0791 | 0.1731 | 0.1791 | 0.0529 | 0.1994 | 0.1570 | 0.0339 | 0.2240 | 0.0939 | 0.0144 |
| -1.2 | 25 | $t_{a}$ |  | 0.2068 | 0.0714 |  | 0.1816 | 0.0490 |  | 0.1601 | 0.0323 |  | 0.0947 | 0.0138 |
| 1.2 | 4 | $z_{\alpha}$ | 0.0837 | 0.1467 | 0.0511 | 0.0891 | 0.1318 | 0.0304 | 0.0864 | 0.0609 | 0.0087 | 0.1141 | 0.0664 | 0.0044 |
| 1.2 | 10 | $t_{\alpha}$ |  | 0.1470 | 0.0453 |  | 0.1348 | 0.0277 |  | 0.0623 | 0.0082 |  | 0.0667 | 0.0043 |
| 1.2 | 4 | $z_{\alpha}$ | 0.0996 | 0.2251 | 0.0668 | 0.1133 | 0.2153 | 0.0450 | 0.1224 | 0.1073 | 0.0170 | 0.2428 | 0.1272 | 0.0119 |
| 1.2 | 25 | $t_{a}$ |  | 0.2269 | 0.0600 |  | 0.2204 | 0.0418 |  | 0.1097 | 0.0160 |  | 0.1274 | 0.0113 |
| -1.2 | 4 | $z_{\alpha}$ | 0.0836 | 0.1474 | 0.0510 | 0.0884 | 0.1302 | 0.0305 | 0.1994 | 0.1570 | 0.0339 | 0.1120 | 0.0682 | 0.0047 |
| -1.2 | 10 | $t_{a}$ |  | 0.1472 | 0.0450 |  | 0.1329 | 0.0276 |  | 0.1601 | 0.0323 |  | 0.0685 | 0.0044 |
| -1.2 | 4 | $z_{\alpha}$ | 0.0993 | 0.2243 | 0.0685 | 0.1123 | 0.2181 | 0.0480 | 0.0998 | 0.1185 | 0.0150 | 0.2419 | 0.1270 | 0.0117 |
| -1.2 | 25 | $t_{\alpha}$ |  | 0.2263 | 0.0618 |  | 0.2228 | 0.0443 |  | 0.1213 | 0.0140 |  | 0.1273 | 0.0111 |
| 1.2 | 4 | $z_{\alpha}$ | 0.1037 | 0.1602 | 0.0628 | 0.1097 | 0.1332 | 0.0419 | 0.1360 | 0.2117 | 0.0283 | 0.1731 | 0.0189 | 0.0084 |
| 3.0 | 14 | $t_{a}$ |  | 0.1585 | 0.0560 |  | 0.1342 | 0.0384 |  | 0.2149 | 0.0267 |  | 0.0191 | 0.0082 |
| 1.2 | 4 | $z_{\alpha}$ | 0.1134 | 0.2081 | 0.0737 | 0.1220 | 0.1919 | 0.0512 | 0.0993 | 0.1181 | 0.0154 | 0.1694 | 0.0871 | 0.0163 |
| 3.0 | 25 | $t_{\alpha}$ |  | 0.2085 | 0.0669 |  | 0.1946 | 0.0475 |  | 0.1210 | 0.0144 |  | 0.0872 | 0.0155 |

3 Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point
5
Table 2 (continued). Comparison of type I error rates ( $n=30, \alpha=0.05$ )

## A NEW TEST FOR CORRELATION

1 Table 2 (continued). Comparison of type I error rates ( $n=30, \alpha=0.05$ )

| $\beta$ | $\boldsymbol{r}$ | CP | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 1.2 | 14 | $z_{a}$ | 0.1364 | 0.1994 | 0.0757 | 0.1533 | 0.1719 | 0.0518 | 0.1363 | 0.2096 | 0.0293 | 0.1960 | 0.0680 | 0.0142 |
| 3.0 | 25 | $t_{a}$ |  | 0.1983 | 0.0683 |  | 0.1739 | 0.0480 |  | 0.2133 | 0.0276 |  | 0.0683 | 0.0137 |
| -1.2 | 4 | $z_{\alpha}$ | 0.1016 | 0.1565 | 0.0623 | 0.1089 | 0.1298 | 0.0405 | 0.1151 | 0.0949 | 0.0212 | 0.1734 | 0.0194 | 0.0083 |
| -3.0 | 14 | $t_{\alpha}$ |  | 0.1550 | 0.0558 |  | 0.1304 | 0.0369 |  | 0.0964 | 0.0199 |  | 0.0196 | 0.0082 |
| -1.2 | 4 | $z_{a}$ | 0.1115 | 0.2082 | 0.0723 | 0.1239 | 0.1912 | 0.0521 | 0.1419 | 0.1719 | 0.0344 | 0.1682 | 0.0845 | 0.0155 |
| -3.0 | 25 | $t_{a}$ |  | 0.2082 | 0.0652 |  | 0.1942 | 0.0481 |  | 0.1742 | 0.0325 |  | 0.0847 | 0.0148 |
| -1.2 | 14 | $z_{\alpha}$ | 0.1365 | 0.1981 | 0.0733 | 0.1534 | 0.1733 | 0.0526 | 0.1759 | 0.1429 | 0.0344 | 0.1931 | 0.0673 | 0.0141 |
| -3.0 | 25 | $t_{a}$ |  | 0.1976 | 0.0659 |  | 0.1747 | 0.0484 |  | 0.1453 | 0.0327 |  | 0.0676 | 0.0136 |
| 3.0 | 25 | $z_{\alpha}$ | 0.1648 | 0.2193 | 0.0833 | 0.1852 | 0.1898 | 0.0588 | 0.1147 | 0.0943 | 0.0216 | 0.2290 | 0.1044 | 0.0156 |
| 3.0 | 25 | $t_{a}$ |  | 0.2183 | 0.0755 |  | 0.1913 | 0.0547 |  | 0.0958 | 0.0205 |  | 0.1049 | 0.0150 |
| -3.0 | 25 | $z_{a}$ | 0.1654 | 0.2186 | 0.0833 | 0.1820 | 0.1922 | 0.0587 | 0.1402 | 0.1696 | 0.0326 | 0.2257 | 0.1068 | 0.0159 |
| -3.0 | 25 | $t_{a}$ |  | 0.2172 | 0.0754 |  | 0.1938 | 0.0543 |  | 0.1723 | 0.0308 |  | 0.1074 | 0.0152 |

## WANG \& SA

1

| $\beta$ | $r$ | CP | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | $z_{\alpha}$ | 0.0107 | 0.0352 | 0.0034 | 0.0107 | 0.0272 | 0.0017 | 0.0103 | 0.0137 | 0.0003 | 0.0103 | 0.0013 | 0.0000 |
| 0.0 | 0 | $t_{\alpha}$ |  | 0.0375 | 0.0027 |  | 0.0297 | 0.0014 |  | 0.0145 | 0.0002 |  | 0.0013 | 0.0000 |
| 0.0 | 1 | $z_{\alpha}$ | 0.0115 | 0.0627 | 0.0056 | 0.0118 | 0.0526 | 0.0034 | 0.0118 | 0.0315 | 0.0008 | 0.0123 | 0.0049 | 0.0001 |
| 0.0 | 1 | $t_{\alpha}$ |  | 0.0671 | 0.0045 |  | 0.0562 | 0.0027 |  | 0.0330 | 0.0007 |  | 0.0050 | 0.0001 |
| 0.0 | 10 | $z_{\alpha}$ | 0.0350 | 0.1768 | 0.0160 | 0.0417 | 0.1615 | 0.0125 | 0.0512 | 0.1263 | 0.0072 | 0.0611 | 0.0501 | 0.0023 |
| 0.0 | 10 | $t_{\alpha}$ |  | 0.1858 | 0.0136 |  | 0.1694 | 0.0108 |  | 0.1295 | 0.0062 |  | 0.0502 | 0.0019 |
| 0.5 | 0 | $z_{\alpha}$ | 0.0125 | 0.0359 | 0.0036 | 0.0126 | 0.0285 | 0.0017 | 0.0135 | 0.0158 | 0.0004 | 0.0138 | 0.0017 | 0.0000 |
| 0.5 | 0 | $t_{\alpha}$ |  | 0.0386 | 0.0030 |  | 0.0308 | 0.0014 |  | 0.0167 | 0.0004 |  | 0.0017 | 0.0000 |
| 1.0 | 0 | $z_{a}$ | 0.0138 | 0.0274 | 0.0025 | 0.0124 | 0.0245 | 0.0012 | 0.0083 | 0.0201 | 0.0003 | 0.0013 | 0.0093 | 0.0000 |
| 1.0 | 0 | $t_{a}$ |  | 0.0293 | 0.0020 |  | 0.0265 | 0.0010 |  | 0.0215 | 0.0002 |  | 0.0097 | 0.0000 |
| 0.5 | 1 | $z_{a}$ | 0.0124 | 0.0612 | 0.0054 | 0.0132 | 0.0513 | 0.0032 | 0.0138 | 0.0310 | 0.0009 | 0.0145 | 0.0057 | 0.0001 |
| 0.5 | 1 | $t_{\alpha}$ |  | 0.0651 | 0.0044 |  | 0.0543 | 0.0026 |  | 0.0325 | 0.0008 |  | 0.0057 | 0.0001 |
| 1.2 | 4 | $z_{a}$ | 0.0246 | 0.1114 | 0.0114 | 0.0258 | 0.1011 | 0.0081 | 0.0296 | 0.0753 | 0.0042 | 0.0338 | 0.0236 | 0.0009 |
| 1.2 | 4 | $t_{a}$ |  | 0.1175 | 0.0095 |  | 0.1060 | 0.0067 |  | 0.0778 | 0.0037 |  | 0.0236 | 0.0008 |
| 1.2 | 10 | $z_{\alpha}$ | 0.0378 | 0.1754 | 0.0173 | 0.0432 | 0.1597 | 0.0126 | 0.0538 | 0.1259 | 0.0083 | 0.0643 | 0.0528 | 0.0028 |
| 1.2 | 10 | $t_{a}$ |  | 0.1834 | 0.0144 |  | 0.1662 | 0.0108 |  | 0.1292 | 0.0073 |  | 0.0530 | 0.0025 |
| 1.2 | 25 | $z_{\alpha}$ | 0.0833 | 0.2346 | 0.0292 | 0.0989 | 0.2168 | 0.0260 | 0.1169 | 0.1826 | 0.0187 | 0.1383 | 0.1010 | 0.0099 |
| 1.2 | 25 | $t_{\alpha}$ |  | 0.2430 | 0.0251 |  | 0.2238 | 0.0228 |  | 0.1866 | 0.0169 |  | 0.1014 | 0.0091 |

3
4
5

[^16]
## A NEW TEST FOR CORRELATION

1 2

|  |  |  | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\gamma$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| -1.2 | 4 | $z_{\alpha}$ | 0.0237 | 0.1116 | 0.0110 | 0.0264 | 0.1010 | 0.0075 | 0.0300 | 0.0749 | 0.0039 | 0.0336 | 0.0243 | 0.0008 |
| -1.2 | 4 | $t_{\alpha}$ |  | 0.1173 | 0.0092 |  | 0.1064 | 0.0064 |  | 0.0774 | 0.0034 |  | 0.0244 | 0.0006 |
| -1.2 | 10 | $z_{\alpha}$ | 0.0381 | 0.1780 | 0.0175 | 0.0439 | 0.1605 | 0.0131 | 0.0531 | 0.1241 | 0.0080 | 0.0628 | 0.0519 | 0.0028 |
| -1.2 | 10 | $t_{\alpha}$ |  | 0.1860 | 0.0146 |  | 0.1676 | 0.0113 |  | 0.1270 | 0.0071 |  | 0.0521 | 0.0025 |
| -1.2 | 25 | $z_{a}$ | 0.0833 | 0.2349 | 0.0295 | 0.0977 | 0.2169 | 0.0258 | 0.1206 | 0.1837 | 0.0196 | 0.1363 | 0.0997 | 0.0091 |
| -1.2 | 25 | $t_{a}$ |  | 0.2437 | 0.0254 |  | 0.2240 | 0.0226 |  | 0.1878 | 0.0178 |  | 0.1002 | 0.0085 |
| 1.2 | 4 | $z_{a}$ | 0.0286 | 0.1740 | 0.0165 | 0.0318 | 0.1676 | 0.0120 | 0.0364 | 0.1403 | 0.0069 | 0.0446 | 0.0689 | 0.0023 |
| 1.2 | 10 | $t_{a}$ |  | 0.1819 | 0.0136 |  | 0.1752 | 0.0101 |  | 0.1434 | 0.0061 |  | 0.0692 | 0.0020 |
| 1.2 | 4 | $z_{a}$ | 0.0400 | 0.2715 | 0.0247 | 0.0457 | 0.2683 | 0.0201 | 0.0567 | 0.2376 | 0.0145 | 0.0995 | 0.1277 | 0.0064 |
| 1.2 | 25 | $t_{\alpha}$ |  | 0.2829 | 0.0206 |  | 0.2774 | 0.0174 |  | 0.2407 | 0.0126 |  | 0.1278 | 0.0057 |
| -1.2 | 4 | $z_{a}$ | 0.0281 | 0.1730 | 0.0161 | 0.0319 | 0.1672 | 0.0125 | 0.0371 | 0.1417 | 0.0072 | 0.0446 | 0.0698 | 0.0023 |
| -1.2 | 10 | $t_{a}$ |  | 0.1813 | 0.0133 |  | 0.1753 | 0.0105 |  | 0.1448 | 0.0060 |  | 0.0699 | 0.0020 |
| -1.2 | 4 | $z_{a}$ | 0.0390 | 0.2699 | 0.0244 | 0.0452 | 0.2660 | 0.0199 | 0.0575 | 0.2371 | 0.0140 | 0.1003 | 0.1284 | 0.0067 |
| -1.2 | 25 | $t_{\alpha}$ |  | 0.2813 | 0.0203 |  | 0.2752 | 0.0172 |  | 0.2400 | 0.0123 |  | 0.1285 | 0.0059 |
| 1.2 | 4 | $z_{\alpha}$ | 0.0414 | 0.1707 | 0.0213 | 0.0436 | 0.1560 | 0.0166 | 0.0481 | 0.1080 | 0.0108 | 0.0818 | 0.0208 | 0.0059 |
| 3.0 | 14 | $t_{\alpha}$ |  | 0.1777 | 0.0177 |  | 0.1623 | 0.0142 |  | 0.1101 | 0.0094 |  | 0.0211 | 0.0054 |
| 1.2 | 4 | $z_{\alpha}$ | 0.0468 | 0.2369 | 0.0275 | 0.0512 | 0.2279 | 0.0234 | 0.0605 | 0.1915 | 0.0182 | 0.0621 | 0.0865 | 0.0089 |
| 3.0 | 25 | $t_{\alpha}$ |  | 0.2463 | 0.0231 |  | 0.2353 | 0.0204 |  | 0.1946 | 0.0161 |  | 0.0866 | 0.0078 |

4 Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

## WANG \& SA

1 Table 3 (continued). Comparison of type I error rates ( $n=30, \alpha=0.01$ )

| $\beta$ | $\boldsymbol{r}$ | CP | $\rho_{0}=0.50$ |  |  | $\rho_{0}=0.60$ |  |  | $\rho_{0}=0.75$ |  |  | $\rho_{0}=0.90$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 1.2 | 14 | $z_{a}$ | 0.0682 | 0.2232 | 0.0290 | 0.0783 | 0.2079 | 0.0248 | 0.0940 | 0.1625 | 0.0187 | 0.1061 | 0.0709 | 0.0090 |
| 3.0 | 25 | $t_{a}$ |  | 0.2311 | 0.0252 |  | 0.2145 | 0.0219 |  | 0.1657 | 0.0165 |  | 0.0712 | 0.0081 |
| -1.2 | 4 | $z_{\alpha}$ | 0.0425 | 0.1756 | 0.0223 | 0.0449 | 0.1566 | 0.0171 | 0.0479 | 0.1088 | 0.0107 | 0.0809 | 0.0205 | 0.0062 |
| -3.0 | 14 | $t_{\alpha}$ |  | 0.1826 | 0.0192 |  | 0.1626 | 0.0150 |  | 0.1108 | 0.0092 |  | 0.0207 | 0.0059 |
| -1.2 | 4 | $z_{a}$ | 0.0466 | 0.2365 | 0.0269 | 0.0531 | 0.2306 | 0.0238 | 0.0593 | 0.1913 | 0.0171 | 0.0607 | 0.0880 | 0.0084 |
| -3.0 | 25 | $t_{a}$ |  | 0.2454 | 0.0225 |  | 0.2375 | 0.0209 |  | 0.1943 | 0.0150 |  | 0.0882 | 0.0075 |
| -1.2 | 14 | $z_{a}$ | 0.0671 | 0.2213 | 0.0283 | 0.0779 | 0.2024 | 0.0248 | 0.0939 | 0.1612 | 0.0179 | 0.1071 | 0.0708 | 0.0093 |
| -3.0 | 25 | $t_{a}$ |  | 0.2291 | 0.0242 |  | 0.2098 | 0.0218 |  | 0.1642 | 0.0163 |  | 0.0711 | 0.0085 |
| 3.0 | 25 | $z_{\alpha}$ | 0.0934 | 0.2429 | 0.0345 | 0.1058 | 0.2221 | 0.0295 | 0.1249 | 0.1873 | 0.0217 | 0.1403 | 0.1080 | 0.0098 |
| 3.0 | 25 | $t_{a}$ |  | 0.2501 | 0.0300 |  | 0.2285 | 0.0262 |  | 0.1912 | 0.0197 |  | 0.1085 | 0.0088 |
| -3.0 | 25 | $z_{\alpha}$ | 0.0936 | 0.2375 | 0.0345 | 0.1060 | 0.2196 | 0.0290 | 0.1241 | 0.1872 | 0.0206 | 0.1406 | 0.1094 | 0.0098 |
| -3.0 | 25 | $t_{a}$ |  | 0.2446 | 0.0301 |  | 0.2262 | 0.0253 |  | 0.1907 | 0.0188 |  | 0.1100 | 0.0090 |

[^17]
## A NEW TEST FOR CORRELATION

1 2

| $\mathrm{ra}=0.0$ |  |  |  |  | $\mathrm{ra}=0.2$ |  |  | $\mathrm{ra}=0.4$ |  |  | $\mathrm{ra}=0.6$ |  |  | $\mathrm{ra}=0.8$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| b | g | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | 0.0498 | 0.0683 | 0.0635 | 0.2778 | 0.3148 | 0.3002 | 0.7231 | 0.7376 | 0.7156 | 0.9775 | 0.9785 | 0.9727 | 1.0000 | 1.0000 | 0.9993 |
| 0.0 | 0 |  | 0.0624 | 0.0576 |  | 0.2976 | 0.2827 |  | 0.7192 | 0.6950 |  | 0.9749 | 0.9674 |  | 1.0000 | 0.9991 |
| 0.0 | 1 | 0.0494 | 0.0739 | 0.0626 | 0.2814 | 0.3321 | 0.2964 | 0.7227 | 0.7489 | 0.6980 | 0.9776 | 0.9798 | 0.9546 | 1.0000 | 1.0000 | 0.9840 |
| 0.0 | 1 |  | 0.0681 | 0.0565 |  | 0.3133 | 0.2778 |  | 0.7314 | 0.6758 |  | 0.9768 | 0.9473 |  | 0.9999 | 0.9815 |
| 0.0 | 10 | 0.0553 | 0.0991 | 0.0485 | 0.3030 | 0.3849 | 0.2641 | 0.7436 | 0.7942 | 0.6416 | 0.9809 | 0.9837 | 0.8472 | 1.0000 | 0.9998 | 0.8527 |
| 0.0 | 10 |  | 0.0933 | 0.0427 |  | 0.3653 | 0.2419 |  | 0.7772 | 0.6128 |  | 0.9811 | 0.8330 |  | 0.9997 | 0.8428 |
| 0.5 | 0 | 0.0502 | 0.0625 | 0.0585 | 0.2821 | 0.3074 | 0.2927 | 0.7201 | 0.7334 | 0.7089 | 0.9766 | 0.9771 | 0.9680 | 1.0000 | 0.9999 | 0.9977 |
| 0.5 | 0 |  | 0.0564 | 0.0528 |  | 0.2888 | 0.2738 |  | 0.7147 | 0.6882 |  | 0.9738 | 0.9622 |  | 0.9998 | 0.9969 |
| 1.0 | 0 | 0.0553 | 0.0550 | 0.0511 | 0.2655 | 0.2531 | 0.2377 | 0.6519 | 0.6278 | 0.5973 | 0.9409 | 0.9341 | 0.9147 | 0.9988 | 0.9981 | 0.9957 |
| 1.0 | 0 |  | 0.0493 | 0.0457 |  | 0.2350 | 0.2193 |  | 0.6049 | 0.5726 |  | 0.9254 | 0.9022 |  | 0.9978 | 0.9948 |
| 0.5 | 1 | 0.0514 | 0.0704 | 0.0602 | 0.2832 | 0.3254 | 0.2907 | 0.7209 | 0.7451 | 0.6952 | 0.9763 | 0.9783 | 0.9497 | 1.0000 | 0.9999 | 0.9808 |
| 0.5 | 1 |  | 0.0650 | 0.0548 |  | 0.3063 | 0.2719 |  | 0.7263 | 0.6721 |  | 0.9753 | 0.9419 |  | 0.9999 | 0.9783 |
| 1.2 | 4 | 0.0570 | 0.0744 | 0.0510 | 0.2927 | 0.3284 | 0.2571 | 0.7218 | 0.7473 | 0.6451 | 0.9755 | 0.9765 | 0.8935 | 1.0000 | 0.9994 | 0.9213 |
| 1.2 | 4 |  | 0.0686 | 0.0452 |  | 0.3090 | 0.2368 |  | 0.7286 | 0.6192 |  | 0.9728 | 0.8809 |  | 0.9994 | 0.9146 |
| 1.2 | 10 | 0.0565 | 0.0919 | 0.0467 | 0.3046 | 0.3716 | 0.2568 | 0.7415 | 0.7841 | 0.6343 | 0.9798 | 0.9808 | 0.8440 | 1.0000 | 0.9997 | 0.8543 |
| 1.2 | 10 |  | 0.0863 | 0.0408 |  | 0.3518 | 0.2342 |  | 0.7672 | 0.6059 |  | 0.9780 | 0.8295 |  | 0.9996 | 0.8442 |
| 1.2 | 25 | 0.0597 | 0.1033 | 0.0359 | 0.3213 | 0.4029 | 0.2392 | 0.7433 | 0.7969 | 0.5898 | 0.9736 | 0.9780 | 0.7636 | 0.9999 | 0.9997 | 0.7636 |
| 1.2 | 25 |  | 0.0988 | 0.0311 |  | 0.3831 | 0.2147 |  | 0.7802 | 0.5592 |  | 0.9743 | 0.7470 |  | 0.9996 | 0.7519 |
| -1.2 | 4 | 0.0556 | 0.0725 | 0.0493 | 0.2916 | 0.3273 | 0.2562 | 0.7208 | 0.7472 | 0.6445 | 0.9755 | 0.9764 | 0.8931 | 1.0000 | 0.9995 | 0.9211 |
| -1.2 | 4 |  | 0.0666 | 0.0436 |  | 0.3078 | 0.2363 |  | 0.7282 | 0.6183 |  | 0.9729 | 0.8809 |  | 0.9994 | 0.9141 |

$3 \frac{1.2}{} \frac{1}{}$ Note: $\beta$ : skewness; $\gamma$ : kurtosis; the "Zf", " $Z \mathrm{~b}$ ", and " $Z \mathrm{c}$ " results are calculated using the critical points $Z_{\alpha}$ and $t_{\alpha}$ as the first and the second number

## WANG \& SA

1 2


3
4
3 Note: $\beta$ : skewness; $\gamma$ : kurtosis; the "Zf", "Kb", and "Kc" results are calculated using the critical points $z_{a}$ and $t_{a}$ as the first and the second number

## A NEW TEST FOR CORRELATION

2

| b | $\mathrm{ra}=0.0$ |  |  |  | $\mathrm{ra}=0.2$ |  |  | $\mathrm{ra}=0.4$ |  |  | $\mathrm{ra}=0.6$ |  |  | $\mathrm{ra}=0.8$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | g | Zf | zb | Zc | Zf | zb | zc | Zf | zb | Zc | Zf | zb | Zc | Zf | zb | zc |
| -1.2 | 4 | 0.0622 | 0.0855 | 0.0388 | 0.3186 | 0.3602 | 0.2388 | 0.7484 | 0.7755 | 0.6131 | 0.9806 | 0.9803 | 0.8225 | 1.0000 | 0.9997 | 0.8325 |
| -3.0 | 25 |  | 0.0802 | 0.0334 |  | 0.3403 | 0.2159 |  | 0.7568 | 0.5834 |  | 0.9772 | 0.8074 |  | 0.9996 | 0.8216 |
| -1.2 | 14 | 0.0618 | 0.0967 | 0.0378 | 0.3200 | 0.3855 | 0.2395 | 0.7459 | 0.7878 | 0.5980 | 0.9776 | 0.9785 | 0.7889 | 1.0000 | 0.9996 | 0.7923 |
| -3.0 | 25 |  | 0.0918 | 0.0323 |  | 0.3662 | 0.2161 |  | 0.7702 | 0.5681 |  | 0.9752 | 0.7738 |  | 0.9996 | 0.7813 |
| 3.0 | 25 | 0.0689 | 0.0935 | 0.0331 | 0.3289 | 0.3661 | 0.2129 | 0.7241 | 0.7498 | 0.5456 | 0.9625 | 0.9629 | 0.7458 | 0.9999 | 0.9991 | 0.7631 |
| 3.0 | 25 |  | 0.0889 | 0.0279 |  | 0.3489 | 0.1918 |  | 0.7311 | 0.5149 |  | 0.9574 | 0.7274 |  | 0.9990 | 0.7509 |
| -3.0 | 25 | 0.0682 | 0.0919 | 0.0334 | 0.3293 | 0.3686 | 0.2146 | 0.7257 | 0.7508 | 0.5431 | 0.9625 | 0.9629 | 0.7451 | 0.9998 | 0.9992 | 0.7676 |
| -3.0 | 25 |  | 0.0875 | 0.0286 |  | 0.3502 | 0.1926 |  | 0.7317 | 0.5121 |  | 0.9573 | 0.7276 |  | 0.9991 | 0.7556 |

Table 4 (continued). Power performance for test $\rho_{0}=0(n=30, \alpha=0.05)$

Note: $\beta$ : skewness; $\gamma$ : kurtosis; the " Zf ", "Zb", and " Zc " results are calculated using the critical points $z_{\alpha}$ and $t_{\alpha}$ as the first and the second number

## WANG \& SA

1 2

| b | g | $\mathrm{ra}=0.0$ |  |  | $\mathrm{ra}=0.2$ |  |  | $\mathrm{ra}=0.4$ |  |  | $\mathrm{ra}=0.6$ |  |  | $\mathrm{ra}=0.8$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | 0.0103 | 0.0271 | 0.0202 | 0.1037 | 0.1508 | 0.1316 | 0.4585 | 0.4838 | 0.4451 | 0.9084 | 0.8966 | 0.8489 | 0.9996 | 0.9995 | 0.9871 |
| 0.0 | 0 |  | 0.0235 | 0.0161 |  | 0.1299 | 0.1095 |  | 0.4305 | 0.3908 |  | 0.8640 | 0.8044 |  | 0.9991 | 0.9797 |
| 0.0 | 1 | 0.0102 | 0.0319 | 0.0180 | 0.1046 | 0.1572 | 0.1185 | 0.4646 | 0.5000 | 0.4061 | 0.9090 | 0.9038 | 0.7928 | 0.9996 | 0.9995 | 0.9313 |
| 0.0 | 1 |  | 0.0283 | 0.0137 |  | 0.1348 | 0.0962 |  | 0.4451 | 0.3492 |  | 0.8728 | 0.7404 |  | 0.9990 | 0.9136 |
| 0.0 | 10 | 0.0147 | 0.0605 | 0.0087 | 0.1247 | 0.2016 | 0.0717 | 0.4893 | 0.5508 | 0.2901 | 0.9128 | 0.9203 | 0.6149 | 0.9997 | 0.9988 | 0.7141 |
| 0.0 | 10 |  | 0.0586 | 0.0061 |  | 0.1801 | 0.0525 |  | 0.4962 | 0.2281 |  | 0.8931 | 0.5508 |  | 0.9983 | 0.6823 |
| 0.5 | 0 | 0.0114 | 0.0226 | 0.0180 | 0.1074 | 0.1410 | 0.1245 | 0.4594 | 0.4744 | 0.4353 | 0.9026 | 0.8925 | 0.8371 | 0.9996 | 0.9992 | 0.9753 |
| 0.5 | 0 |  | 0.0192 | 0.0140 |  | 0.1208 | 0.1030 |  | 0.4210 | 0.3814 |  | 0.8597 | 0.7936 |  | 0.9984 | 0.9647 |
| 1.0 | 0 | 0.0129 | 0.0169 | 0.0130 | 0.1051 | 0.0961 | 0.0850 | 0.4051 | 0.3352 | 0.3022 | 0.8197 | 0.7457 | 0.6739 | 0.9929 | 0.9834 | 0.9398 |
| 1.0 | 0 |  | 0.0144 | 0.0101 |  | 0.0794 | 0.0682 |  | 0.2824 | 0.2524 |  | 0.6781 | 0.6026 |  | 0.9732 | 0.9098 |
| 0.5 | 1 | 0.0115 | 0.0286 | 0.0172 | 0.1047 | 0.1482 | 0.1134 | 0.4621 | 0.4886 | 0.4020 | 0.9055 | 0.8997 | 0.7877 | 0.9995 | 0.9990 | 0.9279 |
| 0.5 | 1 |  | 0.0249 | 0.0130 |  | 0.1256 | 0.0913 |  | 0.4331 | 0.3453 |  | 0.8665 | 0.7364 |  | 0.9986 | 0.9107 |
| 1.2 | 4 | 0.0148 | 0.0338 | 0.0110 | 0.1203 | 0.1512 | 0.0831 | 0.4721 | 0.4806 | 0.3217 | 0.9012 | 0.8896 | 0.6826 | 0.9996 | 0.9974 | 0.8191 |
| 1.2 | 4 |  | 0.0315 | 0.0082 |  | 0.1301 | 0.0637 |  | 0.4247 | 0.2662 |  | 0.8544 | 0.6223 |  | 0.9961 | 0.7932 |
| 1.2 | 10 | 0.0163 | 0.0544 | 0.0076 | 0.1310 | 0.1918 | 0.0676 | 0.4912 | 0.5330 | 0.2835 | 0.9086 | 0.9087 | 0.6116 | 0.9997 | 0.9980 | 0.7199 |
| 1.2 | 10 |  | 0.0527 | 0.0053 |  | 0.1706 | 0.0496 |  | 0.4801 | 0.2235 |  | 0.8782 | 0.5471 |  | 0.9972 | 0.6891 |
| 1.2 | 25 | 0.0215 | 0.0744 | 0.0047 | 0.1591 | 0.2325 | 0.0478 | 0.5077 | 0.5639 | 0.2222 | 0.8869 | 0.8971 | 0.5045 | 0.9989 | 0.9977 | 0.6111 |
| 1.2 | 25 |  | 0.0733 | 0.0032 |  | 0.2128 | 0.0316 |  | 0.5155 | 0.1640 |  | 0.8670 | 0.4369 |  | 0.9965 | 0.5767 |
| -1.2 | 4 | 0.0145 | 0.0328 | 0.0104 | 0.1208 | 0.1503 | 0.0818 | 0.4723 | 0.4783 | 0.3207 | 0.9022 | 0.8890 | 0.6833 | 0.9996 | 0.9973 | 0.8200 |
| -1.2 | 4 |  | 0.0303 | 0.0076 |  | 0.1300 | 0.0630 |  | 0.4228 | 0.2649 |  | 0.8541 | 0.6231 |  | 0.9960 | 0.7939 |

3 Note: $\beta$ : skewness; $\gamma$ : kurtosis; the "Zf", "Zb", and "Zc" results are calculated using the critical points $Z_{\alpha}$ and $t_{\alpha}$ as the first and the second number

## A NEW TEST FOR CORRELATION

1 2


3
4
3 Note: $\beta$ : skewness; $\gamma$ : kurtosis; the " Zf ", " $Z \mathrm{~b}$ ", and " $Z \mathrm{c}$ " results are calculated using the critical points $z_{\alpha}$ and $t_{\alpha}$ as the first and the second number

## WANG \& SA

1
2 3

4
5
4
5
6

Table 5 (continued). Power performance for test $\rho_{0}=0.00(n=30, \alpha=0.01)$

|  |  | $\mathrm{ra}=0.0$ |  |  | $\mathrm{ra}=0.2$ |  |  | $\mathrm{ra}=0.4$ |  |  | $\mathrm{ra}=0.6$ |  |  | $\mathrm{ra}=0.8$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| b | g | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| -1.2 | 4 | 0.0192 | 0.0501 | 0.0055 | 0.1434 | 0.1806 | 0.0550 | 0.5090 | 0.5181 | 0.2614 | 0.9142 | 0.9036 | 0.5878 | 0.9999 | 0.9983 | 0.6962 |
| -3.0 | 25 |  | 0.0483 | 0.0036 |  | 0.1600 | 0.0385 |  | 0.4644 | 0.2026 |  | 0.8726 | 0.5256 |  | 0.9976 | 0.6636 |
| -1.2 | 14 | 0.0216 | 0.0659 | 0.0053 | 0.1499 | 0.2138 | 0.0515 | 0.5060 | 0.5448 | 0.2406 | 0.9001 | 0.9015 | 0.5416 | 0.9994 | 0.9978 | 0.6478 |
| -3.0 | 25 |  | 0.0648 | 0.0035 |  | 0.1941 | 0.0358 |  | 0.4956 | 0.1838 |  | 0.8722 | 0.4775 |  | 0.9968 | 0.6163 |
| 3.0 | 25 | 0.0277 | 0.0645 | 0.0040 | 0.1702 | 0.2099 | 0.0431 | 0.5080 | 0.5167 | 0.2028 | 0.8724 | 0.8581 | 0.4780 | 0.9978 | 0.9945 | 0.6159 |
| 3.0 | 25 |  | 0.0632 | 0.0025 |  | 0.1929 | 0.0293 |  | 0.4710 | 0.1516 |  | 0.8239 | 0.4133 |  | 0.9922 | 0.5829 |
| -3.0 | 25 | 0.0282 | 0.0646 | 0.0044 | 0.1719 | 0.2110 | 0.0424 | 0.5079 | 0.5176 | 0.2031 | 0.8715 | 0.8589 | 0.4801 | 0.9978 | 0.9946 | 0.6142 |
| -3.0 | 25 |  | 0.0633 | 0.0028 |  | 0.1938 | 0.0280 |  | 0.4734 | 0.1533 |  | 0.8239 | 0.4151 |  | 0.9923 | 0.5815 |

Note: $\beta$ : skewness; $\gamma$ : kurtosis; the " $Z \mathrm{f}$ ", " $Z \mathrm{~b}$ ", and " $Z \mathrm{c}$ " results are calculated using the critical points $z_{\alpha}$ and $t_{\alpha}$ as the first and the second number

## A NEW TEST FOR CORRELATION

1
2

|  |  |  | $\rho_{\alpha}=0.55$ |  |  | $\rho_{\alpha}=0.60$ |  |  | $\rho_{\alpha}=0.70$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\boldsymbol{r}$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 0.0 | 0 | $z_{\alpha}$ | 0.0991 | 0.1275 | 0.1139 | 0.3005 | 0.3038 | 0.2861 | 0.6538 | 0.4908 | 0.4779 |
| 0.0 | 0 | $t_{\alpha}$ |  | 0.1115 | 0.0978 |  | 0.2632 | 0.2452 |  | 0.4201 | 0.4073 |
| 0.0 | 1 | $z_{\alpha}$ | 0.1017 | 0.1352 | 0.1142 | 0.2990 | 0.3064 | 0.2795 | 0.6526 | 0.4850 | 0.4678 |
| 0.0 | 1 | $t_{\alpha}$ |  | 0.1219 | 0.0999 |  | 0.2729 | 0.2452 |  | 0.4226 | 0.4043 |
| 0.0 | 10 | $z_{\alpha}$ | 0.1483 | 0.1923 | 0.1358 | 0.3445 | 0.3497 | 0.2934 | 0.6604 | 0.4896 | 0.4611 |
| 0.0 | 10 | $t_{\alpha}$ |  | 0.1855 | 0.1253 |  | 0.3296 | 0.2690 |  | 0.4502 | 0.4190 |
| 0.5 | 0 | $z a$ | 0.1075 | 0.1312 | 0.1164 | 0.3058 | 0.3037 | 0.2853 | 0.6496 | 0.4919 | 0.4789 |
| 0.5 | 0 | $t_{\alpha}$ |  | 0.1153 | 0.1005 |  | 0.2662 | 0.2475 |  | 0.4245 | 0.4114 |
| 1.0 | 0 | $z a$ | 0.0914 | 0.1044 | 0.0918 | 0.2330 | 0.2350 | 0.2156 | 0.4929 | 0.4230 | 0.4030 |
| 1.0 | 0 | $t_{a}$ |  | 0.0938 | 0.0813 |  | 0.2095 | 0.1901 |  | 0.3756 | 0.3561 |
| 0.5 | 1 | $z_{\alpha}$ | 0.1046 | 0.1355 | 0.1142 | 0.3026 | 0.3062 | 0.2798 | 0.6517 | 0.4870 | 0.4698 |
| 0.5 | 1 | $t_{a}$ |  | 0.1220 | 0.0997 |  | 0.2717 | 0.2443 |  | 0.4259 | 0.4082 |
| 1.2 | 4 | $z \alpha$ | 0.1280 | 0.1656 | 0.1259 | 0.3246 | 0.3255 | 0.2843 | 0.6491 | 0.4795 | 0.4556 |
| 1.2 | 4 | $t_{a}$ |  | 0.1556 | 0.1143 |  | 0.3011 | 0.2573 |  | 0.4313 | 0.4057 |
| 1.2 | 10 | $z \alpha$ | 0.1508 | 0.1968 | 0.1387 | 0.3491 | 0.3548 | 0.2969 | 0.6560 | 0.4845 | 0.4546 |
| 1.2 | 10 | $t_{\alpha}$ |  | 0.1889 | 0.1274 |  | 0.3348 | 0.2727 |  | 0.4466 | 0.4144 |
| 1.2 | 25 | $z a$ | 0.2114 | 0.2388 | 0.1621 | 0.3977 | 0.3716 | 0.3024 | 0.6554 | 0.4797 | 0.4407 |
| 1.2 | 25 | $t_{\alpha}$ |  | 0.2341 | 0.1523 |  | 0.3571 | 0.2827 |  | 0.4537 | 0.4114 |

3
4
5
Table 6 Power performance for test $\rho_{0}=0.55(n=30, \alpha=0.10)$

Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

## WANG \& SA

1 2

|  |  |  | $\alpha=0.10$ |  |  | $\alpha=0.05$ |  |  | $\alpha=0.01$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\gamma$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| -1.2 | 4 | $z_{\alpha}$ | 0.1295 | 0.1656 | 0.1266 | 0.3213 | 0.3262 | 0.2836 | 0.6487 | 0.4781 | 0.4545 |
| -1.2 | 4 | $t_{a}$ |  | 0.1558 | 0.1148 |  | 0.2998 | 0.2550 |  | 0.4299 | 0.4046 |
| -1.2 | 10 | $z_{\alpha}$ | 0.1517 | 0.1972 | 0.1394 | 0.3468 | 0.3545 | 0.2967 | 0.6554 | 0.4886 | 0.4593 |
| -1.2 | 10 | $t_{a}$ |  | 0.1899 | 0.1284 |  | 0.3342 | 0.2726 |  | 0.4492 | 0.4176 |
| -1.2 | 25 | $z_{\alpha}$ | 0.2116 | 0.2375 | 0.1610 | 0.3954 | 0.3696 | 0.2989 | 0.6558 | 0.4755 | 0.4379 |
| -1.2 | 25 | $t_{a}$ |  | 0.2328 | 0.1509 |  | 0.3566 | 0.2803 |  | 0.4509 | 0.4097 |
| 1.2 | 4 | $z_{\alpha}$ | 0.1359 | 0.1958 | 0.1373 | 0.3357 | 0.3711 | 0.3103 | 0.6653 | 0.5225 | $0.4913$ |
| 1.2 | 10 | $t_{a}$ |  | 0.1872 | 0.1255 |  | 0.3491 | 0.2842 |  | 0.4801 | 0.4465 |
| 1.2 | 4 | $z_{\alpha}$ | 0.1600 | 0.2497 | 0.1594 | 0.3745 | 0.4387 | 0.3492 | 0.7149 | 0.5731 | 0.5348 |
| 1.2 | 25 | $t_{a}$ |  | 0.2448 | 0.1480 |  | 0.4216 | 0.3255 |  | 0.5386 | 0.4972 |
| -1.2 | 4 | $z_{\alpha}$ | 0.1358 | $0.1953$ | $0.1373$ | 0.3366 | $0.3704$ | $0.3095$ | 0.6645 | $0.5211$ | $0.4904$ |
| -1.2 | 10 | $t_{a}$ |  | 0.1879 | $0.1266$ |  | 0.3486 | 0.2838 |  | 0.4789 | 0.4458 |
| -1.2 | 4 | Z ${ }^{\prime}$ | 0.1570 | 0.2459 | 0.1557 | 0.3746 | 0.4401 | 0.3518 | 0.7164 | 0.5735 | 0.5343 |
| -1.2 | 25 | $t_{a}$ |  | 0.2412 | 0.1447 |  | 0.4239 | 0.3287 |  | 0.5383 | 0.4955 |
| $1.2$ | 4 | $z_{\alpha}$ | 0.1553 | 0.1993 | 0.1414 | 0.3418 | $0.3569$ | $0.3005$ | 0.6380 | 0.4908 | $0.4662$ |
| 3.0 | 14 | $t_{a}$ |  | 0.1937 | 0.1320 |  | 0.3395 | 0.2798 |  | 0.4554 | 0.4288 |
| 1.2 | 4 | Z ${ }^{\prime}$ | 0.1724 | 0.2435 | 0.1626 | 0.3790 | 0.4097 | 0.3344 | 0.6958 | 0.5346 | 0.5011 |
| 3.0 | 25 | $t_{\alpha}$ |  | 0.2378 | 0.1515 |  | 0.3927 | 0.3116 |  | 0.5009 | 0.4644 |

$3 \quad \frac{1}{2} \quad$ Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point
5
Table 6 (continued). Power performance for test $\rho_{0}=0.55(n=30, \alpha=0.10)$

## A NEW TEST FOR CORRELATION

1 Table 6 (continued). Power performance for test $\rho_{0}=0.55(n=30, \alpha=0.10)$

|  |  |  | $\alpha=0.10$ |  |  | $\alpha=0.05$ |  |  | $\alpha=0.01$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\beta$ | $\boldsymbol{r}$ | CP | Zf | Zb | Zc | Zf | Zb | Zc | Zf | Zb | Zc |
| 1.2 | 14 | $z_{\alpha}$ | 0.1942 | 0.2334 | 0.1578 | 0.3843 | 0.3659 | 0.3004 | 0.6628 | 0.4698 | 0.4376 |
| 3.0 | 25 | $t_{a}$ |  | 0.2278 | 0.1479 |  | 0.3510 | 0.2806 |  | 0.4415 | 0.4064 |
| -1.2 | 4 | $z_{\alpha}$ | 0.1586 | 0.2051 | 0.1459 | 0.3404 | 0.3566 | 0.3006 | 0.6378 | 0.4939 | 0.4685 |
| -3.0 | 14 | ta |  | 0.1990 | 0.1363 |  | 0.3395 | 0.2802 |  | 0.4584 | 0.4315 |
| -1.2 | 4 | $z_{\alpha}$ | 0.1700 | 0.2413 | 0.1599 | 0.3790 | 0.4073 | 0.3333 | 0.6958 | 0.5321 | 0.4995 |
| -3.0 | 25 | $t_{a}$ |  | 0.2365 | 0.1498 |  | 0.3901 | 0.3109 |  | 0.4983 | 0.4631 |
| -1.2 | 14 | Za | 0.1955 | 0.2326 | 0.1588 | 0.3865 | 0.3638 | 0.2994 | 0.6620 | 0.4616 | 0.4303 |
| -3.0 | 25 | $t_{\alpha}$ |  | 0.2270 | 0.1484 |  | 0.3495 | 0.2805 |  | 0.4329 | 0.3992 |
| 3.0 | 25 | $z a$ | 0.2214 | 0.2517 | 0.1699 | 0.3995 | 0.3760 | 0.3003 | 0.6430 | 0.4740 | 0.4292 |
| 3.0 | 25 | $t_{a}$ |  | 0.2473 | 0.1604 |  | 0.3630 | 0.2821 |  | 0.4489 | 0.4009 |
| -3.0 | 25 | $z a$ | 0.2210 | 0.2514 | 0.1685 | 0.3996 | 0.3776 | 0.3012 | 0.6411 | 0.4731 | 0.4284 |
| -3.0 | 25 | ta |  | 0.2464 | 0.1584 |  | 0.3644 | 0.2822 |  | 0.4480 | 0.4001 |

4 Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

# Comparison of Some Multivariate Nonparametric Tests in Profile Analysis to Repeated Measurements 

Mehrdad Vossoughi
Shiraz University of Medical
Sciences
Shiraz, Iran

## Shila Shahvali

Bureau of Statistics and Information, Fars Planning and Budget Organization
Shiraz, Iran

Erfan Sadeghi
Isfahan University of Medical
Sciences
Isfahan, Iran

Through Monte Carlo simulations, the performance of six multivariate nonparametric tests for testing the hypothesis of parallelism in profile analysis was studied. In conclusion, the tests based on ranks were as efficient as Hotelling's $T^{2}$ under multivariate normal distribution. For the heavy tailed distribution, the tests based on signs performed best.

Keywords: Monte Carlo simulation, multivariate, nonparametric, profile analysis, heavy tailed

## Introduction

Research in many areas of application frequently involves repeated measurements in which response from each experimental unit is measured repeatedly over different occasions such as time points. The linear mixed model to repeated measurements (Laird \& Ware, 1982; Ware, 1985) was developed to analyze incomplete and unbalanced data. However, the performance of this complex approach is highly sensitive to the choice of model for mean function and correlation structure for errors (Littell, Pendergast, \& Natarajan, 2000; Park, Park, \& Davis, 2001; Vossoughi, Ayatollahi, Towhidi, \& Ketabchi, 2012). Although several nonparametric methods have been developed for non-normal responses (Azzalini \& Bowman, 1991; Singer, Poleto, \& Rosa, 2004; Wernecke \& Kalb, 1999; Wernecke \& Kaufmann, 2000), model building and software implementation of these methods are extremely complicated.

Due to these difficulties, investigators are often interested in using the traditional approaches especially when the circumstances are controlled for

[^18]
## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

obtaining complete data. In this context, the profile analysis method using MANOVA tests makes no assumption regarding the correlation structure and trend of mean model and hence is widely used. Nevertheless, the MANOVA tests perform poorly when the distribution of errors much deviates from multivariate normal (Davis, 1980, 1982; Everitt, 1979; Olson, 1974; Um \& Randles, 1998).

Bhapkar (1984) and Sen (1984) discussed asymptotically distribution-free analogous of profile analysis. Multivariate extensions of Kruskal-Wallis and Brown-Mood median tests based on marginal ranks and signs were discussed in Puri and Sen (1971) but suffer from a lack of invariance with respect to affine transformations. Several authors provided detailed descriptions of affine invariant and non-invariant competitors based on spatial signs and ranks (Hettmansperger, Möttönen \& Oja, 1998; Hettmansperger \& Oja, 1994; Möttönen \& Oja, 1995; Oja, 1999; Oja \& Randles, 2004). The asymptotic efficiency of multivariate spatial sign and rank tests were studied by Möttönen, Oja, and Tienari (1997), Möttönen, Hettmansperger, Oja, and Tienari (1998), Nordhausen, Oja, and Tyler (2006) and Oja and Randles (2004). The theory and software implementation of affine invariant/non-invariant spatial sign and rank tests were well described by Oja (2010).

The aim of this study is to compare the performance of six nonparametric multivariate multi-sample tests with Hotelling's $T^{2}$ in profile analysis for repeated measurements. For this propose, Monte Carlo simulations based on broad spectrum of scenarios are used to study the empirical type I error rates and powers of the tests in testing the hypothesis of parallelism. Affine/non-affine invariant multivariate generalizations of multi-sample tests are compared based on spatial scores discussed in Oja (2010, Ch. 11) and multivariate generalization of multisample tests based on marginal scores discussed in Chapter 5 of Puri and Sen (1971).

Although the test of group main effect or hypothesis that the two groups are at the same level can also be assessed using multivariate multi-sample procedures, it was not included in the simulations for three priori reasons. First, rather than testing the general multivariate hypothesis $\boldsymbol{\mu}_{1}=\boldsymbol{\mu}_{2}=\ldots=\boldsymbol{\mu}_{k}$ to assess group main effect, summarizing the response vector of each subjects using its individual mean and then applying univariate tests is generally implemented in a parametric profile analysis (Davis, 2002; Rencher, 1995). Second, the performance of Hotelling's $T^{2}$ and its nonparametric counterparts were studied to test above general hypothesis (Möttönen et al., 1998; Nordhausen et al., 2006; Um \& Randles, 1998). Finally, group main effect has no direct interpretation in the presence of significant interaction and hence is not the primary hypothesis of interest in profile analysis.

## VOSSOUGHI ET AL.

Although the Monte Carlo comparison of methods for the analysis of repeated measurements has been an active area of research (Bhapkar \& Patterson, 1978; Marcucci, 1986; Mendoza, Toothaker, \& Nicewander, 1974; Park et al., 2001; Schwertman, Flynn, Stein, \& Schenk, 1985; Schwertman, Fridshal, \& Magrey, 1981), this study has been designed to examine some different aspects. First, the performances of recent nonparametric tests based on spatial signs and ranks considered here have not yet been studied in the area of profile analysis. Second, the effect of various correlation structures for errors has not included by most of the previous literature on this subject. Finally, the performance of the non-invariant tests under various transformation matrices widely used in the profile analysis are examined.

## Methodology

## Parametric profile analysis

The structure of profile analysis for the analysis of repeated measurements is now considered. Suppose that repeated measurements have been taken from $k$ groups of subjects at $p$ occasions. Let $\mathbf{y}_{i j}=\left(y_{i j 1}, \ldots, y_{i j p}\right)^{\mathrm{T}}$ represent the response vector from the $j^{\text {th }}$ subject in group $i$ for $j=1, \ldots, n_{k}, i=1, \ldots, k$. The profile analysis model is

$$
\begin{equation*}
\mathbf{y}_{i j}=\boldsymbol{\mu}_{i}+\varepsilon_{i j}, \tag{1}
\end{equation*}
$$

where the vector $\boldsymbol{\varepsilon}_{i j}=\left(\varepsilon_{i j 1}, \ldots, \varepsilon_{i j p}\right)^{T}$ is the vector of errors for the $j^{\text {th }}$ subject in group $I$ and $\boldsymbol{\mu}_{i}=\left(\mu_{i 1}, \ldots, \mu_{i p}\right)^{T}$ is the population mean vector for the $i^{\text {th }}$ group. Error vectors are assumed to be independent and normally distributed with mean vector $\mathbf{0}$ and common covariance matrix $\boldsymbol{\Sigma}$.

Arguably, in the presence of group $\times$ occasion interaction, the tests of main effects are confounded. Therefore, the primary aim in the profile analysis is to test the hypothesis of parallelism of $k$ group profiles. The test of the hypothesis can be constructed as

$$
\begin{equation*}
\mathrm{H}_{0}=\mathbf{C} \boldsymbol{\mu}_{1}=\ldots=\boldsymbol{C} \boldsymbol{\mu}_{k} \text { or } \boldsymbol{\mu}_{1}^{*}=\ldots=\boldsymbol{\mu}_{k}^{*}, \tag{2}
\end{equation*}
$$

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

where $\boldsymbol{\mu}_{h}^{*}$ is the mean of transformed observations, $\mathbf{y}_{i j}^{*}=\mathbf{C} \mathbf{y}_{i j}$. Here, $\mathbf{C}$ is a $p-1 \times p$ transformation matrix with rank $p-1$ satisfying $\mathbf{C}_{\mathbf{1}}=\mathbf{0}$, where $\mathbf{1}$ is the unit matrix. For instance, when $p=3$, three widely-used matrices are:

$$
\mathbf{C}_{1}: \text { Mean difference } \quad \mathbf{C}_{2}: \text { Adjacent difference } \quad \mathbf{C}_{3}: \text { Last-value difference }
$$

$$
\frac{1}{3}\left[\begin{array}{ccc}
2 & -1 & -1 \\
-1 & 2 & -1
\end{array}\right] \quad\left[\begin{array}{ccc}
-1 & 1 & 0 \\
0 & -1 & 1
\end{array}\right] \quad\left[\begin{array}{ccc}
1 & 0 & -1 \\
0 & 1 & -1
\end{array}\right]
$$

For example, the analogous hypothesis of parallelism for $k=2$ and the transformation matrix $\mathbf{C}_{2}$ is

$$
H_{0}:\left(\begin{array}{c}
\mu_{12}-\mu_{11}  \tag{3}\\
\mu_{13}-\mu_{12} \\
\vdots \\
\mu_{1 p}-\mu_{1, p-1}
\end{array}\right)=\left(\begin{array}{c}
\mu_{22}-\mu_{21} \\
\mu_{23}-\mu_{22} \\
\vdots \\
\mu_{2 p}-\mu_{2, p-1}
\end{array}\right)
$$

Then, one-way multivariate analysis of variance (MANOVA) test statistics such as Wilk's $\Lambda$ (if $k>2$ ) or Hotelling's $T^{2}$ (if $k=2$ ) can be used to assess the equality of mean vectors of transformed variables $\mathbf{y}_{i j}^{*}$ or equivalently hypothesis of parallelism. Similarly, nonparametric multivariate tests can be applied on the transformed observations to assess the equality of population locations when the underlying distribution deviates from normality.

## Nonparametric counterparts of MANOVA tests

A brief overview of six nonparametric multivariate multi-sample tests used for profile analysis in the Monte Carlo simulations are now considered. The focus is primarily on recent methods that are supplied in standard statistical software packages. Here, we assume the $p$-dimensional data vectors are generated independently using model

$$
\begin{equation*}
\mathbf{y}_{i j}=\boldsymbol{\theta}_{i}+\boldsymbol{\varepsilon}_{i j}, \tag{4}
\end{equation*}
$$

where $\boldsymbol{\theta}_{i}$ denotes the $p$-dimensional location vector for group $i$ which is not necessary the corresponding mean vector and $\boldsymbol{\varepsilon}_{i j}$ is the vector of errors from an

## VOSSOUGHI ET AL.

elliptical multivariate distribution with location vector $\mathbf{0}$ and scatter matrix $\boldsymbol{\Sigma}$. When measurements are not normally distributed, nonparametric multi-sample multivariate tests can be employed to test the hypothesis of no group $\times$ occasions interaction effect as

$$
\begin{equation*}
H_{0}=\boldsymbol{\theta}_{1}^{*}=\ldots=\boldsymbol{\theta}_{k}^{*}, \tag{5}
\end{equation*}
$$

where $\boldsymbol{\theta}_{i}^{*}$ indicates the location vector of transformed variables from group $i$.

## Tests based on spatial signs

The test statistic based on spatial signs for testing $H_{0}$ is

$$
\begin{equation*}
Q=\sum_{i=1}^{c}\left\{n_{i} \overline{\mathbf{U}}_{i}^{* \prime} \overline{\mathbf{U}}_{i}^{*}\right\} \tag{6}
\end{equation*}
$$

where $\bar{U}_{i}^{*}$ denotes the sample mean vector of spatial signs transformed using inner centering and outer standardization. Although the test is location invariant, it is not affine invariant; that is the condition $Q(\mathbf{A Y})=Q(\mathbf{Y})$ is not satisfied for every nonsingular matrix $\mathbf{A}$ with rank $p$.

The affine invariant test statistic is

$$
\begin{equation*}
Q=p \sum_{i=1}^{c}\left\{n_{i} \overline{\mathbf{U}}_{i}^{*^{\prime}} \overline{\mathbf{U}}_{i}^{*}\right\} \tag{7}
\end{equation*}
$$

where, here, $\overline{\mathbf{U}}_{i}^{*}$ is the sample mean vector of spatial signs transformed using inner centering and inner standardization.

The test statistics are multivariate generalizations of two- and severalsample Mood's median test and are asymptotically distributed as $\chi_{(c-1) p}^{2}$ when $H_{0}$ is true. The spatial sign tests are denoted by $S S$ and $S S I$ for the non invariant and invariant versions in the simulations, respectively. See Oja (2010) regarding the theory and software implementation of spatial sign and rank tests.

## Tests based on spatial ranks

The constructions of tests based on spatial ranks are essentially the same as the spatial sign cases, with the difference that $\overline{\mathbf{U}}_{i}^{*}$ 's are replaced by the corresponding

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

sample mean vector of transformed spatial ranks, $\overline{\mathbf{R}}_{i}^{*}$. Due to the fact that the spatial ranks are naturally centered, one needs only to standardize them using outer or inner approaches to construct non affine or affine invariant versions of test statistic. The test statistics using outer and inner standardization are in the form of

$$
\begin{equation*}
Q=\sum_{i=1}^{c}\left\{n_{i} \overline{\boldsymbol{R}}_{i}^{* \prime} \overline{\boldsymbol{R}}_{i}^{*}\right\} \tag{8}
\end{equation*}
$$

and

$$
\begin{equation*}
Q=p \sum_{i=1}^{c}\left\{n_{i} \overline{\boldsymbol{R}}_{i}^{* \prime} \overline{\boldsymbol{R}}_{i}^{*}\right\}, \tag{9}
\end{equation*}
$$

respectively. The asymptotic null distribution of both test statistics is $\chi_{(c-1) p}^{2}$. The non invariant and affine invariant spatial rank tests are denoted by $S R$ and SRI in the simulations, respectively.

## Tests based on marginal ranks and signs

The multivariate multi-sample rank sum test compares the difference between the sample average rank vector $\overline{\mathbf{r}}_{i}$ and the combined-data average rank vector $\overline{\mathbf{r}}$ as

$$
\begin{equation*}
L_{R}=\sum_{i=1}^{c} n_{i}\left(\overline{\mathbf{r}}_{i}-\overline{\mathbf{r}}\right)^{\prime} \mathbf{V}^{-1}\left(\overline{\mathbf{r}}_{i}-\overline{\mathbf{r}}\right) . \tag{10}
\end{equation*}
$$

The test reduces to the Kruskal-Wallis test when $p=1$ and to Wilcoxon-Mann-Whitney test when $p=1$ and $c=2$.

The multivariate multi-sample median test uses the corresponding average vectors based on sample signs (computed regarding combined-data median vector) to test the null hypothesis as

$$
\begin{equation*}
L_{S}=\sum_{i=1}^{c} n_{i}\left(\overline{\mathbf{s}}_{i}-\overline{\mathbf{s}}\right)^{\prime} \mathbf{V}^{-1}\left(\overline{\mathbf{s}}_{i}-\overline{\mathbf{s}}\right) . \tag{11}
\end{equation*}
$$

Write $\mathbf{V}$ to denote the sample covariance matrix of marginal ranks and signs in $L_{R}$ and $L_{S}$, respectively. The asymptotic null distribution of both statistics is

## VOSSOUGHI ET AL.

$\chi_{(c-1) p}^{2}$. The multivariate multi-sample location tests based on the vector of marginal signs and ranks were discussed in detail by Puri and Sen (1971).

The marginal sign and rank tests are denoted by $M S$ and $M R$, respectively, in the simulation.

## Simulation study

The structure of a Monte Carlo study used to investigate the performances of tests according to empirical type I error rates and powers is now discussed. The profile model (4) with two groups ( $k=2$ ), number of measurements $p=4,8$ and sample sizes $n=10,20$ and 30 for each of the two samples was considered. The performances of MANOVA test (here Hotelling's $T^{2}$ since $k=2$ ) and the six nonparametric counterparts in testing the hypothesis of parallelism were compared under various scenarios. In the simulations, Hotelling's $T^{2}$ test was denoted by $T^{2}$.

Consider three types of correlation structures for errors; compound symmetry (CS) with $\rho=0.2$, first-order autoregressive (AR1) with $\rho=0.5$, and an unstructured model (UN). The UN structure considered here was an arbitrary $p \times p$ correlation matrix producing a positive definite covariance matrix. Errors were generated from multivariate $t$ with 3 degrees of freedom (denoted by $t$ (3)) as a heavy-tailed distribution and multivariate normal distribution with mean vector $\mathbf{0}$ and variances 3 for above correlation structures. Therefore the two distributions had the same mean vector and covariance matrix and differ only by degrees of heaviness of their tails. The MANOVA tests have been shown to have low powers when the underlying distribution is heavy-tailed, in particular (see e.g. (Somorčík, 2006). The reason is that the sample mean vector and covariance matrix would not provide proper estimates of location and variation under the presence of outliers (see, e.g. Um \& Randles, 1998).

Throughout the simulations, $\boldsymbol{\theta}_{1}$ was considered to be a zero vector. To compute the empirical type I error rates, data were simulated under the hypothesis of parallelism, $H_{0}: \mathbf{C} \boldsymbol{\theta}_{1}=\mathbf{C} \boldsymbol{\theta}_{2}$, when $\boldsymbol{\theta}_{2}$ was also considered to be a zero vector. However, the hypothesis of interaction or $H_{1}: \mathbf{C} \boldsymbol{\theta}_{1} \neq \mathbf{C} \boldsymbol{\theta}_{2}$ was simulated when $\boldsymbol{\theta}_{2}=(0,1,1,0)^{T}$ and $(0,0,1,1,1,1,0,0)^{T}$ for $p=4$ and 8 , respectively; so that the empirical powers were computed. Also considered are the three transformation matrices $\mathbf{C}_{1}$ to $\mathbf{C}_{3}$ presented above to evaluate the robustness of non affine-invariant tests.

For each combination of above scenarios, 1000 replications were carried out and significance level was considered to be 0.05 . All simulations performed using

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

R 3.0.1 (R Development Core Team, 2013). In this respect, the multi-sample tests implemented using the R packages MNM, ICSNP, and Hotelling. Multivariate normal and $t$ data were generated using the R packages MASS and mvtnorm, respectively.

## Results

Displayed in Tables 1 and 2 are empirical type I error rates of the tests for errors generated from multivariate normal and multivariate $t(3)$ distributions, respectively. Each value is the proportion of 1000 replications for which the hypothesis of parallelism or null hypothesis was incorrectly rejected. In general, all tests preserved the nominal 5 percent level under all scenarios. However, for $p=8$ and smaller sample size $n=10$, the type I error rates of nonparametric tests were smaller than those of parametric one.

Displayed in Table 3 are empirical powers of the test for multivariate normal distribution. Each power value computed as the proportion of 1000 replications for which the hypothesis of parallelism was correctly rejected. In summary, among the tests, the affine invariant and non-invariant tests based on spatial ranks as well as test based on marginal ranks reached a power level fully close to that of Hotelling $T^{2}$ in which the differences were considerably negligible for all correlation structures. However, for the smaller sample size $n=10$ and larger number of replication $p=8$, the amount of difference somewhat increased. The test based on marginal signs performed unsatisfactorily; that is its powers were much lower than those of other test statistics for all correlation structures and transformation matrices. Interestingly, for all transformation matrices, the competitor based on spatial signs dominated the test based on marginal sign and was comparable to the best tests in the multivariate normal case. The empirical power trends of tests for multivariate normal distribution are visualized in Figure 1.

Shown in Table 4 are empirical powers of the test for data generated from multivariate $t(3)$ as a heavy tailed distribution. The results showed that the tests based on spatial signs and ranks and tests based on marginal ranks fully dominated Hotelling's $T^{2}$ for larger sample sizes $n=20$ and 30 and any given correlation structure. For a fixed sample size, the amount of superiority somewhat decreased as $p$ increased. In summary, the tests based on spatial signs yielded the greater values than the counterparts based on spatial and marginal ranks. Note that for a fixed $p$, the larger the size of sample, the greater the amount of difference in power levels. However, the performance of marginal sign test

## VOSSOUGHI ET AL.

Table 1. The empirical type I error rates of tests under multivariate normal distribution.

|  | Matrix C | Test | Correlation structure |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | CS |  |  | AR1 |  |  | UN |  |  |
|  |  |  | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| $\begin{aligned} & \forall \\ & \text { II } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 047 | 044 | 048 | 061 | 047 | 040 | 044 | 049 | 045 |
|  |  | SRI | 046 | 050 | 051 | 057 | 047 | 041 | 045 | 050 | 046 |
|  |  | SSI | 044 | 040 | 044 | 053 | 045 | 043 | 045 | 052 | 044 |
|  |  | SR | 045 | 048 | 050 | 056 | 046 | 041 | 047 | 049 | 043 |
|  |  | SS | 047 | 040 | 045 | 060 | 045 | 044 | 055 | 047 | 037 |
|  |  | MR | 048 | 045 | 049 | 048 | 037 | 038 | 054 | 053 | 039 |
|  |  | MS | 038 | 046 | 041 | 048 | 037 | 041 | 035 | 040 | 044 |
|  | $\mathrm{C}_{2}$ : | SR | 046 | 046 | 052 | 059 | 048 | 040 | 045 | 047 | 041 |
|  |  | SS | 049 | 047 | 045 | 053 | 045 | 041 | 052 | 049 | 036 |
|  |  | MR | 051 | 047 | 045 | 050 | 046 | 043 | 046 | 047 | 049 |
|  |  | M. ${ }^{\text {S }}$ | 040 | 041 | 044 | 055 | 040 | 038 | 042 | 050 | 049 |
|  | $\mathrm{C}_{3}$ : | SR | 047 | 046 | 050 | 054 | 050 | 042 | 044 | 053 | 047 |
|  |  | SS | 047 | 038 | 043 | 050 | 045 | 043 | 046 | 046 | 043 |
|  |  | MR | 042 | 040 | 047 | 057 | 048 | 038 | 043 | 040 | 046 |
|  |  | MS | 042 | 035 | 040 | 049 | 036 | 036 | 043 | 047 | 049 |
| $\begin{aligned} & \infty \\ & \text { ॥ } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 050 | 051 | 050 | 047 | 052 | 050 | 036 | 054 | 042 |
|  |  | SRI | 018 | 043 | 042 | 016 | 042 | 043 | 015 | 042 | 037 |
|  |  | SSI | 022 | 043 | 045 | 019 | 047 | 046 | 017 | 041 | 035 |
|  |  | SR | 023 | 045 | 045 | 016 | 040 | 043 | 015 | 043 | 040 |
|  |  | SS | 020 | 046 | 046 | 017 | 044 | 044 | 015 | 044 | 034 |
|  |  | MR | 015 | 038 | 048 | 017 | 033 | 045 | 015 | 046 | 043 |
|  |  | MS | 021 | 045 | 044 | 018 | 036 | 047 | 017 | 044 | 038 |
|  | $\mathrm{C}_{2}$ : | SR | 021 | 045 | 041 | 016 | 042 | 042 | 016 | 040 | 040 |
|  |  | SS | 023 | 046 | 045 | 015 | 045 | 040 | 016 | 040 | 042 |
|  |  | MR | 020 | 040 | 048 | 023 | 045 | 047 | 014 | 033 | 039 |
|  |  | MS | 023 | 042 | 038 | 016 | 031 | 036 | 011 | 038 | 036 |
|  | $\mathrm{C}_{3}$ : | SR | 023 | 046 | 048 | 017 | 045 | 045 | 019 | 037 | 038 |
|  |  | SS | 020 | 040 | 043 | 016 | 043 | 045 | 012 | 040 | 038 |
|  |  | MR | 020 | 040 | 042 | 018 | 032 | 042 | 015 | 041 | 041 |
|  |  | MS | 020 | 044 | 040 | 017 | 039 | 037 | 014 | 036 | 033 |

*Note: The entries within table correspond to empirical type I error rates multiplied by 1000.
was unsatisfactory since it was just as efficient as Hotelling $T^{2}$ for some specific choices of C. Surprisingly; even the permutation procedure provided no additional gain in efficiency for Hotelling's $T^{2}$ under the heavy tailed distribution and hence not reported here.

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

Table 2. The empirical type I error rates of tests under multivariate $t$ distribution.

|  | Matrix C | Test | Correlation structure |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | CS |  |  | AR1 |  |  | UN |  |  |
|  |  |  | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| $\begin{aligned} & + \\ & \text { II } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 040 | 036 | 043 | 034 | 045 | 047 | 047 | 046 | 038 |
|  |  | SRI | 051 | 049 | 050 | 048 | 056 | 047 | 057 | 054 | 046 |
|  |  | SSI | 047 | 054 | 051 | 044 | 055 | 047 | 046 | 051 | 061 |
|  |  | SR | 052 | 046 | 056 | 046 | 056 | 050 | 054 | 052 | 050 |
|  |  | SS | 051 | 049 | 060 | 048 | 052 | 050 | 047 | 053 | 054 |
|  |  | MR | 046 | 043 | 049 | 047 | 046 | 050 | 048 | 054 | 041 |
|  |  | MS | 050 | 049 | 049 | 046 | 057 | 057 | 046 | 051 | 039 |
|  | $\mathrm{C}_{2}$ : | SR | 048 | 048 | 051 | 047 | 055 | 048 | 053 | 053 | 051 |
|  |  | SS | 045 | 052 | 050 | 050 | 052 | 046 | 050 | 054 | 057 |
|  |  | MR | 048 | 048 | 055 | 043 | 047 | 050 | 052 | 056 | 052 |
|  |  | MS | 049 | 044 | 049 | 040 | 048 | 062 | 050 | 045 | 043 |
|  | $\mathrm{C}_{3}$ : | SR | 046 | 050 | 046 | 048 | 057 | 050 | 054 | 055 | 050 |
|  |  | SS | 048 | 058 | 044 | 053 | 051 | 042 | 044 | 054 | 062 |
|  |  | MR | 042 | 055 | 051 | 042 | 047 | 047 | 050 | 055 | 055 |
|  |  | MS | 043 | 047 | 046 | 049 | 049 | 041 | 047 | 055 | 059 |
| $\begin{aligned} & \infty \\ & \text { ॥ } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 044 | 033 | 033 | 037 | 034 | 051 | 034 | 028 | 044 |
|  |  | SRI | 022 | 036 | 029 | 013 | 031 | 051 | 021 | 031 | 041 |
|  |  | SSI | 019 | 036 | 038 | 019 | 040 | 044 | 031 | 034 | 046 |
|  |  | SR | 019 | 034 | 027 | 010 | 030 | 049 | 020 | 030 | 045 |
|  |  | SS | 015 | 031 | 039 | 019 | 043 | 041 | 018 | 035 | 049 |
|  |  | MR | 018 | 028 | 038 | 014 | 026 | 029 | 021 | 030 | 045 |
|  |  | MS | 020 | 038 | 038 | 015 | 035 | 038 | 013 | 029 | 036 |
|  | $\mathrm{C}_{2}$ : | SR | 018 | 035 | 027 | 011 | 032 | 053 | 019 | 029 | 042 |
|  |  | SS | 017 | 039 | 038 | 017 | 041 | 046 | 025 | 032 | 041 |
|  |  | MR | 021 | 037 | 032 | 019 | 030 | 044 | 014 | 033 | 047 |
|  |  | MS | 015 | 037 | 039 | 015 | 034 | 045 | 018 | 031 | 049 |
|  | $\mathrm{C}_{3}$ : | SR | 020 | 027 | 026 | 012 | 033 | 046 | 019 | 028 | 042 |
|  |  | SS | 014 | 030 | 040 | 012 | 039 | 044 | 020 | 033 | 047 |
|  |  | MR | 012 | 022 | 022 | 013 | 042 | 048 | 018 | 032 | 038 |
|  |  | MS | 020 | 037 | 038 | 012 | 038 | 043 | 018 | 026 | 044 |

*Note: The entries within table correspond to empirical type I error rates multiplied by 1000.

Although not reported in the tables, additional simulations demonstrated that the superiority of nonparametric tests was not attained until $n$ reached 15 . Figure 2 shows the empirical power trends of tests for the heavy tailed distribution.

## VOSSOUGHI ET AL.

Table 3. The empirical powers of tests under multivariate normal distribution.

|  | Matrix C | Test | Correlation structure |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | CS |  |  | AR1 |  |  | UN |  |  |
|  |  |  | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| $\begin{aligned} & \forall \\ & \text { II } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 178 | 323 | 487 | 226 | 470 | 669 | 183 | 366 | 585 |
|  |  | SRI | 172 | 325 | 478 | 220 | 450 | 659 | 181 | 361 | 562 |
|  |  | SSI | 166 | 283 | 409 | 216 | 400 | 585 | 176 | 323 | 510 |
|  |  | SR | 172 | 326 | 483 | 218 | 450 | 649 | 186 | 359 | 561 |
|  |  | SS | 157 | 283 | 416 | 198 | 402 | 582 | 168 | 313 | 490 |
|  |  | MR | 154 | 311 | 460 | 210 | 422 | 631 | 170 | 323 | 528 |
|  |  | MS | 113 | 174 | 288 | 131 | 282 | 402 | 121 | 187 | 294 |
|  | $\mathrm{C}_{2}$ : | SR | 169 | 332 | 481 | 218 | 455 | 658 | 183 | 367 | 567 |
|  |  | SS | 170 | 292 | 419 | 198 | 412 | 598 | 175 | 329 | 500 |
|  |  | MR | 153 | 308 | 471 | 210 | 439 | 650 | 171 | 357 | 541 |
|  |  | MS | 100 | 185 | 307 | 131 | 276 | 460 | 110 | 226 | 314 |
|  | С3: | SR | 170 | 326 | 473 | 216 | 450 | 656 | 180 | 367 | 554 |
|  |  | SS | 172 | 272 | 409 | 185 | 409 | 578 | 167 | 315 | 502 |
|  |  | MR | 162 | 306 | 458 | 188 | 435 | 617 | 175 | 365 | 551 |
|  |  | MS | 117 | 194 | 291 | 138 | 267 | 372 | 128 | 222 | 349 |
| $\begin{aligned} & \infty \\ & \text { ॥ } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 175 | 421 | 655 | 154 | 388 | 613 | 154 | 398 | 626 |
|  |  | SRI | 90 | 382 | 619 | 82 | 354 | 560 | 78 | 355 | 594 |
|  |  | SSI | 86 | 359 | 577 | 76 | 338 | 535 | 85 | 344 | 567 |
|  |  | SR | 92 | 386 | 618 | 85 | 353 | 564 | 83 | 360 | 605 |
|  |  | SS | 101 | 378 | 592 | 95 | 344 | 522 | 93 | 353 | 578 |
|  |  | MR | 72 | 357 | 592 | 62 | 333 | 558 | 70 | 310 | 575 |
|  |  | MS | 55 | 191 | 332 | 47 | 193 | 333 | 46 | 186 | 351 |
|  | $\mathrm{C}_{2}$ : | SR | 80 | 376 | 612 | 84 | 355 | 566 | 81 | 349 | 590 |
|  |  | SS | 78 | 341 | 579 | 87 | 344 | 535 | 75 | 316 | 553 |
|  |  | MR | 65 | 316 | 556 | 72 | 326 | 537 | 72 | 286 | 532 |
|  |  | MS | 47 | 117 | 216 | 45 | 198 | 307 | 54 | 123 | 236 |
|  | $\mathrm{C}_{3}$ : | SR | 84 | 378 | 611 | 78 | 353 | 567 | 86 | 346 | 587 |
|  |  | SS | 86 | 351 | 569 | 72 | 336 | 503 | 80 | 321 | 539 |
|  |  | MR | 70 | 342 | 567 | 59 | 309 | 517 | 67 | 306 | 546 |
|  |  | MS | 60 | 165 | 302 | 48 | 165 | 247 | 60 | 157 | 274 |

*Note: The entries within table correspond to empirical powers multiplied by 1000.

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

Table 4. The empirical powers of tests under multivariate $t$ distribution.

|  | Matrix C | Test | Correlation structure |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | CS |  |  | AR1 |  |  | UN |  |  |
|  |  |  | 10 | 20 | 30 | 10 | 20 | 30 | 10 | 20 | 30 |
| $\begin{aligned} & + \\ & \text { II } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 261 | 480 | 653 | 344 | 607 | 779 | 273 | 527 | 691 |
|  |  | SRI | 294 | 623 | 822 | 397 | 736 | 919 | 321 | 668 | 872 |
|  |  | SSI | 309 | 654 | 856 | 427 | 782 | 946 | 336 | 700 | 900 |
|  |  | SR | 293 | 620 | 823 | 395 | 728 | 920 | 314 | 667 | 871 |
|  |  | SS | 317 | 648 | 840 | 414 | 771 | 940 | 328 | 689 | 889 |
|  |  | MR | 264 | 587 | 788 | 370 | 713 | 915 | 263 | 601 | 816 |
|  |  | MS | 180 | 410 | 582 | 287 | 511 | 721 | 176 | 391 | 602 |
|  | $\mathrm{C}_{2}$ : | SR | 293 | 616 | 818 | 402 | 735 | 918 | 322 | 674 | 872 |
|  |  | SS | 315 | 653 | 846 | 424 | 782 | 946 | 342 | 691 | 891 |
|  |  | MR | 276 | 591 | 802 | 380 | 720 | 922 | 286 | 643 | 846 |
|  |  | MS | 180 | 401 | 636 | 263 | 537 | 795 | 179 | 387 | 661 |
|  | $\mathrm{C}_{3}$ : | SR | 287 | 635 | 826 | 397 | 733 | 918 | 326 | 670 | 863 |
|  |  | SS | 303 | 658 | 841 | 401 | 756 | 939 | 343 | 695 | 893 |
|  |  | MR | 278 | 591 | 793 | 359 | 711 | 903 | 294 | 630 | 847 |
|  |  | MS | 219 | 432 | 642 | 263 | 544 | 759 | 212 | 495 | 723 |
| $\begin{aligned} & \infty \\ & \text { ॥ } \\ & Q \end{aligned}$ | $\mathrm{C}_{1}$ : | $T^{2}$ | 317 | 632 | 822 | 267 | 590 | 773 | 296 | 607 | 799 |
|  |  | SRI | 194 | 709 | 917 | 169 | 661 | 893 | 184 | 669 | 902 |
|  |  | SSI | 196 | 761 | 953 | 170 | 716 | 934 | 180 | 731 | 940 |
|  |  | SR | 201 | 717 | 920 | 175 | 676 | 900 | 195 | 682 | 911 |
|  |  | SS | 233 | 776 | 954 | 205 | 729 | 931 | 212 | 743 | 945 |
|  |  | MR | 176 | 681 | 893 | 145 | 638 | 868 | 166 | 653 | 878 |
|  |  | MS | 108 | 444 | 718 | 112 | 405 | 695 | 113 | 439 | 722 |
|  | $\mathrm{C}_{2}$ : | SR | 200 | 709 | 916 | 172 | 678 | 899 | 183 | 669 | 906 |
|  |  | SS | 190 | 760 | 947 | 195 | 730 | 931 | 172 | 719 | 939 |
|  |  | MR | 151 | 607 | 870 | 149 | 635 | 866 | 149 | 578 | 840 |
|  |  | MS | 101 | 268 | 504 | 099 | 378 | 632 | 102 | 257 | 496 |
|  | $\mathrm{C}_{3}$ : | SR | 200 | 711 | 917 | 161 | 667 | 901 | 178 | 672 | 902 |
|  |  | SS | 218 | 765 | 946 | 176 | 693 | 923 | 197 | 718 | 930 |
|  |  | MR | 171 | 634 | 890 | 135 | 601 | 836 | 154 | 608 | 859 |
|  |  | MS | 119 | 414 | 694 | 100 | 349 | 606 | 126 | 379 | 645 |

*Note: The entries within table correspond to empirical powers multiplied by 1000.
Except for the test based on marginal sign, the performances of other non invariant tests were relatively robust with respect to different choices of transformation matrix $\mathbf{C}$ to test parallelism. There was not a unique choice for $\mathbf{C}$ which corresponded to the best performance of the tests. Figure 3 illustrates the degree of stability in power values for the 4 non-invariant tests for the three transformation matrices $\mathbf{C}_{1}-\mathbf{C}_{3}$ when $n=30$.

## VOSSOUGHI ET AL.


*Note: For purpose of better illustration, the powers of non-invariant tests are displayed only for the matrix $\mathbf{C}_{2}$.
Figure 1. The empirical powers of tests under multivariate normal distribution.

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS


*Note: For purpose of better illustration, the powers of non-invariant tests are displayed only for the matrix $\mathbf{C}_{2}$.
Figure 2. The empirical powers of tests under multivariate $t(3)$ distribution.

## VOSSOUGHI ET AL.



Figure 3. The empirical powers of non-invariant tests for $n=30$ for various transformation matrices under multivariate normal (a) and $t$ (b) distributions

# MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS 

## Conclusion

The results of the study revealed that the tests based on spatial and marginal ranks could serve as efficient tools for profile analysis since they performed notably better than Hotelling's $T^{2}$ for the heavy tailed distribution and were as efficient as it under normality. Similar results reported in simulation studies by Nordhausen et al. (2006) and Möttönen et al. (1998) only in the context of two sample comparison of locations for normal and $t$ distributions. Interestingly, even for moderate tailed $t$ distributions, the tests based on ranks were superior to Hotelling's $T^{2}$ in both studies. Um and Randles (1998) also reported that the multi sample extensions of multivariate rank tests proposed by Randles and Peters (1990) were more efficient than Lawly-Hotelling's $\mathbf{U}$ for light-tailed and heavytailed distributions. However, the results revealed that when there was sufficient evidence to conclude that the underlying distribution was heavy tailed, the tests based on spatial signs were the best choices to profile analysis. Similarly, this aspect was reported in the study by Nordhausen et al. (2006) and for a different sign test by Um and Randles (1998). It should also be noted that above studies conducted in areas not involving repeated measurements and various correlation structures for errors. The simulations also illustrated that when the number of replication was large (here $p=8$ ) the mentioned nonparametric tests outperformed Hotelling's $T^{2}$ only for larger sample sizes ( $n \geq 10$ ). The panel (b) of Figure 2 illustrated this issue for which Hotelling's $T^{2}$ performed slightly better than any nonparametric counter parts for $p=8$ even if the underlying distribution was heavy-tailed. The effect of sample size relative to the number of measurements has been not reported yet and hence further research in this area is necessary.

In the context of two sample comparison (as our study), Hotelling's $T^{2}$ and all the MANOVA tests (Wilks' $\boldsymbol{\Lambda}$, Pilla's $\mathbf{V}$, Lawley-Hotelling's $\mathbf{U}$ and Roy's $\boldsymbol{\theta}$ ) are functions of each other and give equivalent results; see Rencher (1998). The power of the MANOVA tests has been compared by several authors. However, they are asymptotically equivalent for sufficient sample sizes (Olson, 1974). Therefore it is implied the nonparametric alternatives can be confidently applied in place of MANOVA tests in profile analysis regardless of the nature of underlying distribution. Park et al. (2001) investigated the performance of profile analysis using Hotelling's $T^{2}$ and mixed model approach to test group and interaction effects. Also, Vossoughi et al. (2012) compared the performance of profile analysis, linear mixed model and summary measure approach in repeated measurements generated from a linear mixed model setting. Similarly, both studies showed that the profile analysis preserved the nominal significance level

## VOSSOUGHI ET AL.

and performed relatively robust to the underlying correlation structure but provided less power values than the competitors, in general.

Marcucci (1986) demonstrated that profile analysis using Hotelling's $T^{2}$ and exclusively univariate split-plot analysis with d.f. adjustments gave type I error rates closest to the nominal level, but not one of which was most powerful along various correlation structures and patterns of means. The interested reader is also referred to Schwertman et al. (1985), Boik (1991) and Davidson (1972) for further assessment on this issue.

Thought not reported here, we conducted additional simulations for a variety type of the location trend over occasions such as linear trend as $\boldsymbol{\theta}_{2}=(0.25,0.5,0.75,1)^{\prime}$. The larger number of measurements $p=8$ and sample size $n=50$ in each group were also considered. However, the similar results were yielded and hence not further included in the study.

In conclusion, the findings implied that the use of some nonparametric multivariate tests in place of the parametric counterparts can considerably improve the result of profile analysis for heavy-tailed distributions. Accordingly, the tests based on spatial and marginal ranks are severe competitors for parametric tests in profile analysis since they performed as well as Hotelling's $T^{2}$ under multivariate normal distribution and dominated it under heavy-tailed distribution. Moreover, the simulation results revealed that the tests based on spatial signs under heavy tailed distributions, were more efficient than the MANOVA tests for the analysis of repeated measurements.

## References

Azzalini, A., \& Bowman, A. (1991). Nonparametric regression methods for repeated measurements. In G. Roussas (Ed.), Nonparametric functional estimation and related topics (pp. 377-387). Netherlands: Springer.

Bhapkar, V. P. (1984). Univariate and multivariate multisample location and scale tests. In P. R. Krishnaiah \& P. K. Sen (Eds.), Nonparametric methods (pp. 31-62). Oxford: Elsevier.

Bhapkar, V. P., \& Patterson, K.W. (1978). A Monte Carlo study of some multivariate nonparametric statistics for profile analysis of several samples. Journal of Statistical Computation and Simulation, 6(3/4), 223-237. doi: 10.1080/00949657808810191

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

Boik, R. J. (1991). Scheffés mixed model for multivariate repeated measures: A relative efficiency evaluation. Communications in Statistics - Theory and Methods, 20(4), 1233-1255. doi: 10.1080/03610929108830562

Davidson, M. L. (1972). Univariate versus multivariate tests in repeatedmeasures experiments. Psychological Bulletin, 77(6), 446-452. doi: 10.1037/h0032674

Davis, A. W. (1980). On the effects of moderate multivariate nonnormality on Wilks's likelihood ratio criterion. Biometrika, 67(2), 419-427. doi: 10.1093/biomet/67.2.419

Davis, A. W. (1982). On the effects of moderate multivariate nonnormality on Roy's largest root test. Journal of the American Statistical Association, 77(380), 896-900. doi: 10.1080/01621459.1982.10477904

Davis, C. S. (2002). Statistical methods for the analysis of repeated measurements. New York, NY: Springer-Verlag.

Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one-and two-sample $T^{2}$ tests. Journal of the American Statistical Association, 74(365), 48-51. doi: 10.1080/01621459.1979.10481606

Hettmansperger, T. P., Möttönen, J. \& Oja, H. (1998). Affine invariant multivariate rank tests for several samples. Statistica Sinica, 8(3), 785-800.

Hettmansperger, T. P., \& Oja, H. (1994). Affine invariant multivariate multisample sign tests. Journal of the Royal Statistical Society. Series $B$ (Methodological), 56(1), 235-249.

Laird, N. M., \& Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 38(4), 963-974. doi: 10.2307/2529876

Littell, R. C., Pendergast, J., \& Natarajan, R. (2000). Tutorial in biostatistics: Modelling covariance structure in the analysis of repeated measures data. Statistics in Medicine, 19, 1793-1819.

Marcucci, M. (1986). A comparison of the power of some tests for repeated measurements. Journal of Statistical Computation and Simulation, 26(1/2), 37-53. doi: 10.1080/00949658608810948

Mendoza, J. L., Toothaker, L. E., \& Nicewander, W. A. (1974). A Monte Carlo comparison of the univariate and multivariate methods for the groups by trials repeated measures design. Multivariate Behavioral Research, 9(2), 165-177. doi: 10.1207/s15327906mbr0902_3

## VOSSOUGHI ET AL.

Möttönen, J., Hettmansperger, T. P., Oja, H., \& Tienari, J. (1998). On the efficiency of affine invariant multivariate rank tests. Journal of Multivariate Analysis, 66(1), 118-132. doi: 10.1006/jmva.1998.1740

Möttönen, J., \& Oja, H. (1995). Multivariate spatial sign and rank methods. Journal of Nonparametric Statistics, 5(2), 201-213.
doi: 10.1080/10485259508832643
Möttönen, J., Oja, H., \& Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. The Annals of Statistics, 25(2), 542-552.
doi: 10.1214/aos/1031833663
Nordhausen, K., Oja, H., \& Tyler, D. E. (2006). On the efficiency of invariant multivariate sign and rank tests. In E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen \& G. P. H. Styan (Eds.), Festschrift for Tarmo Pukkila on his 60th birthday (pp.217-231). Tampere, Finland: Dept. of Mathematics, Statistics and Philosophy, University of Tampere.

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. Scandinavian Journal of Statistics, 26(3), 319-343. doi: 10.1111/1467-9469.00152

Oja, H. (2010). Multivariate nonparametric methods with R: An approach based on spatial signs and ranks. New York, NY: Springer.

Oja, H., \& Randles, R. H. (2004). Multivariate nonparametric tests. Statistical Science, 19(4), 598-605. doi: 10.1214/088342304000000558

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 69(348), 894-908.

Park, T., Park, J.-K., \& Davis, C. S. (2001). Effects of covariance model assumptions on hypothesis tests for repeated measurements: Analysis of ovarian hormone data and pituitary-pteryomaxillary distance data. Statistics in Medicine, 20(16), 2441-2453. doi: 10.1002/sim. 859

Puri, M. L., \& Sen, P. K. (1971). Nonparametric methods in multivariate analysis. New York, NY: Wiley.

R Development Core Team. (2013). R: A language and environment for statistical computing. Reference index (Version 3.0.1). Vienna, Austria: R Foundation for Statistical Computing.

Randles, R. H., \& Peters, D. (1990). Multivariate rank tests for the twosample location problem. Communications in Statistics - Theory and Methods, 19(11), 4225-4238. doi: 10.1080/03610929008830439

## MULTIVARIATE NONPARAMETRIC TESTS IN PROFILE ANALYSIS

Rencher, A. C. (1995). Methods of multivariate analysis. New York, NY: Wiley.

Rencher, A. C. (1998). Multivariate statistical inference and applications. New York, NY: Wiley.

Schwertman, N. C., Flynn, W., Stein, S., \& Schenk, K. L. (1985). A Monte Carlo study of alternative procedures for testing the hypothesis of parallelism for complete and incomplete growth curve data. Journal of Statistical Computation and Simulation, 21(1), 1-37. doi: 10.1080/00949658508810794

Schwertman, N. C., Fridshal, D., \& Magrey, J. M. (1981). On the analysis of incomplete growth curve data, a Monte Carlo study of two nonparametric procedures. Communications in Statistics-Simulation and Computation, 10(1), 51-66. doi: 10.1080/03610918108812192

Sen, P. K. (1984). Nonparametric procedures for some miscellaneous problems. In P. R. Krishnaiah \& P. K. Sen (Eds.), Nonparametric methods (pp. 699-739). Oxford: Elsevier.

Singer, J. M., Poleto, F. Z., \& Rosa, P. (2004). Parametric and nonparametric analyses of repeated ordinal categorical data. Biometrical Journal, 46(4), 460-473. doi: 10.1002/bimj. 200310045

Somorčík, J. (2006). Tests using spatial median. Austrian Journal of Statistics, 35(2/3), 331-338. doi: 10.17713/ajs.v35i2\&3.380

Um, Y., \& Randles, R. H. (1998). Nonparametric tests for the multivariate multi-sample location problem. Statistica Sinica, 8(3), 801-812.

Vossoughi, M., Ayatollahi, S. M. T., Towhidi, M., \& Ketabchi, F. (2012). On summary measure analysis of linear trend repeated measures data: performance comparison with two competing methods. BMC Medical Research Methodology, 12(33), 1-12. doi: 10.1186/1471-2288-12-33

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. American Statistician, 39(2), 95-101.

Wernecke, K.-D., \& Kalb, G. (1999). On nonparametric repeated measures analysis of variance. In W. Gaul \& H. Locarek-Junge (Eds.), Classification in the information age (pp. 596-604). Berlin: Springer.

Wernecke, K.-D., \& Kaufmann, J. (2000). A new approach on nonparametric repeated measures analysis of variance. In R. Decker \& W. Gaul (Eds.), Classification and information processing at the turn of the millennium (pp. 430-437). Berlin: Springer.

# Optimal Estimation and Sampling Allocation in Survey Sampling Under a General Correlated Superpopulation Model 

Ioulia Papageorgiou<br>Athens University of Economics and Business<br>Athens, Greece

Sampling from a finite population with correlated units is addressed. The proposed methodology applies to any type of correlation function and provides the sample allocation that ensures optimal efficiency of the population parameters estimates. The expressions of the estimate and its MSE are also provided.

Keywords: $\quad$ Superpopulation, systematic sampling, model-based sampling, sampling strategy, optimality, autocorrelation

## Introduction

In classical sampling theory, the finite population under study is assumed to be a fixed vector of dimension $N$, where $N$ is the number of population members. If $U$ denotes the population set and $Y$ the variable of interest, the population vector can be denoted as $U=\left\{Y_{1}, Y_{2}, \ldots, Y_{N}\right\}$, and is assumed to be fixed but is in general unknown. The superpopulation approach in sampling from a finite population is assumed in this work. According to this approach, the finite set of measurements $U$ is a realization of a sample of size $N$ drawn from an infinite population with common distribution $\xi$.

The superpopulation model was introduced by Cochran (1946 and 1977, 1953) and further developed by Godambe (1955), Cassel et al. (1977), Tam (1984), Blight (1973), Mukerjee \& Sengupta $(1989,1990)$ and Bolfarine \& Zacks (1992), among others. The problem of finding optimum sampling schemes under a superpopulation model is discussed by several authors including Blight (1973), Papageorgiou \& Karakostas (1998), Arnab (1992), Mukerjee \& Sengupta (1989, 1990), Nayak (2003) and Chao (2004). The superpopulation model assumes the population measurements are comprised of a deterministic and a non-

[^19]
## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

deterministic element that can be attached to a variable. More analytically, the superpopulation model in its general form is

$$
\begin{equation*}
Y_{i}={ }_{i}+{ }_{i}, i=1,2, \ldots, N \tag{1}
\end{equation*}
$$

where $\mu_{i}$ is constant, representing the deterministic part, while $\varepsilon_{i}$ are random variables also called errors. The random vector $\varepsilon=\left(\varepsilon_{1}, \varepsilon_{2}, \ldots, \varepsilon_{N}\right)$ is assumed to have zero mean and variance covariance matrix $\boldsymbol{V}$. Various special models to describe more specific or realistic population assumptions can be derived from (1) by making assumptions on matrix $\boldsymbol{V}$ and relationships among $\mu_{i}$. For example, model

$$
E\left(Y_{i}\right)={ }_{i} \text { and } E\left(\begin{array}{ll}
Y_{i} & { }_{i}
\end{array}\right)\left(\begin{array}{l}
Y_{j}
\end{array}\right)=\left\{\begin{array}{cl}
{ }_{j}^{2}, & \text { if } i=j \\
0, & \text { if } i \neq j
\end{array}\right.
$$

where the errors are uncorrelated and with constant variance is the model that describes a finite population with uncorrelated measurements and different superpopulation mean.

Another special case is the model where

$$
E\left(Y_{i}\right)=\text { and } E\left(\begin{array}{ll}
Y_{i} & { }_{i}
\end{array}\right)\left(\begin{array}{ll}
Y_{j} & j
\end{array}\right)= \begin{cases}{ }^{2}, & \text { if } i=j \\
{ }^{2}, & \text { if } i \neq j\end{cases}
$$

according to which the population units are correlated with a constant correlation $\rho$ and constant superpopulation parameter of mean $\mu$.

A more realistic autocorrelated superpopulation model results if one assumes that the degree of correlation among two population units depends on between-unit distance. This is also known as serial correlation. Populations that exhibit this characteristic can be encountered in applications where an order is assigned to each of the population members. The ordering can be according to time, space, magnitude or the serial number in a production line. The model with serial correlation was first introduced by Cochran (1946) and it can be written in mathematical terms as

$$
E\left(Y_{i}\right)=\text { and } E\left(\begin{array}{ll}
Y_{i} & { }_{i}
\end{array}\right)\left(\begin{array}{ll}
Y_{j} & j
\end{array}\right)=2\left(\begin{array}{ll}
i & j \tag{2}
\end{array}\right)
$$

## IOULIA PAPAGEORGIOU

where $\rho(h)$ is the autocorrelation function of the population model for units at distance $h$.

All above models can also be seen as special cases of the more general superpopulation regression model where the deterministic part $\mu$ has been modeled as linear functions of a set of auxiliary nonstochastic variables that may be available for the population vector (Bolfarine and Zacks, 1992).

Madow \& Madow (1944), Cochran (1977), Royall (1970), Blight (1973), Ramakrishnan (1975), Bellhouse \& Rao (1975) and Graubard \& Korn (2002), among others, assumed (2) or a special case of this. The results available from the literature aim to answer two questions: first, to estimate the superpopulation parameter $\mu$; and second, to determine the optimal sampling design. The optimal sampling design is the selection process according to which the sample units are drawn from the population so that the derived estimate will achieve an assumed optimality criterion, such as minimum variance. Sampling strategy is the pair of the sampling design and estimator used towards the estimation problem (see for example Ramakrishnan, 1975). Often in practice, certain properties are attached to the autocorrelation function $\rho(h)$ such as positive, decreasing or convex. An outline of related results from the literature is presented in the following section.

In this current work the assumptions made on function $\rho(h)$ are extended. More specifically, $\rho(h)$ can be the autocorrelation function of any random process with second-order stationarity. The proposed methodology aims to determine the optimal allocation of the sampling units for a sample of size $n$, when the least squared estimator of the superpopulation mean is used as a criterion of optimality. The optimum is defined with respect to the mean squared error (mse) of the estimate. The proposed optimal sampling strategy is completed by providing the statistical inference of the assumed estimate when the sample is selected, according to the derived optimal sampling scheme. Both the derived optimal sample allocation and its $m s e$ depend on $\rho(h)$ and therefore take into account the specific autocorrelation of the population under study.

## General notation and brief review

Denote by $s=\left\{Y_{j_{1}}, Y_{j_{2}}, \ldots, Y_{j_{n}}\right\}$ the sample of size $n$ that is selected from the complete vector $U$. Indexes $j_{i}(i=1,2, \ldots, n)$ in the notation indicate the positions of the selected units in the population $U . \theta=\sum_{i=1}^{N} Y_{i}$, the population sum, is considered as the parameter of interest. $\theta$ is a linear function of the population measurements. Dealing with the estimation of $\theta$ is equivalent with the estimation
of population mean $\bar{Y}=\sum_{i=1}^{N} Y_{i} / N$, as the two quantities differ only by a constant coefficient.

The aim of the sampling procedure is to estimate $\theta$ based on a set of measurements, $s$, selected from $U$. The assumption of selection without replacement is made, but sampling with replacement is equally possible. The sampling design is determined by the probability $p(s)$ that is assigned to each of all possible samples $s$ selected from the population. Let $P_{n}$ denote this set of all possible samples of size $n$. Important probabilities related to the design $p(s)$ are the first and second order inclusion probabilities $\pi_{i}$ and $\pi_{i j}$ respectively, defined as

$$
{ }_{i}=\sum_{s i} p(s) \text { and }_{i j}=\sum_{s i, j} p(s) .
$$

By making use of this notation, simple random sample (that is, the simplest sampling design) is the design that assigns equal probability $p(s)=1 /\binom{N}{n}$ in every sample $s$ that belongs in $P_{n}$, where $P_{n}$ is the selection of all the possible combinations of $n$ measurements chosen from $U$ in this case. For the systematic scheme, the probabilities of selection are also equal, $p(s)=1 / k$, where $k=N / n$. The number of samples that belong in $P_{n}$ is also $k$ in the systematic case and if $s_{i}, i=1,2, \ldots, k$ is a representative sample, then $s_{i}=\left(Y_{i}, Y_{i+k}, Y_{i+2 k}, \ldots, Y_{i+(n-1) k}\right)$, $i=1,2, \ldots, k$ (see for example Cochran, 1977). If $N \neq n k$, a slight complication and need for modification arises, but the effect is negligible (Yates, 1960, 1948 1 st ed.). The samples generated by a systematic procedure are equally spaced, and moreover if the start $Y_{i}$ is chosen with $i$ such that $2 i=N+1-(n-1) k$, the sample is a centrally located systematic sample (Blight, 1973). In this last case $P_{n}$ contains only one sample $s$ with $p(s)=1$.

Blight in the previously mentioned work assumes that the deviations of population values from the superpopulation mean $\mu$ are generated by an autoregressive model of order one, $\operatorname{AR}(1)$, e.g.

$$
\begin{equation*}
Y_{i}=\left(Y_{i 1}\right)+{ }_{i} \tag{3}
\end{equation*}
$$

where $\varepsilon_{i}$ is uncorrelated normally distributed series with zero mean and constant variance $\sigma^{2}$. This yields $\rho(h)=\lambda^{h}$ at lag $h(h=1,2, \ldots, N-1)$. Employing the sample mean as the estimator of the corresponding population mean, the effect of the autocorrelation is studied and the optimal sampling design when $\lambda$ is positive

## IOULIA PAPAGEORGIOU

or negative is obtained. The optimality criterion is the conditional variance $\operatorname{Var}_{\xi}\left(\bar{Y} \mid Y_{j_{i}} \in s\right)$. The sign of $\lambda$ controls the monotonicity shape of $\rho(h)=\lambda^{h}$ and, as expected, the resulting optimal design is remarkably different among the two cases. More specifically, for $\lambda>0$ the optimal sample is the centrally located systematic and for $\lambda<0$ the optimal sampling design is concentrated towards the two ends of the population. This also verifies the fact that the optimal solution for the sampling scheme is not unique, but depends on the specific type of the autocorrelation. However, when the autocorrelation function $\rho(h)$ is not only $\lambda^{h}, h>0$, but in general any positive, decreasing and convex function, the same result holds and the centrally located systematic design is the optimal (Papageorgiou \& Karakostas, 1998).

Function $\rho(h)$ is defined in all positive integer numbers and therefore $\rho(h)$ is decreasing if $\rho(h+1)-\rho(h) \leq 0(\Delta \rho(h) \leq 0)$, while convexity holds when

$$
{ }^{2}(h)=(h+2) \quad 2(h+1)+(h) \geq 0 \text { for } h=0,1,2, \ldots
$$

Denote by $K$ the class of all autocorrelation functions that satisfy the aforementioned properties (positive, decreasing and convex). AR(1) model assumed in (3) has an autocorrelation function that belongs in $K$ when $\lambda>0$ and since the optimality of the centrally located systematic scheme holds for the whole class $K$ it also holds for this occasion as a special case. In fact, class $K$ includes a wide range of correlation functions (Bellhouse, 1984).

Although the question about the optimum sampling scheme seems to have a unique answer when $\rho(h) \in K$ and it is closely related to the systematic scheme, under almost any combination of estimators and optimality criterions considered, the problem remains when $\rho(h)$ does not belong in $K$. The optimum sampling scheme in this case can be quite far from the systematic and it varies depending on the specific type of $\rho(h)$. In other words, there is no uniquely defined optimum sampling scheme that can cover any random process with respect to the sampling problem. In this direction, a practical and easy-to-implement methodology, that suggests the optimum sampling procedure once the specific type of $\rho(h)$ or $\boldsymbol{V}$ is provided, is proposed in this paper.

A related work is provided by Chao (2004), where a general known matrix $\boldsymbol{V}$ is assumed, and a similar to principal component analysis method is suggested in order to obtain the sampling procedure. More specifically, the idea is to choose as sampled units those population units pointed from the $n$ most important components or the largest eigenvalues of matrix $\boldsymbol{V}$. Two algorithms are proposed,
called Design I and Design II, with the second being a slight modification of the former. Design I makes use of the $n$ eigenvectors $\boldsymbol{e}_{1}, \boldsymbol{e}_{2}, \ldots, \boldsymbol{e}_{n}$ of $\boldsymbol{V}$ that correspond to the $n$ first-in-magnitude eigenvalues of the matrix. If $\boldsymbol{e}_{i}=\left(e_{i 1}, e_{i 2}, \ldots, e_{i N}\right)$ is such an eigenvector and $j,(j=1, \ldots, N)$ is the index with the largest-in-absolute-magnitude component in $\boldsymbol{e}_{i}$, the population unit that corresponds in position $j$ is the one selected in the sample according to this design. If the unit is already in the sample, the second-in-absolute-magnitude component is selected. Design II works in the same principle, with the difference that the sign of the components is also taken into account. From each eigenvector two components are selected, the largest-in-absolute-magnitude and the second-largest with opposite sign of the first. The approach for both designs is rather intuitive and the resulting designs do not hold any optimality criterion. Their performance is measured by the relative efficiency over the simple random sample as a general sampling scheme. They indicate improved efficiency with respect to the simple random most of the times, but their performance is not stable and the simple random sample itself is usually far from the optimal when a correlation exists.

Before dealing with the problem and proposing the solution of the optimal sampling design, a list of possible applications is provided. The range of applications is wide, and they cover any scientific area where the framework includes correlated measurements and a sample is selected from the population. A typical application of sampling from autocorrelated populations where the autocorrelation is not necessarily decreasing and convex is seen in the context of statistical process control in monitoring manufacture and industrial production lines. A variety of control charts or other statistical instruments can be constructed based on a set of measurements selected from the process, and help practitioners to derive information or warning if the process is out of control. Traditionally the statistical theory behind the control charts is based on the assumption that the sample measurements are independent. It is however quite common in practiceand especially in continuous manufacture or production lines-that this assumption is violated, and this produces misleading and unreliable control charts (Alwan, 1992; Montgomery and Mastrangelo, 1991) with tighter control limits than the true ones. A lot of attention has been drawn lately to this area of research; see for example Alwan and Roberts (1988), Harris and Ross (1991), Mastrangelo and Montgomery (1995), Apley and Lee (2003) and Lu and Reynolds (1999, 2001), and all proposed approaches make use of the present autocorrelation to either modify the existing control limits, or to model the process, identify the autocorrelation, and use the independent errors instead of the measurements for constructing any statistical tool. The models that have been assumed are $\operatorname{AR}(1)$

## IOULIA PAPAGEORGIOU

(Autogregressive), MA(1) (Moving Average) and ARMA(1,1) (Autoregressive Moving Average) (Wardell et al., 1992) and efficiency in sampling and therefore construction of the control limits provided can be improved further if the specific type of correlation is taken into account.

Similarly, geostatistical data in spatial statistics very often exhibit a smallscale variation, typically a strong correlation between data at neighboring locations (Watson, 1972). If the population mean is the parameter of interest, failure to realize the presence of positive correlation in the data leads to very narrow confidence interval (Cressie, 1993), a result similar with this in quality control charts. The superpopulation model is therefore extensively used in modeling geostatistical data in order to accommodate this correlation (Cressie, 1993). In this context, let $\boldsymbol{s} \in \mathbb{R}^{d}$ be the data location in $d$-dimensional Euclidean space and $\boldsymbol{Y}(\boldsymbol{s})$ the measured data, assumed random, at location $\boldsymbol{s}$. Assuming that $\boldsymbol{s}$ takes values over an index set $D \subset \mathbb{R}^{d}$, the superpopulation model results as a realization $\{y(\boldsymbol{s}): \boldsymbol{s} \in D\}$ from the multivariate random field $\{\boldsymbol{Y}(\boldsymbol{s}): \boldsymbol{s} \in \mathrm{D}\}$. Land and agricultural surveys, ground-water monitoring, environmental statistics and socio-economic habitat surveys are some of the sampling applications in two dimensions with spatial dependence among population units.

Other applications of sampling from correlated populations include genetics and ecological statistics. In particular, the superpopulation model is often used to explain genetic or ecological patterns where the covariance in the genetic makeup of individuals or in the growth of populations can be assumed to be a function of the spatial distance separating the units (Lande, 1991; Bjørnstad et al., 1999).

Clustered data, often found in social, educational, psychometric and behavioral studies, also represent an application of sampling from correlated populations. Clustered data may result either because of repeated measurements in time such as in longitudinal studies or because of sub-sampling from a large primary unit: for instance, sampling graduates from the same educational institute or the same region/country for a large scale study. The existing intra-class correlation has to be taken into account during the analysis and the statistical inference in order to produce valid results (Neuhaus and Kalbfleisch, 1998). Moreover the knowledge of the intra-class correlation can contribute at the selection stage of the sub-sampling.

Another application of sampling in time series, apart from the serial correlation and the typical applications described already, is the use of composite marginal likelihoods in order to estimate the parameters of the model (Cox and Reid, 2004; Varin, 2008). Pairwise likelihoods, based only on the bivariate joined distributions of the measurements, produce estimates very close to those under the

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

full likelihood with respect to the dimension. The benefit of pairwise likelihood is on the computational demand that is required for the optimization. Moreover, further improvement in this direction can be achieved if not all possible pairs but only a selection of them will be used instead. Current work in this context shows that the same accurate estimates can be obtained if the correlation between observations is taken into account towards the selection procedure: for example, pairwise likelihood of order $m$ (Hjort and Varin, 2008).

## The general problem

Model (2) describes the population and $\rho(h)$ is assumed to be any autocorrelation function. Moreover, ${ }^{\wedge}$, the least squared estimator for the parameter $\theta$, is assumed as the optimality criterion. The aim is to determine the sampling design $p$ or the sample $s$ that minimizes the mean square error of ${ }^{\wedge}$ under this model. The least squared estimator of the population mean is the sample mean and it is unbiased under model (2) (see Karakostas, 1984), which yields that

$$
\begin{equation*}
\hat{\theta}=\frac{N}{n} \sum_{i=1}^{n} Y_{j_{i}} \tag{4}
\end{equation*}
$$

The mean square error of ${ }^{\wedge}$ when a sample $s=\left\{Y_{j_{1}}, Y_{j_{2}}, \ldots, Y_{j_{n}}\right\}$ is provided is given by

$$
\begin{aligned}
\operatorname{mse}\left(\wedge^{\wedge} \mid s\right) & =E\left[\left(\begin{array}{ll}
\wedge & )^{2} \mid s\right] \\
& =\operatorname{Var}()+\operatorname{Var}(\wedge) \quad 2 \operatorname{Cov}(\wedge,) \\
& =\operatorname{Var}\left(\sum_{i=1}^{N} Y_{i}\right)+\frac{N^{2}}{n^{2}} \operatorname{Var}\left(\sum_{i=1}^{n} Y_{j_{i}}\right) \quad 2 \frac{N}{n} \operatorname{Cov}\left(\sum_{i=1}^{n} Y_{j_{i}}, \sum_{i=1}^{N} Y_{i}\right)
\end{array} r=\frac{1}{n}\right)\right.
\end{aligned}
$$

Let $\boldsymbol{V}$ be the variance covariance matrix of the complete population vector under (2). The partition of matrix $\boldsymbol{V}$ according to the sampled part, $s$, is considered next and let $\boldsymbol{V}_{s}$ denote the part of $\boldsymbol{V}$ that corresponds to the sampled units and $\boldsymbol{V}_{s, U}$ the $n \times N$ matrix of $\boldsymbol{V}$ where its rows correspond to the sampled units while the columns to the whole population $U$. Under this notation the mse can be written as

$$
\begin{equation*}
\operatorname{mse}\left(\wedge^{\wedge} \mid S\right)=\mathbf{1}_{N} \boldsymbol{V} \mathbf{1}_{N}+\frac{N^{2}}{n^{2}} \mathbf{1}_{n} \boldsymbol{V}_{S} \mathbf{1}_{n} \quad 2 \frac{N}{n} \mathbf{1}_{n} \boldsymbol{V}_{s, U} \mathbf{1}_{N} \tag{5}
\end{equation*}
$$

## IOULIA PAPAGEORGIOU

where $\mathbf{1}_{j}^{\prime}$ stands for the $j$-dimension vector of units. For the sampling problem it is necessary to minimize $m s e\left({ }^{\wedge}\right)$ with respect to the sample $s$, or equivalently to find the minimum

$$
\begin{equation*}
\min _{s}\left\{\frac{N}{n} \mathbf{1}_{n}^{\prime} \boldsymbol{V}_{s, U} \mathbf{1}_{n}-2 \rightarrow \mathbf{1}_{n}^{\prime} \boldsymbol{V}_{s, U} \mathbf{1}_{n}\right\} \tag{6}
\end{equation*}
$$

For any sample $s=\left\{Y_{j_{1}}, Y_{j_{2}}, \ldots, Y_{j_{n}}\right\}$ let $h_{i}=j_{i+1}-j_{i}, 1,2, \ldots, n-1$ denote the distances of two successive sampled units with moreover $h_{0}=j_{1}-1$ and $h_{N}=N-j_{n}$ the two end distances. Under this notation any sample $s$ can also take the form $s=\left\{Y_{h_{0}+1}, Y_{h_{1}+h_{0}+1}, \ldots, Y_{h_{n-1}+\cdots+h_{1}+h_{0}+1}\right\}$ and uniquely represented by the vector of distances $h=\left(h_{0}, h_{1}, h_{2}, \ldots, h_{n}\right)$ with $h_{i}, i=0,1, \ldots, n$ to be integers with $h_{0}+h_{1}+\ldots+h_{n}=N-1$. Using this equivalent notation for the sample $s$ the minimization expression can finally be written as

$$
\min _{h_{i}}\left[\begin{array}{c}
N-2 n+\frac{2 N}{n}\left(\begin{array}{r}
\sum_{i=1}^{n-1} \rho\left(h_{i}\right)+\sum_{i=1}^{n-2} \rho\left(h_{i}+h_{i+1}\right) \\
\\
+\sum_{i=1}^{n-3} \rho\left(h_{i}+h_{i+1}+h_{i+2}\right)+\cdots+\rho\left(h_{1}+\cdots+h_{n-1}\right)
\end{array}\right] \\
-2\left(\sum_{i=1}^{h_{0}} \rho(i)+\sum_{i=1}^{h_{0}+h_{1}} \rho(i)+\cdots+\sum_{i=1}^{h_{0}++h_{n-1}} \rho(i)+\sum_{i=1}^{h_{1}+\cdots+h_{n}} \rho(i)+\cdots+\sum_{i=1}^{h_{n}} \rho(i)\right)
\end{array}\right]
$$

or equivalently

$$
\min _{h_{i}} Q\left(h_{0}, h_{1}, h_{2}, \cdots, h_{n}\right)
$$

where

$$
\begin{align*}
Q & =\frac{N}{n}\binom{\sum_{i=1}^{n-1} \rho\left(h_{i}\right)+\sum_{i=1}^{n-2} \rho\left(h_{i}+h_{i+1}\right)}{+\sum_{i=1}^{n-3} \rho\left(h_{i}+h_{i+1}+h_{i+2}\right)+\cdots+\rho\left(h_{1}+\cdots+h_{n-1}\right)}  \tag{7}\\
& -\left(\sum_{i=1}^{h_{0}} \rho(i)+\sum_{i=1}^{h_{0}+h_{1}} \rho(i)+\cdots+\sum_{i=1}^{h_{0}+\cdots+h_{n-1}} \rho(i)+\sum_{i=1}^{h_{1}+\cdots+h_{n}} \rho(i)+\cdots+\sum_{i=1}^{h_{n}} \rho(i)\right)
\end{align*}
$$

Finding the optimum sample ( $s$ ) is now a constrained minimization problem of minimizing (7) with respect to the unknowns $h_{i}, i=0,1, \ldots, n$. The parameter constrains, that mainly result from their definition, are

$$
\begin{align*}
& 0 \leq h_{0} \leq N-1 \\
& 0<h_{i} \leq N-1, i=1,2, \ldots, n-1 \\
& 0 \leq h_{n} \leq N-1 \text { and } \\
& h_{0}+h_{1}+\ldots+h_{n}=N-1 \tag{8}
\end{align*}
$$

Therefore, the sampling problem is mathematically formulated as a constrained minimization problem. However the mathematical solution is not straightforward, due to the unknown integer function $\rho(h)$ involved in $Q$. Unless certain properties are assumed for $\rho(h)$ the problem cannot be solved in its general case. The difficulty is mainly caused from the upper bounds of the summations in the second parenthesis of $Q$ that depend on the unknowns $h_{i}$ and make the number of the terms in those summations a variable itself.

## Methodology

## A Solution Based On The Continuous Approximation

The objective function $Q$ given by (7) is in general a sum of values of the function $\rho(h)$. Function $\rho(h)$ on the other hand represents the autocorrelation function of the population series and takes values at lag $h, h=0,1,2, \ldots, N-1$, being therefore an integer defined function. The integer feature of $\rho(h)$ leads to the summations appearing in Q that in turn prevent from its minimization.

The idea is to use an approximate, but approachable towards its minimization expression instead of $Q$. The approximation consists of two stages, first to approximate every sum that appears in the second parenthesis of $Q$ by an

## IOULIA PAPAGEORGIOU

integral and secondly to approximate the integer function $\rho(h)$ with a continuous function. Approximating a sum with an integral is a known practice in literature and departs from the Euler-Maclaurin formula. The aim is to use Euler-Maclaurin formula in order to obtain a continuous approximation of the objective function and provide bounds for the error in the approximation. Note however that the derivation of the point(s) ( $h_{0}, h_{1}, h_{2}, \ldots, h_{n}$ ) where the minimum is attained will suffice the sampling problem and will provide with the optimal sample. Once the optimal sample is determined the corresponding for the estimate exact mse under the optimal sample can be calculated by a single substitution in (5) and not through its continuous approximation. In other words, the approximate and the true versions of $Q$ need only to share the same monotonicity and not coincide. The second condition is stronger and guarantees the first.

Euler-Maclaurin formula is a mathematical tool, an equality, where a finite sum of values of a function $f$ at the left side part is expressed as a finite integral of the same function $f$ plus an error term at the right side part. The error term involves all consecutive derivatives of $f$, the Bernoulli numbers and Bernoulli polynomials. More analytically, it holds

$$
\begin{equation*}
\sum_{k=a}^{b-1} f(k)=\int_{a}^{b} f(x) d x+\left.\sum_{k=1}^{m} \frac{B_{k}}{k!} f^{(k-1)}(x)\right|_{a} ^{b}+R_{m} \tag{9}
\end{equation*}
$$

where

$$
R_{m}=(-1)^{m-1} \int_{a}^{b} \frac{B_{m}(\{x\})}{m!} f^{m}(x) d x, \text { for integers } a \leq b, m \geq 1
$$

$B_{k}, k=1,2, \ldots$ stands for the Bernoulli numbers and $B_{m}(\{x\})$ is the Bernoulli polynomial with $\{x\}=x-\lfloor x\rfloor$ the fractional part of $x . R_{m}$ is the remainder and $m$ is chosen accordingly. Euler-Maclaurin expression is a fundamental result in algebra providing a link between a sum and the corresponding integral. A number of other important results can be derived from this formula. For more details see Graham et al, 1994, p. 469.

The integer number $m$ that can be chosen accordingly in (9) will affect the remainder and consequently the error in this continuous approximation. The Bernoulli numbers are closely related with this choice. Recall the first few values:

$$
\begin{gathered}
B_{0}=1, B_{1}=-\frac{1}{2}, B_{2}=\frac{1}{6}, B_{4}=-\frac{1}{30}, B_{6}=\frac{1}{42}, B_{8}=-\frac{1}{30} \\
\text { and } B_{3}=B_{5}=B_{7}=B_{9}=\cdots=0
\end{gathered}
$$

For $m=3$, for example, the Euler-Maclaurin equation (9) for a function $f$ studied in the interval $[a, b]$ is

$$
\begin{equation*}
\sum_{k=a}^{b 1} f(k)=\int_{a}^{b} f(x) d x+\left.\left\{\frac{1}{2} f(x)+\frac{1}{12} f^{1}(x)\right\}\right|_{a} ^{b}+\int_{a}^{b} \frac{B_{3}(\{x\})}{6} f^{(3)}(x) d x \tag{10}
\end{equation*}
$$

The remainder in general must always be considered, as often it diverts, depending on function $f$ (Graham et al, 1994). Function $\rho(h)$ is playing the role of function $f$ in this present application of Euler-Maclaurin. Consequently, the second stage of approximation in $Q$ consisted of a continuous approximation of $\rho(h)$, and is also related to the remainder calculation. Such an approximation is needed because of the integer nature of $\rho(h)$ and the presence of integrals at the right hand side of formula (9).

Because equation (9) involves all the successive-in-order derivatives of $f$, a continuous extension of $\rho(h)$ through a spline interpolating technique is proposed. If the spline is selected within the broad group of cubic polynomial splines, the third derivative can always be constant and the fourth or higher equal to zero. There are a few alternative splines that preserve the cubic characteristics, with most popular (i) the piecewise cubic shape-preserving hermite interpolation and (ii) the cubic spline, both implemented in Matlab with routines pchip and csaps respectively. The characteristics that these two alternatives share in common are that they both produce a polynomial which passes though the provided data points, they are piecewise three degree polynomials and they have continuous first derivatives. The differences between them is that the pchip produces a function that in order to reserve the shape of the data has discontinuous second derivatives, while csaps leads to a smoother function with continuous second derivatives. Moreover, csaps allows a smoothing parameter $p$ to be chosen, either manually or by default, which controls the smoothness of the resulting curve in contrast with how close this curve will be to the data points to which it will be fitted.

Let $r(h)$ denote a continuous piecewise cubic interpolation of $\rho(h)$, obtained by either pchip or csaps. Applying next Euler formula for $m=4$ to a typical summation of those contained in $Q$, it can take the form

## IOULIA PAPAGEORGIOU

$$
\begin{align*}
\sum_{k=a}^{b 1}(k) & =\int_{a}^{b} r(x) d x+\left.\sum_{k=1}^{4} \frac{B_{k}}{k!} r^{(k 1)}(x)\right|_{a} ^{b}+R_{4} \\
& =\int_{a}^{b} r(x) d x+\left.\left\{\frac{1}{2} r(x)+\frac{1}{12} r^{(1)}(x)\right\}\right|_{a} ^{b} \tag{11}
\end{align*}
$$

This last expression is an equality and not an approximation because $R_{4}=0$ since $r(x)$ is a polynomial of third order and therefore $r^{(4)}(x)$. Moreover $B_{3}=0$ and also $r^{(3)}(x)$ is constant, not depending on $x$, and therefore it adds to zero when evaluated at the two ends of the interval. The only limitation for the exact equivalent and not an approximate expression is $r(x)=\rho(x)$ for all the discrete points between $a$ and $b$. In other words, $r$ has to be a continuous extension of $\rho(x)$. Under these conditions the error term in Euler-Maclaurin formula is zero and the two functions $Q$ and the corresponding continuous will coincide for all possible points of $\left(h_{0}, h_{1}, h_{2}, \ldots, h_{n}\right)$.

Summarizing, the steps of the proposed methodology in order to determine the optimal sampling allocation and inference about the population parameter are

Step 1. Use (11) for every summation in the second parenthesis of $Q$ in (7) and obtain the continuous equivalent expression given by

$$
\left.\begin{array}{rl}
Q^{c}\left(h_{0}, \ldots, h_{n}\right)= & \frac{N}{n}\left(\sum_{i=1}^{n-1} r\left(h_{i}\right)\right.
\end{array}+\sum_{i=1}^{n-2} r\left(h_{i}+h_{i+1}\right)+\cdots+r\left(h_{1}+\cdots+h_{n-1}\right)\right) ~ 子 \begin{aligned}
& h_{1} \\
&-\left(\int_{1}^{h_{0}+1} r(x) d x\right.+\int_{1}^{h_{0}+h_{1}+1} r(x) d x+\cdots+\int_{1}^{h_{0}+\cdots+h_{n-1}+1} r(x) \\
&+\int_{1}^{h_{1}+\cdots+h_{n}+1} r(x) d x+\int_{1}^{h_{2}+\cdots+h_{n}+1} r(x) d x \\
&+\cdots+\int_{1}^{h_{n}+1} r(x) d x+n r(1) \\
&-\frac{1}{2}\left[r\left(h_{0}+1\right)\right.+r\left(h_{0}+h_{1}+1\right) \\
&+\cdots+r\left(h_{0}+\cdots+h_{n-1}+1\right) \\
&+r\left(h_{n}+1\right)+r\left(h_{n}+h_{n-1}+1\right) \\
&\left.+\cdots+r\left(h_{n}+h_{n-1}+\cdots+h_{1}+1\right)\right]
\end{aligned}
$$

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

$$
\begin{aligned}
& +r^{(1)}\left(h_{0}+\cdots+h_{n-1}+1\right) \\
& +r^{(1)}\left(h_{n}+1\right)+r^{(1)}\left(h_{n}+h_{n-1}+1\right) \\
& \left.+\cdots+r^{(1)}\left(h_{n}+\cdots+h_{1}+1\right)\right]
\end{aligned}
$$

Step 2. Minimize $Q^{c}$ with respect to ( $h_{0}, h_{1}, h_{2}, \ldots, h_{n}$ ) and constrains (8), where $n$ is the sample size. Numerical constrained optimization can be used since function $Q^{c}$ is easily programmed. Function $r(h)$ is calculated by a cubic interpolation on the original discrete function $\rho(h)$.

Step 3. If $\boldsymbol{h}^{*}=\left(h_{0}^{*}, h_{1}^{*}, \ldots, h_{n}^{*}\right)$ is the vector where the minimum in step 2 is attained and $\left(\bar{h}_{0}, \bar{h}_{1}, \ldots, \bar{h}_{n}\right)$ is its closest integer vector, the optimal sample is the collection of units at positions

$$
\bar{l}_{1}=\bar{h}_{0}, \bar{l}_{2}=\bar{h}_{0}+\bar{h}_{1}+1, \bar{\iota}_{3}=\bar{h}_{0}+\bar{h}_{1}+\bar{h}_{2}+1, \ldots, \bar{l}_{n}=\bar{h}_{0}+\bar{h}_{1}+\ldots+\bar{h}_{n-1}+1
$$

Step 4. The mse of the population mean estimate calculated on the optimal sample $s=\left(Y_{\bar{l}_{1}}, Y_{\bar{l}_{2}}, \ldots, Y_{\bar{l}_{n}}\right)$ is derived from (5) by single substitution.

For a small numerical example, a set of simulated $N$ observations from a Moving Average (MA) process of order 2, with parameters -0.4 and 0.5 are assumed to represent the population. The autocorrelation function of the assumed MA model within the population range is listed in the first part of Table 1. The resulting set of values forming the population is $U=(-0.52,-1.33,0.19,1.70$, $-1.37,-1.35,-0.22,-0.16)$ and let the aim of the experiment to be the selection of a sample of size $n=3$ that minimizes the $m s e$. The set of all possible samples $P_{n}$ contains 56 samples, and in order to obtain the optimal $s=\left(h_{0}, h_{1}, h_{2}, h_{3}\right)$, the quantity $Q$ needs to be minimized with respect to ( $h_{0}, h_{1}, h_{2}, h_{3}$ ). Function $Q$ given by (7) for this example is

## IOULIA PAPAGEORGIOU

$$
\begin{aligned}
Q\left(h_{0}, h_{1}, h_{2}, h_{3}\right) & =\frac{N}{n}\left(\rho\left(h_{1}\right)+\rho\left(h_{2}\right) \rho\left(h_{1}+h_{2}\right)\right) \\
& -\left(\sum_{i=1}^{h_{0}} \rho(i)+\sum_{i=1}^{h_{0}+h_{1}} \rho(i)+\sum_{i=1}^{h_{0}+h_{1}+h_{2}} \rho(i)+\sum_{i=1}^{h_{1}+h_{2}+h_{3}} \rho(i)+\sum_{i=1}^{h_{2}+h_{3}} \rho(i)+\sum_{i=1}^{h_{3}} \rho(i)\right) \\
& +\int_{1}^{h_{1}+\ldots+h_{n}+1} r(x) d x+\int_{1}^{h_{2}+\ldots+h_{n}+1} r(x) d x+\cdots+\int_{1}^{h_{n}+1} r(x) d x
\end{aligned}
$$

and the corresponding $Q^{c}$ is

$$
\begin{aligned}
& Q^{c}\left(h_{0}, h_{1}, h_{2}, h_{3}\right)= \\
& =\frac{8}{3}\left(r\left(h_{1}\right)+r\left(h_{2}\right)+r\left(h_{1}+h_{2}\right)\right) \\
& \quad-\left(\int_{1}^{h_{0}+1} r(x) d x+\int_{1}^{h_{0}+h_{1}+1} r(x) d x+\int_{1}^{h_{0}+h_{2}+h_{3}+1} r(x) d x\right. \\
& \\
& +\int_{1}^{h_{1}+h_{2}+h_{3}+1} r(x) d x+\int_{1}^{h_{2}+h_{3}+1} r(x) d x+\int_{1}^{h_{3}+1} r(x) d x \\
& + \\
& +3 r(1)-\frac{1}{2}\left[r\left(h_{0}+1\right)+r\left(h_{0}+h_{1}+1\right)+r\left(h_{0}+h_{1}+h_{2}+1\right)\right. \\
& \\
& \left.\quad+r\left(h_{3}+1\right)+r\left(h_{3}+h_{2}+1\right)+r\left(h_{3}+h_{2}+h_{1}+1\right)\right] \\
& +\frac{1}{12}\left[r^{(1)}\left(h_{0}+1\right)+r^{(1)}\left(h_{0}+h_{1}+1\right)+r^{(1)}\left(h_{0}+h_{1}+h_{2}+1\right)\right. \\
& \\
& \quad+r^{(1)}\left(h_{3}+1\right)+r^{(1)}\left(h_{3}+h_{2}+1\right)+\cdots \\
& \\
& \\
& \left.\left.+\left(h_{3}+h_{2}+h_{1}+1\right)\right]-\frac{1}{2} r^{(1)}(1)\right)
\end{aligned}
$$

Note the complexity of $Q^{c}$ does not depend on $N$. The number of the unknowns and consequently the efficiency of the numerical minimization depends only on $n$. Sizes $N$ and $n$ have been chosen small in order to proceed in an exhaustive enumeration of all samples in $P_{n}$ and confirm both the approximation of $Q$ and its minimum. Minimizing $Q^{c}\left(h_{0}, h_{1}, h_{2}, h_{3}\right)$ yields $\left(h_{0}, h_{1}, h_{2}\right)=(0,1.80,1.91)$ and $h_{3}=N-1-\left(h_{0}, h_{1}, h_{2}\right)$. The closest discrete solution is $\left(h_{0}, h_{1}, h_{2}\right)=(0,2,2)$ and this corresponds to the sample $s=\left(Y_{1}, Y_{3}, Y_{5}\right)$.

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

Table 1a. Numerical example for a population with $N=8$ generated from MA(2)

| $\boldsymbol{i}$ | $\boldsymbol{Y} \boldsymbol{i}$ | $\boldsymbol{l a g} \boldsymbol{h}$ | $\boldsymbol{\rho}(\boldsymbol{h})$ | $\boldsymbol{r}(\boldsymbol{h})$ |
| :--- | ---: | ---: | ---: | ---: |
| 1 | -0.52 | 0 | 1.00 | 1.00 |
| 2 | -1.33 | 1 | -0.06 | -0.06 |
| 3 | 0.19 | 2 | -0.66 | -0.66 |
| 4 | 1.70 | 3 | 0.02 | 0.02 |
| 5 | -1.37 | 4 | 0.21 | 0.21 |
| 6 | -1.35 | 5 | 0.02 | 0.02 |
| 7 | -0.22 | 6 | -0.03 | -0.03 |
| 8 | -0.16 | 7 | 0.00 | 0.00 |

Table 1b. Numerical example for a population with $N=8$ generated from MA(2)

| sample | $\boldsymbol{Q}$ | Hermite $\boldsymbol{Q}^{\boldsymbol{c}}$ Spline $\boldsymbol{Q}^{\boldsymbol{c}}$ | sample | $\boldsymbol{Q}$ | Hermite $\boldsymbol{Q}^{\boldsymbol{c}}$ | Spline $\boldsymbol{Q}^{\boldsymbol{c}}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $(0,1,1)$ | 2.4533 | 2.4308 | 2.4531 | $(1,2,3)$ | 3.3185 | 3.2853 | 3.3069 |
| $(0,1,2)$ | 2.8326 | 2.8135 | 2.8284 | $(1,2,4)$ | 3.8756 | 3.8564 | 3.8714 |
| $(0,1,3)$ | 6.9570 | 6.9379 | 6.9528 | $(1,3,1)$ | 8.1837 | 8.1634 | 8.1842 |
| $(0,1,4)$ | 6.9570 | 6.9345 | 6.9568 | $(1,3,2)$ | 3.3185 | 3.2853 | 3.3069 |
| $(0,1,5)$ | 4.6993 | 4.6638 | 4.6870 | $(1,3,3)$ | 6.1985 | 6.1794 | 6.1943 |
| $(0,1,6)$ | 4.4978 | 4.4764 | 4.4929 | $(1,4,1)$ | 7.0637 | 7.0271 | 7.0561 |
| $(0,2,1)$ | 3.9526 | 3.9464 | 3.9605 | $(1,4,2)$ | 3.8756 | 3.8530 | 3.8754 |
| $(0,2,2)$ | 2.0089 | 2.0027 | 2.0168 | $(1,5,1)$ | 4.6993 | 4.6638 | 4.6870 |
| $(0,2,3)$ | 4.3319 | 4.3223 | 4.3437 | $(2,1,1)$ | 4.8000 | 4.7960 | 4.8085 |
| $(0,2,4)$ | 3.8756 | 3.8530 | 3.8754 | $(2,1,2)$ | 5.1793 | 5.1719 | 5.1918 |
| $(0,2,5)$ | 3.0104 | 3.0020 | 3.0176 | $(2,1,3)$ | 8.1837 | 8.1634 | 8.1842 |
| $(0,3,1)$ | 8.0770 | 8.0742 | 8.0809 | $(2,1,4)$ | 8.0770 | 8.0709 | 8.0849 |
| $(0,3,2)$ | 4.3319 | 4.3257 | 4.3397 | $(2,2,1)$ | 5.1793 | 5.1719 | 5.1918 |
| $(0,3,3)$ | 6.1985 | 6.1794 | 6.1943 | $(2,2,2)$ | 2.1156 | 2.0953 | 2.1160 |
| $(0,3,4)$ | 7.1348 | 7.1298 | 7.1380 | $(2,2,3)$ | 4.3319 | 4.3257 | 4.3397 |
| $(0,4,1)$ | 8.0770 | 8.0709 | 8.0849 | $(2,3,1)$ | 8.1837 | 8.1600 | 8.1182 |
| $(0,4,2)$ | 3.8756 | 3.8564 | 3.8714 | $(2,3,2)$ | 4.3319 | 4.3223 | 4.3437 |
| $(0,4,3)$ | 7.1348 | 7.1298 | 7.1380 | $(2,4,1)$ | 6.9570 | 6.9345 | 6.9568 |
| $(0,5,1)$ | 5.8193 | 5.7967 | 5.8191 | $(3,1,1)$ | 4.8000 | 4.7960 | 4.8085 |
| $(0,5,2)$ | 3.0104 | 3.0020 | 3.0176 | $(3,1,2)$ | 4.0593 | 4.0424 | 4.0557 |
| $(0,6,1)$ | 4.4978 | 4.4764 | 4.4929 | $(3,1,3)$ | 8.0770 | 8.0742 | 8.0809 |
| $(1,1,1)$ | 3.6800 | 3.6597 | 3.6805 | $(3,2,1)$ | 4.0593 | 4.0390 | 4.0597 |
| $(1,1,2)$ | 4.0593 | 4.0390 | 4.0597 | $(3,2,2)$ | 2.0089 | 2.0027 | 2.0168 |
| $(1,1,3)$ | 8.1837 | 8.1600 | 8.1882 | $(3,3,1)$ | 6.9570 | 6.9379 | 6.9528 |
| $(1,1,4)$ | 7.0637 | 7.0271 | 7.0561 | $(4,1,1)$ | 3.6800 | 3.6597 | 3.6805 |
| $(1,1,5)$ | 5.8193 | 5.7967 | 5.8191 | $(4,1,2)$ | 3.9526 | 3.9464 | 3.9605 |
| $(1,2,1)$ | 4.0593 | 4.0424 | 4.0557 | $(4,2,1)$ | 2.8326 | 2.8135 | 2.8284 |
| $(1,2,2)$ | 2.1156 | 2.0953 | 2.1160 | $(5,1,1)$ | 2.4533 | 2.4308 | 2.4531 |

Table 1 b provides a comparison of the arithmetic values of $Q$ and $Q^{c}$ for every sample in $P_{n}$. The 56 samples of $P_{n}$ consist of all possible vectors

## IOULIA PAPAGEORGIOU

( $h_{0}, h_{1}, h_{2}$ ) that fulfill constrains (8); $h_{3}=7-h_{0}-h_{1}-h_{2}$ and is not given. Two versions of $Q^{c}$ are presented for the example, the first one using piecewise cubic hermite interpolation to construct $r$, noted as Hermite $Q^{c}$, and the second using smooth spline, noted as Spline $Q^{c}$. The differences compared to the true function $Q$ are in the second decimal place, while the range of values is between 2.0089 and 8.1837. The differences among the function values are due to the use of numerical instead of analytical integration. It is also verified that the minimum $m s e$ value is achieved for the same sample $s=(0,2,2)$ for all methods, and agrees with the one derived from the numerical minimization. Since the optimal sample is found, the exact $m s e$ can be calculated from (5) and is 6.0267.

The smoothness characteristic of the spline $r(\cdot)$ improves the performance of the numerical integration and produces numerical values closer to the true ones. The smoothing parameter for the csaps routine, which has been used for this example, was chosen as 1 . This means that a priority to the exact matching of the spline values with the initial was given, rather than the smoothness.

## Experiments and Applications

## Experiments with Simulated Data

Three numerical examples follow, with simulated data generated from three different ARMA models to represent the population values under study. The justification for the ARMA model is that its autocorrelation function is general enough to cover a wide range of types for the serial correlation, depending on the specific values of their order and parameters. The aim of the experiment is to obtain the optimal sampling allocation following the proposed methodology, and compare its efficiency with other competitive sampling designs, chosen for either their broad use, because they are standard sampling designs, or because the literature suggests their application is appropriate to the case of correlated populations. More specifically, the sampling designs chosen are simple random sampling (srs), systematic sampling (sy), an optimal design for correlated populations with positive correlation, and Designs I and II proposed by Chao (2004) for correlated populations.

A range of values for the sample size is taken in every population case for a more complete view of the sampling design performance. The corresponding mean square errors of the estimates are calculated for all examined sampling designs by simulation and assuming normality. More specifically, if $K$ realizations from each population model are generated, and $\hat{\theta}_{j}^{d}$ is the estimate for the

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

population total at the $j^{\text {th }}$ realization according to the sampling design $d$, the mse for the estimate will be calculated by

$$
E_{d}(\theta-\hat{\theta})^{2}=\frac{1}{K} \sum_{j=1}^{K}\left(\theta_{j}-\hat{\theta}_{j}^{d}\right)^{2}
$$

The number of iterations for each experiment is 15,000 , while the common variance $\sigma^{2}$ is assumed unity. The optimal allocation of samples is derived by implementing the proposed methodology as previously described in Steps 1 to 5. For the numerical optimization, twenty different starting values have been used for each application and the smooth spline with $p=1$ has been used as the interpolation function of $\rho$. The performance of the examined sampling designs is evaluated by the relative efficiency to the srs, defined as the ratio of the mse obtained with a sampling design to that obtained with the srs. Values of relative efficiency greater than one indicate efficiency of the examined design.

Model 1. The population measurements are generated from an ARMA $(1,1)$ model with autocorrelation function plotted in Figure 1a. The degree of correlation is moderate for the assumed population occasion with the sign to alternate because of the negative sign of the parameter $\varphi$, the autoregressive part of the process. For population size $N=80$ and sample size that ranges from $n=3$ to $n=12$ the calculated efficiencies of the examined designs compared to srs are plotted in Figure 1b. For better illustration the reciprocal of the design effect is plotted in Figure 1b. Systematic, Design I and Design II are comparable with srs with respect to their mse, while the optimal allocation derived by the proposed methodology is clearly more efficient.

As a specific example, for $n=12$ the optimal sample determined from the solution of the numerical minimization problem is $s=\left[\begin{array}{llllllllll}1 & 2 & 3 & 59 & 60 & 61 & 62 & 63 & 78 & 79 \\ 80\end{array}\right]$. Sample $s$ has the sampling units separated into three groups, two groups located at the two ends of the population and one in the middle. Moreover neighbor units of the population are selected within the groups. Its mean square error by using simulation is 113.79 , and its exact value from expression (5) is 113.87 . Design II, the second best with respect to the mean square error in this example, has mse of 441.87.

## IOULIA PAPAGEORGIOU



Figure 1. Relative efficiencies using the empirical autocorrelation function on Model 1


Figure 2. Relative efficiencies using the theoretical autocorrelation function on Model 1

The empirical autocorrelation function calculated from the population of size $N=80$ has been used for implementing the methodology in this first example. If not the empirical but the exact autocorrelation function according to the assumed $\operatorname{ARMA}(1,1)$ model is used, the two resulting plots (corresponding to those in Figure 1) are presented in Figure 2. The sampling allocation according to the proposed method remains efficient. The assumed theoretical ARMA model has a negative $\varphi$ parameter as it can be seen from Figure 2a and the sign of the

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

ACF alternates. For such cases, the systematic sampling is far from the optimal, and plots in Figures 1b and 2b verify this result from the literature.

Model 2. $N=80$ is assumed for this example. The population vector, $U$, is generated from an $\operatorname{ARMA}(2,1)$ model with autocorrelation function plotted in Figure 3a. The serial correlation is not strong in this model, but the sign alternates and therefore it cannot be characterized as a positive, convex function. Following the same steps as in Model 1, the corresponding plot that presents the relative efficiencies of the sampling schemes under study with srs are presented in Figure 3b.

The optimal sample derived by the proposed methodology implemented here is the most efficient sample along all examined sample sizes, as shown in Figure 3 b . The three other samples proposed from the remaining techniques exhibit similar performance, faintly better if not comparable with the srs. The comparable to srs performance of sy is explained from the fact that the correlation between observations is low. It is known in sampling literature that sy and srs are equivalent with respect to accuracy when the population measurements do not present a trend or correlation (Cochran, 1977). Figure 3b demonstrates that the efficiency of the optimal sample is increased as the sample size increases.


Figure 3. Relative efficiencies using the autocorrelation function on Model 2

Model 3.
Assume $N=50$ and population values generated from an ARMA(2,4) model with autocorrelation function plotted in Figure 4a. The

## IOULIA PAPAGEORGIOU

autocorrelation in this model is not strong and alternates in sign. Again for sample size ranging between $n=3$ and $n=12$, the relative efficiencies are plotted in Figure 4b. The optimal sample as derived by the proposed methodology is clearly the sample with the minimum mse. Its relative efficiency is between 0.207 and 0.48 , indicating a significant improvement in accuracy with respect to srs sampling scheme. Among the remaining competitive designs, Design II compares better than the srs, although not consistently, followed by Design I and systematic sampling, which produce higher than the srs $m s e$ and are not appropriate for this population case.


Figure 4. Relative efficiencies using the autocorrelation function on Model 3

## An application in Statistical Process Control

Consider an application of the proposed methodology in Statistical Process Control (SPC) based on a real data set. The data include 204 consecutive measurements of electrical resistance of insulation in megaohms and was first presented in Shewhart (1931, p. 20). The data set often serves as a typical example in SPC, where the existing autocorrelation can lead to incorrect conclusions about a process if it has not been detected or handled properly. The implementation of sampling in SPC happens during the construction of the statistical charts, which aim to provide some warning limits for the production line and detect a deviation in mean or variance of the process. Many forms of statistical charts are available, but the common basis for any chart is a sample
taken at an initial stage from the production line. Shewhart's control chart is one of the best known statistical control charts, and its basic components are presented during this application. Any statistical control chart can be evaluated by calculating the expected probability of false positive or negative alarms.

Shewhart's control chart of the $\bar{X}$, the mean of a sample taken from the process, was originally constructed for the electrical measurements and presented by taking successive groups of four. The resulting 51 subsamples were used to estimate the mean and the variance of the population towards the construction of the upper and lower control limits. The control limits provide a reference interval for a mean of a sample of four selected from the process if this is in control. The two limits are in mathematical terms

$$
\left[\bar{x}-3 * \frac{\hat{\sigma}}{\sqrt{k}}, \bar{x}+3 * \frac{\hat{\sigma}}{\sqrt{k}}\right]
$$

where $\bar{x}$ is the mean of the means of the 51 subsamples and is used as an estimate of the population mean, $\hat{\sigma}$ is the estimate of $\sigma$, the square root of the process units variance, and $k$ is the size of the sub-samples. For $k=4$ and the total of 51 subsamples in the data, the resulting control limits for the mean of the process are plotted in Figure 5, solid line. The means of the 51 samples are plotted together in the same Figure, and a large percentage of those means are outside the limits an indication that the process is not in control. The process is however in statistical control, as subsequent analyses of the same data set concluded (see for example, Alwan and Roberts, 1995). The variation that the data exhibit can be explained from the present autocorrelation not taken into account in the first application, and is not due to a special cause.

Yang and Hancock (1990) introduced the autocorrelation into the calculation of the control limits for Shewhart's control chart. The new control limits suggested by their methodology are given by

$$
\rho=\frac{\sum_{i \neq j} r_{i j}}{k(k-1)}
$$

where $r_{i j}$ are the $i, j$ elements of matrix $R$ if assumed that the variance covariance matrix $\boldsymbol{V}$ of a sample can be written as $\boldsymbol{V}=\sigma^{2} R$. Implementing this approach for the electrical measurements and using all 51 samples of four, the

## IOULIA PAPAGEORGIOU

resulting control limits are wider, as expected, and include all 51 sub-sample means, as can be seen in Figure 5 (dotted line).

Alternatively, the control limits can be calculated by implementing the methodology proposed in Steps 1-4. The implementation is possible in both stages of sampling. For the first stage of sub-samples of four, formula (5) can be used to estimate the variance of the sample mean. The resulting estimate is more accurate than the average correlation $\rho$ because the exact matrix $\boldsymbol{V}$ according to the model, and not an average $\rho$, is used. For the second stage of the 51 subsamples, either complete enumeration or sampling is possible. Sampling is more realistic in practice and can also be applied to continuous processes. Both scenarios are presented here using the proposed methodology to choose the sample in the case of sampling at the level of sub-samples. Moreover, in a real situation application, a sample instead of successive measurements could also be the case for the first stage of SPC.

A first-order autoregressive model has been fitted to the data with parameter $\varphi$ equal to 0.549 (Alwan \& Roberts, 1995), and this is the model used for the implementation. When all sub-samples are taken into account and the variance of the mean with sub-sample is calculated by (5), the resulting control limits are plotted in Figure 5 (dashed line). The use of the exact form of the model that describes the population units allows control limits that are wider than in the first analysis, but not as much as according to Yang and Hancock methodology. Note that too wide control limits lead to an increase of the probability of falsity in control conclusions.

If a sample of seven sub-samples is selected according to the proposed methodology, and the estimates of the mean as well as their standard errors in both stages are calculated from expression (4) and (5) respectively, the resulting control limits are plotted in Figure 5 (dash-dot line), and compare closely to the ones derived from the complete population of $N=51$ sub-samples.

Therefore, identifying the model correctly and fully incorporating this information in the selection of samples procedure and the statistical inference allows us to accurately construct control limits using only 28 measurements instead of 204.

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS



Figure 5. Control limits for electrical measurements.

The optimal sample of size seven for this application was found not to be equally spaced. An equally spaced scheme in SPC, also called fixed distance sampling, corresponds to a systematic design and is often the choice for selecting the sub-samples during the second stage for SPC applications, especially in cases where a positive correlation is detected. However, it has been verified that variable distance outperforms fixed distance sampling. The comparison has been conducted with simulation studies that calculate the average time to signal (ATS), a measure of efficiency of control charts. The advantage of variable distance sampling depends on the degree and type of correlation (Prybutok, et al., 2007).

Within the same framework, other models, more general than the AR, can also be treated with the proposed methodology.

## Conclusion

A continuous approach was proposed for an intractable otherwise discrete optimization problem with primary application in sampling. During the process of

## IOULIA PAPAGEORGIOU

sampling from correlated populations, the specific type of the autocorrelation function $\rho(h)$ among the populations' units affects both the choice of the sample and the inference about the population parameters. If $\rho(h)$ has certain properties, such as constant, positive, decreasing, convex, etc., it is possible to derive conclusions about the optimal sampling designs even if $\rho(h)$ is not known in its exact form. In cases of a more general type of correlation (for example, a realization of a time series process), characterizing the optimal sampling designs or the class of the optimal samples is not possible and the results depend closely on the specific type of $\rho(h)$. A feasible and accurate way of deriving a sample that belongs to the class of optimal samples in such cases is proposed here. The estimate with its mse is also provided. The proposed technique uses continuous approximation of a finite sum from an integral. A continuous interpolation function $r(h)$ based on $\rho(h)$ is an important component for its implementation, and when $r(h)$ holds certain properties it is shown that the proposed approach is not an approximation but exact.

The method can be used in any case of correlated population, or not. It is fast, easily programmed and implemented, and computationally efficient. The dimensionality coincides with the sample size and therefore the computational efficiency remains unaffected from the population size. As a general approach, it can find applications in other than sampling context and facilitate the solution of a mathematical problem that depends on a function with a discrete nature.

The benefit for estimation is significant. Ignoring or incorrectly specifying the existing correlation within a population set can lead to misleading results, especially regarding the accuracy of the derived parameter estimate. The proposed methodology suggests a more sophisticated and informative sampling procedure, specialized for the population under study. This specialization has been incorporated into the mse calculation of the assumed estimator and the minimum $m s e$ is the criterion for the sampling procedure derivation. Therefore, the suggested sample is optimal with respect to the accuracy of the resulting estimate, and the improvement in mse is significant when compared to other known and widely used sampling schemes. Moreover, the simulation experiments suggest that the inclusion of the population model towards the correct calculation of the $m s e$ is necessary, and has a considerable impact on efficiency even if a small degree correlation occurs. Finally, the actual arithmetic value of both the estimate and its exact $m s e$ implemented for the optimal sampling allocation are provided.

The extension of the proposed methodology to continuous stationary processes is straightforward. The assumption of other than the least squared estimator is also possible. The least squared estimator for the population

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

parameter has been considered here because of its frequent use in practice, due to its simplicity and ease of implementation. Assumption of the best linear unbiased or the best unbiased estimators are some possible extensions, along with the assumptions of model (2). The constant mean parameter $\mu$ may be assumed dependent on population unit $i$. Under this model the least squared estimator (4) is not unbiased for the population total. The bias depends on the sample $s$, but not on the type of autocorrelation. The new expression of the estimator's mse needs to be minimized following a procedure similar to that proposed here.

## References

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. Communications in Statistics - Theory and Methods, 21(4), 1025-1049. doi: 10.1080/03610929208830829

Alwan, L. C. \& Roberts, H. V. (1995). The problem of misplaced control limits. Applied Statistics, 44(3), 269-278.

Alwan, L. C. \& Roberts, H. V. (1988). Time-series modeling for statistical process control. Journal of Business and Economic Statistics, 6(1), 87-95. doi: 10.1080/07350015.1988.10509640

Apley, D. W. \& Lee H. C. (2003). Design of exponentially weighted moving average control charts for autocorrelated processes With model uncertainty. Technometrics, 45(3), 187-198. doi: 10.1198/0040170030000000014

Arnab, R. (1992). Estimation of a finite population mean under superpopulation models. Communications in Statistics - Theory and Methods, 21(6), 1717-1724. doi: 10.1080/03610929208830874

Bellhouse, D. R. (1984). A review of optimal designs in survey sampling. Canadian Journal of Statistics, 12(1), 53-65. doi: 10.2307/3314724

Bellhouse, D. R. \& Rao, J. N. K. (1975). Systematic sampling in the presence of a trend. Biometrika, 62(3), 694-697. doi: 10.1093/biomet/62.3.694

Bjørnstad, O. N., Ims, R. A., \& Lambin, X. (1999). Spatial population dynamics: Analysing patterns and processes of population synchrony. Trends in Ecology and Evolution, 14(11), 427-431. 10.1016/S0169-5347(99)01677-8

Bolfarine, H. \& Zacks, S. (1992). Prediction theory for finite population. NY: Springer Varlag.

Blight, B. J. N. (1973). Sampling from an autocorrelated finite population. Biometrika, 60(2), 375-385. doi: 10.1093/biomet/60.2.375

## IOULIA PAPAGEORGIOU

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). Foundations of inference in survey sampling. New York: Wiley.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. Annals of Mathematical Statistics, 17(2), 164-177. doi: 10.1214/aoms/1177730978

Cochran, W. G. (1953). Sampling techniques, $1^{\text {st }}$ Ed. NY: Wiley \& Sons
Cochran, W. G. (1977). Sampling techniques, $3^{\text {rd }}$ Ed. NY: Wiley \& Sons.
Cox, D. R. \& Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. Biometrika, 91(3), 729-737. doi: 10.1093/biomet/91.3.729

Chao, C-T. (2004). Selection of sampling units under a correlated population based on the eigensystem of the population covariance matrix. Environmetrics, 15(8), 757-775. doi: 10.1002/env. 655

Cressie, N. A. (1993). Statistics for spatial data, Revised version. NY: Wiley.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 17(2), 269-278.

Graham, R. L., Knuth, D. E. \& Patashnik, O. (1994) Concrete mathematics, 2nd Ed. Boston, MA: Addison-Wesley.

Graubard, B. I. \& Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. Statistical Science, 17(1), 73-96.

Harris, T. J. \& Ross, W. H. (1991). Statistical process control procedures for correlated observations, Canadian Journal of Chemical Engineering, 69(1), 48-57. doi: 10.1002/cjce. 5450690106

Hjort, N. L. \& Varin, C. (2008). ML, PL and QL for Markov chain models. Scandinavian Journal of Statistics, 35(1), 64-82. doi: 10.1111/j.14679469.2007.00559.x

Karakostas, K. X. (1984). Optimum systematic sampling for autocorrelated superpopulations. (Unpublished thesis). Imperial College, University of London.

Lande, R. (1991). Isolation by distance in a quantitative trait. Genetics, 128(2), 443-453.

Lu, C. W. \& Reynolds, M. R. (1999). Control charts for monitoring the mean and variance of autocorrelated processes. Journal of Quality Technology, 31(3), 259-274.

Lu, C. W. \& Reynolds, M. R. (2001). CUSUM control charts with variable sample sizes and sampling internals. Journal of Quality Technology, 33(1), 66-81.

## OPTIMAL SAMPLE ALLOCATION FOR CORRELATED POPULATIONS

Madow, L. H. \& Madow, W. G. (1944). On the theory of systematic sampling, I. Annals of Mathematical Statistics, 15(1), 1-24. doi: 10.1214/aoms/1177731312

Mastrangelo, C. M. \& Montgomery, D. C. (1995). SPC with correlated observations for the chemical and process industries. Quality and Reliability Engineering International, 11(2), 79-89. doi: 10.1002/qre. 4680110203

Montgomery, D. C. \& Mastrangelo, C. M. (1991) Some statistical process control chart methods for autocorrelated data. Journal of Quality Technology, 23(3), 179-193.

Mukerjee, R. \& Sengupta, S. (1989). Optimal estimation of finite population total under a general correlated model. Biometrika, 76(4), 789-794. doi: 10.1093/biomet/76.4.789

Mukerjee, R. \& Sengupta, S. (1990). Optimal estimation of finite mean in the presence of linear trend. Biometrika, 77(3), 625-630. doi: 10.1093/biomet/77.3.625

Nayak, T. K. (2003). Finding optimal estimators in survey sampling using unbiased estimators of zero. Journal of Statistical Planning and Inference, 114(12), 21-30. doi: 10.1016/S0378-3758(02)00460-3

Neuhaus, J. M. \& Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics, 54(2), 638-645. doi: 10.2307/3109770

Papageorgiou, I., \& Karakostas, K. X. (1998). On optimal sampling designs for autocorrelated finite populations. Biometrika, 85(2), 482-486. doi: 10.1093/biomet/85.2.482

Prybutok, V. R., Clayton H. R., \& Harvey, M. M. (1997). Comparison of fixed versus variable sampling interval Shewhart $\bar{X}$ control charts in the presence of positively autocorrelated data. Communications in Statistics - Simulation and Computation, 26(1), 83-106. doi: 10.1080/03610919708813369

Ramakrishnan, M. K. (1975). Choice of an optimum sampling strategy-1. Annals of Statistics, 3(3), 669-679. doi: 10.1214/aos/1176343129

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. Biometrika, 57(2), 377-387. doi:
10.1093/biomet/57.2.377

Shewhart, W. A. (1931). The economic control of manufactured products. NY: Van Nostrand.

## IOULIA PAPAGEORGIOU

Tam, S. M. (1984). Optimal estimation in survey sampling under a regression superpopulation model. Biometrika, 71(3), 645-647. doi:
10.1093/biomet/71.3.645

Varin, C. (2008). On composite marginal likelihoods. Advances in Statistical Analysis, 92(1), 1-28. doi: 10.1007/s10182-008-0060-7

Wardell, D. G., Moskowitz, H. \& Plante, R. D. (1992) Control charts in presence of data correlation. Management Science, 38(8), 1084-1105. doi: 10.1287/mnsc.38.8.1084

Watson, G. S. (1972). Trent-surface analysis and spatial correlation. Geological Society of America Special Papers, 146, 39-46. doi: 10.1130/SPE146p39

Yang, K., \& Hancock, W. M. (1990). Statistical quality control for correlated samples. International Journal of Production Research, 28(3), 595-608. doi: 10.1080/00207549008942738

Yates, F. (1960). Sampling methods for censuses and surveys, 3rd Ed. London, England: Griffin.

# Doubly Censored Data from Two-Component Mixture of Inverse Weibull Distributions: Theory and Applications 

Tabassum Naz Sindhu<br>Quaid-i-Azam University Islamabad, Pakistan

Navid Feroze<br>Ripha International University<br>Islamabad, Pakistan

Muhammad Aslam<br>Ripha International University Islamabad, Pakistan

Finite mixture distributions consist of a weighted sum of standard distributions and are a useful tool for reliability analysis of a heterogeneous population. They provide the necessary flexibility to model failure distributions of components with multiple failure models. The analysis of the mixture models under Bayesian framework has received sizable attention in the recent years. However, the Bayesian estimation of the mixture models under doubly censored samples has not yet been introduced in the literature. The main objective of this paper is to discuss the Bayes estimation of the inverse Weibull mixture distributions under doubly censoring. Different priors and loss functions were assumed for the posterior estimation. The performance of the different estimators has been compared in terms of posterior risks.

Keywords: Inverse transformation method, mixture model, doubly censoring, loss functions, Bayes estimator

## Introduction

In survival analysis, data are subject to censoring. The most common type of censoring is right censoring, in which the survival time is larger than the observed right censoring time. In some cases, however, data are subject to left as well as right censoring. When left censoring occurs, the only information available to an analyst is that the survival time is less than or equal to the observed left censoring time. A more complex censoring scheme is found when both initial and final times are interval-censored. This situation is referred as double censoring, and the data with both right and left censored observations are known as doubly censored data.

Tabassum Sindhu is in the Department of Statistics. Email them at:
sindhuqau @gmail.com. Navid Feroze is in the Department of Statistics and is a lecturer in the Department of Statistics at the Government Post Graduate College Muzaffarabad, Azad Kashmir, Pakistan. Email them at: navidferoz@gmail.com. Muhammad Aslam is in the Department of Statistics. Email him at: aslamsdqu @yahoo.com.

## SINDHU ET AL.

Analysis of doubly censored data for simple (single) distribution has been studied by many authors. Fernandez (2000) investigated maximum likelihood prediction based on type-II doubly censored exponential data. Fernandez (2006) has discussed Bayesian estimation based on trimmed samples from Pareto populations. Khan, Provost, and Singh (2010) studied predictive inference from a two-parameter Rayleigh life model given a doubly censored sample. Kim and Song (2010) have discussed Bayesian estimation of the parameters of the generalized exponential distribution from doubly censored samples. Khan, Albatineh, Alshahrani, Jenkins, and Ahmed (2011) studied sensitivity analysis of predictive modeling for responses from the three-parameter Weibull model with a follow-up doubly censored sample of cancer patients. Pak, Parham, and Saraj (2013) proposed the estimation of Rayleigh scale parameter under doubly type-II censoring from imprecise data.

A mixture distribution is signified as a convex fusion of other probability distributions. It can be used to model a statistical population with subpopulations, where the constituents of mixture probability densities are the densities of the subpopulations. Mixture distribution may appropriately be used for certain data sets where the subsets of the whole data set possess different properties that can best be modeled separately. They can be more mathematically manageable, because the individual mixture components are dealt with more ease than the overall mixture density. The families of mixture distributions have a wider range of applications in different fields such as fisheries, agriculture, botany, economics, medicine, psychology, electrophoresis, finance, communication theory, geology, and zoology.

Soliman (2006) derived estimators for the finite mixture of Rayleigh model based on progressively censored data. Sultan, Ismail, and Al-Moisheer (2007) have discussed some properties of the mixture of two inverse Weibull distributions. Saleem and Aslam (2008) presented a comparison of the Maximum Likelihood (ML) estimates with the Bayes estimates assuming the Uniform and the Jeffreys priors for the parameters of the Rayleigh mixture. Kundu and Howalder (2010) considered the Bayesian inference and prediction of the inverse Weibull distribution for type-II censored data. Saleem, Aslam, and Economou (2010) considered the Bayesian analysis of the mixture of Power function distribution using the complete and the censored sample. Shi and Yan (2010) studied the case of the two parameter exponential distribution under type-I censoring to get empirical Bayes estimates. Eluebaly and Bouguila (2011) have presented a Bayesian approach to analyze finite generalized Gaussian mixture models which incorporate several standard mixtures, widely used in signal and image processing applications, such as Laplace and Gaussian. Sultan and Al-Moisheer (2012)

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

developed approximate Bayes estimation of the parameters and reliability function of mixture of two inverse Weibull distributions under type-II censoring.

## Model and Likelihood Function

If the probability density function (pdf) of the Weibull distribution is

$$
\mathrm{f}_{i}\left(y_{i j}, \theta, \tau\right)=\theta_{i} \tau_{i} y_{i j}^{\theta_{i j}-1} \exp \left(-\tau_{i} y_{i j}^{\theta_{i}}\right)
$$

with $y_{i j}>0, i=1,2$, and $j=1,2, \ldots, n_{i}$, then the random variable $x_{i j}=1 / y_{i j}$ has the inverse Weibull distribution with pdf

$$
\begin{equation*}
\mathrm{f}_{i}\left(x_{i j}, \theta, \tau\right)=\theta_{i} \tau_{i} x_{i j}^{-\left(\theta_{i}+1\right)} \exp \left(-\tau_{i} x_{i j}^{-\theta_{i}}\right) \tag{1}
\end{equation*}
$$

with $x_{i j}>0, i=1,2$, and $j=1,2, \ldots, n_{i}$, and where $\theta_{i}>0$ and $\tau_{i}>0$ are shape and scale parameters, respectively.

The cumulative distribution function (cdf) of the distribution is

$$
\begin{equation*}
\mathrm{F}_{i}\left(x_{i j}, \theta, \tau\right)=\exp \left(-\tau_{i} x_{i j}^{-\theta_{i}}\right), \quad x_{i j}, \theta_{i}, \tau_{i}>0, i=1,2, j=1,2, \ldots, n_{i} \tag{2}
\end{equation*}
$$

A density function for the mixture of two components densities with mixing weights $\left(p_{1}, 1-p_{1}\right)$ is given by

$$
\begin{equation*}
\mathrm{f}(x)=p_{1} \mathrm{f}_{1}(x)+\left(1-p_{1}\right) \mathrm{f}_{2}(x), \quad 0<p_{1}<1 \tag{3}
\end{equation*}
$$

The cdf for the mixture model is:

$$
\begin{equation*}
\mathrm{F}(x)=p_{1} \mathrm{~F}_{1}(x)+\left(1-p_{1}\right) \mathrm{F}_{2}(x) \tag{4}
\end{equation*}
$$

Consider a random sample of size $n$ from the inverse Weibull distribution, and let $x_{r}, x_{r+1}, \ldots, x_{s}$ be the ordered observations that can only be observed. The remaining $r-1$ smallest observations and the $n-s$ largest observations have been assumed to be censored. Now based on causes of failure, the failed items are assumed to come either from subpopulation 1 or from subpopulation 2; so the $x_{1 r_{1}}, \ldots x_{1 s_{1}}$ and $x_{2 r_{2}}, \ldots x_{2 s_{2}}$ failed items come from first and second subpopulations, respectively.

## SINDHU ET AL.

The rest of the observations which are less than $x_{r}$ and greater than $x_{s}$ have been assumed to be censored from each component, where $x_{s}=\max \left(x_{1, s_{1}}, x_{2, s_{2}}\right)$ and $x_{r}=\min \left(x_{1, r_{1}}, x_{2, r_{2}}\right)$. Therefore, $m_{1}=s_{1}-r_{1}+1$ and $m_{2}=s_{2}-r_{2}+1$ number of failed items can be observed from first and second subpopulation, respectively. The remaining $n-(s-r+2)$ items are assumed to be censored observations, and $s-r+2$ are the uncensored items, where $r=r_{1}+r_{2}, s=s_{1}+s_{2}$, and $m=m_{1}+m_{2}$. Then the likelihood function for the type-II doubly censored sample $\mathbf{x}=\left\{\left(x_{1 r_{1}}, \ldots, x_{1 s_{1}}\right),\left(x_{2 r_{2}}, \ldots, x_{2 s_{2}}\right)\right\}$, assuming the causes of the failure of the left censored items are identified, can be written as

$$
\begin{align*}
\mathrm{L}\left(\tau_{1}, \tau_{2}, p_{1} \mid \mathbf{x}\right) \propto\{ & \left.\mathrm{F}_{1}\left(x_{\left(r_{1}\right)}, \tau_{1}\right)\right\}^{r_{1}-1}\left\{\mathrm{~F}\left(x_{\left(r_{2}\right)}, \tau_{2}\right)\right\}^{r_{2}-1}\left\{1-\mathrm{F}\left(x_{s}, \tau_{1}, \tau_{2}\right)\right\}^{n-2} \\
\times & \left.\times \prod_{i=r_{1}}^{s_{1}} p_{1} \mathrm{f}_{1}\left(x_{1(i)}, \tau_{1}\right)\right\}\left\{\prod_{i=r_{2}}\left(1-p_{1}\right) \mathrm{f}_{2}\left(x_{2(i)}, \tau_{2}\right)\right\}  \tag{5}\\
\mathrm{L}\left(\tau_{1}, \tau_{2}, p_{1} \mid \mathbf{x}\right) \propto & \left\{\exp \left(-\tau_{1} x_{\left(r_{1}\right)}^{-\theta_{1}}\right)\right\}^{s_{2}-1}\left\{\exp \left(-\tau_{2} x_{\left(r_{2}\right)}^{-\theta_{2}}\right)\right\}^{r_{2}-1} \\
& \times\left\{1-p_{1} \exp \left(-\tau_{1} x_{\left(r_{1}\right)}^{-\theta_{1}}\right)-p_{2} \exp \left(-\tau_{2} x_{\left(r_{2}\right)}^{-\theta_{2}}\right)\right\}^{n-2} \\
& \times\left\{\prod_{i=r_{1}}^{s_{1}} p_{1} \theta_{1} \tau_{1} x_{1(i)}^{-\left(\theta_{1}+1\right)} \exp \left(-\tau_{1} x_{1(i)}^{-\theta_{1}}\right)\right\}  \tag{6}\\
& \times\left\{\prod_{i=r_{2}}^{s_{2}}\left(1-p_{1}\right) \theta_{2} \tau_{2} x_{2(i)}^{-\left(\theta_{2}+1\right)} \exp \left(-\tau_{2} x_{2(i)}^{-\theta_{2}}\right)\right\}
\end{align*}
$$

Assuming the shape parameter to be known, the likelihood function (6) reduces to

$$
\begin{align*}
\mathrm{L}\left(\tau_{1}, \tau_{2}, p_{1} \mid \mathbf{x}\right) \propto & \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} p_{1}^{m_{1}+k_{1}-k_{2}}\left(1-p_{1}\right)^{m_{2}+k_{2}}  \tag{7}\\
& \times \tau_{1}^{m_{1}} \tau_{2}^{m_{2}} \exp \left\{-\tau_{1}\left(\gamma_{1}\left(x_{1 j}\right)\right)\right\} \exp \left\{-\tau_{2}\left(\gamma_{2}\left(x_{2 j}\right)\right)\right\}
\end{align*}
$$

where

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

$$
\begin{gathered}
\gamma_{1}\left(x_{1 j}\right)=\sum_{i=r_{1}}^{s_{1}} x_{1(i)}^{-\theta_{1}}+\left(k_{1}-k_{2}\right) x_{(s)}^{-\theta_{1}}+\left(r_{1}-1\right) x_{\left(r_{1}\right)}^{-\theta_{1}} \\
\gamma_{2}\left(x_{2 j}\right)=\sum_{i=r_{2}}^{s_{2}} x_{2(i)}^{-\theta_{2}}+\left(k_{2}\right) x_{(s)}^{-\theta_{2}}+\left(r_{2}-1\right) x_{\left(r_{2}\right)}^{-\theta_{2}}
\end{gathered}
$$

## Bayes Estimation

The simple estimation of the scale parameter often pre-assumes the knowhow of the shape parameter (for more detail, see Panaitescu, George, Cozma, \& Popa, 2010; Zanakis, 1979; Kundu \& Howaldar, 2010; Shi \& Yan, 2010; etc.). For the Bayesian estimation, let us assume that the parameters $\tau_{i}, i=1,2$, and $p_{1}$ are independent random variables, and then consider the following priors for different parameters.

## Bayesian Estimation using Conjugate Prior

The prior for the rate parameters $\tau_{i}$ for $i=1,2$, is assumed to be the gamma distribution, with the hyperparameters $a_{i}$ and $b_{i}$ given by

$$
\begin{equation*}
\mathrm{f}_{\tau_{i}}\left(\tau_{i}\right)=\frac{b_{i}^{a_{i}}}{\Gamma\left(a_{i}\right)} \tau_{i}^{a_{i}-1} \exp \left(-\tau_{i} b_{i}\right), \quad a_{i}, b_{i}>0 \tag{8}
\end{equation*}
$$

The prior for $p_{1}$ is the beta distribution, whose density is given by

$$
\begin{equation*}
\mathrm{f}_{p}\left(p_{1}\right)=\frac{\Gamma\left(c_{1}+d_{1}\right)}{\Gamma\left(c_{1}\right) \Gamma\left(d_{1}\right)} p_{1}^{c_{1}-1}\left(1-p_{1}\right)^{d_{1}-1}, \quad c_{1}, d_{1}>0 \tag{9}
\end{equation*}
$$

From equations (8)-(9), the following joint prior density of the vector $\boldsymbol{\Theta}=\left(\tau_{1}, \tau_{2}, p_{1}\right)$ is proposed:

$$
\begin{equation*}
\mathrm{g}(\boldsymbol{\Theta}) \propto \tau_{i}^{a_{i}-1} \exp \left(-\tau_{i} b_{i}\right) p_{1}^{c_{1}-1}\left(1-p_{1}\right)^{d_{1}-1}, \quad 0<p_{1}<1, a_{i}, b_{i}, c_{1}, d_{1}>0 \tag{10}
\end{equation*}
$$

By multiplying equation (10) and equation (7), the joint posterior density for the vector $\Theta$, given the data, becomes

## SINDHU ET AL.

$$
\begin{align*}
\pi(\boldsymbol{\Theta} \mid x)= & \Omega_{1}^{-1} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} p_{1}^{m_{1}+k_{1}-k_{2}+c_{1}-1}\left(1-p_{1}\right)^{m_{2}+k_{2}+d_{1}-1}  \tag{11}\\
& \times \exp \left\{-\tau_{i}\left(\xi_{i}\left(x_{i j}\right)\right)\right\}
\end{align*}
$$

where

$$
\begin{aligned}
\Omega_{1}= & \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \\
& \times \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{\xi_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}
\end{aligned}
$$

and $\xi_{i}\left(x_{i j}\right)=\gamma_{i}\left(x_{i j}\right)+b_{i}$ for $i=1$, 2. Marginal distributions of $\tau_{i}, i=1,2$, and $p_{1}$ can be obtained by integrating the nuisance parameters.

## Bayesian Estimation using Inverse Levy Prior

The prior for the rate parameters $\tau_{i}$ for $i=1,2$, is assumed to be the inverse Levy distribution, with hyperparameter $v_{i}$, given by

$$
\begin{equation*}
\mathrm{f}_{\tau_{i}}\left(\tau_{i}\right)=\sqrt{\frac{v_{i}}{2 \pi}} \tau_{i}^{-1 / 2} \exp \left(\frac{-\tau_{i} v_{i}}{2}\right), \quad v_{i}>0 \tag{12}
\end{equation*}
$$

The prior for $p_{1}$ is the beta distribution, whose density is given by

$$
\begin{equation*}
\mathrm{f}_{p}\left(p_{1}\right)=\frac{\Gamma\left(c_{2}+d_{2}\right)}{\Gamma\left(c_{2}\right) \Gamma\left(d_{2}\right)} p_{1}^{c_{2}-1}\left(1-p_{1}\right)^{d_{2}-1}, \quad c_{2}, d_{2}>0 \tag{13}
\end{equation*}
$$

From equation (12)-(13), we propose the following joint prior density of the vector $\boldsymbol{\Theta}=\left(\tau_{1}, \tau_{2}, p_{1}\right)$ :

$$
\begin{equation*}
\mathrm{g}(\boldsymbol{\Theta}) \propto \tau_{i}^{-1 / 2} \exp \left(\frac{-\tau_{i} v_{i}}{2}\right) p_{1}^{c_{2}-1}\left(1-p_{1}\right)^{d_{2}-1}, \quad 0<p_{1}<1, v_{i}, c_{2}, d_{2}>0 \tag{14}
\end{equation*}
$$

By multiplying equation (14) with equation (7), the joint posterior density for the vector $\boldsymbol{\Theta}$, given the data, becomes

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

$$
\begin{align*}
\pi(\boldsymbol{\Theta} \mid x)= & \Omega_{2}^{-1} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} p_{1}^{m_{1}+k_{1}-k_{2}+c_{2}-1}\left(1-p_{1}\right)^{m_{2}+k_{2}+d_{2}-1}  \tag{15}\\
& \times \exp \left\{-\tau_{i}\left(\psi_{i}\left(x_{i j}\right)\right)\right\}
\end{align*}
$$

where

$$
\begin{aligned}
\Omega_{2}= & \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{2}, m_{2}+k_{2}+d_{2}\right) \\
& \times \prod_{i=1}^{2} \frac{\Gamma\left(m_{i}+1 / 2\right)}{\left\{\psi_{i}\left(x_{i j}\right)\right\}^{\left(m_{i}+1 / 2\right)}}
\end{aligned}
$$

and $\psi_{i}\left(x_{i j}\right)=\gamma_{i}\left(x_{i j}\right)+v_{i} / 2$. Marginal distributions of $\tau_{i}, i=1,2$, and $p_{1}$ can be obtained by integrating the nuisance parameters.

## Bayes Estimation of the Vector of Parameters ©

The Bayesian point estimation is connected to a loss function in general, signifying the loss is induced when the estimate $\hat{\theta}$ differs from the true parameter $\theta$. Because there is no specific rule that helps to identify the appropriate loss function to be used, we can use the K-loss function (KLF), which is particularized as

$$
1(\hat{\theta}, \theta)=\frac{(\hat{\theta}-\theta)^{2}}{\hat{\theta} \theta}
$$

is proposed by Wasan (1970), and is well-fitted for a measure of inaccuracy for an estimator of a scale parameter of a distribution defined on $\mathbb{R}^{+}=(0, \infty)$. The Bayes estimator and posterior risk under KLF are $\hat{\theta}=\left\{\mathrm{E}(\theta \mid \mathbf{x}) / \mathrm{E}\left(\theta^{-1} \mid \mathbf{x}\right)\right\}^{-1 / 2}$ and $\rho(\hat{\theta})=2\left\{\mathrm{E}(\theta \mid \mathbf{x}) \mathrm{E}\left(\theta^{-1} \mid \mathbf{x}\right)-1\right\}$, respectively. In Bayesian analysis, a widely used loss function is the quadratic loss function given by $1(\hat{\theta}, \theta)=w(\hat{\theta}-\theta)^{2}$; if $w=1$, it reduces to the squared error loss function (SELF) and, for $w=\theta^{-2}$, it becomes $1(\hat{\theta}, \theta)=(\hat{\theta}-\theta)^{2} / \theta^{2}$. This is known as the minimum expected loss function

## SINDHU ET AL.

(MELF), and is introduced by Tummala and Sathe (1978) in their study. The Bayes estimator and posterior risk under MELF are $\hat{\theta}=\mathrm{E}\left(\theta^{-1} \mid \mathbf{x}\right) / \mathrm{E}\left(\theta^{-2} \mid \mathbf{x}\right)$ and $\rho(\hat{\theta})=1-\left\{\mathrm{E}\left(\theta^{-1} \mid \mathbf{x}\right)\right\}^{2} / \mathrm{E}\left(\theta^{-2} \mid \mathbf{x}\right)$, respectively.

The respective marginal distribution of each parameter is used to derive the Bayes estimators and posterior risks of $\tau_{1}, \tau_{2}$, and $p_{1}$ under KLF and MELF. The Bayes estimators and their posterior risks of the parameters $\tau_{1}, \tau_{2}$, and $p_{1}$ for the conjugate (gamma and beta) priors using the KLF and MELF functions are given in the Appendix. Thus, expressions for Bayes estimators and their posterior risks under the inverse Levy can be obtained with little alteration.

## Elicitation

The elicitation of opinion is a crucial step. It helps to make it easy for us to understand what the experts believe in, and what their opinions are. In statistical inference, the characteristics of a certain predictive distribution proposed by an expert determine the hyperparameters of a prior distribution. In this article, we focused on a method of elicitation based on prior predictive distribution. The elicitation of hyperparameters from the prior $p(\lambda)$ is a difficult task. The prior predictive distribution is used for the elicitation of the hyperparameters, which are compared with the experts' judgment about this distribution and then the hyperparameters are chosen in such a way so as to make the judgment agree as closely as possible with the given distribution. Readers desiring more detail may refer to: Grimshaw, Collings, Larsen, and Hurt (2001), O’Hagan et al. (2006), Jenkinson (2005) and Leon, Vazquez-Polo, and Gonzalez (2003). According to Aslam (2003), the preferred method of elicitation is to compare the prior predictive distribution with experts' assessment about this distribution, and then to choose the hyperparameters that make the assessment agree closely with the member of the family. The prior predictive distributions under all the priors are derived using the following formula:

$$
\mathrm{p}(y)=\int_{\boldsymbol{\Theta}} \mathrm{p}(y \mid \boldsymbol{\Theta}) \mathrm{p}(\boldsymbol{\Theta}) d \boldsymbol{\Theta}
$$

## Elicitation under Gamma Distribution

The prior predictive distribution using gamma prior is

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

$$
\begin{equation*}
\mathrm{p}(y)=\frac{\theta_{1} y^{-\left(\theta_{1}+1\right)} a_{1} b_{1}^{a_{1}} c_{1}}{\left(c_{1}+d_{1}\right)\left(y^{-\theta_{1}}+b_{1}\right)^{\left(a_{1}+1\right)}}+\frac{\theta_{2} y^{-\left(\theta_{2}+1\right)} a_{2} b_{2}^{a_{2}} d_{1}}{\left(c_{1}+d_{1}\right)\left(y^{-\theta_{2}}+b_{2}\right)^{\left(a_{2}+1\right)}}, \quad y>0 \tag{16}
\end{equation*}
$$

Assume $\left(\theta_{1}, \theta_{2}\right)=(1,1)$ for convenience in calculations. For the elicitation of the six hyperparameters, six different intervals are considered. From equation (16), the experts' probabilities/assessments are supposed to be 0.10 for each case. The six integrals for equation (16) are considered with the following limits of the values of random variable $\mathrm{Y}:(0,10),(10,20),(20,30),(30,40),(40,50)$, and $(50,60)$ respectively. For the elicitation of hyperparameters, $a_{1}, a_{2}, b_{1}, b_{2}, c_{1}$, and $d_{1}$, these six integrals are solved simultaneously through computer program developed in SAS package using the command of PROC SYSLIN. Thus the values of hyperparameters obtained by applying this methodology are: $a_{1}=4.982587$, $a_{2}=3.356211, b_{1}=0.987542, b_{2}=0.46523, c_{1}=1.45987$, and $d_{1}=0.05690$.

## Elicitation under Inverse Levy Prior

$$
\mathrm{p}(y)=\frac{\theta_{1} y^{-\left(\theta_{1}+1\right)}\left(\frac{v_{1}}{2}\right)^{\frac{1}{2}} c_{1}}{2\left(c_{1}+d_{1}\right)\left(\frac{y^{-\theta_{1}}+v_{1}}{2}\right)^{\frac{v_{1}+1}{2}+1}}+\frac{\theta_{2} y^{-\left(\theta_{2}+1\right)}\left(\frac{v_{2}}{2}\right)^{\frac{1}{2}} d_{2}}{2\left(c_{1}+d_{1}\right)\left(\frac{y^{-\theta_{2}}+v_{2}}{2}\right)^{\frac{v_{2}}{2}+1}}, \quad y>0
$$

Now, to elicit four hyperparameters, consider the four integrals. The expert probabilities are assumed to 0.15 for each integral with the following limits of the values of random variable $\mathrm{Y}:(0,15),(15,30),(30,45)$, and $(45,60)$. Using a similar kind of program as discussed above, we have the following values of the hyperparameters: $v_{1}=0.062138, v_{2}=0.19136, c_{2}=0.895777$, and $d_{2}=0.63889$.

## Simulation Study and Comparisons

A simulation study was conducted to compare the performance of the discussed estimators on the basis of generated samples from the inverse Weibull mixture distribution using doubly censored data. Assume $\left(\theta_{1}, \theta_{2}\right)=(1,1)$ for convenience in calculations. Take random samples of sizes $n=20,40$, and 80 from the two component mixture of inverse Weibull distributions with following choice of parametric values: $\left(\tau_{1}, \tau_{2}\right) \in\{(0.1,0.15),(10,15),(0.1,15),(10,0.15)\}, p_{1}=0.45$ and 0.6. To develop a mixture data, we adopt the probabilistic mixing model with

## SINDHU ET AL.

probability $p_{1}$ and $\left(1-p_{1}\right)$. A uniform number $u$ is generated $n$ times and, if $u<p_{1}$, the observation is taken randomly from $F_{1}$ (the inverse Weibull distribution with parameter $\tau_{1}$ ), and is otherwise taken from $F_{2}$ (the inverse Weibull with parameter $\tau_{2}$ ). Hence, the parameters to be estimated are known to be ( $\tau_{1}, \tau_{2}$ ) and $p_{1}$. The choice of the censoring time is made in such a way that the censoring rate in the resultant sample is to be approximately $20 \%$. The simulated data sets have been obtained using following steps:

Step 1: Draw samples of size $n$ from the mixture model
Step 2: Generate a uniform random number $u$ for each observation
Step 3: If $u \leq \pi$, take the observation from first subpopulation; otherwise, take the observation from the second subpopulation

Step 4: Determine the test termination points on left and right, that is, determine the values of $x_{r}$ and $x_{s}$

Step 5: The observations which are less than $x_{r}$ and greater than $x_{s}$ have been considered to be censored from each component

Step 6: Use the remaining observations from each component for the analysis

To avoid an extreme sample, simulate 10,000 data sets, each of size $n$. The Bayes estimates and posterior risks (in parenthesis) are computed using Mathematica 8.0. The average of these estimates and corresponding risks are reported in Tables 1-8. The abbreviations used in the tables are: BEs: Bayes estimators; PRs: Posterior risks; GP: Gamma prior; ILP: Inverse Levy prior.

The simulation study has revealed some interesting properties of the Bayes estimates. It is worth mentioning that in each case the posterior risks of estimates of lifetime parameters are decreasing as the sample size increases. The posterior risks of the estimates of $\tau_{1}, \tau_{2}$ have been assessed to be quite large when the values of the parameters are large, and entirely small for rather smaller values of $\tau_{1}, \tau_{2}$. Another interesting point regarding the posterior risks of the estimates of parameters $\tau_{1}, \tau_{2}$ is that by increasing (decreasing) the proportion of the component in mixture reduces (increases) the posterior risk of the concerned $\tau$ parameter's estimate.

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

Table 1. BEs and their PRs under GP for $\left(T_{1}, T_{2}, p_{1}\right)=(0.10,0.15,0.45)$ and (0.10, 0.15, 0.60)

K-Loss Function

|  | $n$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | 0.153042 | 0.217287 | 0.488886 | 0.149162 | 0.243873 | 0.652455 |
| 20 | $(0.161568)$ | $(0.166297)$ | $(0.118884)$ | $(0.127883)$ | $(0.227595)$ | $(0.060474)$ |
| 40 | 0.130631 | 0.181089 | 0.461140 | 0.126142 | 0.188513 | 0.635182 |
|  | $(0.101929)$ | $(0.091817)$ | $(0.069768)$ | $(0.076355)$ | $(0.131665)$ | $(0.034437)$ |
| 80 | 0.113720 | 0.171546 | 0.449263 | 0.115099 | 0.182363 | 0.627186 |
|  | $(0.074710)$ | $(0.063162)$ | $(0.049079)$ | $(0.054635)$ | $(0.092332)$ | $(0.024224)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| 20 | 0.152631 | 0.194781 | 0.445046 | 0.136376 | 0.200732 | 0.621717 |
|  | $(0.080798)$ | $(0.083651)$ | $(0.066112)$ | $(0.064041)$ | $(0.114609)$ | $(0.033749)$ |
| 40 | 0.123116 | 0.167329 | 0.447311 | 0.118357 | 0.168123 | 0.618551 |
|  | $(0.051022)$ | $(0.046079)$ | $(0.036942)$ | $(0.038252)$ | $(0.066219)$ | $(0.018267)$ |
| 80 | 0.113790 | 0.161134 | 0.447937 | 0.113935 | 0.162226 | 0.610625 |
|  | $(0.037331)$ | $(0.031706)$ | $(0.025748)$ | $(0.027277)$ | $(0.046496)$ | $(0.012683)$ |

Table 2. BEs and their PRs under GP for $\left(\tau_{1}, \tau_{2}, p_{1}\right)=(10,15,0.45)$ and $(10,15,0.60)$

|  | K-Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ |
| 20 | 7.2322300 | 11.9032000 | 0.4851680 | 7.8576700 | 10.4070000 | 0.6564870 |
|  | $(0.1628030)$ | $(0.165620)$ | $(0.1206010)$ | $(0.1274380)$ | $(0.2306880)$ | $(0.0594520)$ |
| 40 | 8.0121000 | 13.7528000 | 0.4556190 | 8.7621200 | 12.0339000 | 0.6369690 |
|  | $(0.1029490)$ | $(0.0908861)$ | $(0.0709860)$ | $(0.0763040)$ | $(0.1328160)$ | $(0.0342290)$ |
| 80 | 8.4481100 | 14.0172700 | 0.4465120 | 8.7865800 | 12.9782000 | 0.6284630 |
|  | $(0.0750960)$ | $(0.0628280)$ | $(0.0493037)$ | $(0.0546180)$ | $(0.0929030)$ | $(0.0217830)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
|  | $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ |  |

## SINDHU ET AL.

Table 3. BEs and their PRs under GP for $\left(\tau_{1}, \tau_{2}, p_{1}\right)=(0.10,15,0.45)$ and $(0.10,15,0.60)$
K-Loss Function

| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 20 | 0.1533410 | 12.5483000 | 0.4483640 | 0.1397540 | 11.7884000 | 0.5905870 |
|  | $(0.1669040)$ | $(0.1504220)$ | $(0.1273990)$ | $(0.1334790)$ | $(0.1951910)$ | $(0.0685210)$ |
| 40 | 0.1193940 | 14.5209000 | 0.4489700 | 0.1107460 | 13.6061000 | 0.5978991 |
|  | $(0.1053590)$ | $(0.0823310)$ | $(0.0740830)$ | $(0.0800540)$ | $(0.1096810)$ | $(0.0388500)$ |
| 80 | 0.1114640 | 15.0405000 | 0.4511250 | 0.1057960 | 14.6865000 | 0.5986610 |
|  | $(0.0771020)$ | $(0.0565920)$ | $(0.0432290)$ | $(0.0580370)$ | $(0.0777460)$ | $(0.0048650)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\boldsymbol{1}}$ |  |
| 20 | 0.140090 | 11.354900 | 0.404051 | 0.133802 | 10.261200 | 0.567283 |  |
|  | $(0.083452)$ | $(0.075235)$ | $(0.070580)$ | $(0.066740)$ | $(0.097732)$ | $(0.037961)$ |  |
| 40 | 0.112806 | 13.171700 | 0.419673 | 0.109543 | 12.681400 | 0.567551 |  |
|  | $(0.052679)$ | $(0.041176)$ | $(0.039065)$ | $(0.040027)$ | $(0.054874)$ | $(0.020488)$ |  |
| 80 | 0.108045 | 14.175500 | 0.429351 | 0.103915 | 13.796700 | 0.587920 |  |
|  | $(0.038552)$ | $(0.028369)$ | $(0.031335)$ | $(0.028531)$ | $(0.038394)$ | $(0.022886)$ |  |

Table 4. BEs and their PRs under GP for $\left(\tau_{1}, T_{2}, p_{1}\right)=(10,0.15,0.45)$ and $(10,0.15,0.60)$
K-Loss Function

|  | K-Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| 20 | 7.789440 | 0.206080 | 0.544287 | 8.052190 | 0.224464 | 0.695652 |
|  | $(0.144196)$ | $(0.176093)$ | $(0.086478)$ | $(0.118203)$ | $(0.239330)$ | $(0.044531)$ |
| 40 | 8.918560 | 0.166512 | 0.522777 | 8.909610 | 0.175136 | 0.681657 |
|  | $(0.087638)$ | $(0.098245)$ | $(0.049039)$ | $(0.069216)$ | $(0.139309)$ | $(0.024854)$ |
| 80 | 9.274560 | 0.155907 | 0.515036 | 9.687610 | 15.652800 | 0.652686 |
|  | $(0.062971)$ | $(0.068130)$ | $(0.033541)$ | $(0.049070)$ | $(0.098428)$ | $(0.001594)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ |
| 20 | 7.007170 | 0.175346 | 0.507474 | 7.329860 | 0.187362 | 0.671009 |
|  | $(0.072105)$ | $(0.088049)$ | $(0.047976)$ | $(0.059115)$ | $(0.119666)$ | $(0.024700)$ |
| 40 | 8.392620 | 0.148824 | 0.503235 | 8.052989 | 0.155773 | 0.668620 |
|  | $(0.043817)$ | $(0.049123)$ | $(0.025863)$ | $(0.034608)$ | $(0.069655)$ | $(0.013108)$ |
| 80 | 8.850450 | 0.151859 | 0.495015 | 9.424450 | 0.151359 | 0.661397 |
|  | $(0.031496)$ | $(0.034063)$ | $(0.017918)$ | $(0.024634)$ | $(0.049023)$ | $(0.024360)$ |

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

Table 5. BEs and their PRs under ILP for $\left(T_{1}, T_{2}, p_{1}\right)=(0.10,0.15,0.45)$ and (0.10, 0.15, 0.60)

K-Loss Function

|  | $n$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | 0.107446 | 0.174887 | 0.454348 | 0.108630 | 0.172018 | 0.620560 |
| 20 | $(0.256954)$ | $(0.215458)$ | $(0.136335)$ | $(0.180738)$ | $(0.336696)$ | $(0.069258)$ |
| 40 | 0.104352 | 0.164206 | 0.44143 | 0.104008 | 0.154266 | 0.618750 |
|  | $(0.133538)$ | $(0.104548)$ | $(0.075409)$ | $(0.133544)$ | $(0.104585)$ | $(0.036939)$ |
| 80 | 0.098973 | 0.158185 | 0.436433 | 0.102810 | 0.151531 | 0.617152 |
|  | $(0.090341)$ | $(0.068909)$ | $(0.045525)$ | $(0.062284)$ | $(0.106620)$ | $(0.025056)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| 20 | 0.092775 | 0.147587 | 0.407706 | 0.097211 | 0.134592 | 0.586746 |
|  | $(0.128863)$ | $(0.108963)$ | $(0.075846)$ | $(0.090791)$ | $(0.171009)$ | $(0.038768)$ |
| 40 | 0.096182 | 0.147725 | 0.416953 | 0.099533 | 0.142006 | 0.600375 |
|  | $(0.066809)$ | $(0.052609)$ | $(0.039818)$ | $(0.046432)$ | $(0.081512)$ | $(0.019661)$ |
| 80 | 0.096554 | 0.149812 | 0.429139 | 0.102130 | 0.146210 | 0.600153 |
|  | $(0.045175)$ | $(0.034586)$ | $(0.027256)$ | $(0.031126)$ | $(0.053686)$ | $(0.014347)$ |

Table 6. BEs and their PRs under ILP for $\left(\tau_{1}, \tau_{2}, p_{1}\right)=(10,15,0.45)$ and $(10,15,0.60)$
K-Loss Function

| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 20 | 9.985290 | 14.431300 | 0.456402 | 10.696500 | 13.981300 | 0.625004 |
|  | $(0.255990)$ | $(0.216752)$ | $(0.135572)$ | $(0.179341)$ | $(0.341694)$ | $(0.068025)$ |
| 40 | 10.643800 | 14.798400 | 0.443693 | 10.480710 | 14.656400 | 0.620334 |
|  | $(0.132816)$ | $(0.105192)$ | $(0.074823)$ | $(0.092251)$ | $(0.163078)$ | $(0.036711)$ |
| 80 | 10.122700 | 14.845100 | 0.453762 | 10.174900 | 14.854300 | 0.617783 |
|  | $(0.090007)$ | $(0.069232)$ | $(0.051825)$ | $(0.062220)$ | $(0.106884)$ | $(0.025106)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{1}$ |
| 20 | 9.277410 | 11.474100 | 0.412225 | 9.321530 | 9.642290 | 0.593112 |
|  | $(0.127923)$ | $(0.110266)$ | $(0.074962)$ | $(0.089785)$ | $(0.173529)$ | $(0.037767)$ |
| 40 | 9.637820 | 14.223600 | 0.417554 | 9.502120 | 12.713300 | 0.601931 |
|  | $(0.066774)$ | $(0.052724)$ | $(0.039748)$ | $(0.046313)$ | $(0.081862)$ | $(0.019531)$ |
| 80 | 9.729790 | 14.560200 | 0.428610 | 9.999100 | 13.616000 | 0.601586 |
|  | $(0.045118)$ | $(0.034705)$ | $(0.027784)$ | $(0.031143)$ | $(0.053736)$ | $(0.013452)$ |

## SINDHU ET AL.

Table 7. BEs and their PRs under ILP for $\left(T_{1}, \tau_{2}, p_{1}\right)=(0.10,15,0.45)$ and (0.10, 15, 0.60)

K-Loss Function

|  | K-Lsser |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ |
| 20 | 0.0995060 | 15.8176000 | 0.4191220 | 0.1019610 | 16.6349000 | 0.5696700 |
|  | $(0.2666630)$ | $(0.1911270)$ | $(0.1435660)$ | $(0.1904710)$ | $(0.2689990)$ | $(0.0772190)$ |
| 40 | 0.0957830 | 15.7349000 | 0.4315360 | 0.0972550 | 15.9842000 | 0.5698800 |
|  | $(0.1379300)$ | $(0.0932430)$ | $(0.0788790)$ | $(0.0975590)$ | $(0.1298730)$ | $(0.0413270)$ |
| 80 | 0.0925177 | 15.3503000 | 0.4450500 | 0.0931070 | 15.5486000 | 0.5765170 |
|  | $(0.0929980)$ | $(0.0616730)$ | $(0.0448390)$ | $(0.0655840)$ | $(0.0856100)$ | $(0.0268270)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
|  | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ |  |
| $\mathbf{n}$ | 0.078252 | 14.237100 | 0.372592 | 0.089637 | 12.991000 | 0.535129 |  |
|  | $(0.133332)$ | $(0.095557)$ | $(0.079529)$ | $(0.095236)$ | $(0.134672)$ | $(0.042776)$ |  |
| 40 | 0.084029 | 14.501970 | 0.380824 | 0.090933 | 14.232200 | 0.545625 |  |
|  | $(0.068965)$ | $(0.046618)$ | $(0.041602)$ | $(0.048779)$ | $(0.064944)$ | $(0.021795)$ |  |
| 80 | 0.086764 | 14.687900 | 0.403030 | 0.091828 | 15.480600 | 0.548745 |  |
|  | $(0.046508)$ | $(0.030839)$ | $(0.030867)$ | $(0.032782)$ | $(0.042834)$ | $(0.019410)$ |  |

Table 8. BEs and their PRs under ILP for $\left(T_{1}, T_{2}, p_{1}\right)=(10,0.15,0.45)$ and (10, 0.15, 0.60)

K-Loss Function

|  | K-Loss Function |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{\boldsymbol{r}}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ |
| 20 | 11.639900 | 0.143830 | 0.516704 | 10.857400 | 0.143796 | 0.667442 |
|  | $(0.212306)$ | $(0.235279)$ | $(0.096294)$ | $(0.160525)$ | $(0.363626)$ | $(0.050583)$ |
| 40 | 11.407900 | 0.144247 | 0.508186 | 10.697180 | 0.148710 | 0.665794 |
|  | $(0.108873)$ | $(0.114281)$ | $(0.051895)$ | $(0.081860)$ | $(0.173910)$ | $(0.026553)$ |
| 80 | 10.967200 | 0.143250 | 0.501879 | 10.568890 | 0.149423 | 0.636676 |
|  | $(0.073551)$ | $(0.075455)$ | $(0.014772)$ | $(0.054934)$ | $(0.114285)$ | $(0.017568)$ |


|  | Minimum Expected Loss Function |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
|  | $\boldsymbol{n}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{2}}$ | $\hat{\boldsymbol{p}}_{\mathbf{1}}$ | $\hat{\boldsymbol{r}}_{\mathbf{1}}$ | $\hat{\boldsymbol{f}}_{\mathbf{2}}$ |  |

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

It was observed that for the relatively smaller value of $\tau$, i.e. $(0.10,0.15)$, the performance of the minimum expected loss function and the gamma prior is better than their counterparts, as the amounts of posterior risks are smaller than those in case of their counterparts. However, the inverse Levy prior produces some closer estimates to the true value of parameters. Estimates of mixing proportion are found to be underestimated using inverse Levy prior when $p_{1}=0.45$, but they are pretty good under gamma prior. When we consider the estimation of comparatively larger value of $\tau$, i.e. $(10,15)$, again under estimation is observed of the estimates of parameters under both priors and loss functions. But the extent of underestimation is higher under the minimum expected loss function using gamma prior. Nonetheless, this underestimation is due to the random procedure and is tolerable.

Further, this problem can be faced off by using lager sample sizes. As far as the efficiency of the prior is concerned, gamma is found to be the efficient than inverse Levy prior. Moreover, on assessing the behavior of estimates, in the case of the extremely different value of the parameters $\left(\tau_{1}<\tau_{2}\right.$ and $\left.\tau_{1}>\tau_{2}\right)=(0.10,15$ and $10,0.15$ ), i.e. one is small and other is hundred fold large, it is noticed that the parameters are once again underestimated, and this underestimation is higher at every point using the minimum expected loss function under both priors. However, the use of the K-loss function has exhibited pretty good estimates with few exceptions (in terms of convergence). In general, the estimates under gamma prior using the minimum expected loss function are the best, as the amounts of posterior risks associated with these estimates are the least in almost all cases.

## Real Data Analysis

Real data sets are considered to illustrate the methodology discussed in previous sections. In order to show the usefulness of the proposed mixture model, consider survival times (in days) of guinea pigs, injected with different doses of tubercle bacilli, in Table 9. This data set was discussed by Kundu and Howlader (2010). Singh, Singh, and Sharma (2013) also analyzed this data set. The regimen number is the common logarithm of the number of bacillary units in 0.5 mL of challenge solution; e.g., regimen 6.6 corresponds to $4.0 * 10^{6}$ bacillary units per 0.5 mL . Corresponding to regimen 6.6, there are 72 observations listed below. Further, the Kolmogorov-Smirnov and chi-square tests are used to see if the data follow the inverse Weibull distribution. These tests say that the data follow the inverse Weibull distribution at $5 \%$ level of significance with $p$-values 0.1361 and 0.1290 , respectively. We have assumed $\left(\theta_{1}, \theta_{2}\right)=(1,1)$ for convenience in calculations.

## SINDHU ET AL.

Table 9. Survival times (in days) of guinea pigs injected with different doses of tubercle bacilli

| 12 | 15 | 22 | 24 | 24 | 32 | 32 | 33 | 34 | 38 | 38 | 43 | 44 | 48 | 52 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 53 | 54 | 54 | 55 | 56 | 57 | 58 | 58 | 59 | 60 | 60 | 60 | 60 | 61 | 62 |
| 63 | 65 | 65 | 67 | 68 | 70 | 70 | 72 | 73 | 75 | 76 | 76 | 81 | 83 | 84 |
| 85 | 87 | 91 | 95 | 96 | 98 | 99 | 109 | 110 | 121 | 127 | 129 | 131 | 143 | 146 |
| 146 | 175 | 211 | 233 | 99 | 258 | 258 | 263 | 297 | 341 | 341 | 376 |  |  |  |

Consider the case when the data are doubly Type II censored. Data are randomly grouped into two sets when $p_{1}=0.45$. It is assumed that we observe 33 data points belonging to population I and 39 data points belonging to population II. To implement censored samplings, the $x_{1 r_{1}}, \ldots, x_{1 s_{1}}$ and $x_{2 r_{2}}, \ldots, x_{2 s_{2}}$ failed items come from the first and second subpopulations, respectively. The rest of the observations, which are less than $x_{r}$ and greater than $x_{s}$, have been assumed to be censored from each component. Here, $m_{1}=s_{1}-r_{1}+1$ and $m_{2}=s_{2}-r_{2}+1$ numbers of failed items can be observed from the first and second subpopulations, respectively. The remaining $n-(s-r+2)$ items are assumed to be censored observations, and $s-r+2$ are the uncensored items, where $r=r_{1}+r_{2}, s=s_{1}+s_{2}$, and $m=m_{1}+m_{2}$. The detail of the censored mixture data can be found in Table 10.

The following characteristics are extracted from the censored data for the analysis of the mixture model:

$$
\begin{aligned}
& p_{1}=0.45, n=72, r=8, r_{1}=4, r_{2}=43, s=64, s_{1}=29, s_{2}=35, n_{1}=33, n_{2}=39 \\
& \theta_{1}=\theta_{2}=0.5, x_{r_{1}}=32, x_{s_{1}}=233, x_{r_{2}}=33, x_{s_{2}}=258 \\
& \sum_{i=r_{1}}^{s_{1}} x_{1(i)}^{-\theta_{1}}=3.21314, \sum_{i=r_{2}}^{s_{2}} x_{2(i)}^{-\theta_{2}}=3.85409
\end{aligned}
$$

Similar methodology was employed when

$$
\begin{aligned}
& p_{1}=0.60, n=72, r=8, r_{1}=5, r_{2}=3, s=64, s_{1}=39, s_{2}=25, n_{1}=44, n_{2}=28 \\
& \theta_{1}=\theta_{2}=0.5, x_{r_{1}}=33, x_{s_{1}}=211, x_{r_{2}}=32, x_{s_{2}}=175 \\
& \sum_{i=r_{1}}^{s_{1}} x_{1(i)}^{-\theta_{1}}=4.16450, \sum_{i=r_{2}}^{s_{2}} x_{2(i)}^{-\theta_{2}}=3.21392
\end{aligned}
$$

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

Table 10. Doubly-censored mixture real life data

| Population I |  |  |  |  |  |  | Population II |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 61 | 12 | 24 | 60 | 24 | 32 | 65 | 15 | 131 | 87 | 143 | 91 | 95 | 175 |
| 34 | 68 | 38 | 43 | 67 | 72 | 48 | 110 | 121 | 127 | 297 | 341 | 60 | 62 |
| 54 | 73 | 76 | 55 | 81 | 83 | 58 | 65 | 63 | 70 | 96 | 211 | 98 | 258 |
| 84 | 233 | 341 | 263 | 146 | 175 | 129 | 258 | 70 | 75 | 76 | 59 | 60 | 57 |
| 146 | 109 | 99 | 35 | 376 |  |  | 56 | 58 | 53 | 54 | 44 | 52 | 43 |
|  |  |  |  |  |  |  | 38 | 33 | 32 | 22 |  |  |  |

Table11. BEs and their PRs under minimum expected loss function and K-loss function for the real data set

| Priors | K-loss function |  |  | Minimum expected loss function |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $p_{1}=0.45$ | $\boldsymbol{r}_{1}$ | $\hat{\boldsymbol{r}}_{2}$ | $\hat{\boldsymbol{p}}_{1}$ | $\hat{r}_{1}$ | $\boldsymbol{r}_{2}$ | $\hat{p}_{1}$ |
| Gamma | 7.023900 | 7.914180 | 0.453725 | 6.699360 | 7.600860 | 0.439455 |
|  | (0.062637) | (0.053542) | (0.041482) | (0.031384) | (0.026819) | (0.021459) |
| Inverse | 7.613170 | 7.918130 | 0.446087 | 7.206180 | 7.583200 | 0.431593 |
| Levy | (0.072641) | (0.058103) | (0.042864) | (0.036424) | (0.029113) | (0.022179) |
| $p_{1}=0.60$ | $\hat{r}_{1}$ | $\boldsymbol{r}_{2}$ | $\hat{p}_{1}$ | $\boldsymbol{r}_{1}$ | $\boldsymbol{r}_{2}$ | $\hat{p}_{1}$ |
| Gamma | 7.400650 | 6.984160 | 0.610524 | 7.142880 | 6.603080 | 0.600336 |
|  | (0.047031) | (0.074187) | (0.021878) | (0.023548) | (0.037188) | (0.011324) |
| Inverse | 7.923470 | 6.899140 | 0.602689 | 7.616030 | 6.478070 | 0.592309 |
| Levy | (0.052462) | (0.083158) | (0.022581) | (0.026276) | (0.041710) | (0.011689) |

The results in Table 11 indicate that the Bayes estimates under gamma prior are better than those under inverse levy prior under both loss functions. Similarly, in the comparison of the loss functions, it has been assessed that the performance of the minimum expected loss function is better than the K-loss function. The larger values of the mixing parameter $\left(p_{1}\right)$ impose a positive impact on the performance of the estimation of the first component of the mixture. Hence the analysis of reallife data endorsed the findings of the simulation study, suggesting the preference of gamma prior along with minimum expected loss function.

## SINDHU ET AL.

## Graphical Representation of Posterior Risks under Different Loss Functions, Various Priors

Risks of the estimators are empirically evaluated based on a Monte-Carlo simulation study of samples. A number of values of unknown parameters are considered. Sample size is varied to observe the effect of small and large samples on the estimators. Different combinations of parameters are considered in studying the change in the estimators and their risks. The results are summarized in Figures $1-4$. The risk of the estimators will be a function of sample size, population parameters, and hyperparameters of the prior distribution. After an extensive study of the results, the conclusions are drawn regarding the behavior of the estimators, which are summarized below. (Due to space restrictions, all results are not shown in the graphs.) As sample size increases, the risk of all the estimators decrease, as indicated in Figures 1-4. The effect of variation of parameters on the risks of the estimator has also been studied. The risk of the estimators increases when the value of parameters increases.


Figure 1. Posterior risks of $T_{1}$ for $\left(\tau_{1}, T_{2}, p_{1}\right)=(0.10,0.15,0.45)$

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS



Figure 2. Posterior risks of $T_{2}$ for $\left(\tau_{1}, T_{2}, p_{1}\right)=(0.10,0.15,0.45)$


Figure 3. Posterior risks of $T_{1}$ for $\left(\tau_{1}, T_{2}, p_{1}\right)=(10,15,0.45)$

## SINDHU ET AL.



Figure 4. Posterior risks of $T_{2}$ for $\left(T_{1}, T_{2}, p_{1}\right)=(10,15,0.45)$

## Conclusion

The Bayesian inference of inverse Weibull mixture distribution based on doubly type-II censored data was considered. The prior belief of the model is represented by the independent gamma, beta priors and inverse Levy, beta priors on the scale, and mixing proportion parameters. Numerical results of the simulation study presented in Tables 1-8 exposed salient properties of the proposed Bayes estimators. The parameters of the mixture distributions have been over/under estimated in different cases. In general, the larger values of the parameters have been overestimated and smaller values of the parameters have been underestimated in the majority of cases. However, it is nice to observe that the estimated values converge to the true values and the amounts of the posterior risks tend to decrease by increasing the sample size.

This indicates that the proposed estimators are consistent. The smaller (larger) values of the parameter representing one component of the mixture impose a positive (negative) impact on the estimation of the parameter representing the other component of the mixture distribution. The larger values of the mixing parameter $\left(p_{1}\right)$ impose a positive impact on the performance of the estimation of the first component of the mixture. This may be due to the fact that the lager values of the mixing parameter incorporate more values for the analysis of the first component.

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

Bayes estimators performed better under the minimum expected loss function than under the K-loss function under both priors. In addition, the performance of the estimates under gamma prior is better than those under inverse levy prior using both loss functions. However, in the case of gamma prior, the estimates under both loss functions are comparatively more underestimated, though this problem is less severe in the larger samples. Therefore, on the basis of the above discussion, we can recommend the use of the gamma prior under minimum expected loss function for the analysis of the inverse Weibull mixture distribution under the Bayesian framework.

However, when such a mixture model was used in real-life, the prior may be chosen as well as the loss function according to the need. In case of loss functions, if lower posterior risk is desired than in the present scenario, the minimum expected loss function should be given importance. If compromise on risk is affordable then one can easily select to use the K-loss function. Also, the informative gamma prior can easily be preferred over the other informative prior as shown by results. It may be mentioned here that, because of space restriction, only selected results are included and presented graphically. The findings of real life example are in accordance with the simulation study. The findings of the paper are useful for the analysts (from different fields) in dealing with the Bayesian analysis of the time to failure data when causes of the failure are more than one, and the data is doubly censored.

## References

Aslam, M. (2003). An application of prior predictive distribution to elicit the prior density. Journal of Statistical Theory and Applications, 2(1), 70-83.

Eluebaly, T., \& Bougila, N. (2011). Bayesian learning of finite generalized Gaussian mixture models on images. Signal Processing, 91(4), 801-820. doi: 10.1016/j.sigpro.2010.08.014

Fernandez, A. J. (2000). On maximum likelihood prediction based on type II doubly censored exponential data. Metrika, 50(3), 211-220. doi: 10.1007/s001840050046

Fernandez, A. J. (2006). Bayesian estimation based on trimmed samples from Pareto populations. Computational Statistics \& Data Analysis, 51(2), 11191130. doi: 10.1016/j.csda.2005.11.010

## SINDHU ET AL.

Grimshaw, S. D., Collings, B. J., Larsen, W. A., \& Hurt, C. R. (2001). Eliciting factor importance in a designed experiment. Technometrics, 43(2), 133146. doi: 10.1198/004017001750386251

Jenkinson, D. (2005). The elicitation of probabilities: A review of the statistical literature. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.6173\&rep=rep1\&t ype=pdf

Khan, H. M., Albatineh, A., Alshahrani, S., Jenkins, N., \& Ahmed, N. U. (2011). Sensitivity analysis of predictive modeling for responses from the threeparameter Weibull model with a follow-up doubly censored sample of cancer patients. Computational Statistics \& Data Analysis, 55(12), 3093-3103. doi: 10.1016/j.csda.2011.05.017

Khan, H. M., Provost, S. B., \& Singh, A. (2010). Predictive inference from a two-parameter Rayleigh life model given a doubly censored sample. Communications in Statistics - Theory and Methods, 39(7), 1237-1246. doi: 10.1080/03610920902871453

Kim, C., \& Song, S. (2010). Bayesian estimation of the parameters of the generalized exponential distribution from doubly censored samples. Statistical Papers, 51(3), 583-597. doi: 10.1007/s00362-008-0142-3

Kundu, D., \& Howlader, H. (2010). Bayesian inference and prediction of the inverse Weibull distribution for type-II censored data. Computational Statistics \& Data Analysis, 54(6), 1547-1558. doi: 10.1016/j.csda.2010.01.003

Leon, C. J., Vazquez-Polo, J. F., \& Gonzalez, L.R. (2003). Elicitation of expert opinion in benefit transfer of environmental goods. Environmental and Resource Economics, 26(2), 199-210. doi: 10.1023/A:1026307420804

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. E., Garthwaite, P. H., Jenkinson, D. J.,...Rakow, T. (2006). Uncertain judgements: Eliciting expert probabilities. Hoboken, NJ: Wiley.

Pak, A., Parham, G. A., \& Saraj, M. (2013). On estimation of Rayleigh scale parameter under doubly type-II censoring from imprecise data. Journal of Data Science, 11(2), 305-322. Retrieved from http://www.jdsruc.org/Article_Detail.aspx?id=312

Panaitescu, E., George, P., Cozma, P., \& Popa, M. (2010). Bayesian and non-Bayesian estimators using record statistics of the modified-inverse Weibull distribution. Proceedings of the Romanian Academy, Series A, 11(3), 224-231.

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

Retrieved from http://www.acad.ro/sectii2002/proceedings/doc2010-3/05Panaitescu.pdf

Saleem, M., \& Aslam, M. (2008). Bayesian analysis of the two component mixture of the Rayleigh distribution assuming the uniform and the Jeffreys prior from censored data. Journal of Applied Statistical Science, 16(4), 105-113.

Saleem, M., Aslam, M., \& Economou, P. (2010). On the Bayesian analysis of the mixture of power function distribution using the complete and the censored sample. Journal of Applied Statistics, 37(1), 25-40. doi:
10.1080/02664760902914557

Shi, Y., \& Yan, W. (2010). The EB estimation of scale parameter for two parameter exponential distribution under the type-I censoring life test. Journal of Physical Sciences, 14, 25-30. Retrieved from http://14.139.211.206:8080/jspui/handle/123456789/134

Singh, S. K., Singh, U., \& Sharma, V. K. (2013). Bayesian analysis for Type-II hybrid censored sample from inverse Weibull distribution. International Journal of System Assurance Engineering and Management, 4(3), 241-248. doi: 10.1007/s13198-013-0172-7

Soliman, A. A. (2006). Estimators for the finite mixture of Rayleigh model based on progressively censored data. Communications in Statistics - Theory and Methods, 35(5), 803-820. doi: 10.1080/03610920500501379

Sultan, K. S., \& Al-Moisheer, A. S. (2012). Approximate Bayes estimation of the parameters and reliability function of a mixture of two inverse Weibull distributions under Type-2 censoring. Journal of Statistical Computation and Simulation, 83(10), 1900-1914. doi: 10.1080/00949655.2012.673614

Sultan, K. S., Ismail, M. A., \& Al-Moisheer, A. S. (2007). Mixture of two inverse Weibull distributions: Properties and estimation. Computational Statistics \& Data Analysis, 51(11), 5377-5387. doi: 10.1016/j.csda.2006.09.016

Tummala, V. M., \& Sathe, P. T. (1978). Minimum expected loss functions of reliability and parameters of certain lifetime distributions. IEEE Transactions on Reliability, 27(4), 283-285. doi: 10.1109/TR.1978.5220373

Wasan, M. T. (1970). Parametric estimation. New York, NY: McGraw-Hill.
Zanakis, S. H. (1979). A simulation study of some simple estimators for the three-parameter Weibull distribution. Journal of Statistical Computation and Simulation, 9(2), 101-116. doi: 10.1080/00949657908810302

## SINDHU ET AL.

## Appendix

The Bayes estimators of $\tau_{1}, \tau_{2}$, and $p_{1}$ under KLF, assuming gamma prior are:

$$
\begin{aligned}
& \hat{\tau}_{1(\mathrm{KLF})}=\left(\frac{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}+1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}+1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}}}{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}-1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}-1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}}}\right)^{\frac{1}{2}} \\
& \hat{\tau}_{2(\mathrm{KLF})}=\left(\frac{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}+1\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}+1\right)}}}{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-1\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-1\right)}}}\right)^{\frac{1}{2}} \\
& \hat{p}_{1(\mathrm{KLF})}=\left(\frac{\left.\left.\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}+1, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}\right)^{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}-1, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}}\right)^{\frac{1}{2}}}{m^{2}}\right.
\end{aligned}
$$

## TWO-COMPONENT MIXTURE OF INVERSE WEIBULL DISTRIBUTIONS

The posterior risks of $\tau_{1}, \tau_{2}$, and $p_{1}$ under KLF using gamma prior are:

$$
\begin{aligned}
\rho\left(\hat{\tau}_{1(\text { KLF })}\right)= & 2 \Omega_{1}^{-2} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}+1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}+1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}} \\
& \times \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}-1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}-1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}-1} \\
\rho\left(\hat{\tau}_{2(\text { KLF })}\right)= & 2 \Omega_{1}^{-2} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}+1\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}+1\right)}} \\
& \times \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-1\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-1\right)}-1} \\
\rho\left(\hat{\tau}_{2(\text { KLF })}\right)= & 2 \Omega_{1}^{-2} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}_{i}^{\left(a_{i}+m_{i}\right)}} \\
& \times \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}-1}
\end{aligned}
$$

## SINDHU ET AL.

The Bayes estimators of $\tau_{1}, \tau_{2}$, and $p_{1}$ under MELF, assuming gamma prior are:

$$
\begin{aligned}
& \hat{\tau}_{1(\mathrm{MELF})}=\left(\frac{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}-1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}-1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}}}{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}-2\right) \Gamma\left(a_{2}+m_{2}\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}-2\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}}}\right)^{\frac{1}{2}} \\
& \hat{\tau}_{2(\mathrm{MELF})}=\left(\frac{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-1\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-1\right)}}}{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right) \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-2\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-2\right)}}}\right)^{\frac{1}{2}} \\
& \hat{p}_{1(\mathrm{SELF})}=\left(\frac{\left.\left.\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}-1, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}\right)_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}-2, m_{2}+k_{2}+d_{1}\right) \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}\right)^{\frac{1}{2}}}{\sum^{2}}\right.
\end{aligned}
$$

The posterior risks of $\tau_{1}, \tau_{2}$, and $p_{1}$ under KLF using gamma prior are:

$$
\begin{aligned}
& \times \frac{\Gamma\left(a_{1}+m_{1}-1\right) \Gamma\left(a_{2}+m_{2}\right)}{\left.\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}\right\}_{1}^{\left(a_{1}-m_{1}-1\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}\right)}} \\
& \left.\left.\rho\left(\hat{\tau}_{2(\mathrm{SELF})}\right)=1-\Omega_{1}^{-1} \frac{\left(\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right)}\right.}{\times \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-1\right)}{\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{\left.b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-1\right)}}\binom{k_{1}}{k_{2}} \operatorname{Beta}\left(m_{1}+k_{1}-k_{2}+c_{1}, m_{2}+k_{2}+d_{1}\right)}}\right\}^{2}\right) \\
& \times \frac{\Gamma\left(a_{1}+m_{1}\right) \Gamma\left(a_{2}+m_{2}-2\right)}{\left\{b_{1}+\gamma_{1}\left(x_{1 j}\right)\right\}^{\left(a_{1}+m_{1}\right)}\left\{b_{2}+\gamma_{2}\left(x_{2 j}\right)\right\}^{\left(a_{2}+m_{2}-2\right)}}
\end{aligned}
$$

## SINDHU ET AL

$$
\begin{aligned}
& \rho\left(\hat{p}_{1(\text { SLLF })}\right)=1-\Omega_{1}^{-1} \frac{\left\{\begin{aligned}
\sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}} & \binom{k_{1}}{k_{2}} \\
& \mathrm{~B}\left(m_{1}+k_{1}-k_{2}+c_{1}-1, m_{2}+k_{2}+d_{1}\right) \\
& \times \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}
\end{aligned} \sum_{k_{1}=0}^{n-s} \sum_{k_{2}=0}^{k_{1}}(-1)^{k_{1}}\binom{n-s}{k_{1}}\binom{k_{1}}{k_{2}} \mathrm{~B}\left(m_{1}+k_{1}-k_{2}+c_{1}-2, m_{2}+k_{2}+d_{1}\right)\right.}{} \\
& \times \prod_{i=1}^{2} \frac{\Gamma\left(a_{i}+m_{i}\right)}{\left\{b_{i}+\gamma_{i}\left(x_{i j}\right)\right\}^{\left(a_{i}+m_{i}\right)}}
\end{aligned}
$$

# Misspecification of Variants of Autoregressive GARCH models and Effect on In-Sample Forecasting 

Olusanya E. Olubusoye<br>University of Ibadan Ibadan, Nigeria

OlaOluwa S. Yaya<br>University of Ibadan Ibadan, Nigeria

Oluwadare O. Ojo<br>Federal University of Technology<br>Akure, Nigeria


#### Abstract

Generally, in empirical financial studies, the determination of the true conditional variance in GARCH modelling is largely subjective. In this paper, we investigate the consequences of choosing a wrong conditional variance specification. The methodology involves specifying a true conditional variance and then simulating data to conform to the true specification. The estimation is then carried out using the true specification and other plausible specification that are appealing to the researcher, using model and forecast evaluation criteria for assessing performance. The results show that GARCH model could serve as better alternative to other asymmetric volatility models.


Keywords: Forecasts, GARCH, misspecification, specification

## Introduction

Since the seminal articles of Engle (1982) and Bollerslev (1986), the class of Generalized Autoregressive Conditionally Heteroscedasticity (GARCH) models has been a key model in financial industries. Due to wide applications of this model in financial industries and related areas, Lee and Hansen (1994) referred to the model as the workhouse of the industry. Considered here is the misspecification of variants of GARCH models. The variants include the GARCH model of Bollerslev (1986), Exponential GARCH model of Nelson (1991), Glosten Jagannathan and Runkle-GARCH (GJR-GARCH) model of Glosten, Jagannathan and Runkle (1993) and Asymmetric Power ARCH (APARCH) model of Ding, Granger and Engle (1993). Using model and forecast evaluation

[^20]
## OLUBUSOYE ET AL.

criteria, both correctly specified and the misspecified model performances are judged.

Specification of the form of GARCH model depends on the behavior and properties of the series. For example, there are asymmetric GARCH specifications which are preferred for the asymmetric series. Unlike Smooth Transition Autoregressive (STAR) model of Teräsvirta (1994) which allows selection of model between the Exponential STAR (ESTAR) and Logistic STAR (LSTAR) model based on the model specification tests, GARCH model is yet to develop such tests which selects among many alternatives. A particular GARCH model is often considered on the asset returns/residuals based on the properties of the series. The GARCH specification is a parametric model in which a particular structure is imposed at a time, and therefore, it is important to perform misspecification tests to check for the consequence of choosing a wrong model structure. Engle and Ng (1993) and Li and Mak (1994) proposed an adequacy test using the squared standardized error process. Recently, Lundbergh and Teräsvirta (2002) proposed tests for remaining ARCH in standardized errors, linearity and parameter constancy. None of the specification tests were designed to select or reject a particular GARCH specification.

Misspecification of GARCH model may pose serious problem to forecast values hence it deserves to be investigated. Wang (2002) affirmed that spurious and inefficient inference is expected when pure GARCH models are misspecified, this as well may affect the Quasi Maximum Likelihood Estimates (QMLEs) of the misspecified model. The QMLE of a pure $\operatorname{GARCH}(1,1)$ model indicates that the ARCH parameter is small, GARCH parameter is close to unity and the sum of both parameters approaches unity as the sampling frequency increases (Engle \& Bollerslev, 1986; Bollerslev \& Engle, 1993; Baillie, Bollerslev, \& Mikkelsen, 1996; Ding \& Granger, 1996; Andersen \& Bollerslev, 1997 and Engle \& Patton, 2001). This fact is reflected in the Integrated GARCH (IGARCH) model of Engle and Bollerslev (1986). More recent paper by Jansen and Lange (2010) shows that in a $\operatorname{GARCH}(1,1)$ model, the estimates of $\hat{\alpha}_{1}$ and $\hat{\beta}_{1}$ tend to 0 and 1 , respectively as the sampling frequency increases, which is an IGARCH effect.

In a situation whereby the GARCH series is fitted to any other variants of the model, particularly those ones with asymmetric effect, do we still expect this IGARCH convergence? This paper therefore considers the misspecification of GARCH models using simulation approach. The model and forecast evaluation criteria are used to judge the alternative models.

## MISSPECIFICATION OF STATIONARY GARCH VARIANTS

## Variants of GARCH model

The (GARCH) model predicts the volatility in the residuals $\varepsilon_{t}$ of the mean equation

$$
\begin{equation*}
y_{t}=\phi_{0}+\phi_{1} y_{t-1}+\varepsilon_{t} \tag{1}
\end{equation*}
$$

where $y_{t}$ is the time series or returns series under investigation, $\phi_{0}$ and $\phi_{i}$ are the constant and Autoregressive (AR) parameters of the model. In volatility modelling, autoregressive order is usually less than 3 and in some cases autoregression as well as constant may not be significant, which is the case of a pure GARCH process. The residuals of this model often violate normality assumption and are serially correlated. In that case, the non-normal residuals $\varepsilon_{t}$ are modelled using variance equation.

Engle (1982) proposed the first variance equation for predicting volatility in the asset returns/innovations $\varepsilon_{t}$, and this has been the origin of other volatility models in the literature. Bollerslev (1986) proposed using lags of the conditional volatility in the model specification. The $\operatorname{GARCH}(1,1)$ model, proposed in Bollerslev (1986) is,

$$
\begin{equation*}
\sigma_{t}^{2}=\omega+\alpha_{1} \varepsilon_{t-1}^{2}+\beta_{1} \sigma_{t-1}^{2} \tag{2}
\end{equation*}
$$

where $\varepsilon_{t}$ are the log-returns series of the financial asset. The residuals relates to the volatility as $\varepsilon_{t}=\sigma_{t} z_{t}$ with $z_{t} \approx N(0,1)$. The $\sigma_{t}$ is the unconditional standard deviation expressed by the variance equation (GARCH model). The parameter is conditioned as $\mathrm{w}>0, \alpha_{1} \geq 0$ and $\beta_{1} \geq 0$ in order to ensure positive definite variance. These $\alpha_{1}$ and $\beta_{1}$ are the ARCH and GARCH parameters for the ARCH term $\varepsilon_{t-1}^{2}$ and GARCH term $\sigma_{t-1}^{2}$, respectively while the stationarity imposition on the $\operatorname{GARCH}(1,1)$ is that the sum of the ARCH and GARCH parameters should be less than unity, that is $\alpha_{1}+\beta_{1}<1$. Then, combining the AR model in (1) with GARCH model in (2) gives AR (1)-GARCH ( 1,1 ) model.

The Exponential GARCH (EGARCH) model is given in Nelson (1991). This model was developed based on the fact that GARCH $(1,1)$ model of Bollerslev (1986) uses the magnitude of the innovations to predict future volatility but do not consider the effect of the positivity or negativity of the innovations on the volatility. The positive constraint imposed on the intercept $\omega$ often poses serious estimation problems. In that case, Nelson (1991) considered the

## OLUBUSOYE ET AL.

$\operatorname{GARCH}(1,1)$ model as symmetric type while the EGARCH $(1,1)$ is asymmetric in the sense that it assumes different conditional volatility responses for either positive or negative innovations. The simplest EGARCH $(1,1)$ specification is

$$
\begin{equation*}
\log \sigma_{t}^{2}=\omega+\alpha_{1}\left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right|+\beta_{1} \log \sigma_{t-1}^{2}-\gamma_{1}\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) \tag{3}
\end{equation*}
$$

This model can also be re-specified as,

$$
\begin{equation*}
\log \sigma_{t}^{2}=\omega+\alpha_{1}\left|z_{t-1}\right|+\beta_{1} \log \sigma_{t-1}^{2}-\gamma_{1}\left(z_{t-1}\right) \tag{4}
\end{equation*}
$$

because $\varepsilon_{t}=\sigma_{t} z_{t}$. Here, there is good news if $\varepsilon_{t-1}>0$ and bad news if $\varepsilon_{t-1}<0$ which have different effect on the conditional variance. The response of either good news or bad news on the conditional volatility is then measured by the asymmetric parameter, $\gamma_{1}$.

The Asymmetric Power ARCH (APARCH) model is proposed in Ding et al. (1993) with power specification $\delta$. The proposition was based on modelling standard deviation instead of the variance as in the case of GARCH and EGARCH models. This ideas was earlier considered in Taylor (1986) and Schwert (1989). The power parameter is estimated simultaneously with other parameters in the model. The specification of the $\operatorname{APARCH}(1,1)$ model is,

$$
\begin{equation*}
\sigma_{t}^{\delta}=\omega+\alpha_{1}\left(\left|\varepsilon_{t-i}\right|-\gamma_{1} \varepsilon_{t-1}\right)^{\delta}+\beta_{1} \log \sigma_{t-1}^{\delta} \tag{5}
\end{equation*}
$$

where $\delta>0$ and $\left|\gamma_{1}\right| \leq 1$. At $\delta=2$ and $\gamma_{1}=0$, the $\operatorname{APARCH}(1,1)$ model reduces to GARCH (1,1) model.

## Estimation and Forecasts Evaluation

Estimation of GARCH model is often carried out by numerical derivatives. Numerical derivatives are used in GARCH estimation since the model lacks closed form estimation (Xekalaki \& Degiannakis, 2010). The derivatives simplifies and maximises the QML log likelihood function

$$
\begin{equation*}
L_{t}=-\frac{1}{2}\left[N \log (2 \pi)+\sum_{t=1}^{N} \frac{\varepsilon_{t}^{2}}{\sigma_{t}^{2}}+\sum_{t=1}^{N} \log \sigma_{t}^{2}\right] \tag{6}
\end{equation*}
$$

## MISSPECIFICATION OF STATIONARY GARCH VARIANTS

where $\varepsilon_{t}$ are the innovations from the initial AR model, $\sigma_{t}^{2}$ are the conditional volatility realized from the variance equation and $N$ is the sample size. Berndt, Hall, Hall and Hausman (BHHH) algorithm of Berndt, Hall, Hall and Hausman (1974) is often preferred to other numerical derivatives such as Marquadt and Gauss Newton, since it uses only the first derivatives of the likelihood function to estimate the parameter values. The algorithm is

$$
\begin{equation*}
\psi^{(i+1)}=\psi^{(i)}-\left(\sum_{t=1}^{N} \frac{\partial L_{t}^{(i)}}{\partial \psi} \cdot \frac{\partial L_{t}^{(i)}}{\partial \psi^{\prime}}\right)^{-1} \frac{\partial L_{N}^{(i)}}{\partial \psi} \tag{7}
\end{equation*}
$$

with initial parameter set as $\psi^{(0)}$, the parameter set which maximizes the likelihood function is denoted as $\psi^{(i+1)}$ and the log-likelihood $L_{t}$ as given in (6) above. The number of iteration is denoted by $i$, and the iteration stops once there is no further improvement in the likelihood function. Ideally, EViews software allows setting the number of iteration and the level of precision for the estimation.

Forecast evaluation criteria considered are the Root Mean Squares Forecast Error (RMSFE), Mean Absolute Forecast Error (MAFE), Mean Absolute Percentage Forecast Error (MAPFE) and Theil Inequality of Theil (1961; 1966). The MSFE is defined as,

$$
\begin{equation*}
M S F E=\frac{1}{m} \sum_{t=1}^{m}\left(\hat{\sigma}_{t}^{2}-\sigma_{t}^{2}\right)^{2} \tag{8}
\end{equation*}
$$

where $\hat{\sigma}_{t}^{2}$ is the predicted in-sample conditional variances, and this depends on the scale of the variance series, $\sigma_{t}^{2}$. The square root of MSFE is the RMSFE,

$$
\begin{equation*}
R M S F E=\sqrt{\frac{1}{m} \sum_{t=1}^{m}\left(\hat{\sigma}_{t}^{2}-\sigma_{t}^{2}\right)^{2}} \tag{9}
\end{equation*}
$$

The MAFE and MAPFE are obtained by taking the absolute differences of the predicted conditional volatilities and the observed volatilities as,

$$
\begin{equation*}
M A F E=\frac{1}{m} \sum_{t=1}^{m}\left|\hat{\sigma}_{t}^{2}-\sigma_{t}^{2}\right| \tag{10}
\end{equation*}
$$

## OLUBUSOYE ET AL.

$$
\begin{equation*}
M A P F E=100 \sum_{t=1}^{m}\left|\frac{\hat{\sigma}_{t}^{2}-\sigma_{t}^{2}}{\sigma_{t}^{2}}\right| \tag{11}
\end{equation*}
$$

The Theil inequality is given as,

$$
\begin{equation*}
T I=\frac{\sqrt{\frac{1}{m} \sum_{t=1}^{m}\left(\hat{\sigma}_{t}^{2}-\sigma_{t}^{2}\right)^{2}}}{\sqrt{\frac{1}{m} \sum_{t=1}^{m} \hat{\sigma}_{t}^{2}}+\sqrt{\frac{1}{m} \sum_{t=1}^{m} \sigma_{t}^{2}}} \tag{12}
\end{equation*}
$$

The inequality coefficient is time invariant and always lies between 0 and unity. The smaller these forecast evaluation criteria, the better the candidate model represents well the data.

## Monte Carlo Experiment, Results and Discussion

The Monte Carlo experiment is set up using the Data Generating Processes (DGPs) in (13)-(16) below. The AR (1) DGP in (12) is the mean equation, with $\phi_{0}=0.15$ and $\phi_{1}=0.5$, setting the process at the stationarity level.

$$
\begin{equation*}
y_{t}=0.15+0.5 y_{t-1}+\varepsilon_{t} \tag{13}
\end{equation*}
$$

The error distribution $\varepsilon_{t}=\sigma_{t} z_{t}, z_{t} \sim N(0,1)$ for each of the variance equations,

$$
\begin{gather*}
\sigma_{t}^{2}=0.02+0.25 \varepsilon_{t-1}^{2}+0.60 \sigma_{t-1}^{2}  \tag{14}\\
\log \sigma_{t}^{2}=0.02+0.25\left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right|+0.60 \log \sigma_{t-1}^{2}-0.10\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right)  \tag{15}\\
\sigma_{t}^{1.2}=0.02+0.25\left(\left|\varepsilon_{t-1}\right|+0.10 \varepsilon_{t-1}\right)^{1.2}+0.60 \log \sigma_{t-1}^{1.2} \tag{16}
\end{gather*}
$$

representing GARCH (1,1), EGARCH (1,1) and APARCH (1,1) models, respectively. The parameters of the models were generated by arbitrarily fixing values for them making sure the parameters of the ARCH and GARCH terms are in stationarity range, and this realizes positive definite stationary non-explosive

## MISSPECIFICATION OF STATIONARY GARCH VARIANTS

conditional variance. These parameter values are fixed in the three models. The AR (1) DGP is combined with each variance equations in (14), (15) and (16) giving AR (1)-GARCH (1,1), AR (1)-EGARCH (1,1) and AR (1)-APARCH (1,1) DGPs, respectively. The asymmetric parameter in EGARCH and APARCH models are fixed at $\gamma_{1}=-0.10$ and the power parameter in APARCH model fixed at 1.2. The misspecification of each model is considered and the behaviour of the realized conditional variance is examined using the model and forecast evaluation criteria. Sample sizes are $N=2000,4000$ and 6000 each with $25 \%$ of samples as in-sample forecasts.

The results of the Monte Carlo experiment are presented here as Scenarios $1-3$, where both parameter and forecasts evaluation estimates are given.

## Scenario 1: When the true model is GARCH

Table 1. Model parameter estimates

| Sample | Estimated Model | $\hat{\phi}_{0}(0.15)$ | $\hat{\phi}_{1}(0.50)$ | $\hat{w}(0.02)$ | $\hat{\alpha}_{1}(0.25)$ | $\hat{\beta}_{1}(0.60)$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | 0.1480 | 0.4839 | 0.0169 | 0.2110 | 0.6596 |
| 4000 | GARCH | 0.1517 | 0.4724 | 0.0173 | 0.2049 | 0.6590 |
| 6000 | GARCH | 0.1475 | 0.4750 | 0.0180 | 0.2052 | 0.6503 |
| 2000 | EGARCH | 0.1480 | 0.4839 | 0.0169 | 0.2110 | 0.6596 |
| 4000 | EGARCH | 0.1517 | 0.4724 | 0.0173 | 0.2049 | 0.6590 |
| 6000 | EGARCH | 0.1475 | 0.4750 | 0.0180 | 0.2052 | 0.6503 |
| 2000 | APARCH | 0.1464 | 0.4875 | 0.0405 | 0.1988 | 0.0184 |
| 4000 | APARCH | 0.1488 | 0.4741 | 0.0412 | 0.1947 | 0.0633 |
| 6000 | APARCH | NA | NA | NA | NA | NA |

The results presented in Scenario 1 is when GARCH simulated series is used to estimate EGARCH, APARCH as well as GARCH model and the parameter and in-sample forecasts estimates presented in Tables 1 and 2, respectively. The parameter estimates for the three models are very close to the real values but these are not consistent with sample sizes. This is expected since we do not expect the least squares estimates to be consistent in the presence of serial correlation and heteroscedasticity of the residuals. We also noted the similarity in the results obtained for GARCH and EGARCH models, across the sample sizes. The APARCH estimation posed serious problem at very high sample sizes due to tendencies of the simulator to realize some non-positive volatility and the power estimates of these cannot be obtained.

## OLUBUSOYE ET AL.

Table 2. Forecast evaluation estimates

| Sample size | Estimated Model | RMSFE | MAFE | MAPFE | Theil |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | 0.000271 | 0.011228 | 48.09666 | 0.0289 |
| 4000 | GARCH | 0.000174 | 0.010415 | 44.88254 | 0.0257 |
| 6000 | GARCH | 0.000139 | 0.010104 | 49.58554 | 0.0287 |
| 2000 | EGARCH | 0.002067 | 0.085482 | 617.9674 | 0.1772 |
| 4000 | EGARCH | 0.001301 | 0.078142 | 645.6404 | 0.1688 |
| 6000 | EGARCH | 0.001165 | 0.084255 | 639.2407 | 0.1798 |
| 2000 | APARCH | 0.000520 | 0.021537 | 94.48200 | 0.1077 |
| 4000 | APARCH | 0.000344 | 0.020411 | 93.54733 | 0.0974 |
| 6000 | APARCH | NA | NA | NA | NA |

From the forecasts evaluation results in Table 2 of Scenario 1, the estimates obtained for GARCH and EGARCH models are different. Actually the RMSFE and MAFE for the models across different sample sizes are very low but the MAPFE vary significantly. The RMSFE, MAFE, MAPFE and Theil inequality coefficient for the GARCH models are the lowest, followed by that of APARCH models. This is expected since the DGP is GARCH. The MAPFE estimates vary significantly, about $50 \%$ for GARCH, $600 \%$ for EGARCH and $90 \%$ for APARCH models. It is clear to see that GARCH model forecasts are better than EGARCH and APARCH model forecasts in terms of RMSPE and Theil inequality when GARCH model is the DGP.

## Scenario 2: When the true model is EGARCH

Table 3. Model parameter estimates

| Sample | Estimated <br> Size | $\hat{\phi}_{0}(0.15)$ | $\hat{\phi}_{1}(0.50)$ | $\hat{w}(0.02)$ | $\hat{\alpha}_{1}(0.25)$ | $\hat{\beta}_{1}(0.60)$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | 0.1320 | 0.4998 | 0.2037 | 0.0078 | 0.7817 |
| 4000 | GARCH | 0.1441 | 0.5067 | 0.4191 | 0.0058 | 0.5268 |
| 6000 | GARCH | 0.1212 | 0.4737 | 0.1355 | -0.0101 | 0.8559 |
| 2000 | EGARCH | 0.1320 | 0.4998 | 0.2037 | 0.0078 | 0.7817 |
| 4000 | EGARCH | 0.1293 | 0.4806 | 0.1915 | -0.0094 | 0.7933 |
| 6000 | EGARCH | 0.1212 | 0.4737 | 0.1355 | -0.0101 | 0.8559 |
| 2000 | APARCH | 0.1679 | 0.4765 | 0.5952 | -0.0823 | 1.0000 |
| 4000 | APARCH | NA | NA | NA | NA | NA |
| 6000 | APARCH | NA | NA | NA | NA | NA |

In Scenario 2 of Table 3, the true series follows EGARCH process. The parameter estimates are not consistent with sample sizes. Here, both the estimates

## MISSPECIFICATION OF STATIONARY GARCH VARIANTS

of the mean and variance equations are very far from the real values. Even with EGARCH DGP to estimate EGARCH model, the estimates seem not to improve when compared with that of the misspecified GARCH model. Estimates of APARCH model for all the sampled points in the simulations could not be computed except for sample 2000, the estimation was very slow at samples 4000 and 6000 and the estimation process crashed unexpectedly.

Table 4. Forecast evaluation estimates

| Sample size | Estimated Model | RMSFE | MAFE | MAPFE | Theil |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | 0.009880 | 0.412570 | 79.28039 | 0.2282 |
| 4000 | GARCH | 0.006603 | 0.389932 | 69.61645 | 0.2152 |
| 6000 | GARCH | 0.005286 | 0.381419 | 69.51897 | 0.2158 |
| 2000 | EGARCH | 0.028755 | 1.224082 | 210.8467 | 0.4251 |
| 4000 | EGARCH | 0.010901 | 0.642821 | 138.9766 | 0.3230 |
| 6000 | EGARCH | 0.006889 | 0.495946 | 107.6174 | 0.2687 |
| 2000 | APARCH | 0.065849 | 2.763889 | 372.1500 | 0.9589 |
| 4000 | APARCH | NA | NA | NA | NA |
| 6000 | APARCH | NA | NA | NA | NA |

In Table 4 of Scenario 2, forecast estimates for the three models are different, with estimated GARCH models presenting better forecasts than the estimated EGARCH and APARCH model at corresponding sample sizes.

## Scenario 3: When the true model is APARCH

Table 5. Model parameter estimates

| Sample | Estimated <br> Model | $\hat{\phi}_{0}(0.15)$ | $\hat{\phi}_{1}(0.50)$ | $\hat{w}(0.02)$ | $\hat{\alpha}_{1}(0.25)$ | $\hat{\beta}_{1}(0.60)$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | 0.1514 | 0.4795 | 0.0037 | 0.2240 | 0.6188 |
| 4000 | GARCH | 0.1459 | 0.5028 | 0.0046 | 0.2567 | 0.5399 |
| 6000 | GARCH | 0.1526 | 0.4721 | 0.0039 | 0.2110 | 0.6157 |
| 2000 | EGARCH | 0.1514 | 0.4795 | 0.0037 | 0.2240 | 0.6188 |
| 4000 | EGARCH | 0.1438 | 0.5052 | 0.0046 | 0.2514 | 0.5526 |
| 6000 | EGARCH | 0.1526 | 0.4721 | 0.0039 | 0.2110 | 0.6157 |
| 2000 | APARCH | 0.1476 | 0.5164 | 0.0260 | 0.2753 | -0.0012 |
| 4000 | APARCH | 0.1395 | 0.5345 | 0.0188 | 0.2371 | -0.0800 |
| 6000 | APARCH | $N A$ | $N A$ | $N A$ | $N A$ | $N A$ |

## OLUBUSOYE ET AL.

Scenario 3, Table 5 presents the case where APARCH series is assumed. The APARCH model is more complex in structure than the GARCH and EGARCH models, therefore estimating APARCH model from the series at very high sample size posed serious problems. For samples 2000 and 4000, estimates of parameters were computed.

Table 6. Forecast evaluation estimates

| Sample size | Estimated Model | RMSFE | MAFE | MAPFE | Theil |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 2000 | GARCH | $8.20 \mathrm{E}-06$ | 0.000341 | 54.08614 | 0.0053 |
| 4000 | GARCH | $5.98 \mathrm{E}-06$ | 0.000401 | 56.20084 | 0.0057 |
| 6000 | GARCH | $4.25 \mathrm{E}-06$ | 0.000307 | 54.36182 | 0.0054 |
| 2000 | EGARCH | 0.000347 | 0.014360 | 3191.426 | 0.0925 |
| 4000 | EGARCH | 0.000480 | 0.028411 | 5535.569 | 0.1344 |
| 6000 | EGARCH | 0.000187 | 0.013544 | 31268.40 | 0.0910 |
| 2000 | APARCH | $2.33 E-05$ | 0.000955 | 85.89381 | 0.0182 |
| 4000 | APARCH | $9.13 E-06$ | 0.000538 | 94.78748 | 0.0181 |
| 6000 | APARCH | NA | NA | NA | NA |

In Scenario 3, Table 6, the simulated forecasts for GARCH, EGARCH and APARCH models from APARCH DGP are presented. Closer look still showed that GARCH forecasts are the best in terms of forecast evaluation criteria. Followed after GARCH is the APARCH model and EGARCH is the least.

## Conclusion

The misspecification of some GARCH models were considered using parameter and forecast evaluation estimates as criteria. It was found that a correctly specified EGARCH and APARCH models actually, in the real sense, did not give better parameter estimates and forecasts when compared with that of GARCH model. These results are not consistent with sample sizes. The results obtained in this paper therefore support the seminal work of Hansen and Lunde (2005) titled: "A Forecast Comparison of Volatility Models: Does Anything Beat a $\operatorname{GARCH}(1,1)$ ", which was their argument with Andersen and Bollerslev (1998). Great care should be taken wherever volatility model are being specified for assets returns, since misspecified model could cause great loss in model information criteria and forecasts. This work, therefore re-popularize the use of symmetric GARCH (1,1) model of Bollerslev (1986) and Taylor (1986) in empirical analysis and simulations.

## MISSPECIFICATION OF STATIONARY GARCH VARIANTS

## References

Andersen, T. G. \& Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. Journal of Empirical Finance, 4(2/3), 115-158. doi: 10.1016/S0927-5398(97)00004-2

Andersen, T. G., \& Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4), 885-905. doi: 10.2307/2527343

Baillie, R. T., Bollerslev, T., \& Mikkelsen, H. O. (1996). Fractional integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 74(1), 3-30. doi: 10.1016/S0304-4076(95)01749-6

Berndt, E., Hall, B., Hall, R., \& Hausman, J. (1974). Estimation and inference in nonlinear structural models. Annals of Economic and Social Measurement, 3(4), 653-665.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327. doi: 10.1016/0304-4076(86)90063-1

Bollerslev, T. \& Engle, R. F. (1993). Common persistence in conditional variances. Econometrica, 61(1), 167-186. doi: 10.2307/2951782

Ding, Z. \& Granger, C. W. J. (1996). Modeling volatility persistence of speculative returns: A new approach. Journal of Econometrics, 73(1), 185-215. doi: 10.1016/0304-4076(95)01737-2

Ding, Z., Granger, C. W. J., \& Engle, R. F. (1993). A long memory property of stock market returns and a new model. Journal of Empirical Finance, 1(1), 83-106. doi: 10.1016/0927-5398(93)90006-D

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1008. doi: 10.2307/1912773

Engle, R. F. \& Bollerslev, T. (1986). Modelling the persistence of conditional variances. Econometric Reviews, 5(1), 1-50.
doi: 10.1080/07474938608800095
Engle, R. F. \& Ng, V. K. (1993). Measuring and testing the impact of news on volatility. Journal of Finance, 48(5), 1749-1777. doi: 10.1111/j.1540-6261.1993.tb05127.x

Engle, R. F. \& Patton, A. J. (2001). What good is a volatility model? Quantitative Finance, 1(2), 237-245. doi: 10.1088/1469-7688/1/2/305

## OLUBUSOYE ET AL.

Glosten, L. R., Jagannathan, R., \& Runkle, D. E. (1993). On the relation between the expected value and the volatility of the normal excess return on stocks. Journal of Finance, 48(5), 1779-1801.
doi: 10.1111/j.1540-6261.1993.tb05128.x
Hansen, P. R., \& Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Journal of Applied Econometrics, 20(7) 873-889. doi: 10.1002/jae. 800

Jensen, A. T. \& Lange, T. (2010). On convergence of the QMLE for misspecified GARCH models. Journal of Time Series Econometrics, 2(1), 1941-1928. doi: 10.2202/1941-1928.1034

Lee, S.-W. \& Hansen, B. E. (1994). Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. Econometric Theory, 10(1), 29-52. doi: 10.1017/S0266466600008215

Li, W. K. \& Mak, T. K. (1994). On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. Journal of Time Series Analysis, 15(6), 627-636. doi: 10.1111/j.1467-9892.1994.tb00217.x

Lundbergh, S. \& Teräsvirta, T. (2002). Evaluating GARCH models. Journal of Econometrics, 110(2), 417-435. doi: 10.1016/S0304-4076(02)00096-9

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. Econometrica, 59(2), 347-370. doi: 10.2307/2938260

Schwert, G. W. (1989). Stock volatility and crash of '87. Review of Financial Studies, 3(1), 77-102. doi: 10.1093/rfs/3.1.77

Taylor, S. (1986). Modelling financial time series. New York, NY: John Wiley \& Sons.

Teräsvirta, T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association, 89(425), 208-218. doi: 10.1080/01621459.1994.10476462

Theil, H. (1961). Economic forecasts and policy. Amsterdam: NorthHolland

Theil, H. (1966). Applied Economic forecasting. Amsterdam: North-Holland Wang, Y. (2002). Asymptotic nonequivalence of GARCH models and diffusions. Annals of Statistics, 30(3), 754-783. doi: 10.1214/aos/1028674841

Xekalaki, E., \& Degiannakis, S. 2010. ARCH models for financial applications. John Wiley \& Sons Ltd.

# Improved Ridge Estimator in Linear Regression with Multicollinearity, Heteroscedastic Errors and Outliers 

A. V. Dorugade<br>Y C Mahavidyalaya, Halkarni<br>Maharashtra, India

This paper introduces a new estimator, of ridge parameter $k$ for ridge regression and then evaluated by Monte Carlo simulation. We examine the performance of the proposed estimators compared with other well-known estimators for the model with heteroscedastics and/or correlated errors, outlier observations, non-normal errors and suffer from the problem of multicollinearity. It is shown that proposed estimators have a smaller MSE than the ordinary least squared estimator (LS), Hoerl and Kennard (1970) estimator (RR), jackknifed modified ridge (JMR) estimator, and Jackknifed Ridge M-estimator (JRM).

Keywords: Heteroscedasticity, mean squared error, multicollinearity, outlier, ridge estimator

## Introduction

In multiple linear regressions the estimation of parameters is a common interest for many users. It is well known that an LS estimator has been treated as the best unbiased estimator for a long time since it has minimum variance. Multicollinearity, linear or near-linear dependency among the explanatory variables in the regression model, is an important problem faced in applications. If multicollinearity or the ill-conditioned design matrix in linear regression model is present, the LS estimator is sensitive to number 'errors', namely, there is an 'explosion' of the sampling variance of the estimators. Moreover, some of the regression coefficients may be statistically insignificant with wrong sign and meaningful statistical inference becomes impossible for practitioners.

To overcome multicollinearity various biased estimators were put forward in the literature. The Ridge Regression (RR) estimator proposed by Hoerl and

[^21]
## A V DORUGADE

Kennard (1970) is the most popular biased estimator. However, RR estimator has some disadvantages; mainly it is a nonlinear function of the ridge parameter (or biasing constant) $k$. Currently there are various methods for determination and much of the discussions on ridge regression concern the problem of finding or selecting good empirical value of $k$. Our primary aim in this article is overcome this problem by suggesting new estimator for ridge parameter and then evaluate its performance when model defined in linear regression exhibits with not only multicollinearity but also heteroscedastics and/or correlated errors, non-normal errors and outliers, respectively.

Much of the discussions on ridge regression concern the problem of finding better alternative to the LS estimator. Some popular numerical techniques to deal with multicollinearity are the ridge regression due to Singh and Chaubey (1987), Batah, Ramnathan, and Gore (2008), Yang and Chang (2010) and others. Most of the estimation procedures are obtained results when specific assumptions like elements of the random vector $\varepsilon$ were independent and identically distributed random variables are achieved. But if these assumptions are violated, these methods do not assure the desirable results. Involving such problems as heteroscedasticity and autocorrelation few methods including Trenkler (1984), Firinguetti (1989), Bayhan and Bayhan (1998), Özkale (2008), Alheety and Kibria (2009) are available in the present literature. Recently, Li and Yang (2011) suggested Jackknifed Modified Ridge Estimator (JMRE) and show that it superior to the generalized least squares estimate, the generalized modified ridge estimator and the generalized jackknifed ridge estimator, to overcome multicollinearity in the presence of a linear regression model with correlated or heteroscedastic errors.

Apart from the problem of multicollinearity in real life situation, outliers and departure from the normality assumption are common problems in regression. These also produce undesirable effects on the LS estimator. This fact is pointed out by many researchers. Many researchers have pointed out that M-estimator is better than LS estimator in the presence of outliers (Huber, 1981; Rousseeuw \& Leroy, 1987; Birkes \& Dodge, 1993). In standard text like Birkes and Dodge (1993) and Montgomery, Peck, and Vining (2001) have given detail description. Recently, Jadhav and Kashid (2011) gives Jackknifed Ridge M-estimator (JRM) and show that it performs better than LS, ridge and M-estimator in the presence of both outliers and multicollinearity. Hence our secondary aim in this article is to provide an alternative method to combat both the problem of outliers and heteroscedastics and/or correlated errors, respectively in linear regression model in the presence of multicollinearity.

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

## Model and Estimators

Consider, widely used linear regression model

$$
\begin{equation*}
Y=X \beta+\varepsilon \tag{1}
\end{equation*}
$$

where $Y$ is an $n \times 1$ random vector of response variables, $\boldsymbol{X}$ is a known $n \times p$ matrix with full column rank, $\varepsilon$ is the vector of errors $E(\varepsilon)=0$ and $\operatorname{Cov}(\varepsilon)=\sigma^{2} \operatorname{In}$., $\beta$ is a $p \times 1$ vector of unknown regression parameters and $\sigma^{2}$ is the unknown variance parameter. For the sake of convenience, we assume that the matrix $X$ and response variable $Y$ are standardized in such a way that $X^{\prime} X$ is a non-singular correlation matrix and $X^{\prime} Y$ is the correlation between $X$ and $Y$.

Let $\wedge$ and $T$ be the matrices of eigenvalues and eigenvectors of $X^{\prime} X$, respectively, satisfying $T^{\prime} X^{\prime} X T=\wedge=\operatorname{diagonal}\left(\lambda_{1}, \lambda_{2}, \ldots, \lambda_{p}\right)$ where $\lambda_{i}$ being the $i^{\text {th }}$ eigenvalue of $X^{\prime} X$ and $T^{\prime} T=T T^{\prime}=I p$. We obtain the equivalent model

$$
\begin{equation*}
Y=Z \alpha+\varepsilon, \tag{2}
\end{equation*}
$$

where $Z=X T$, it implies that $Z^{\prime} Z=\wedge$, and $\alpha=T^{\prime} \beta$ (see Montgomery et al., 2001).

Then LS estimator of $\alpha$ is given by

$$
\begin{equation*}
\hat{\alpha}_{L S}=\left(Z^{\prime} Z\right)^{-1} Z^{\prime} Y=\wedge^{-1} Z^{\prime} Y \tag{3}
\end{equation*}
$$

Therefore, LS estimator of $\beta$ is given by

$$
\hat{\beta}_{L S}=T \hat{\alpha}_{L S}
$$

## Ridge Regression Estimator (RR)

To overcome multicollinearity under ridge regression, Hoerl and Kennard (1970) suggested an alternative estimate by adding a ridge parameter $k$ to the diagonal elements of the least square estimator. It is given as:

$$
\begin{equation*}
\hat{\alpha}_{R R}=\left\lfloor I-k(\wedge+k I)^{-1}\right\rfloor \hat{\alpha}_{L S} \tag{4}
\end{equation*}
$$

## A V DORUGADE

Therefore, RR estimator of $\beta$ is given by

$$
\hat{\beta}_{R R}=T \hat{\alpha}_{R R}
$$

and mean square error of $\hat{\alpha}_{R R}$ is

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\alpha}_{R R}\right)=\hat{\sigma}^{2} \sum_{i=1}^{p} \lambda_{i} /\left(\lambda_{i}+k\right)^{2}+k^{2} \sum_{i=1}^{p} \hat{\alpha}_{i}^{2} /\left(\lambda_{i}+k\right)^{2} \tag{5}
\end{equation*}
$$

Observe when $k=0$ in (5), MSE of LS estimator of $\alpha$ is recovered. Hence,

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\alpha}_{L S}\right)=\hat{\sigma}^{2} \sum_{i=1}^{p} 1 / \lambda_{i} \tag{6}
\end{equation*}
$$

## Jackknifed Modified Ridge Estimator (JMR)

Li and Yang (2011) introduced a jackknifed modified ridge (JMR) estimator an alternative method to overcome multicollinearity in the presence of a linear regression model with correlated or heteroscedastic errors. With the assumptions $E(\varepsilon)=0$ and $\operatorname{Cov}(\varepsilon)=\sigma^{2} V$, where $V$ is a known $n \times n$ symmetric positive definite (pd) matrix there exists a nonsingular symmetric matrix $P$ such that $V^{-1}=P^{\prime} P$ and $\sigma^{2}>0$ is the unknown variance parameter. Then the linear model given in (2) can be written as

$$
\begin{equation*}
\tilde{Y}=\tilde{Z} \alpha+\tilde{\varepsilon}, \tag{7}
\end{equation*}
$$

where, $\tilde{Y}=P Y, \tilde{\varepsilon}=P \varepsilon$ and $\tilde{Z}=P Z$ with a prior mean $c=\sum_{i=1}^{p} \hat{\alpha}_{i L S} / p$ and $k>0$ is the ridge parameter (see Li \& Yang, 2011). The JMR estimator of $\alpha$ is given as

$$
\begin{gather*}
\hat{\alpha}_{J M R}=\left\lfloor I-k^{2}\left(\tilde{Z}^{\prime} \tilde{Z}+k I\right)^{-2}\right\rfloor\left(\tilde{Z}^{\prime} \tilde{Z}\right)^{-1} \tilde{Z}^{\prime} \tilde{Y}+k^{2}\left(\tilde{Z}^{\prime} \tilde{Z}+k I\right)^{-2} c  \tag{8}\\
\hat{\beta}_{J M R}=T \hat{\alpha}_{J M R}
\end{gather*}
$$

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

Hoerl, Kennard, and Baldwin (1975) suggested the value of ' $k$ ' should be chosen small enough so the mean squared error of ridge estimator is less than the mean squared error of LS estimator. Among the various methods here, the ridge parameter was used to compute $\hat{\alpha}_{R R}$ and $\hat{\alpha}_{J M R}$ given by Hoerl et al. (1975), given as

$$
\begin{equation*}
k_{H K B}=p \hat{\sigma}^{2} / \sum_{i=1}^{p} \hat{\alpha}_{i}^{2} \tag{9}
\end{equation*}
$$

where $\hat{\alpha}_{i}$ is the $i^{\text {th }}$ element of $\hat{\alpha}_{L S}, i=1,2, \ldots, p$ and $\hat{\sigma}^{2}$ is the LS estimator of $\sigma^{2}$ i.e.

$$
\hat{\sigma}^{2}=\frac{Y^{\prime} Y-\hat{\alpha}_{L S}^{\prime} Z^{\prime} Y}{n-p-1} .
$$

## Jackknifed Ridge M-Estimator (JRM)

Jadhav and Kashid (2011) gave Jackknifed Ridge M-estimator (JRM) which takes into account the presence of both multicollinearity and outlier problems simultaneously.

It is given as:

$$
\begin{equation*}
\hat{\alpha}_{J R M}=\left(I-k^{* 2} A^{-2}\right) \hat{\alpha}_{M} \tag{10}
\end{equation*}
$$

Where, $\hat{\alpha}_{M}$ is an M-estimator of $\alpha$, which is obtained by solving the following equations

$$
\sum_{i=1}^{n} v_{i j} \psi\left(Y_{i}-z_{i}^{\prime} \hat{\alpha}_{M}\right)=0
$$

where $\psi($.$) is some function (see Huber (1981), Hampel, Ronchetti, Rousseeuw,$ and Stahel (1986)). Therefore, JRM estimator of $\beta$ is given by

$$
\hat{\beta}_{J R M}=T \hat{\alpha}_{J R M},
$$

## A V DORUGADE

where, ridge parameter says $k^{*}$ can be estimated using robust estimates of $\sigma^{2}$ and $\alpha$ respectively given by $s$ and $\hat{\alpha}_{M}$, it is obtained by

$$
k^{*}=\frac{p s^{2}}{\hat{\alpha}_{M}^{\prime} \hat{\alpha}_{M}},
$$

where $s=1.4826$ median $\mid e_{i}-$ median $\left(e_{1}\right) \mid$ and $e_{i}$ is $i^{\text {th }}$ residual obtained by using LS estimator.

## Proposed Ridge Parameter

The existence of multicollinearity may cause to have wide confidence interval for individual parameters or linear combination of the parameters, may give estimates with wrong signs. Ridge regression is a concept proposed in the sixties to combat the multicollinearity in regression problems. After then, many new versions of this method have been studied to extended Hoerl and Kennard (1970) original estimator. It has been made a more definite comparison of these various versions of the biased estimators versus the unbiased LS estimator. The constant, $k>0$ is known as ridge parameter which plays an important role in ridge regression. As $k$ increases from 0 and continues upto $\infty$ the regression estimates tend towards 0 . In ridge regression our interest lies in finding a value of $k$ such that the reduction in the variance term is greater than the increase in the squared bias. Though these estimators result in biased, for certain value of $k$, they yield minimum mean squared error (MMSE) compared to the LS estimator (see Hoerl \& Kennard, 1970). In the last decades, researchers concentrated on estimating the shrinkage ridge parameter $k$ in different ways and under different situations, and then compared the results with those obtained by applying the LS estimators. Much of the discussions on ridge regression concern the problem of finding good empirical value of $k$.

Ridge regression estimator of Hoerl and Kennard (1970) was proposed as alternative to the least squares estimator in the presence of multicollinearity. It depends on the biasing parameter $k$ which is the Lagrange multiplier used in the objective function although proposing the estimator. To compute the ridge regression estimator, the analyst must know the value of $k$. Therefore, various estimators of $k$ were proposed. Many different techniques for estimating $k$ have been proposed or suggested by different researchers Hoerl et al. (1975), Lawless and Wang (1976), Kibria (2003), Khalaf and Shukur (2005), Alkhamisi and

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

Shukur (2007), Muniz and Kibria (2009), Dorugade and Kashid (2010), Al-Hassan (2010), Muniz, Kibria, Mansson, and Shukur (2012) to mention a few. The well known of them is obtained by minimizing the mean square error of the ridge regression estimator and it depends on the variance of the regression model, $\sigma^{2}$, and the parameter vector $\beta$. Since $\sigma^{2}$ and $\beta$ are unknown, the analyst have to use estimates of these parameters. In this article we propose estimator of $k$ depends on the variance of the regression model, $\sigma^{2}$ only. Even though this approach is quite straightforward and simple, to the best of our knowledge, it has not been considered in the literature at all.

We denote our ridge parameter by $k_{R}$ and given by

$$
\begin{equation*}
k_{R}=\sigma, \tag{11}
\end{equation*}
$$

where error variance $\sigma^{2}$, replaced by its LS estimator $\hat{\sigma}^{2}$ i.e.

$$
\hat{\sigma}^{2}=\frac{Y^{\prime} Y-\hat{\alpha}_{L S}^{\prime} Z^{\prime} Y}{n-p-1}
$$

The RR estimator based on $k_{R}$ is given as

$$
\begin{equation*}
\hat{\alpha}_{R R}^{*}=\left\lfloor I-k_{R}\left(\wedge+k_{R} I\right)^{-1}\right\rfloor \hat{\alpha}_{L S} \tag{12}
\end{equation*}
$$

Therefore, RR estimator of $\beta$ is given by

$$
\hat{\beta}_{R R}^{*}=T \hat{\alpha}_{R R}^{*}
$$

and using (5) mean square error of $\hat{\alpha}_{R R}^{*}$ is

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\alpha}_{R R}^{*}\right)=\frac{\hat{\sigma}^{2}}{\left(\lambda_{i}+\hat{\sigma}\right)^{2}} \sum_{i=1}^{p}\left(\lambda_{i}+\hat{\alpha}_{i}^{2}\right) \tag{13}
\end{equation*}
$$

## Comparison Between the $\hat{\alpha}_{R R}^{*}$ and $\hat{\alpha}_{R R}$

Using (5) and (13), consider the following difference

## A V DORUGADE

$$
\begin{aligned}
\operatorname{MSE}\left(\hat{\alpha}_{R R}\right)-\operatorname{MSE}\left(\hat{\alpha}_{R R}^{*}\right) & =\sum_{i=1}^{p}\left[\frac{\left(\hat{\sigma}^{2} \lambda_{i}+k^{2} \hat{\alpha}_{i}^{2}\right)}{\left(\lambda_{i}+k\right)^{2}}\right]-\hat{\sigma}^{2} \sum_{i=1}^{p}\left[\frac{\left(\lambda_{i}+\hat{\alpha}_{i}^{2}\right)}{\left(\lambda_{i}+\hat{\sigma}\right)^{2}}\right] \\
& =\sum_{i=1}^{p} \frac{\left(k^{2} \hat{\alpha}_{i}^{2}\right)}{\left(\lambda_{i}+k\right)^{2}}+\hat{\sigma}^{2} \sum_{i=1}^{p}\left[\frac{\left(\lambda_{i}+\hat{\sigma}\right)^{2} \lambda_{i}-\left(\lambda_{i}+k\right)^{2}\left(\lambda_{i}+\hat{\alpha}_{i}^{2}\right)}{\left(\lambda_{i}+k\right)^{2}\left(\lambda_{i}+\hat{\sigma}\right)^{2}}\right]
\end{aligned}
$$

From above equation, the difference $\operatorname{MSE}\left(\hat{\alpha}_{R R}\right)-\operatorname{MSE}\left(\hat{\alpha}_{R R}^{*}\right)$ can be positive because

$$
\sum_{i=1}^{p}\left(\lambda_{i}+\hat{\sigma}\right)^{2} \lambda_{i} \geq \sum_{i=1}^{p}\left(\lambda_{i}+k\right)^{2}\left(\lambda_{i}+\hat{\alpha}_{i}^{2}\right) .
$$

Thus, $\operatorname{MSE}\left(\hat{\alpha}_{R R}\right) \geq \operatorname{MSE}\left(\hat{\alpha}_{R R}^{*}\right)$

## Simulation Study

Consider the behavior of the proposed parameter estimators via a simulation study. Most of the researchers compare the performance of their suggested ridge parameter in the sense of smaller MSE compared to LS and other well-known existing ridge parameters via ridge regression estimators. But, we evaluate the performance of our suggested ridge parameter by considering following different situations in linear regression when data exhibits with multicollinearity.

Case I. Data generated using normal errors.
Case II. Data generated using heteroscedastic errors.
Case III. Data generated using outlier observations.
Case IV. Data generated using outlier observations and heteroscedastic errors.
Case V. Data generated using non-normal errors.
Consider the average MSE (AMSE) of the $\hat{\beta}_{L S}, \hat{\beta}_{R R}, \hat{\beta}_{J M R}, \hat{\beta}_{J R M}$ and $\hat{\beta}_{R R}^{*}$ estimators for different degrees of multicollinearity. We consider the true model as $Y=X \beta+\varepsilon$. Following McDonald and Galarneau (1975) the explanatory variables are generated by

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

$$
x_{i j}=\left(1-\rho^{2}\right)^{1 / 2} u_{i j}+\rho u_{i p}, \quad i=1,2, \ldots, n \quad j=1,2, \ldots, p
$$

where uij are independent standard normal pseudo-random numbers and $\rho$ is specified so that the theoretical correlation between any two explanatory variables is given by $\rho 2$. In this study, to investigate the effects of different degrees of multicollinearity on the estimators, we consider two different correlations, $\rho=0.85,0.90,0.95$ and $0.99 . \beta$ parameter vectors are chosen arbitrarily for $p=4$, respectively. We assumed samples of size of 25,60 and 100 . Estimators $\hat{\beta}_{L S}, \hat{\beta}_{R R}$, $\hat{\beta}_{J M R}, \hat{\beta}_{J R M}$ and $\hat{\beta}_{R R}^{*}$ are computed and obtained the average MSE (AMSE) of estimators. The experiment is repeated 1500 times using the following expression.

$$
\operatorname{AMSE}(\hat{\beta})=\frac{1}{1500} \sum_{i=1}^{p} \sum_{j=1}^{1500}\left(\hat{\beta}_{i j}-\beta_{i}\right)^{2}
$$

where, $\hat{\beta}_{i j}$ denote the estimator of the $i^{\text {th }}$ parameter in the $j^{\text {th }}$ replication and $\beta_{i}$, $i=1,2, \ldots, p$ are the true parameter values. Consider the method that leads to the minimum AMSE to the best from the MSE point of view.

Case I. Here $\varepsilon$ follows a normal distribution $N\left(0, \sigma^{2} I_{n}\right)$. The variance of the error terms is taken as $\sigma^{2}=1,5,10$ and 25 . Firstly, we computed the AMSE values for $\hat{\beta}_{L S}, \hat{\beta}_{R R}, \hat{\beta}_{J M R}, \hat{\beta}_{J R M}$ and $\hat{\beta}_{R R}^{*}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$ for $p=4$ and reported in Table 1.

## A V DORUGADE

Table 1. AMSE of LS and various ridge estimators ( $p=4$ and $\left.\beta=(5,3,4,2)^{\prime}\right)$

| $\rho$ | Estimator | $\mathrm{n}=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{\text {LS }}$ | 3.38 | 93.73 | 394.86 | 2451.82 | 0.59 | 12.36 | 46.85 | 297.86 |
|  | $\hat{\beta}_{R 8}$ | 2.77 | 33.84 | 125.85 | 726.46 | 0.57 | 7.86 | 20.38 | 98.82 |
|  | $\hat{\beta}_{\text {sun }}$ | 2.50 | 45.32 | 203.00 | 1202.44 | 0.68 | 7.34 | 24.22 | 152.46 |
|  | $\hat{\beta}_{\text {man }}$ | 3.00 | 50.99 | 192.95 | 1146.93 | 0.58 | 9.87 | 28.98 | 153.65 |
|  | $\hat{\beta}_{R 8}{ }^{\text {en }}$ | 2.03 | 14.21 | 34.35 | 138.08 | 0.50 | 5.27 | 12.93 | 54.74 |
| 0.90 | $\hat{\beta}_{\text {Ls }}$ | 19.20 | 526.30 | 1694.80 | 12498.50 | 2.90 | 90.30 | 281.50 | 2202.30 |
|  | $\hat{\beta}_{R R}$ | 10.66 | 172.23 | 468.04 | 3571.36 | 2.54 | 37.88 | 86.29 | 717.45 |
|  | $\hat{\beta}_{\text {sun }}$ | 11.01 | 273.95 | 810.47 | 6158.55 | 2.49 | 47.82 | 132.76 | 1196.13 |
|  | $\hat{\beta}_{\text {sen }}$ | 13.65 | 241.76 | 740.49 | 5369.88 | 2.78 | 52.35 | 140.80 | 1046.13 |
|  | $\hat{\beta}_{R R}$ | 4.14 | 12.58 | 22.72 | 82.23 | 1.80 | 8.29 | 11.38 | 38.4 |
| 0.95 | $\hat{\beta}_{\text {LS }}$ | 163.00 | 5120.00 | 19151.00 | 129866.0 | 31.00 | 968.00 | 3581.00 | 17759.00 |
|  | $\hat{\beta}_{R 8}$ | 47.40 | 1659.60 | 4884.20 | 34858.20 | 15.20 | 308.10 | 1047.60 | 4582.80 |
|  | $\hat{\beta}_{\text {sun }}$ | 68.60 | 2273.70 | 7997.70 | 51354.00 | 19.80 | 447.40 | 1576.00 | 7899.10 |
|  | $\hat{\beta}_{\text {sen }}$ | 74.20 | 2619.40 | 8863.60 | 65176.40 | 16.50 | 507.60 | 1777.80 | 8105.60 |
|  | $\hat{\beta}_{R 8}$ | 4.72 | 7.17 | 12.51 | 49.62 | 4.03 | 5.98 | 6.76 | 17.48 |
| 0.99 | $\hat{\beta}_{\text {LS }}$ | 1804.00 | 48622.00 | $\begin{array}{r} 220824.0 \\ 0 \end{array}$ | $\begin{array}{r} 1400590 . \\ 00 \end{array}$ | 320.00 | 7592.00 | 30461.00 | 213251.0 0 |
|  | $\hat{\beta}_{R R}$ | 470.00 | 14008.00 | 76100.00 | $\begin{array}{r} 387524.0 \\ 0 \end{array}$ | 84.00 | 2015.00 | 8000.00 | 61836.00 |
|  | $\hat{\beta}_{\text {sun }}$ | 717.00 | 21406.00 | $\begin{array}{r} 105437.0 \\ 0 \end{array}$ | $573785.0$ | 144.00 | 3306.00 | 12875.00 | 90558.00 |
|  | $\hat{\beta}_{\text {ren }}$ | 840.00 | 23485.00 | $\begin{array}{r} 119862.0 \\ 0 \end{array}$ | $722312.0$ | 141.00 | 3514.00 | 14124.00 | $104016.0$ |
|  | $\hat{\beta}_{R R}$ | 5.05 | 6.94 | 10.52 | 47.23 | 4.93 | 5.28 | 6.75 | 13.24 |

From Table 1, we observe that performance of our proposed estimator $\hat{\beta}_{R R}^{*}$ is better than $\hat{\beta}_{L S}, \hat{\beta}_{R R}, \hat{\beta}_{J M R}$ and $\hat{\beta}_{J R M}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$. Because $\hat{\beta}_{R R}, \hat{\beta}_{J M R}$ gives better performance than $\hat{\beta}_{L S}$ and $\hat{\beta}_{J R M}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$. Particularly for increasing degree of multicollinearity, $\hat{\beta}_{R R}^{*}$ gives significantly smaller AMSE values as compare to other estimators.

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

Case II. Consider the problem of heteroscedasticity. Assume that the elements of the random vector $\varepsilon$ were not independent and identically distributed random variables. To introduce heteroscedastics and/or correlated errors in the model given in (2) and converted into model given in (7) matrix $V$ will be estimated by method suggested by Firinguetti (1989). In the present study we choose $\rho=0.95$ and consider matrix $V$ is estimated as below

Table 2. AMSE of LS and various ridge estimators ( $p=4$ and $\left.\beta=(5,7,3,1)^{\prime}\right)$

| $\rho$ | Estimator | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{L S}$ | 20.10 | 596.90 | 2180.80 | 12215.10 | 8.60 | 297.80 | 928.00 | 6097.90 |
|  | $\hat{\beta}_{R R}$ | 17.31 | 451.55 | 1648.97 | 9146.44 | 8.48 | 264.17 | 802.37 | 5347.04 |
|  | $\hat{\beta}^{\text {JMR }}$ | 18.34 | 481.14 | 1786.08 | 9709.46 | 8.53 | 270.53 | 835.75 | 5524.28 |
|  | $\hat{\beta}_{J R M}$ | 17.30 | 501.90 | 1819.30 | 10198.20 | 8.83 | 275.20 | 846.90 | 5609.70 |
|  | $\hat{\beta}_{R R}^{*}$ | 13.62 | 333.51 | 1186.55 | 6739.84 | 7.40 | 181.94 | 568.38 | 3582.63 |
| 0.90 | $\hat{\beta}_{L S}$ | 78.80 | 1723.70 | 6451.40 | 46119.30 | 26.60 | 693.50 | 2406.00 | 18390.80 |
|  | $\hat{\beta}_{R R}$ | 55.80 | 1002.10 | 3739.50 | 27348.40 | 24.20 | 545.10 | 1839.00 | 14318.50 |
|  | $\hat{\beta}^{\text {JMR }}$ | 61.50 | 1139.30 | 4336.70 | 32280.90 | 24.70 | 580.30 | 1974.20 | 15252.90 |
|  | $\hat{\beta}_{J R M}$ | 61.80 | 1263.50 | 4713.30 | 34289.70 | 26.80 | 607.50 | 2074.60 | 16102.30 |
|  | $\hat{\beta}_{R R}^{*}$ | 20.50 | 334.41 | 1069.86 | 7534.85 | 12.05 | 153.74 | 695.38 | 4277.38 |
| 0.95 | $\hat{\beta}_{L S}$ | 669.00 | 15446.00 | 68939.00 | 393609.00 | 176.00 | 4798.00 | 20237.00 | 135777.00 |
|  | $\hat{\beta}_{R R}$ | 378.00 | 8964.00 | 37083.00 | 216822.00 | 130.00 | 3306.00 | 13954.00 | 95154.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 448.00 | 10472.00 | 43461.00 | 255218.00 | 141.00 | 3630.00 | 15264.00 | 105690.00 |
|  | $\hat{\beta}^{\text {JRM }}$ | 490.00 | 11207.00 | 50079.00 | 279251.00 | 142.00 | 3893.00 | 16606.00 | 112258.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 21.14 | 339.32 | 1065.06 | 8295.32 | 12.41 | 125.99 | 492.23 | 2897.20 |
| 0.99 | $\hat{\beta}_{L S}$ | 5874.00 | 156068.00 | 599010.00 | 4373664.00 | 2221.00 | 56796.00 | 240541.00 | 1323948.00 |
|  | $\hat{\beta}_{R R}$ | 3093.00 | 77901.00 | 333494.00 | 2335520.00 | 1635.00 | 41391.00 | 164114.00 | 958398.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 3773.00 | 93485.00 | 394200.00 | 2770590.00 | 1777.00 | 45134.00 | 185561.00 | 1038487.00 |
|  | $\hat{\beta}_{J R M}$ | 4072.00 | 106393.00 | 441980.00 | 3111729.00 | 1881.00 | 48157.00 | 198360.00 | 1115227.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 17.44 | 277.09 | 1282.84 | 6875.21 | 9.27 | 143.95 | 531.16 | 3250.81 |

$$
V=\frac{1}{1-\rho^{2}}\left[\begin{array}{cccc}
1 & \rho & \cdots & \rho^{n-1} \\
\rho & 1 & \cdots & \rho^{n-2} \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
\rho^{n-1} & \rho^{n-2} & \cdots & 1
\end{array}\right]
$$

The AMSE was computed for various values of combination $\left(\rho, n, V, \sigma^{2}\right)$ for $p=4$ and reported in

From the results reported in Table 2, observe $\hat{\beta}_{R R}, \hat{\beta}_{J M R}$ gives equivalently better performance than $\hat{\beta}_{L S}$ and $\hat{\beta}_{J R M}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$. However performance of $\hat{\beta}_{R R}^{*}$ is better than $\hat{\beta}_{L S}, \hat{\beta}_{R R}, \hat{\beta}_{J M R}$ and $\hat{\beta}_{J R M}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$.

Case III. The same simulation experiment is repeated for 1500 times for all combinations of $\rho, n$, and $\sigma^{2}$ by introducing one and two outliers for different model specifications. For $p=4$ we computed the AMSE for various values of combination ( $\rho, n, V, \sigma^{2}$ ) for one and two outliers and reported in Table 3 and 4 respectively.

Results in Table 3 and 4, shows that $\hat{\beta}_{J R M}$ gives better performance than $\hat{\beta}_{L S}, \hat{\beta}_{R R}$ and $\hat{\beta}_{J M R}$ for all combinations of $\rho, n$, and $\sigma^{2}$. However, proposed estimator $\hat{\beta}_{R R}^{*}$ gives better performance than $\hat{\beta}_{L S}, \hat{\beta}_{R R}$ and $\hat{\beta}_{J M R}$ including $\hat{\beta}_{J R M}$ for various values of triplet $\left(\rho, n, \sigma^{2}\right)$.

Case IV. With respect to our secondary aim of the proposed work in this article, here we evaluate the performance of proposed estimator $\hat{\beta}_{R R}^{*}$ against $\hat{\beta}_{L S}$, $\hat{\beta}_{R R}, \hat{\alpha}_{J M R}$ and $\hat{\beta}_{J R M}$ for the simulated data exits with one or multiple outliers and heteroscedastics and/or correlated errors, in linear regression model in the presence of multicollinearity. We introduce respectively one and two outliers in the simulated data with heteroscedastics and/or correlated errors, where heteroscedastics and/or correlated errors are introduced using the same method as given in Case II with the help of matrix $V$ at $\rho=0.95$. The simulation experiment is repeated for 1500 times for all combinations of $\rho, n, \sigma^{2}$ and $V$ and computed

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

AMSE values for one and two outliers respectively, for $p=4$ are reported in Tables 5 and 6.

Table 3. AMSE of LS and various ridge estimators (With one outlier, $p=4$ and $\left.\beta=(10,3,4,1)^{\prime}\right)$

| $\rho$ | Estimator | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{L S}$ | 63093.50 | 51479.90 | 56809.70 | 52816.70 | 1780.00 | 1744.00 | 1845.80 | 2164.60 |
|  | $\hat{\beta}_{R R}$ | 19398.90 | 13650.00 | 16861.10 | 14144.10 | 510.70 | 519.20 | 540.70 | 569.40 |
|  | $\hat{\beta}_{\text {JMR }}$ | 28805.10 | 22122.40 | 28174.90 | 22578.70 | 858.20 | 861.20 | 864.90 | 951.70 |
|  | $\hat{\beta}_{J R M}$ | 6261.57 | 4513.43 | 5729.10 | 4750.19 | 159.57 | 167.09 | 177.22 | 174.30 |
|  | $\hat{\beta}_{R R}^{*}$ | 3544.36 | 3922.95 | 4084.13 | 3994.31 | 299.85 | 268.90 | 283.91 | 283.74 |
| 0.90 | $\hat{\beta}_{L S}$ | 324191.00 | 308106.00 | 304710.00 | 335486.00 | 9945.00 | 9611.00 | 9304.00 | 11664.00 |
|  | $\hat{\beta}_{R R}$ | 95866.80 | 91106.00 | 78611.80 | 90283.00 | 2893.40 | 2593.80 | 2465.90 | 3214.60 |
|  | $\hat{\beta}^{\text {JMR }}$ | 146348.00 | 138431.00 | 119726.00 | 138150.00 | 4469.00 | 4023.00 | 3873.00 | 4911.00 |
|  | $\hat{\beta}_{\text {JRM }}$ | 28132.50 | 25922.10 | 20062.80 | 24722.10 | 768.80 | 636.50 | 603.70 | 873.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 2163.99 | 2379.11 | 1851.32 | 2123.19 | 165.03 | 169.74 | 171.49 | 208.08 |
| 0.95 | $\hat{\beta}_{L S}$ | 3563590.00 | 3754264.00 | 3331604.00 | 3317982.00 | 101439.00 | 109183.00 | 102304.00 | 129443.00 |
|  | $\hat{\beta}_{R R}$ | 1013175.00 | 1025041.00 | 818990.00 | 941995.00 | 26462.00 | 30554.00 | 25880.00 | 38298.00 |
|  | $\hat{\beta}_{J M R}$ | 1569621.00 | 1638275.00 | 1360205.00 | 1409372.00 | 41486.00 | 45617.00 | 39268.00 | 58030.00 |
|  | $\hat{\beta}_{\text {JRM }}$ | 277101.00 | 276193.00 | 214808.00 | 251716.00 | 6206.00 | 7271.00 | 6163.00 | 10128.00 |
|  | $\hat{\beta}_{R R}$ | 1244.35 | 1248.52 | 1174.22 | 1319.63 | 65.37 | 62.37 | 63.64 | 75.89 |
| 0.99 | $\hat{\beta}_{L S}$ | 36409135.00 | 33502778.00 | 28002957.00 | 33771993.00 | 1177860.00 | 1021218.00 | 968594.00 | 1107336.00 |
|  | $\hat{\beta}_{R R}$ | 9715560.00 | 8705724.00 | 6824202.00 | 8129113.00 | 295078.00 | 249096.00 | 230598.00 | 295623.00 |
|  | $\hat{\beta}_{J M R}$ | 15072802.00 | 13295260.00 | 10408403.00 | 12343872.00 | 473153.00 | 397830.00 | 375101.00 | 482420.00 |
|  | $\hat{\beta}_{J R M}$ | 2534630.00 | 2280994.00 | 1725382.00 | 1839086.00 | 62366.00 | 53576.00 | 50324.00 | 71652.00 |
|  | $\hat{\beta}_{R R}$ | 1121.53 | 1109.93 | 1258.06 | 1247.55 | 50.70 | 47.03 | 48.06 | 59.10 |

## A V DORUGADE

Table 4. AMSE of LS and various ridge estimators (With two outlier, $p=4$ and $\left.\beta=(3,3,8,1)^{\prime}\right)$

| $\rho$ | Estimator | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{L S}$ | 96023.00 | 103032.00 | 140414.00 | 114534.00 | 3048.00 | 3267.00 | 3679.00 | 3899.00 |
|  | $\hat{\beta}_{\text {RR }}$ | 29822.20 | 31245.50 | 51089.80 | 34497.80 | 916.70 | 1170.50 | 1264.30 | 1230.90 |
|  | $\hat{\beta}^{\text {JMR }}$ | 46684.60 | 50507.30 | 71186.90 | 51338.00 | 1479.80 | 1809.00 | 1828.50 | 1904.30 |
|  | $\hat{\beta}^{\text {JRM }}$ | 11177.50 | 10682.10 | 19345.00 | 12368.40 | 329.40 | 463.90 | 454.20 | 431.70 |
|  | $\hat{\beta}_{R R}^{*}$ | 6494.94 | 7297.01 | 8536.97 | 7910.10 | 493.91 | 571.06 | 563.61 | 626.35 |
| 0.90 | $\hat{\beta}_{L S}$ | 589009.00 | 649515.00 | 531514.00 | 565691.00 | 16516.00 | 15254.00 | 17831.00 | 20356.00 |
|  | $\hat{\beta}_{R R}$ | 170930.00 | 201753.00 | 154311.00 | 162987.00 | 4617.00 | 3778.00 | 4867.00 | 5477.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 265753.00 | 297287.00 | 221441.00 | 254970.00 | 7301.00 | 6493.00 | 7623.00 | 8377.00 |
|  | $\hat{\beta}^{\text {JRM }}$ | 47232.60 | 54401.30 | 46217.10 | 43368.10 | 1252.50 | 956.60 | 1282.50 | 1436.80 |
|  | $\hat{\beta}_{R R}^{*}$ | 5099.35 | 4651.04 | 4997.34 | 4731.34 | 340.98 | 329.05 | 324.26 | 354.44 |
| 0.95 | $\hat{\beta}_{L S}$ | 6581935.00 | 6082993.00 | 6188108.00 | 5363832.00 | 188136.00 | 201837.00 | 167767.00 | 193081.00 |
|  | $\hat{\beta}_{R R}$ | 1936566.00 | 1645660.00 | 1869023.00 | 1084286.00 | 51291.00 | 58167.00 | 37254.00 | 48946.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 2791415.00 | 2505398.00 | 2695831.00 | 1824673.00 | 76196.00 | 90786.00 | 68680.00 | 78656.00 |
|  | $\hat{\beta}_{J R M}$ | 510121.00 | 462091.00 | 545257.00 | 221990.00 | 12362.00 | 15340.00 | 8067.00 | 11273.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 3329.23 | 3610.92 | 3343.09 | 3315.94 | 160.60 | 164.00 | 159.19 | 169.48 |
| 0.99 | $\hat{\beta}_{L S}$ | 57371428.00 | 59252077.00 | 56364293.00 | 63635854.00 | 1909329.00 | 1609903.00 | 2052240.00 | 2219103.00 |
|  | $\hat{\beta}_{R R}$ | 14545081.00 | 16779217.00 | 15804173.00 | 16298348.00 | 466721.00 | 354639.00 | 503548.00 | 615258.00 |
|  | $\hat{\beta}_{\text {JMR }}$ | 23238145.00 | 25256656.00 | 25038333.00 | 25397750.00 | 733936.00 | 639763.00 | 798444.00 | 963890.00 |
|  | $\hat{\beta}_{J R M}$ | 3842991.00 | 4420227.00 | 4316709.00 | 4018464.00 | 106047.00 | 76237.00 | 108351.00 | 145024.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 3300.96 | 3350.57 | 3304.98 | 3206.13 | 135.45 | 134.47 | 140.39 | 154.87 |

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

Table 5. AMSE of LS and various ridge estimators (With one outlier, $p=4$ and $\left.\beta=(7,3,1,2)^{\prime}\right)$

| $\rho$ | Estimator | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{L S}$ | 9600.90 | 8815.80 | 10990.80 | 18945.10 | 164.90 | 382.60 | 1120.40 | 6778.00 |
|  | $\hat{\beta}_{R R}$ | 3220.50 | 3195.20 | 4573.30 | 10069.10 | 92.40 | 279.70 | 881.30 | 5778.20 |
|  | $\hat{\beta}_{J M R}$ | 4746.10 | 4343.20 | 5937.50 | 11893.60 | 108.80 | 305.20 | 937.50 | 6059.80 |
|  | $\hat{\beta}^{\text {JRM }}$ | 1690.05 | 1820.80 | 2773.62 | 7195.41 | 61.33 | 225.06 | 673.48 | 4722.61 |
|  | $\hat{\beta}_{R R}^{*}$ | 1325.28 | 1417.28 | 2279.12 | 6087.94 | 66.49 | 218.87 | 580.02 | 3822.50 |
| 0.90 | $\hat{\beta}_{L S}$ | 30584.90 | 33646.90 | 44180.40 | 69210.60 | 391.30 | 1197.60 | 2950.30 | 19979.50 |
|  | $\hat{\beta}_{R R}$ | 8051.80 | 9919.30 | 13989.70 | 27391.40 | 116.60 | 632.70 | 1839.80 | 15375.60 |
|  | $\hat{\beta}_{J M R}$ | 11883.10 | 14706.20 | 19874.50 | 35823.80 | 171.30 | 738.40 | 2094.40 | 16489.40 |
|  | $\hat{\beta}^{\text {JRM }}$ | 2702.40 | 3636.80 | 5393.60 | 13457.10 | 48.60 | 335.00 | 1032.20 | 10209.30 |
|  | $\hat{\beta}_{R R}^{*}$ | 897.59 | 1639.73 | 2314.36 | 6054.40 | 39.43 | 167.23 | 444.88 | 3598.33 |
| 0.95 | $\hat{\beta}_{L S}$ | 392845.00 | 443548.00 | 444798.00 | 792065.00 | 4038.00 | 11303.00 | 29083.00 | 167873.00 |
|  | $\hat{\beta}_{R R}$ | 97097.00 | 124982.00 | 123019.00 | 313791.00 | 917.00 | 5086.00 | 16915.00 | 117076.00 |
|  | $\hat{\beta}_{J M R}$ | 156817.00 | 186855.00 | 183490.00 | 412848.00 | 1495.00 | 6407.00 | 19801.00 | 127461.00 |
|  | $\hat{\beta}_{\text {JRM }}$ | 25807.00 | 31123.00 | 36186.00 | 110867.00 | 242.00 | 1862.00 | 8229.00 | 65067.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 1150.97 | 1179.11 | 1994.33 | 8665.58 | 40.57 | 149.42 | 635.75 | 3323.51 |
| 0.99 | $\hat{\beta}_{L S}$ | 3271226.00 | 3741131.00 | 3216915.00 | 7165352.00 | 43951.00 | 95120.00 | 223294.00 | 1713269.00 |
|  | $\hat{\beta}_{R R}$ | 735792.00 | 978962.00 | 840373.00 | 2303451.00 | 10190.00 | 42328.00 | 123189.00 | 1177989.00 |
|  | $\hat{\beta}_{\text {JMR }}$ | 1132082.00 | 1446732.00 | 1262678.00 | 3179624.00 | 17044.00 | 52675.00 | 144926.00 | 1303063.00 |
|  | $\hat{\beta}_{J R M}$ | 178739.00 | 238836.00 | 220975.00 | 701483.00 | 2432.00 | 17802.00 | 55416.00 | 649813.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 1058.07 | 1119.13 | 2085.34 | 4880.42 | 45.91 | 129.29 | 431.40 | 2756.36 |

## A V DORUGADE

Table 6. AMSE of LS and various ridge estimators (With two outlier, $p=4$ and $\left.\beta=(5,8,4,1)^{\prime}\right)$

| $\rho$ Estimator |  | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\sigma^{2}=1$ | 5 | 10 | 25 | 1 | 5 | 10 | 25 |
| 0.85 | $\hat{\beta}_{L S}$ | 41261.00 | 41176.10 | 48149.30 | 43909.90 | 772.00 | 1011.90 | 1660.60 | 6055.50 |
|  | $\hat{\beta}_{R R}$ | 14889.10 | 13985.10 | 17189.10 | 17160.50 | 435.00 | 598.70 | 1099.70 | 4683.20 |
|  | $\hat{\beta}^{\text {JMR }}$ | 20563.80 | 20032.50 | 24093.60 | 23708.50 | 497.90 | 698.30 | 1243.60 | 4979.50 |
|  | $\hat{\beta}^{\text {JRM }}$ | 7908.00 | 7977.00 | 10424.80 | 12221.80 | 341.00 | 457.20 | 823.40 | 3597.20 |
|  | $\hat{\beta}_{R R}^{*}$ | 6332.80 | 6256.90 | 9736.60 | 11913.90 | 320.10 | 436.00 | 773.30 | 3101.60 |
| 0.90 | $\hat{\beta}_{L S}$ | 145383.00 | 152150.00 | 171664.00 | 216624.00 | 3410.00 | 3734.00 | 6045.00 | 20448.00 |
|  | $\hat{\beta}_{R R}$ | 33303.10 | 33452.10 | 47627.60 | 67027.20 | 1361.90 | 1626.20 | 3080.40 | 13200.30 |
|  | $\hat{\beta}^{\text {JMR }}$ | 52887.70 | 58909.10 | 70394.90 | 93015.60 | 1760.80 | 2070.50 | 3715.10 | 14821.50 |
|  | $\hat{\beta}^{\text {JRM }}$ | 11866.90 | 12645.60 | 17357.20 | 28180.90 | 569.60 | 799.60 | 1567.10 | 7411.50 |
|  | $\hat{\beta}_{R R}^{*}$ | 5511.50 | 6470.50 | 8104.90 | 14333.30 | 291.80 | 477.70 | 817.90 | 2695.30 |
| 0.95 | $\hat{\beta}_{L S}$ | 1980295.00 | 1955972.00 | 1546191.00 | 2190932.00 | 27950.00 | 29444.00 | 45744.00 | 178431.00 |
|  | $\hat{\beta}_{R R}$ | 509210.00 | 459140.00 | 326485.00 | 605908.00 | 8348.00 | 9357.00 | 18715.00 | 107883.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 792571.00 | 745312.00 | 548092.00 | 918717.00 | 12725.00 | 13522.00 | 24832.00 | 124177.00 |
|  | $\hat{\beta}_{J R M}$ | 137224.00 | 105464.00 | 76545.00 | 176526.00 | 2448.00 | 2937.00 | 6655.00 | 50663.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 6753.40 | 5063.10 | 7185.50 | 10095.80 | 285.20 | 389.60 | 917.90 | 3859.40 |
| 0.99 | $\hat{\beta}_{L S}$ | 15458210.00 | 15216186.00 | 15121157.00 | 18711248.00 | 289717.00 | 365075.00 | 439706.00 | 1939908.00 |
|  | $\hat{\beta}_{R R}$ | 3941721.00 | 3719049.00 | 3121535.00 | 4240048.00 | 81054.00 | 127113.00 | 181203.00 | 1214587.00 |
|  | $\hat{\beta}_{\text {JMR }}$ | 6116418.00 | 6002584.00 | 5272954.00 | 6857110.00 | 116202.00 | 181722.00 | 236596.00 | 1377073.00 |
|  | $\hat{\beta}_{J R M}$ | 1104532.00 | 1048186.00 | 735478.00 | 1104213.00 | 20517.00 | 38537.00 | 63692.00 | 611190.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 5809.60 | 6557.80 | 9078.90 | 15027.20 | 284.50 | 419.00 | 660.70 | 3859.90 |

From AMSE values reported in Tables 5 and $6, \hat{\beta}_{J R M}$ gives better performance than $\hat{\beta}_{L S}, \hat{\beta}_{R R}$ and $\hat{\beta}_{J M R}$. However, proposed estimator $\hat{\beta}_{R R}^{*}$ gives better performance than $\hat{\beta}_{L S}, \hat{\beta}_{R R}$ and $\hat{\beta}_{J M R}$ including $\hat{\beta}_{J R M}$ for all combinations of $\rho, n, \sigma^{2}$ and $V$. Particularly, $\hat{\beta}_{R R}^{*}$ having the significantly less AMSE values as compare to other estimators for all combinations of $\rho, n, \sigma^{2}$ and $V$.

## IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION

Case V. As with heteroscedastic and/or correlated errors, departure from the normality assumption is also one of the common problems in regression. Assume $\varepsilon$ follows a non-normal distribution. To examine the robustness of all estimators under consideration, random numbers are generated for the error terms ( $\varepsilon$ ) from each of the $t, F$, Chi-square and exponential distributions respectively. The AMSE was computed for various values of triplet ( $\rho, n$, distribution of $\varepsilon$ ) for $p=4$ and reported in Table 7.

Table 7. AMSE of LS and various ridge estimators ( $p=4$ and $\left.\beta=(6,3,5,3)^{\prime}\right)$

| $\rho$ | Estimator | $n=25$ |  |  |  | 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\varepsilon \sim \operatorname{Chi}(6)$ | F(2,3) | T(8) | Exp(4) | Chi (6) | F(2,3) | T(8) | Exp(4) |
| 0.85 | $\hat{\beta}_{L S}$ | 76.24 | 89.45 | 4.61 | 108.34 | 52.91 | 93.15 | 62.91 | 871.96 |
|  | $\hat{\beta}_{\text {RR }}$ | 64.27 | 35.53 | 3.59 | 38.46 | 20.91 | 68.10 | 60.16 | 330.95 |
|  | $\hat{\beta}_{\text {JMR }}$ | 64.32 | 45.86 | 3.37 | 51.85 | 25.30 | 71.55 | 59.56 | 463.09 |
|  | $\hat{\beta}_{\text {JRи }}$ | 68.43 | 49.71 | 4.05 | 56.37 | 32.26 | 75.46 | 61.46 | 486.09 |
|  | $\hat{\beta}_{\text {RR }}$ | 58.16 | 17.56 | 2.58 | 15.95 | 8.35 | 56.71 | 57.78 | 173.11 |
| 0.90 | $\hat{\beta}_{L S}$ | 189.40 | 494.30 | 30.00 | 607.00 | 254.00 | 290.20 | 97.60 | 4609.90 |
|  | $\hat{\beta}_{R R}$ | 99.40 | 159.70 | 14.30 | 157.20 | 72.80 | 124.30 | 71.70 | 1286.50 |
|  | $\hat{\beta}_{\text {JMR }}$ | 122.50 | 255.90 | 15.90 | 263.70 | 108.90 | 168.50 | 75.10 | 2153.00 |
|  | $\hat{\beta}_{\text {JRM }}$ | 119.90 | 233.70 | 19.60 | 257.90 | 124.40 | 162.30 | 78.90 | 2053.30 |
|  | $\hat{\beta}_{R R}$ | 59.43 | 16.93 | 4.59 | 13.71 | 10.56 | 57.36 | 59.62 | 170.73 |
| 0.95 | $\hat{\beta}_{L S}$ | 1597.00 | 4623.00 | 325.00 | 6546.00 | 3147.00 | 2834.00 | 609.00 | 48742.00 |
|  | $\hat{\beta}_{R R}$ | 516.00 | 1162.00 | 89.00 | 1678.00 | 722.00 | 876.00 | 224.00 | 12090.00 |
|  | $\hat{\beta}^{\text {JMR }}$ | 862.00 | 2160.00 | 157.00 | 3054.00 | 1344.00 | 1506.00 | 325.00 | 22501.00 |
|  | $\hat{\beta}_{\text {Jкn }}$ | 737.00 | 1830.00 | 138.00 | 2679.00 | 1223.00 | 1238.00 | 296.00 | 19425.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 57.66 | 12.49 | 5.11 | 9.28 | 5.74 | 54.17 | 59.05 | 121.59 |
| 0.99 | $\hat{\beta}_{L S}$ | 19447.00 | 47712.00 | 4916.00 | 64387.00 | 39836.00 | 27943.00 | 5447.00 | 526282.00 |
|  | $\hat{\beta}_{R R}$ | 6080.00 | 13278.00 | 1596.00 | 15450.00 | 11031.00 | 6690.00 | 1575.00 | 136786.00 |
|  | $\hat{\beta}_{\text {JMR }}$ | 10382.00 | 23410.00 | 2852.00 | 27852.00 | 19415.00 | 12269.00 | 2660.00 | 243775.00 |
|  | $\hat{\beta}_{\text {JRN }}$ | 8360.00 | 20304.00 | 2286.00 | 23779.00 | 16420.00 | 11426.00 | 2313.00 | 210271.00 |
|  | $\hat{\beta}_{R R}^{*}$ | 57.04 | 11.86 | 5.05 | 9.27 | 5.41 | 53.30 | 58.91 | 113.13 |

## A V DORUGADE

From Table 7, $\hat{\beta}_{R R}^{*}$ gives better performance than $\hat{\beta}_{L S}, \hat{\beta}_{R R}, \hat{\beta}_{J M R}$ and $\hat{\beta}_{J R M}$ for various values of triplet ( $\rho, n$, distribution of $\varepsilon$ ). This indicates $\hat{\beta}_{R R}^{*}$ is not sensitive to the departure from the normality assumption of error terms.

From Case I to V in above simulation study, it is observed that $\hat{\beta}_{L S}$ is sensitive in each above case and produces unreliable results. Among estimators $\hat{\beta}_{R R}, \hat{\beta}_{J M R}$ and $\hat{\beta}_{J R M}$ no any estimator is better in each of the above cases. But $\hat{\beta}_{R R}^{*}$ is superior than other estimators in each of the above cases for different combinations of size of the sample ( $n$ ), level of multicollinearity $(\rho)$, variance of the error term $\left(\sigma^{2}\right)$, number of predictors $(p)$, matrix $V$ and number of outliers. The novel feature of the proposed estimator is that it can be used without any modification in the proposed estimator it is better alternative to combat one or more than one problems among multicollinearity, outliers, heteroscedastics and/or correlated errors and departure from the normality assumption.

## Conclusion

A new estimation method for the ridge parameter and hence the ridge regression estimator $\hat{\beta}_{R R}^{*}$ was introduced. A simulation study indicated $\hat{\beta}_{R R}^{*}$ gave better performance than other estimators used when the model defined in linear regression exhibits multicollinearity and heteroscedastic and/or correlated errors, non-normal errors, and outliers. The proposed estimator performed well compared with the alternatives considered, and should be useful for practitioners.

## References

Alheety, M. I., \& Kibria, B. M. G. (2009). On the Liu and almost unbiased Liu estimators in the presence of multicollinearity with heteroscedastic or correlated errors. Surveys in Mathematics and its Applications, 4, 155-167.

Alkhamisi, M. A., \& Shukur, G. (2007). Monte Carlo study of recent ridge parameters. Communications in Statistics - Simulation and Computation, 36(3), 535-547. doi: 10.1080/03610910701208619

Al-Hassan, Y. M. (2010). Performance of a new ridge regression estimator. Journal of the Association of Arab Universities for Basic and Applied Sciences, 9(1), 23-26. doi: 10.1016/j.jaubas.2010.12.006

# IMPROVED RIDGE ESTIMATOR IN LINEAR REGRESSION 

Batah, F. S. M., Ramnathan, T. V., \& Gore, S. D. (2008). The efficiency of modified jackknife and ridge type regression estimators: A comparison. Surveys in Mathematics and its Applications, 3, 111-122.

Bayhan, G. M., \& Bayhan, M. (1998). Forecasting using autocorrelated errors and multicollinear predictor variables. Computers \& Industrial Engineering, 34(2), 413-421. doi: 10.1016/S0360-8352(97)00278-7

Birkes, D., \& Dodge, Y. (1993). Alternative Methods of Regression. New York, NY: John Wiley and Sons.

Dorugade, A. V., \& Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9), 447-456.

Firinguetti, L. (1989). A simulation study of ridge regression estimators with autocorrelated errors. Communications in Statistics - Simulation and Computation, 18(2), 673-702. doi: 10.1080/03610918908812784

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., \& Stahel, W. A. (1986). Robust statistics: The approach based on influence function. New York, NY: John Wiley.

Hoerl, A. E., \& Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67. doi: 10.1080/00401706.1970.10488634

Hoerl, A. E., Kennard, R. W., \& Baldwin, K. F. (1975). Ridge regression: Some simulations. Communications in Statistics, 4(2), 105-123.

Huber, P. J. (1981). Robust Statistics. New York, NY: John Wiley.
Jadhav, N. H., \& Kashid, D. N. (2011). A jackknifed ridge m-estimator for regression model with multicollinearity and outliers. Journal of Statistical Theory and Practice, 5(4), 659-673. doi: 10.1080/15598608.2011.10483737

Khalaf, G., \& Shukur, G. (2005). Choosing ridge parameter for regression problems. Communications in Statistics - Theory and Methods, 34(5), 1177-1182. doi: 10.1081/STA-200056836

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics - Simulation and Computation, 32(2), 419-435.doi: 10.1081/SAC-120017499

Lawless, J. F., \& Wang, P. (1976). Simulation study of ridge and other regression estimators. Communications in Statistics - Theory and Methods, 5(4), 307-323. doi: 10.1080/03610927608827353

Li, Y., \& Yang, H. (2011). On the performance of the jackknifed modified ridge estimator in the linear regression model with correlated or heteroscedastic

## A V DORUGADE

errors. Communications in Statistics - Theory and Methods, 40(15), 2695-2708. doi: 10.1080/03610926.2010.491589

McDonald, G. C., \& Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association, 70(350), 407-416. doi: 10.1080/01621459.1975.10479882

Montgomery, D. C., Peck, E. A., \& Vining, G. G. (2001). Introduction to Linear Regression Analysis (3 ${ }^{\text {rd }}$ ed.). New York, NY: John Wiley.

Muniz, G., \& Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparisons. Communications in Statistics - Simulation and Computation, 38(3), 621-630. doi: 10.1080/03610910802592838

Muniz, G., Kibria, B. M. G., Mansson, K., \& Shukur, G. (2012). On developing ridge regression parameters: A graphical investigation. SORT: Statistics and Operations Research, 36(2), 115-138.

Özkale, M. R. (2008). A jackknifed ridge estimator in the linear regression model with heteroscedastic or correlated errors. Statistics \& Probability Letters, 78(18), 3159-3169. doi: 10.1016/j.spl.2008.05.039

Rousseeuw, P. J., \& Leroy, A. M. (1987). Robust regression and outlier detection. New York, NY: John Wiley and Sons.

Singh, B., \& Chaubey, Y. P. (1987). On some improved ridge estimators. Statistische Hefte, 28(1), 53-67. doi: 10.1007/bf02932590.

Trenkler, G. (1984). On the performance of biased estimators in the linear regression mode with correlated or heteroscedastic errors. Journal of Econometrics, 25(1/2), 179-190. doi: 10.1016/0304-4076(84)90045-9

Yang, H., \& Chang, X. (2010). A new two-parameter estimator in linear regression. Communications in Statistics - Theory and Methods, 39(6), 923-934. doi: 10.1080/03610920902807911

# Some Tests for Seasonality in Time Series Data 

Eleazar Chukwunenye Nwogu<br>Federal University of Technology, Owerri<br>Owerri, Imo State, Nigeria

Iheanyi Sylvester Iwueze

Federal University of Technology, Owerri
Owerri, Imo State, Nigeria

Valentine Uchenna Nlebedim<br>Federal University of Technology, Owerri<br>Owerri, Imo State, Nigeria

This paper presents some tests for seasonality in a time series data which considers the model structure and the nature of trending curve. The tests were applied to the row variances of the Buys Ballot table. The Student $t$-test and Wilcoxon Signed-Ranks test have been recommended for detection of seasonality.

Keywords: Model structure, trending curves and seasonal indices, Buys-Ballot table, row variances, paired sample data

## Introduction

Decision making is paramount to any organization. Making a good decision depends largely on predicting future events and conditions. The basic assumption made when forecasting is that there is always an underlying pattern which describes the event and conditions, and that it repeats in the future. A time series is a chronological sequence of observations on a particular variable. Hence, there are two major goals of time series analysis: (1) identifying the nature of the phenomenon represented by the sequence of observations; and (2) forecasting (predicting future values of the time series variable). Identification of the pattern and choice of model in time series data is critical to facilitate forecasting. Thus, both of these goals of time series analysis require that the pattern of observed time series data is identified and described. Two patterns that may be present are trend and seasonality. In order to understand the effectiveness of identification of patterns of observed time series data, it is important to first identify what a time series

[^22]
## NWOGU ET AL.

consists of. In time series analysis, it is assumed that the data consists of a systematic pattern (usually a set of identifiable components) and random noise (error). Most time series patterns can be described in terms of four basic classes of components: The systematic pattern includes the trend (denoted as $T_{t}$ ), seasonal (denoted as $S_{t}$ ), and cyclical (denoted as $C_{t}$ ) components. The irregular component is denoted as $I_{t}$ or $e_{t}$, where $t$ stands for the particular point in time. These four classes of time series components may or may not coexist in real-life data.

The two main goals of a time series analysis are better achieved if the correct model is used. The specific functional relationship among these components can assume different forms. However, the possibilities are that they are combined in an additive (additive seasonality) or a multiplicative (multiplicative seasonality) fashion, but can also take other forms such as pseudo-additive/mixed (combining the elements of both the additive and multiplicative models) model.

The additive model (when trend, seasonal and cyclical components are additively combined) is given as:

$$
\begin{equation*}
X_{t}=T_{t}+S_{t}+C_{t}+I_{t}, \quad t=1,2, \ldots, n \tag{1}
\end{equation*}
$$

The multiplicative model (when trend, seasonal and cyclical components are multiplicatively combined) is given as:

$$
\begin{equation*}
X_{t}=T_{t} \times S_{t} \times C_{t} \times I_{t}, \quad t=1,2, \ldots, n \tag{2}
\end{equation*}
$$

and the Pseudo-Additive/Mixed Model (combining the elements of both the additive and multiplicative models) is given as:

$$
\begin{equation*}
X_{t}=T_{t} \times S_{t} \times C_{t}+I_{t}, \quad t=1,2, \ldots, n \tag{3}
\end{equation*}
$$

Cyclical variation refers to the long term oscillation or swings about the trend, and only long period sets of data will show cyclical fluctuation of any appreciable magnitude. If short periods of time are involved (which is true of all examples in this study), the cyclical component is superimposed into the trend (Chatfield, 2004) and then the trend-cycle component is denoted by $M_{t}$. In this case, (1), (2), and (3) may, respectively, be written as:

$$
\begin{equation*}
X_{t}=M_{t}+S_{t}+I_{t}, \quad t=1,2, \ldots, n \tag{4}
\end{equation*}
$$

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

$$
\begin{array}{ll}
X_{t}=M_{t} \times S_{t} \times I_{t}, & t=1,2, \ldots, n \\
X_{t}=M_{t} \times S_{t}+I_{t}, & t=1,2, \ldots, n \tag{6}
\end{array}
$$

The pseudo-additive model is used when the original time series contains very small or zero values. However, this work will discuss only the additive and multiplicative models.

As long as the trend is monotonous (consistently increasing or decreasing), the identification of the trend component is not very difficult. Tests for trend are given in Kendall and Ord (1990). The cyclical component exhibits variation at periods that may be fixed or not fixed, but which are predictable. Many time series exhibit a variation which repeats itself in systematic intervals over time and this behavior is known as seasonal dependency (seasonality). The seasonal component, $S_{t}$, is associated with the property that $S_{(i-j) s+j}=S_{j}, i=1,2, \ldots$. The difference between a cyclical and a seasonal component is that the latter occurs at regular (seasonal) intervals, although cyclical factors have usually a longer duration that varies from cycle to cycle.

In some time series data, the presence of a seasonal effect in a series is quite obvious and the seasonal periods are easy to find (e.g., 4 for quarterly data, 12 for monthly data, etc.). Seasonality can be visually identified in the series as a pattern that repeats every $k$ elements. The following graphical techniques can be used to detect seasonality: (1) a run sequence plot (Chambers, Cleveland, Kleiner, \& Tukey, 1983); (2) a seasonal sub-series plot (Cleveland, 1993); (3) multiple box plots (Chambers et al., 1983); and (4) the autocorrelation plot (Box, Jenkins, \& Reinsel, 1994). Both the seasonal subseries plot and the box plot assume that the seasonal periods are known. If there is significant seasonality, the autocorrelation plot should show spikes at multiples of lags equal to the period, the seasonal lag (Box et al., 1994). For quarterly data, we would expect to see significant spikes at lag $4,8,12$, 16, and so on. Iwueze, Nwogu, Ohakwe, and Ajaraogu (2011) pointed out that seasonality in time series can be identified from the time plot of the entire series by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every period/year, relative to the trend.

In some cases the presence of a seasonal effect in a series is not quite obvious and, therefore, testing is required in order to confirm the presence of the seasonal effect in a series. Davey and Flores (1993) proposed a method which adds statistical tests of seasonal indexes for the multiplicative model that helps identify seasonality with greater confidence. Tests for seasonality are also given in Kendall and Ord
(1990). Chatfield (2004) suggested the use of the Buys Ballot table for inspecting time series data for the presence of trend and seasonal effects. Fomby (2008) presented various graphs suggested by the Buys Ballot table for inspecting time series data for the presence of seasonal effects. Fomby (2010), in his study of Stable Seasonal Pattern (SSP) models, gave an adaptation of Friedman's two-way analysis of variance by ranks test for seasonality in time series data. Several statistics have also been proposed to test for seasonality. They can be broken down into three groups: the Chi-Square ( $\chi^{2}$ ) Goodness-of-Fit test and the Kolmogorov-Smirnov type statistic, the Harmonic analyses based on the Edwards' type statistic (Edwards, 1961), and the Nonparametric Tests.

The $\chi^{2}$ goodness-of-fit test is relatively popular for detecting seasonality because of its simple mathematical theory, which makes it easy to calculate and understand (Hakko, 2000). The test is on whether the empirical data can be a sample of a certain distribution with sampling error as the only source of variability (McLaren, Legler, \& Brittenham, 1994). This test requires a sample from a population with an unknown distribution function $\mathrm{F}(x)$ and a certain theoretical distribution function $\mathrm{F}_{0}(x)$. Although there is no restriction on the underlying distribution, usually the hypothetical distribution is a uniform distribution.

For seasonality studies, the frequency $O_{i}, i=1,2, \ldots, k$ is the observed value at the $i^{\text {th }}$ season, while the frequency $E_{i}, i=1,2, \ldots, k$ is the expected cell frequency at the $i^{\text {th }}$ season. Under the null hypothesis that there is no seasonal effect (i.e., $\mathrm{F}_{0}(x)$ is a uniform distribution), then $E_{1}=E_{2}=\ldots=E_{k}$ and the statistic

$$
\begin{equation*}
T=\sum_{i=1}^{k}\left[\frac{\left(O_{i}-E_{i}\right)^{2}}{E_{i}}\right] \tag{7}
\end{equation*}
$$

is asymptotically distributed as $\chi^{2}$ with $v=k-1$ degrees of freedom (Horn, 1977). The $\chi^{2}$ goodness-of-fit test for seasonality has been recently used for the analysis of seasonality in suicide, myocardial infarction, diarrhoea, pneumonia, and overall mortality (Flisher, Parry, Bradshaw, \& Juritz, 1996; Herring \& Hoppa, 1997; Rihmer, Rutz, Pihlgren, \& Pestiality, 1998; Sheth, Nair, Muller, \& Yusuf, 1999; Underwood, 1991; Villa, Guisecafré, Martinez, \& Muñoz, 1999).

In his article, Edwards (1961) explicitly mentions the possibility to estimate cyclic trends by considering the ranking order of the events which are above or below the median number. This idea was used by Hewitt, Milner, Csima, and Pakula (1971) but did not use a binary indicator as suggested by Edwards (1961), instead using all of the ranking information. Rogerson (1996) made an attempt to

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

generalize this test by relaxing the relatively strict assumption of Hewitt et al. (1971) that seasonality is only present if a six-month peak period is followed by a six-month trough period. Rogerson (1996) allowed that the peak period can also last three, four, or five months. Rau (2005) further relaxed these assumptions and allows total flexibility for the basic time duration as well as for the length of the peak period.

The Kolmogorov-Smirnov goodness-of-fit test (KS-Test) is comparable to the $\chi^{2}$ goodness-of-fit test because both approaches are designed to test if a sample drawn from a population fits a specified distribution. However, the KS-Test does not compare observed and expected frequencies at each season, but rather the cumulative distribution functions between the ordered observed and expected values (Rau, 2005).

For seasonality studies, if $\mathrm{F}_{N}(t), t=1,2, \ldots, s$, is the empirical distribution function based on the observed frequencies at each season and $\mathrm{F}_{0}(t)$ is the corresponding distribution function under the null hypothesis that there is no seasonal effect, the test-statistic used is:

$$
\begin{equation*}
T=V_{N} \sqrt{N}=\left[\max _{1 \leq t \leq 12}\left(\mathrm{~F}_{N}(t)-\mathrm{F}_{0}(t)\right)+\max _{1 \leq t \leq 12}\left(\mathrm{~F}_{N}(t)-\mathrm{F}_{0}(t)\right) \mid\right] \tag{8}
\end{equation*}
$$

The statistic $T$ does not follow any of the known distributions (e.g. $\chi^{2}, \mathrm{~N}\left(\mu, \sigma^{2}\right)$, etc.). The distribution of $T$ was determined empirically by Freedman (1979) using Monte Carlo simulations and tabulated in Freedman's article. Freedman's modified KS-Type Test has been used for the study of seasonality (Verdoux, Takei, Cassou de Saint-Mathurin, \& Bourgeois, 1997).

In all these tests for the presence of seasonal effect in a time series data, the model structure (i.e. whether Additive or Multiplicative models) and nature of the trending curve (Linear, Quadratic, Exponential, etc.) were not taken into consideration. However, Iwueze and Nwogu (2014) have shown that, for precise detection of presence of seasonal effect in a series the model, structure and trending curves are necessary. Some of the questions that come to mind are: "How does the model structure affect the detection of presence of seasonal effect in a time series data?"; "How does the nature of the trending curves affect the test for presence of seasonal effect in a series?" These and other related questions are what this study intends to address.

Therefore, the ultimate objective of this study is to develop tests for seasonality in a series which take into account the nature of the model structure and
trending curves for precise detection of the presence of seasonal effect in a series where it exists. The specific objectives are to:
(a) Review the Buys Ballot procedure for selected trending curves,
(b) Construct test(s) for the detection of presence of seasonal effect in a series using the row, column, overall means and variances of the Buys-Ballot table, and
(c) Assess the performance of the developed test statistics in detection of the presence of seasonal effects in a series using empirical examples.

Based on the results, recommendations are made.
The rationale for this study is to fill the gap in the existing tests for seasonality by providing analyst with objective test for the detection of the presence of seasonal effect in a series when it exists.

The Buys-Ballot procedure was developed by Iwueze and Nwogu (2004) for short period data in which trend and cyclical components are jointly estimated; the tests developed in this study are based on this assumption. In their results, on the basis of which the proposition for choice of appropriate model was made, Iwueze and Nwogu (2014) showed that, for the selected trending curves, the column variances depend only on the trend parameters for the additive model and on both trend parameters and seasonal indices for the multiplicative model. Therefore, if the seasonal/column variances are functions of the trend parameters, only then is Additive the appropriate model. However, if the seasonal/column variances are functions of both the trend parameters and seasonal indices, then the appropriate model is Multiplicative. It is the presence of the seasonal effect in the seasonal/column variances that makes the model multiplicative. In other words, once the seasonal/column variances indicate that the appropriate model is Multiplicative, it also indicates that the series contains seasonal effects. Therefore, in this study, tests for detection of the presence of seasonal effect in a time series data are developed for the additive model only.

For the additive model and all trending curves studied, the row variances contain both the trending parameters and the seasonal component, while the column variances do not contain the seasonal component. Therefore, the parameters of the trending curves have been varied in order to see their effects on the powers of the tests. In particular, the slope parameter $b$ of the linear trend has been assigned the values $b=0.02,0.20$, and 2.00 to check its effect on the test(s).

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

Furthermore, the power of the tests will be measured by considering the percentages of the total simulations in which the test correctly detected the presence of seasonal effect when it exists.

## Methodology

The summary of the row variances for the additive model derived by Iwueze and Nwogu (2014) are shown in Table 1 for the selected trending curves, with

$$
C_{1}=\sum_{j=1}^{s} j S_{j}, \quad C_{2}=\sum_{j=1}^{s} j^{2} S_{j}
$$

Tests for seasonality in the Additive model are constructed by applying the tests for the matched pairs of data to the row variances of the Buys-Ballot table.

Table 1. Summary of row variances of the Buys-Ballot table for the additive model and the selected trending curves

| Trending Curve | Row Variance ( $\sigma_{i}^{2}$ ) |
| :---: | :---: |
| Linear: $a+b t$ | $b^{2}\left(\frac{s(s+1)}{12}\right)+\left(\frac{2 b}{s-1}\right) \sum_{j=1}^{s} j S_{j}+\frac{1}{s-1} \sum_{j=1}^{s} S_{j}^{2}$ |
| Quadratic:$a+b t+c t^{2}$ | $\frac{s(s+1)}{180}\left\{(2 s-1)(8 s-11) c^{2}-30(s-1) b c+15 b^{2}\right\}$ |
|  | $+\frac{1}{s-1}\left\{\sum_{j=1}^{s} s_{j}^{2}+2[b-2 c s] C_{1}+2 c C_{2}\right\}$ |
|  | $+\left\{\frac{s^{2}(s+1)}{3}\left[b c-c^{2}(s-1)+\frac{4 c s C_{1}}{s-1}\right]\right\} i+\left[\frac{s^{3}(s+1) c^{2}}{3}\right] i$ |
| Exponential: $b e^{c t}$ | $b^{2} \mathrm{e}^{2 c[(i-1) s+1]}\left[\left(\frac{1-\mathrm{e}^{2 c s}}{1-\mathrm{e}^{2 c}}\right)-\frac{1}{s}\left(\frac{1-\mathrm{e}^{c s}}{1-\mathrm{e}^{c}}\right)\right]+\sum_{j=1}^{s} S_{j}^{2}+2 b \mathrm{e}^{c(i-1) s} \sum_{j=1}^{s} \mathrm{e}^{c j} S_{j}$ |

[^23]For the matched pairs of data, $\left(U_{i}, V_{i}\right), i=1,2, \ldots, n$, define

$$
\begin{equation*}
d_{i}=U_{i}-V_{i} \tag{9}
\end{equation*}
$$

## NWOGU ET AL.

where, for the $i^{\text {th }}$ observation unit, $U_{i}$ and $V_{i}$ denote measures on two characteristics. For the variable $d_{i}$, any of these test statistics: (a) the Student's $t$-distribution; (b) the sign test; or (c) the Wilcoxon Signed-Rank; can be used to test the null hypothesis that the two characteristics have the same mean or median.

## Student $\boldsymbol{t}$-Distribution

The statistic

$$
\begin{equation*}
t_{c}=\frac{\bar{d}-d_{0}}{S_{d} / \sqrt{n}} \tag{10}
\end{equation*}
$$

is known to follow the Student's $t$-distribution with $n-1$ degrees of freedom under the null hypothesis that the two characteristics have the same mean or median (or are drawn from a population with the same median), where

$$
\bar{d}=\frac{1}{n} \sum_{i=1}^{n} d_{i}, \quad S_{d}=\sqrt{\frac{1}{n-1} \sum_{i=1}^{n}\left(d_{i}-\bar{d}\right)^{2}}
$$

and $d_{0}$ (usually assumed zero under $\mathrm{H}_{0}: d=d_{0}$ ) is the value of the man or median of the deviations under the null hypothesis. The null hypothesis $\left(\mathrm{H}_{0}\right)$ is rejected at $\alpha$ level of significance if $\left|t_{c}\right|>t_{1-\alpha / 2}$, where $t_{1-\alpha / 2}$ is the $100(1-\alpha)$ percentile of the Student's $t$-distribution with $n-1$ degrees of freedom.

## Sign Test

The test statistic for the sign test is $k$, the smaller of the number of positive signs $n_{+}$ and the number of negative signs $n$.. That is

$$
\begin{equation*}
k=\min \left(n_{+}, n_{-}\right) \tag{11}
\end{equation*}
$$

Under the null hypothesis that the medians of the two variates are equal, the random variable $k$ follows the binomial distribution with parameters $n$ and $p=0.5$. That is, the number of positive signs ( $n_{+}$) and negative ( $n-$ ) signs are expected to be equal.

For smaller sample sizes (i.e., $0<n<25$ ), the observed value of $k$ is compared with the critical value $\left(k_{\alpha}\right)$ and the null hypothesis $\left(\mathrm{H}_{0}\right)$ is rejected at $\alpha$ level of significance if $k<k_{\alpha}$, where $k_{\alpha}$ is computed from the binomial probability function as

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

$$
\begin{equation*}
\mathrm{F}\left(k_{\alpha} ; n, p\right)=\operatorname{Pr}\left(X \leq k_{\alpha}\right)=\sum_{x=1}^{\left\lfloor k_{\alpha}\right\rfloor} \frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x} \leq \alpha \tag{12}
\end{equation*}
$$

where $p=0.5$ and $\left\lfloor k_{\alpha}\right\rfloor$ is the "floor" under $k_{\alpha}$, i.e., the greatest integer less than or equal to $k_{\alpha}$ (Corder \& Foreman, 2014).

For larger sample sizes (i.e., $n \geq 25$ ), Corder and Foreman (2014) recommended the use of $z_{c}$, given as

$$
\begin{equation*}
z_{c}=\frac{\left(k^{\prime}-0.5\right)-0.5 n}{0.5 \sqrt{n}} \tag{13}
\end{equation*}
$$

where $k^{\prime}=\max \left(n_{+}, n_{-}\right)$. This approximately follows the standard normal distribution under the null hypothesis. The null hypothesis $\left(\mathrm{H}_{0}\right)$ is rejected at $\alpha$ level of significance if $k^{\prime}>z_{\alpha}$ and accepted otherwise.

## Wilcoxon Signed-Ranks Test

For small sample sizes (i.e., $n \leq 30$ ), the Wilcoxon Signed-Ranks test statistic is given by

$$
\begin{equation*}
T_{c}=\operatorname{Min}\left\{\sum_{i=1}^{n_{+}} R_{d_{i}^{+}}, \sum_{i=1}^{n}\left|R_{d_{i}}\right|\right\} \tag{14}
\end{equation*}
$$

where $\sum_{i=1}^{n_{+}} R_{d_{i}^{+}}$is the sum of the positive ranks of non-zero differences and $\left|\sum_{i=1}^{n_{-}} R_{d_{i}^{-}}\right|$is the sum of the absolute values of the negative ranks of non-zero differences. If the null hypothesis $\left(\mathrm{H}_{0}\right)$ is true, these sums are expected to be equal.

For large sample sizes (i.e., $n>30$ ), the Wilcoxon Signed-Ranks test statistic is given by Corder and Foreman (2014) as

$$
\begin{equation*}
z_{c}=\frac{T_{c}-\mu_{T}}{\sigma_{T}} \tag{15}
\end{equation*}
$$

where $n$ is the number of matched pairs of data for which their differences is not zero and

## NWOGU ET AL.

$$
\mu_{T}=\frac{n(n+1)}{4}, \quad \sigma_{T}=\sqrt{\frac{n(n+1)(2 n+1)}{24}}
$$

i.e., the mean and standard deviation of $T_{c}$, respectively, under the null hypothesis. The null hypothesis $\left(\mathrm{H}_{0}\right)$ is rejected at $\alpha$ level of significance if $z_{c}<z_{\alpha}$ and accepted otherwise.

When the usual parametric assumptions (difference scores are normally and identically distributed in the population from which the sample was drawn and that they are measured on at least an interval scale) are met, the Student's $t$-distribution is used. The sign test and the Wilcoxon signed-ranks test are used when the usual assumptions of parametric tests are not met. It is important to note that the sign test and Wilcoxon signed-ranks test require only that the distribution of the study data be symmetric.

For detection of the presence of seasonal effect in a time series data, we let $U_{i}$ denote the row variance in the presence of the seasonal effect and $V_{i}$ denote row variance in the absence of the seasonal effect.

For example:
(a) For the linear trend-cycle component, in the presence of seasonal effect, the row variance is

$$
\begin{equation*}
U_{i}(L)=\hat{\sigma}_{i}^{2}(L)=b^{2}\left(\frac{s(s+1)}{12}\right)+\left(\frac{2 b}{s-1}\right) \sum_{j=1}^{s} j S_{j}+\frac{1}{s-1} \sum_{j=1}^{s} S_{j}^{2} \tag{16}
\end{equation*}
$$

When there is no seasonal effect, $S_{j}=0 \forall j=1,2, \ldots, s$, and so

$$
\begin{equation*}
V_{i}(L)=b^{2}\left(\frac{s(s+1)}{12}\right) \tag{17}
\end{equation*}
$$

and

$$
\begin{equation*}
d_{i}(L)=U_{i}(L)-V_{i}(L)=\left(\frac{2 b}{s-1}\right) \sum_{j=1}^{s} j S_{j}+\frac{1}{s-1} \sum_{j=1}^{s} S_{j}^{2} \tag{18}
\end{equation*}
$$

which is zero under null hypothesis $\left(\mathrm{H}_{0}: S_{j}=0\right)$.
(b) For the Quadratic trend-cycle component, in the presence of seasonal effect, the row variance is

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

$$
\begin{align*}
U_{i}(Q) & =\hat{\sigma}_{i}^{2}(Q) \\
& =\left\{\begin{array}{l}
\frac{s(s+1)}{180}\left\{(2 s-1)(8 s-11) c^{2}-30(s-1) b c+15 b^{2}\right\} \\
+\frac{1}{s-1}\left\{\sum_{j=1}^{s} S_{j}^{2}+2[b-2 c s] C_{1}+2 c C_{2}\right\} \\
+\left\{\frac{s^{2}(s+1)}{3}\left[b c-c^{2}(s-1)+\frac{4 c s C_{1}}{s-1}\right]\right\} i+\left[\frac{s^{3}(s+1) c^{2}}{3}\right] i^{2}
\end{array}\right\} \tag{19}
\end{align*}
$$

When there is no seasonal effect, $S_{j}=0 \forall j=1,2, \ldots, s, C_{1}=C_{2}=\sum_{j=1}^{s} S_{j}^{2}=0$. Hence

$$
\begin{align*}
& V_{i}(Q)=\left\{\begin{array}{l}
\frac{s(s+1)}{180}\left\{(2 s-1)(8 s-11) c^{2}-30(s-1) b c+15 b^{2}\right\} \\
+\frac{s^{2}(s+1)}{3}\left[b c-c^{2}(s-1)\right] i+\left(\frac{s^{3}(s+1) c^{2}}{3}\right) i^{2}
\end{array}\right\}  \tag{20}\\
& d_{i}(Q)=U_{i}(Q)-V_{i}(Q) \\
& =\frac{1}{s-1}\left[\sum_{j=1}^{s} S_{j}^{2}+2[b-2 c s] C_{1}+2 c C_{2}\right]+\frac{s^{2}(s+1)}{3}\left(\frac{4 c s C_{1}}{s-1}\right) i \tag{21}
\end{align*}
$$

which is zero under null hypothesis $\left(\mathrm{H}_{0}: S_{j}=0\right)$.
(c) For the Exponential trend-cycle component, in the presence of seasonal effect, the row variance is

$$
\begin{align*}
U_{i}(E) & =\sigma_{i}^{2}(E) \\
& =b^{2} \mathrm{e}^{2 c[(i-1) s+1]}\left[\left(\frac{1-\mathrm{e}^{2 c s}}{1-\mathrm{e}^{2 c}}\right)-\frac{1}{s}\left(\frac{1-\mathrm{e}^{c s}}{1-\mathrm{e}^{c}}\right)\right]+\sum_{j=1}^{s} S_{j}^{2}+2 b \mathrm{e}^{c(i-1) s} \sum_{j=1}^{s} \mathrm{e}^{c j} S_{j} \tag{22}
\end{align*}
$$

When there is no seasonal effect, $S_{j}=0 \forall j=1,2, \ldots, s, \sum_{j=1}^{s} S_{j}^{2}=\sum_{j=1}^{s} \mathrm{e}^{c j} S_{j}=0$. Hence

$$
\begin{align*}
& V_{i}(E)=\left\{b^{2} \mathrm{e}^{2 c[(i-1) s+1]}\left[\left(\frac{1-\mathrm{e}^{2 c s}}{1-\mathrm{e}^{2 c}}\right)-\frac{1}{s}\left(\frac{1-\mathrm{e}^{c s}}{1-\mathrm{e}^{c}}\right)\right]\right\}  \tag{23}\\
& d_{i}(E)=U_{i}(E)-V_{i}(E)=\sum_{j=1}^{s} S_{j}^{2}+2 b \mathrm{e}^{c(i-1) s} \sum_{j=1}^{s} \mathrm{e}^{c j} S_{j} \tag{24}
\end{align*}
$$

Which again is zero under the null hypothesis ( $\mathrm{H}_{0}: S_{j}=0$ ).
It is clear from $d_{i}=U_{i}-V_{i}$ (see (18), (21), and (24)) that when the trend dominates the series, the presence of the seasonal effect in a series will be difficult to detect. Therefore, it is advisable to isolate the trend before embarking on test for presence of seasonal effect in a series. It is important to note that the $d_{i}$ represented by (18), (21), and (24) for linear, quadratic, and exponential curves respectively are functions of the seasonal components only when the trend is removed.

## Empirical Examples

This section presents some empirical examples to illustrate the application of the tests for seasonality in time series data discussed previously, and to compare the powers of the tests in the detection of the presence of seasonal effects in a series. The data used consists of 106 data sets of 120 observations each, simulated using the MINITAB software from: (a) $X_{t}=(a+b t)+S_{t}+e_{t}$ with $a=1$ and $b=0.02$, 0.20 , and 2.00, for the linear trend-cycle component; (b) $X_{t}=\left(a+b t+c t^{2}\right)+S_{t}+e_{t}$ with $\mathrm{a}=1, b=2.0$, and $c=3$ for the Quadratic trend-cycle component; and (c) $X_{t}=\left(b \mathrm{e}^{c t}\right)+S_{t}+e_{t}$ with $b=10$ and $c=0.02$ for exponential trend-cycle component. In each case it is assumed that $e_{t} \sim \mathrm{~N}(0,1)$ and $S_{j}, j=1,2, \ldots, 12$ are as shown in Table 2. Meteorological data were collected from the meteorlogical station in Owerri, southeastern Nigeria, for the period of 1990-2010 with the assistance of the computer unit of the Federal Meteorological Centre Oshodi, Lagos. The weather parameters collected are mean monthly values of air temperature, relative humidity, and rainfall. Data on monthly U.S. male ( 16 to 19 years) unemployment figures (in thousands) for the period 1948 to 1981, monthly gasoline demand Ontario (gallon millions) for the period 1960 to 1975, monthly production of Portland cement (thousands of tons) for the period 1956 to 1970, and monthly milk production (pounds per cow) for the period 1962 to 1975, sourced from Hyndman (2014), were used to further illustrate the application of the proposed tests for seasonality in real life time series data.

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

Table 2. Seasonal indices used for simulation

| $\boldsymbol{j}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\mathbf{8}$ | $\mathbf{9}$ | $\mathbf{1 0}$ | $\mathbf{1 1}$ | $\mathbf{1 2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $S_{j}$ | -0.89 | -1.22 | 0.10 | -0.15 | -0.09 | 1.16 | 2.34 | 1.95 | 0.64 | -0.73 | -2.14 | -0.97 |

Table 3. Summary results of tests for seasonality when $b=0.02,0.20$, and 2.00 for linear trend curve

| Slope | Test Statistic | 1\%(0.01) |  | 5\% (0.05) |  | 10\% (0.10) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | \%Pass | \%Fail | \%Pass | \%Fail | \%Pass | \%Fail |
| $b=0.02$ | $t$-test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
|  | S-R test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
|  | Sign test | 84.91 | 15.09 | 99.06 | 0.04 | 99.06 | 0.04 |
| $b=0.20$ | $t$-test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
|  | S-R test | 74.53 | 25.47 | 100.00 | 0.00 | 100.00 | 0.00 |
|  | Sign test | 76.41 | 23.59 | 99.06 | 0.04 | 99.06 | 0.04 |
| $b=2.00$ | $t$-test | 67.92 | 32.62 | 74.53 | 25.47 | 82.08 | 17.92 |
|  | S-R test | 60.38 | 39.62 | 74.53 | 25.47 | 80.11 | 19.81 |
|  | Sign test | 47.17 | 52.83 | 65.09 | 34.91 | 65.09 | 34.91 |

The summary of the results of the application of the three tests for the presence of seasonal effects in the simulated series are shown in Table 3 when the trend-cycle component is present for linear trend curve and Table 4 when trendcycle component is absent for linear, quadratic, and exponential trend curves.

As Table 3 shows, when the slope $b$ is 0.02 , the $t$-test and Wilcoxon signedranks test performed equally well ( $100 \%$ of the time) in detecting the presence of seasonal effect at the three chosen levels of significance (i.e. $1 \%, 5 \%$, and $10 \%$ ). The sign test was able to detect the presence of seasonal effect from at least $84.91 \%$ of the time at $1 \%$ level of significance to about $99.06 \%$ of the time at both $5 \%$ and $10 \%$ levels of significance. When the slope $b$ is increased to 0.20 , the $t$-test was able to detect the presence of seasonality $100 \%$ of the times at the three chosen levels of significance. The Wilcoxon signed-ranks test was able to detect the presence of seasonality $100 \%$ of the time at $5 \%$ and $10 \%$ levels of significance and less than $75 \%$ of the time at $1 \%$ levels of significance. The sign test, on the other hand, was able to detect the presence of seasonal effect about $99.06 \%$ of the time at both $5 \%$ and $10 \%$ levels of significance but at about $76.41 \%$ of the time at $1 \%$ level of significance. For $b=2.00$, all three tests did not perform as well in detection of the presence of seasonal effects in a series.

Table 4. Summary results of tests for seasonality for the de-trended series for linear, quadratic, and exponential trend curves

| Trend Component | Test Statistic | 1\% |  | 5\% |  | 10\% |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | \%Pass | \%Fail | \%Pass | \%Fail | \%Pass | \%Fail |
| Linear: | $t$-test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
| $a=1.0$, | S-R test | 85.85 | 14.15 | 100.00 | 0.00 | 100.00 | 0.00 |
| $b=2.0$ | Sign test | 85.85 | 14.15 | 99.06 | 0.04 | 99.06 | 0.04 |
| Quadratic: | $t$-test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
| $a=1.0, b=2.0$, | S-R test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
| $c=3.0$ | Sign test | 84.91 | 15.09 | 99.06 | 0.04 | 99.06 | 0.04 |
| Exponential: | $t$-test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
| $b=10$ | S-R test | 100.00 | 0.00 | 100.00 | 0.00 | 100.00 | 0.00 |
| $c=0.02$ | Sign test | 96.23 | 4.77 | 100.00 | 0.00 | 100.00 | 0.00 |

The best, the $t$-test, was able to detect the presence of seasonal effects at most $82 \%$ of the time at $10 \%$ level of significance and less than $75 \%$ of the time at $1 \%$ and $5 \%$ levels of significance.

In summary, the performances of all the tests ( $t$-test, Wilcoxon signed-ranks test, and sign test) appear to have decreased with increasing dominance of the trendcycle component in the simulated series and increased with increasing levels of significance. The $t$-test was observed to have performed better than the other two statistical tests applied while the Sign test appears to be trailing behind others.

The results also appear to support the claim made by Iwueze and Nwogu (2014) that it is necessary to de-trend time series data before conducting test for seasonality. This claim was supported by results of (18), (21), and (24). In other to assess the authenticity of this claim, the three tests ( $t$-test, Wilcoxon signed-ranks test, and sign test) were applied to the de-trended series from the simulated series with $b=2.0$. The results of these are shown in Table 4.

The results in Table 4 show that the $t$-test and Wilcoxon signed-ranks test are equal and perfect in performance ( $100 \%$ all through) in detecting the presence of seasonal effects, although the sign test has performance percentages of about $85.85 \%$ at $1 \%$ level of significance and $99.06 \%$ at both $5 \%$ and $10 \%$ significance levels. These are in line with the results obtained when the slope $b=0.02$, and supports the claim that dominance of a series by trend can obscure the presence of seasonal effect in a series.

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

Table 5. Results of tests for seasonality using real life time series data

|  | $t$-Test | Wilcoxon S-R Test | Sign Test |
| ---: | :---: | :---: | :---: |
| Weather Parameter |  | P-Value Sig. (2-tailed) |  |
| Air Temperature | 0.003 | 0.001 | 0.000 |
| Relative Humidity | 0.000 | 0.000 | 0.000 |
| Rain Fall | 0.000 | 0.000 | 0.000 |
| US Male (16-19 years) Unemployment | 0.000 | 0.003 | 0.006 |
| Gasoline Demand | 0.000 | 0.002 | 0.000 |
| Production of Portland Cement | 0.004 | 0.008 | 0.006 |
| Milk Production | 0.000 | 0.002 | 0.000 |

The summary of the results of the application of the three tests for presence of seasonal effect in the real life time series are shown in Table 5. The three proposed tests ( $t$-test, Wilcoxon signed-ranks test, and sign test) performed well in the detection of the presence of seasonal effects in all the real life time series data used, even at $1 \%$ significance level.

## Concluding Remark

In this study, three tests ( $t$-test, Wilcoxon signed-ranks test, and sign test for paired sample data) for detection of seasonal effects in a time series data have been proposed. The tests were developed using the row variances of the Buys-Ballot table when the model structure is additive, and for selected trending curves. The performances of the tests were assessed using simulated series with different trending curves and at different levels of significance, and with real life time series data.

The results of the analysis from the simulated series show that the performances of all three tests to have decreased with increasing dominance of the trend-cycle component in the simulated series, and increased with increasing levels of significance. The $t$-test was observed to have performed better than the other two statistical tests applied, while the sign test appears to be trailing behind others.

When the tests were applied to the de-trended series from a trend dominated series (simulated series with $b=2.00$ ), the results are in line with the results obtained when the slope $b$ is 0.02 . This supports the claim by Iwueze and Nwogu (2014) that dominance of a series by trend can obscure the presence of seasonal effect in a series and that it is necessary to de-trend a time series data before conducting test for seasonality.

In view of these results, it has been recommended that the Student's $t$-test and Wilcoxon signed-ranks test be used for the detection of the presence of seasonal effects in time series data when the model structure is additive until further studies prove otherwise. It has also been recommended that the tests be applied to the detrended series when a series is dominated by trend. Preliminary assessments like the time plot of the study series can offer a guide to determining when a series is dominated by the trend.

Furthermore, when real life time series data were used, the three proposed tests ( $t$-test, Wilcoxon signed-ranks test, and sign test) performed well in detection of the presence of seasonal effect even at $1 \%$ significance level.

## References

Box, G. E. P., Jenkins, G. M., \& Reinsel, G. C. (1994). Time series analysis: Forecasting and control. Englewood Cliffs, NJ: Prentice Hall.

Chambers, J. M., Cleveland, W. S., Kleiner, B., \& Tukey, P. A. (1983). Graphical methods for data analysis. Bellmont, CA: Wadsworth.

Chatfield, C. (2004). The analysis of time series: An introduction. Boca Raton, FL: Chapman \& Hall/CRC.

Cleveland, W. S. (1993). Visualizing data. Murray Hill, NJ: AT \& T Bell Laboratories.

Corder, G. W., \& Foreman, D. I. (2014). Nonparametric statistics: A step-by-step approach (2nd ed.). Hoboken, NJ: John Wiley \& Sons.

Davey, A. M., \& Flores, B. E. (1993). Identification of seasonality in time series: A note. Mathematical and Computer Modelling, 18(6), 73-81. doi: 10.1016/0895-7177(93)90126-J

Edwards, J. H. (1961). The recognition and estimation of cyclic trends. Annals of Human Genetics, 25(1), 83-86. doi: 10.1111/j.1469-
1809.1961.tb01501.x

Flisher, A. J. C., Parry, D. H., Bradshaw, D., \& Juritz. J. M. (1996).
Seasonal variation of suicide in South Africa. Biological Psychiatry, 39(1), 522523. doi: 10.1016/0006-3223(96)84029-7

Fomby, T. B. (2008). Buys ballot plots: Graphical methods for detecting seasonality in time series (Unpublished manuscript). Retrieved from http://faculty.smu.edu/tfomby/eco5375/data/season/buys\ ballot\ plots.pdf

## SOME TESTS FOR SEASONALITY IN TIME SERIES DATA

Fomby, T. B. (2010). Stable seasonal pattern (SSP) model (Unpublished manuscript). Retrieved from http://faculty.smu.edu/tfomby/eco5375/data/notes/Stable\ Seasonal\ Pattern \%20Model_v2.pdf

Freedman, L. S. (1979). The use of a Kolmogorov-Smirnov type statistic in testing hypotheses about seasonal variation. Journal of Epidemiology and Community Health, 33(3), 223-228. doi: 10.1136/jech.33.3.223

Hakko, H. (2000). Seasonal variation of suicides and homicides in Finland, with special attention to statistical techniques used in seasonality studies (Unpublished doctoral thesis). University of Oulu, Fnland.

Herring, D. A., \& Hoppa. R. D. (1997). Changing patterns of mortality seasonality among the western James Bay Cree. International Journal of Circumpolar Health, 56(4), 121-133.

Hewitt, D., Milner, J., Csima, A., \& Pakula, A. (1971). On Edwards' criterion of seasonality and a non-parametric alternative. British Journal of Preventive Social Medicine, 25, 174-176.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. Biometrics, 33(1), 237-247. doi: 10.2307/2529319

Hyndman, R. (2014). Time series data library: Data on monthly U.S. male (16 to 19 years) unemployment figures (in thousands), monthly gasoline demand Ontario (gallon millions), monthly production of Portland cement (thousands of tonnes), and monthly milk production (pounds per cow) [Data files]. Retrieved from http://datamarket.com/data/list/?q=provider:tsdl

Iwueze, I. S., \& Nwogu, E. C. (2004). Buys-Ballot estimates for time series decomposition. Global Journal of Mathematics, 3(2), 83-98. doi:
10.4314/gjmas.v3i2.21356

Iwueze, I. S., \& Nwogu, E. C. (2014). Framework for choice of models and detection of seasonal effect in time series. Far East Journal of Theoretical Statistics, 48(1), 45-66.

Iwueze, I. S., Nwogu, E. C., Ohakwe, J, \& Ajaraogu, J. C. (2011). Uses of the Buys-Ballot table in time series analysis. Applied Mathematics, 2, 633-645. doi: 10.4236/am. 2011.25084

Kendall, M., \& Ord, J. K. (1990). Time series (3rd ed.). London: Griffin.

## NWOGU ET AL.

McLaren, C. E., Legler, J. M., \& Brittenham, G. M. (1994). The generalized $\chi^{2}$ goodness-of-fit test. Journal of the Royal Statistical Society. Series D (The Statistician), 43(2), 247-258. doi: 10.2307/2348342

Rau, R. (2005). Seasonality in human mortality. A demographic approach. (Unpublished doctoral thesis). Universität Rostock, Rostock, Germany.

Rihmer, Z., Rutz, W., Pihlgren, H., \& Pestiality, P. (1998). Decreasing tendency of seasonality in suicide may indicate lowering rate of depressive suicides in the population. Psychiatry Research, 81(2), 233-240. doi: 10.1016/S0165-1781(98)00106-1

Rogerson, P. A. (1996). A Generalization of Hewitt's test for seasonality. International Journal of Epidemiology, 25(3), 644-648. 10.1093/ije/25.3.644

Sheth, T., Nair, C., Muller, J., \& Yusuf, S. (1999). Increased winter mortality from acute myocardial infarction and stroke: The effect of age. Journal of the American College of Cardiology, 33(7), 1916-1919. 10.1016/S0735-1097(99)00137-0

Underwood, J. H. (1991). Seasonality of vital events in a Pacific island population. Social Biology, 38(1-2), 113-126. doi:
10.1080/19485565.1991.9988775

Verdoux, H., Takei, N., Cassou de Saint-Mathurin, R., \& Bourgeois, M. (1997). Analysis of the seasonal variation of schizophrenic births using a Kolmogorov-Smirnov type statistic. European Psychiatry, 12(3), 111-116. doi: 10.1016/S0924-9338(97)80198-8

Villa, S., Guisecafré, H., Martinez, H., \& Muñoz, O. (1999). Seasonal diarrhoeal mortality among Mexican children. Bulletin of the World Health Organization, 77(5), 375-380.

# Multicollinearity and a Ridge Parameter Estimation Approach 

Ghadban Khalaf<br>King Khalid University<br>Abha, Saudi Arabia

Mohammed Iguernane<br>King Khalid University<br>Abha, Saudi Arabia

One of the main goals of the multiple linear regression model, $Y=X \beta+u$, is to assess the importance of independent variables in determining their predictive ability. However, in practical applications, inference about the coefficients of regression can be difficult because the independent variables are correlated and multicollinearity causes instability in the coefficients. A new estimator of ridge regression parameter is proposed and evaluated by simulation techniques in terms of mean squares error (MSE). Results of the simulation study indicate that the suggested estimator dominates ordinary least squares (OLS) estimator and other ridge estimators with respect to MSE.

Keywords: OLS, ridge regression, multicollinearity, simulation; MSC 62J07, 62J05

## Introduction

Consider the general linear regression model

$$
\begin{equation*}
Y=\beta_{0} 1+X \beta+u \tag{1}
\end{equation*}
$$

where $Y$ is an $(n \times 1)$ vector of observations on the dependent variable, $\beta_{0}$ is a scalar intercept, 1 is an $(n \times 1)$ vector with all components equal to unity, $X$ is an $(n \times p)$ matrix of regression variables of full rank $p, \beta$ is the unknown parameter vector of regression coefficients, and $u \sim N\left(0, \sigma^{2} I\right)$ is an $(n \times 1)$ vector of unobservable errors. Because the interest is in estimating $\beta$, omit the constant term $\beta_{0}$ in order to keep the notation simple.

The OLS estimator for the regression parameters is given by

$$
\begin{equation*}
\hat{\beta}=\left(X^{\prime} X\right)^{-1} X^{\prime} Y \tag{2}
\end{equation*}
$$

Dr. Khalaf is a Professor in the Department of Mathematics. Email him at albadran50@yahoo.com. Dr. Iguernane is an Assistant Professor in the Department of Mathematics. Email him at Mohamed.iguernane @gmail.com.

## KHALAF \& IGUERNANE

If any $X$ 's are highly correlated (or, multicollinear), the matrix becomes nonorthogonal, the inversion unstable and the inverse or estimated fractions highly sensitive to random error, and therefore, the OLS solution in (2) has inflated values of the coefficients of regression. Such a regression can be used for prediction, but is worthless in the analysis and interpretation of the individual predictors role in the model. In practice, multicollinearity almost always exists but is typically overlooked or ignored. The following overview stages the later proposed approaches.

## Multicollinearity

Multicollinearity is a high degree of correlation among several independent variables. It commonly occurs when a large number of independent variables are incorporated in a regression model. Only existence of multicollinearity is not a violation of the OLS assumptions. However, a perfect multicollinearity violates the assumption that the $X$ matrix is full ranked, making OLS, given by (2), impossible, because when the model, defined by (1), is not full ranked, then the inverse of $X$ cannot be defined, there can be an infinite number of least squares solutions. Symptoms of multicollinearity may be observed in the following situations:

1. Small changes in the data produce wide swings in the parameters estimates.
2. Coefficients may have very high standard errors and low significance levels even though they are jointly significant and the $R^{2}$ for the regression is high.
3. Coefficients may have the wrong sign or implausible magnitude, Green (2000).

The consequences of multicollinearity are that the variance of the model (i.e. the error sum of squares) and the variances of coefficients are inflated. As a result, any inference is not reliable and the confidence interval becomes wide. Hence, even though the OLS estimator of $\beta$ is the minimum variance unbiased estimator, its MSE will still be large if multicollinearity exists among the independent variables.

To detect multicollinearity, in fact there is no clear-cut criterion for evaluating multicollinearity of linear regression models. We may compute

## MULTICOLLINEARITY AND A RIDGE PARAMETER ESTIMATION

correlation coefficients of independent variables. But high correlation coefficients do not necessarily imply multicollinearity. We can make a judgment by checking related statistics, such as variance inflation factor (VIF) and condition number (CN), where

## Variance Inflation Factor

The VIF is given by

$$
\begin{equation*}
V I F=\frac{1}{1-R_{i}^{2}}, i=1,2, \ldots, p \tag{3}
\end{equation*}
$$

and $R_{i}^{2}$ represents the squared multiple correlation coefficients when $X_{i}$ (the $i^{\text {th }}$ column of $X$ ) is regressed on the remaining $(p-1)$ regressor variables.

The VIF shows how multicollinearity has increased the instability of the coefficient estimates (Freund and Littell, 2000). In other words, it tells us how inflated the variance of the coefficient is, compared to what it would be if the variable were uncorrelated with any other variable in the model (Allison, 1999). However, there is no formal criterion for determining the bottom line of the VIF. Some argue that VIF greater than 10 roughly indicates significant multicollinearity. Others insist that magnitude of model's $R^{2}$ be considered determining significance of multicollinearity. Klein (1962) suggested an alternative criterion that $R_{i}^{2}$ (the coefficient of determination for regression of the $i^{\text {th }}$ independent variable) exceeds $R^{2}$ of the regression model. In this vein, if VIF is greater than $1 /\left(1-R^{2}\right)$, then multicollinearity can be considered statistically significant.

## Condition Number

To quantify the seriousness of multicollinearity, computation of the eigenvalues, $\lambda_{i}$, of the matrix $X^{\prime} X$ is recommended, because the degree of collinearity of any data set is indicated the CN , which is given by

$$
\begin{equation*}
C N=\frac{\lambda_{1}}{\lambda_{p}} \tag{4}
\end{equation*}
$$

where $\lambda_{1}$ is the largest eigenvalue of the matrix $X^{\prime} X$ and $\lambda_{p}$ is the smallest eigenvalue of $X^{\prime} X$.

## KHALAF \& IGUERNANE

A set of eigenvalues of relatively equal magnitudes indicates that there is little multicollinearity (Freund and Littell, 2000). A zero eigenvalue means perfect collinearity among independent variables and very small eigenvalues implies severe multicollinearity. In other words, an eigenvalue close to zero (less than 0.01 , say) or CN greater than 50 indicates significant multicollinearity. Belsley et al. (1980) insist 10 to 100 as a beginning, and maintains that collinearity affects estimates.

There are several ways to solve the problem of multicollinearity. Some of them are

1. Changing specification by omitting or adding independent variables.
2. Obtaining more data (observations) if problems arise because of a shortage of information.
3. Transforming independent variables by taking logarithmic or exponential.
4. Trying biased estimated methods such as ridge regression estimation. The ridge regression estimator has a covariance matrix smaller than that of OLS (Judge, et al., 1985)

## Ridge Regression and a New Proposed Ridge Parameter

Although the OLS estimator is BLUE, it is not necessarily closest to $\beta$, because linearity and unbiasedness are not irrelevant for closeness, particularly when the input matrix of the design is multicollinear. For orthogonal data, the OLS estimator for $\beta$ in the linear regression model is strongly efficient (getting estimates with minimum MSE). But in the presence of multicollinearity, the OLS efficiency can be reduced and hence an improvement upon it would be necessary and desirable. Thus it is natural to look at biased estimator for an improvement over the OLS estimator because it is meaningful to focus on small MSE as the relevant criterion, if a major reduction in variance can be obtained as a result of allowing a little bias. This is precisely what the ridge regression estimator can accomplish.

Ridge regression, due to Hoerl and Kennard (1970), amounts to adding a small positive quantity, say $k$, to each of the diagonal elements of the matrix $X^{\prime} X$. The resulting estimator is

## MULTICOLLINEARITY AND A RIDGE PARAMETER ESTIMATION

$$
\begin{equation*}
\hat{\beta}(k)=\left(X^{\prime} X+k I\right)^{-1} X^{\prime} Y \tag{5}
\end{equation*}
$$

where $k$ is a positive scalar. When $k=0$, (5) reduces to the unbiased OLS estimator given by (2).

Considering $\hat{\beta}(k)$ with regards to MSE

$$
\operatorname{MSE}(\hat{\beta}(k))=\operatorname{Var}(\hat{\beta}(k))+\operatorname{Bias}^{2}(\hat{\beta}(k))=\sigma^{2} \sum_{i=1}^{p} \frac{\lambda_{i}}{\left(\lambda_{i}+k\right)^{2}}+\sum_{i=1}^{p} \frac{k^{2} \beta_{i}^{2}}{\left(\lambda_{i}+k\right)^{2}}
$$

It is known that, as $k$ increases from zero, the MSE initially decreases to a minimum, and then increases with increasing $k$. Hence, there always exists a minimum. Thus it is quite helpful allowing a small bias in order to achieve the main criterion of keeping the MSE small.

When using ridge estimates, the choice of $k$ in (5) is important and several methods have been proposed for this purpose (see, e.g., Hoerl \& Kennard, 1970; McDonald \& Galarneau, 1975; Nomura, 1988; Hag \& Kibria, 1996; Khalaf \& Shukur, 2005; Muniz \& Kibria, 2009; Khalaf, 2011; Khalaf, 2013; Khalaf \& Iguernane, 2014).

Hoerl and Kennard (1970) suggested that the best method for achieving an improved estimate (with respect to MSE) is by choosing

$$
\begin{equation*}
\hat{k}=\frac{\hat{\sigma}^{2}}{\hat{\beta}_{\max }^{2}} \tag{6}
\end{equation*}
$$

where $\hat{\beta}_{\text {max }}$ denote the maximum of $\beta_{i}$ and $\sigma^{2}$ is the usual estimate of $\sigma^{2}$, defined by

$$
\hat{\sigma}^{2}=\frac{(Y-X \hat{\beta})^{\prime}(Y-X \hat{\beta})}{n-p-1}
$$

and referred to henceforth as the HK estimator. They proved that there exists a $k>0$ such that the sum of the MSEs of all $\hat{\beta}_{i}(k)$ is smaller than the corresponding term of $\hat{\beta}_{i}$, the OLS estimator, i.e.

## KHALAF \& IGUERNANE

$$
\operatorname{MSE}(\hat{\beta}(k))<\operatorname{MSE}(\hat{\beta})=\hat{\sigma}^{2} \sum_{i=1}^{p} \lambda_{i}^{-1}
$$

Khalaf and Shukur (2005) suggested a new method of estimating $k$ as a modification of equation (6), as follows

$$
\begin{equation*}
\hat{k}_{K S}=\frac{\lambda_{\max } \hat{\sigma}^{2}}{(n-p) \hat{\sigma}^{2}+\lambda_{\max } \hat{\beta}_{\max }^{2}} \tag{7}
\end{equation*}
$$

where $\lambda_{\text {max }}$ is the largest eigenvalue of the matrix $X^{\prime} X$. They concluded the ridge estimator using (7) performed very well and was substantially better than any estimators included in their study.

In the light of above, which indicates the satisfactory performance of $\hat{k}_{K S}$ with the potential for improvement, modification of the ridge estimator using $\hat{k}_{K S}$ (the KS estimator) by taking its square root is suggested. This proposed estimator (the KSM estimator) is

$$
\begin{equation*}
\hat{k}_{K S M}=\sqrt{\hat{k}_{K S}} \tag{8}
\end{equation*}
$$

To investigate the performance, relative to the OLS and other ridge estimators given by (6) and (7), of the new ridge estimator given by (8), we calculate the MSE using the following equation

$$
\begin{equation*}
M S E=\frac{\sum_{i=1}^{R}(\hat{\beta}-\beta)_{i}^{\prime}(\hat{\beta}-\beta)}{R} \tag{9}
\end{equation*}
$$

where $\hat{\beta}$ is the estimator of $\beta$ obtained from OLS or other ridge estimators, and $R$ equals 5000 which corresponds to the number of replicates used in the simulation.

## Simulations

Consider the true model $Y=X \beta+u$. Here $u \sim N\left(0, \sigma^{2} I\right)$ and the independent variables are generated from

## MULTICOLLINEARITY AND A RIDGE PARAMETER ESTIMATION

$$
\begin{equation*}
x_{i j}=\left(1-\rho^{2}\right)^{\frac{1}{2}} z_{i j}+\rho z_{i p}, i=1,2, \ldots, n, j=1,2, \ldots, p \tag{10}
\end{equation*}
$$

where $z_{i j}$ are generated using the standard normal distribution. Here, we consider four values of $\rho$ corresponding to $0.7,0.9,0.95$ and 0.99 . The dependent variable is then determined by

$$
\begin{equation*}
y_{i}=\beta_{0}+\beta_{1} x_{i 1}+\ldots+\beta_{p} x_{i p}+u_{i}, i=1,2, \ldots, n \tag{11}
\end{equation*}
$$

where $n$ is the number of observations, $u_{i}$ are i.i.d. pseudo-random numbers, and $\beta_{0}$ is taken to be zero. Parameter values are chosen such that $\sum_{j=1}^{p} \beta_{j}^{2}=1$, which is a common restriction in simulation studies (McDonald and Galarneau, 1975; Muniz and Kibria, 2009). Sample sizes selected are $n=10,25,50,85,200$ and 1000 , with 4 or 7 independent variables. The variance of the error terms is taken as $\sigma^{2}=0.01,0.1$, and 0.5 . Ridge estimates are computed using the different ridge parameters given in (6) and (7). Because the proposed estimator (8) is a modification of (7), this estimator is included for purposes of comparison. The MSE of the ridge regression parameters is obtained using (9). This experiment is repeated 5000 times.

## Result

All factors chosen to vary in the design of the experiment affect the estimated MSE. As expected, increasing the degree of correlation leads to a higher estimated MSE, especially when $n$ is small and $\sigma^{2}=0.01$. This increase is much greater for OLS than for ridge regression estimators.

Table 1a. Estimated MSE when $p=4$ and $\rho=0.7$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 16114 | 5236 | 6140 | 31 | 156.00 | 52.00 | 60.00 | 7.00 | 6.320 | 3.030 | 3.220 | 1.850 |
| 25 | 3799 | 1242 | 2153 | 27 | 39.00 | 15.00 | 23.00 | 5.90 | 1.560 | 1.170 | 1.240 | 0.990 |
| 50 | 1722 | 597 | 1248 | 32 | 17.00 | 7.00 | 12.00 | 5.00 | 0.690 | 0.600 | 0.620 | 0.560 |
| 85 | 988 | 344 | 806 | 36 | 9.70 | 4.60 | 8.00 | 4.10 | 0.390 | 0.360 | 0.370 | 0.340 |
| 200 | 399 | 141 | 363 | 42 | 4.00 | 2.40 | 3.60 | 2.60 | 0.161 | 0.156 | 0.157 | 0.153 |
| 1000 | 77 | 28 | 76 | 35 | 0.77 | 0.67 | 0.75 | 0.70 | 0.032 | 0.031 | 0.031 | 0.031 |

## KHALAF \& IGUERNANE

Table 1b. Estimated MSE when $p=4$ and $\rho=0.9$

|  | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 46391 | 14512 | 15254 | 41 | 478.0 | 149.0 | 156.0 | 8.0 | 18.000 | 6.700 | 7.000 | 2.500 |
| 25 | 11854 | 3692 | 4695 | 29 | 114.0 | 37.0 | 46.0 | 5.7 | 4.700 | 2.500 | 2.700 | 1.600 |
| 50 | 5179 | 1678 | 2607 | 27 | 52.0 | 18.0 | 27.0 | 5.3 | 2.120 | 1.480 | 1.560 | 1.170 |
| 85 | 2967 | 969 | 1778 | 25 | 29.0 | 11.0 | 18.0 | 4.9 | 1.190 | 0.950 | 0.990 | 0.820 |
| 200 | 1184 | 380 | 885 | 26 | 12.0 | 5.1 | 9.2 | 4.0 | 0.482 | 0.439 | 0.446 | 0.410 |
| 1000 | 233 | 75 | 216 | 36 | 2.3 | 1.6 | 2.2 | 1.7 | 0.094 | 0.092 | 0.093 | 0.090 |

Table 1c. Estimated MSE when $p=4$ and $\rho=0.95$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | $K S$ | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 99744 | 29610 | 30311 | 51 | 957.00 | 282.00 | 289.00 | 9.00 | 39.000 | 12.000 | 13.000 | 3.000 |
| 25 | 24979 | 7538 | 8527 | 32 | 240.00 | 74.00 | 84.00 | 6.00 | 9.000 | 4.100 | 4.400 | 2.000 |
| 50 | 10642 | 3290 | 4305 | 26 | 108.00 | 36.00 | 46.00 | 5.40 | 4.330 | 2.380 | 2.570 | 1.570 |
| 85 | 6109 | 1945 | 2925 | 23 | 60.00 | 20.00 | 29.00 | 5.00 | 2.480 | 1.650 | 1.760 | 1.250 |
| 200 | 2498 | 802 | 1543 | 22 | 24.00 | 9.00 | 15.00 | 4.60 | 1.010 | 0.830 | 0.858 | 0.724 |
| 1000 | 494 | 163 | 426 | 31 | 4.82 | 2.60 | 4.21 | 2.64 | 0.192 | 0.185 | 0.186 | 0.179 |

Table 1d. Estimated MSE when $p=4$ and $\rho=0.99$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | KS | KSM | OLS | HK | $K S$ | KSM | OLS | HK | KS | KSM |
| 10 | 533881 | 156406 | 157056 | 84 | 5352.0 | 1605.0 | 1612.0 | 12.0 | 218.0 | 67.0 | 67.3 | 5.0 |
| 25 | 130105 | 39322 | 40154 | 46 | 1325.0 | 417.0 | 425.0 | 7.4 | 54.0 | 16.0 | 17.0 | 3.0 |
| 50 | 59142 | 18290 | 19221 | 32 | 593.0 | 189.0 | 199.0 | 6.5 | 23.0 | 8.0 | 8.4 | 2.5 |
| 85 | 33685 | 10461 | 11481 | 25 | 330.0 | 105.0 | 160.0 | 5.7 | 13.0 | 5.1 | 5.4 | 2.1 |
| 200 | 13727 | 4394 | 5464 | 17 | 137.0 | 43.0 | 54.0 | 5.1 | 5.4 | 2.7 | 3.0 | 1.6 |
| 1000 | 2637 | 814 | 1575 | 16 | 26.0 | 9.0 | 16.0 | 4.4 | 1.0 | 0.8 | 0.9 | 0.7 |

Table 2a. Estimated MSE when $p=7$ and $\rho=0.7$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 74818 | 24592 | 25042 | 110 | 768.00 | 238.00 | 242.00 | 19.00 | 29.00000 | 10.00000 | 11.00000 | 4.20000 |
| 25 | 8804 | 3457 | 4423 | 46 | 89.00 | 37.00 | 46.00 | 10.00 | 3.54000 | 2.76000 | 2.81000 | 2.13000 |
| 50 | 3618 | 1508 | 2367 | 48 | 36.00 | 17.00 | 24.00 | 8.70 | 1.44000 | 1.31000 | 1.32000 | 1.17000 |
| 85 | 1998 | 848 | 1506 | 52 | 19.00 | 10.00 | 15.00 | 7.40 | 0.78300 | 0.74400 | 0.74800 | 0.69900 |
| 200 | 795 | 337 | 691 | 63 | 7.90 | 5.50 | 7.00 | 4.80 | 0.31700 | 0.31100 | 0.31200 | 0.30300 |
| 1000 | 152 | 67 | 148 | 60 | 1.52 | 1.39 | 1.48 | 1.35 | 0.06110 | 0.06094 | 0.06096 | 0.06060 |

## MULTICOLLINEARITY AND A RIDGE PARAMETER ESTIMATION

Table 2b. Estimated MSE when $p=7$ and $\rho=0.9$

|  | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 235966.0 | 68291.0 | 68644.0 | 136.0 | 2224.0 | 658.0 | 661.0 | 27.0 | 91.0000 | 28.1000 | 28.2000 | 6.4000 |
| 25 | 26871.0 | 10240.0 | 11090.0 | 49.0 | 273.0 | 105.0 | 113.0 | 12.0 | 10.0000 | 6.2000 | 6.3000 | 3.5000 |
| 50 | 10990.0 | 4275.0 | 5224.0 | 39.0 | 110.0 | 45.0 | 54.0 | 10.0 | 4.3800 | 3.2900 | 3.3400 | 2.3900 |
| 85 | 6112.0 | 2430.0 | 3321.0 | 38.1 | 59.0 | 25.0 | 33.0 | 8.8 | 2.4200 | 2.0500 | 2.0700 | 1.6700 |
| 200 | 2430.0 | 966.0 | 1624.0 | 40.0 | 23.0 | 11.0 | 16.0 | 7.0 | 0.9790 | 0.9120 | 0.9170 | 0.8300 |
| 1000 | 466.0 | 185.0 | 410.0 | 57.0 | 4.6 | 3.5 | 4.2 | 3.1 | 0.1878 | 0.1852 | 0.1854 | 0.1816 |

Table 2c. Estimated MSE when $p=7$ and $\rho=0.95$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 516796 | 152429 | 152764 | 171 | 4818.0 | 1430.0 | 1434.0 | 35.0 | 192.000 | 62.400 | 62.600 | 9.300 |
| 25 | 57214 | 21072 | 21887 | 55 | 582.0 | 219.0 | 227.0 | 15.0 | 23.000 | 10.000 | 11.000 | 4.500 |
| 50 | 22961 | 8791 | 9736 | 41 | 231.0 | 91.0 | 100.0 | 12.0 | 9.200 | 5.600 | 5.800 | 3.300 |
| 85 | 12508 | 4916 | 5857 | 35 | 126.0 | 50.0 | 59.0 | 10.0 | 5.000 | 3.600 | 3.700 | 2.500 |
| 200 | 5037 | 1977 | 2795 | 34 | 50.0 | 21.0 | 29.0 | 8.4 | 2.010 | 1.730 | 1.740 | 1.430 |
| 1000 | 985 | 396 | 771 | 49 | 9.8 | 6.1 | 8.0 | 4.7 | 0.389 | 0.377 | 0.378 | 0.361 |

Table 2d. Estimated MSE when $p=7$ and $\rho=0.99$

| $n$ | $\sigma^{2}=0.01$ |  |  |  | $\sigma^{2}=0.1$ |  |  |  | $\sigma^{2}=0.5$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | OLS | HK | KS | KSM | OLS | HK | KS | KSM | OLS | HK | KS | KSM |
| 10 | 2501132 | 764126 | 764446 | 235 | 25773 | 7976 | 7979 | 62 | 1019.0 | 289.3 | 289.4 | 18.0 |
| 25 | 314693 | 115277 | 116046 | 72 | 3077 | 1107 | 1115 | 21 | 126.0 | 48.4 | 48.7 | 8.7 |
| 50 | 128529 | 48265 | 49173 | 48 | 1259 | 475 | 484 | 17 | 50.0 | 20.4 | 20.7 | 6.0 |
| 85 | 67913 | 25511 | 26492 | 38 | 691 | 262 | 272 | 15 | 28.0 | 12.8 | 13.0 | 5.0 |
| 200 | 27914 | 10645 | 11673 | 31 | 271 | 102 | 112 | 11 | 11.0 | 6.3 | 6.5 | 3.6 |
| 1000 | 5479 | 2117 | 2922 | 32 | 53 | 22 | 29 | 8 | 2.1 | 1.7 | 1.8 | 1.4 |

## Conclusion

Based on the result from the simulation study, some recommendations are warranted. The KSM is usually among the estimators with the lowest estimated MSE, especially when $\rho=0.95$ and $p=7$. Also, regardless of the degree of correlations, KSM is the best among the considered ridge estimators, followed by HK, and then KS, specifically when the sample size is high, $n=1000$, and $\sigma^{2}=0.5$.

## KHALAF \& IGUERNANE

Several procedures for constructing ridge estimators have been proposed in the literature. These procedures aim at establishing a rule for selecting the constant $k$ in equation (5). Nevertheless, to date there is no rule for choosing $k$ that assures that the corresponding ridge estimator is better than OLS estimator.

The proposed choice of $k$, the ridge regression parameter defined by (8), was shown through simulation to yield a lower MSE than $\hat{\beta}$ for all $\beta$, as noted in Tables 1 and 2. The estimators HK and KS, which were evaluated in other simulation studies, also performed well. However, the superiority of the suggested estimator KSM over the estimators HK and KS was observed, especially at the large values of $n$ and $\sigma^{2}$. In general, the OLS estimator has larger estimated MSE values than all estimators considered, and the proposed estimator given by (8) performs very well and has the lowest MSE when compared with the other ridge estimators. This is to say that ridge estimators are more helpful when high multicollinearity exists, especially when $\sigma^{2}$ is not too small.

## References

Allison, P. D. (1999). Logistic regression using the SAS System: Theory and application. Cary, NC: SAS Institute.

Belsley, D. A., Edwin, K. and Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley and Sons.

Freund, R. J. and Littell, R. C. (2000). SAS System for regression (3rd ed.). Cary, NC: SAS Institute.

Green, W. H. (2000). Econometric Analysis (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hag, M. S. and Kibria, B. M. G. (1996). A shrinkage estimator for the restricted linear regression ridge regression approach. Journal of Applied Statistical Science, 3, 301-316.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12(1), 55-67. doi: 10.1080/00401706.1970.10488634

Judge, G. G., Griffiths, W. E., Hill, R. C. and Lee, T. C. (1985). The theory and practice of econometrics (2nd ed.) New York: John Wiley and Sons.

Klein, L. (1962). An introduction to econometrics. New York: Prentice Hall.

## MULTICOLLINEARITY AND A RIDGE PARAMETER ESTIMATION

Khalaf, G. (2011). Ridge regression: An evaluation to some new modifications. International Journal of Statistics and Analysis, 1(4), pp. 325-342.

Khalaf, G. (2013). A comparison between biased and unbiased estimators. Journal of Modern Applied Statistical Methods, 12(2), 293-303.

Khalaf, G. and Iguernane, M. (2014). Ridge regression and ill-conditioning. Journal of Modern Applied Statistical Methods, 13(2), 355-363.

Khalaf, G. and Shukur, G. (2005). Choosing ridge parameter for regression problems. Communication in Statistics - Theory and Methods, 34(5), 1177-1182. doi: 10.1081/STA-200056836

McDonald, G. C., and Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association, 70(350), 407-416. doi: 10.1080/01621459.1975.10479882

Nomura, M. (1988). On the almost unbiased ridge regression estimation. Communication in Statistics - Simulation and Computation, 17(3), 729-743. doi: 10.1080/03610918808812690

Muniz, G. and Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. Communication in Statistics - Simulation and Computation, 38(3), 621-630. doi: 10.1080/03610910802592838

# An Adjusted Network Information Criterion for Model Selection in Statistical Neural Network Models 

Christopher Godwin<br>Udomboso<br>University of Ibadan<br>Ibadan, Nigeria

Godwin Nwazu<br>Amahia<br>University of Ibadan Ibadan, Nigeria

Isaac K. Dontwi<br>Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

In this paper, we derived and investigated the Adjusted Network Information Criterion (ANIC) criterion, based on Kullback's symmetric divergence, which has been designed to be an asymptotically unbiased estimator of the expected Kullback-Leibler information of a fitted model. The ANIC improves model selection in more sample sizes than does the NIC.

Keywords: Statistical neural network, network information criterion, adjusted network information criterion, transfer function

## Introduction

In choosing an appropriate model to characterize the sample data, it is ideal to be guided by scientific theory, as well as be well served by a data-driven selection method. Akaike (1973, 1974) introduced the Akaike information criterion, AIC, which endeavors discern the closeness of a fitted model is to the generating or true model. Akaike's work stimulated many other approaches to model selection, leading to the development of criteria such as SIC (Schwarz, 1978), BIC (Akaike, 1978), and HQ (Hannan, \& Quinn 1979). Sugiura (1978) extended Akaike's original work by proposing AICc, a corrected version of AIC justified in the context of linear regression with normal errors.

The development of AICc was motivated by the need to adjust for AIC's propensity to favor high-dimensional models when the sample size is small relative to the maximum order of the models in the candidate class. Hurvich and

[^24]
## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

Tsai (1989) show that AICc dramatically outperforms AIC in small-sample regression settings, and further extend AICc to include univariate Gaussian autoregressive models. Hurvich, Shumway, and Tsai (1990) generalize AICc to encompass univariate Gaussian autoregressive moving-average models, and Hurvich and Tsai (1993) handle the vector Gaussian autoregressive case.

The purpose of this study is to consider the selection of Statistical Neural Network model using the proposed method by Murata, Yoshizawa, and Amari (1994), which is the NIC. The NIC is observed to be sample biased, as it does not account for sample sizes. The selection of a model from a set of fitted candidate models requires objective data-driven criteria. The criterion we shall use in this study is that designed to be an asymptotically unbiased estimator of the expected Kullback-Leibler information of a fitted model (Akaike, 1973).

## Methodology

## Adjusted Network Information Criterion (ANIC):

We note that

$$
\begin{array}{ll}
\mathbf{Y}^{*}=\mathbf{H W}+\boldsymbol{U} & \text { (true model) } \\
\mathbf{Y}^{*}=\mathbf{H W}+\boldsymbol{e} & \text { (estimated model) } \tag{2}
\end{array}
$$

Anders (1996) noted that should the network exactly map the true function $F$, then the asymptotic relationship, $G=2 B \sigma^{2}$, so that $\operatorname{tr}\left(G B^{-1}\right)=2 \sigma^{2} \operatorname{tr}(I)=2 \sigma^{2} k$. Thus, NIC becomes AIC as proposed by Amemiya (1980):

$$
\begin{equation*}
\mathrm{AIC}=M S E+2 \sigma^{2} \frac{k}{n} \tag{3}
\end{equation*}
$$

Therefore, in deriving an alternative NIC, we assume that the estimates network model includes the true network model, and the approach shall use the corrected AIC based on Kullback's systematic divergence as used by Hafidi and Mkhadri (2006).

We recall that

$$
\begin{equation*}
\mathrm{NIC} \equiv D[q, p(\mathbf{W})] \tag{4}
\end{equation*}
$$

## UDOMBOSO ET AL.

$$
\begin{equation*}
\cong D\left[q, p\left(\mathbf{W}_{o p t}\right)\right]+\frac{1}{2}\left(\mathbf{W}-\mathbf{W}_{o p t}\right)^{\prime}\left(\mathbf{W}-\mathbf{W}_{o p t}\right) \frac{\partial^{2}}{\partial \mathbf{W}_{o p t}^{2}} D\left[q, p\left(\mathbf{W}_{o p t}\right)\right] . \tag{5}
\end{equation*}
$$

Kullback (1968) defined the discrepancy between the true model and the estimated model as

$$
\begin{equation*}
J\left(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}\right)=\left[D\left(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}\right)-D\left(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}_{0}\right)\right]-\left[D\left(\theta, \theta_{0}\right)-D(\theta, \theta)\right] \tag{6}
\end{equation*}
$$

where $\boldsymbol{\theta}_{\boldsymbol{0}}$ is the true and unknown parameter vector, $\boldsymbol{\theta}$ is the parameter vector of the candidate model. Also, $f\left(\mathbf{Y} \mid \boldsymbol{\theta}_{\mathbf{0}}\right)$ and $f(\mathbf{Y} \mid \boldsymbol{\theta})$ denote the densities for the true and estimates models.

Note that the second term does not depend on $\boldsymbol{\theta}$. Thus, Cavanaugh (1997, 1999), in order to discriminate among various models, proposed another form of Kullback's symmetric divergence as

$$
\begin{equation*}
K\left(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}\right)=\left[D\left(\boldsymbol{\theta}_{0}, \boldsymbol{\theta}\right)\right]+\left[D\left(\boldsymbol{\theta}, \boldsymbol{\theta}_{0}\right)-D(\boldsymbol{\theta}, \boldsymbol{\theta})\right] \tag{7}
\end{equation*}
$$

Given that the estimated model includes the true model, we can define the improved NIC as

$$
\begin{equation*}
\mathrm{ANIC}=D(\mathbf{W}, \mathbf{W})+T \tag{8}
\end{equation*}
$$

which is asymptotically an unbiased estimator of

$$
\begin{equation*}
\Omega(d, \mathbf{W})=E_{\mathbf{W}}[N(\mathbf{W}, \mathbf{W})] \tag{9}
\end{equation*}
$$

where $T$ is some value that improves the NIC, $d$ is the dimension of $\mathbf{W}$, and is given as

$$
\begin{equation*}
d=p+1 \tag{10}
\end{equation*}
$$

and $N(\mathbf{W}, \mathbf{W})$ is the NIC.

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

## Proof:

$$
\begin{equation*}
\Omega(d, \mathbf{W})=E_{\mathbf{w}}\{D(\mathbf{W}, \mathbf{w})+[D(\mathbf{W}, \mathbf{w})-D(\mathbf{W}, \mathbf{w})]\} \tag{11}
\end{equation*}
$$

But the true model is given as

$$
\begin{equation*}
\mathbf{Y}^{*}=\mathbf{H W}+\mathbf{U} \quad \mathbf{U} \sim N\left(0, \sigma^{2} I_{n}\right), \tag{12}
\end{equation*}
$$

and the estimated model is

$$
\begin{equation*}
\mathbf{Y}^{*}=\mathbf{H W}+\boldsymbol{e} \tag{13}
\end{equation*}
$$

where $\mathbf{Y}^{*}$ is an $n \times 1$ observation, $\mathbf{H}$ is an $n \times p$ observations, $\mathbf{W}=\mathbf{W}^{*}$ is an $p \times 1$ observation. Assume that $\mathbf{H}$ is twice continuously differentiable in $\mathbf{W}$. Let $t(\lambda)=\mathbf{H W}$. Then, the log-likelihood of the estimated model is given as

$$
\begin{equation*}
\ln f\left(\mathbf{Y}^{*} \mid \mathbf{W}\right)=\frac{n}{2} \ln 2 \pi \sigma^{2}-\frac{1}{2 \sigma^{2}}\left(\mathbf{Y}^{*}-t(\lambda)\right)^{\prime}\left(\mathbf{Y}^{*}-t(\lambda)\right) \tag{14}
\end{equation*}
$$

Approach the second term of (1) by considering two hypothetical estimators $w_{1}$ and $w_{2}$, such that

$$
\begin{align*}
& D\left(w_{1}, w_{2}\right)=E_{w_{1}}\left[\ln f\left(Y^{*} \mid w_{2}\right)\right]  \tag{15}\\
&=E_{w_{1}}\left[-\frac{n}{2} \ln 2 \pi \sigma_{2}^{2}-\frac{1}{2 \sigma_{2}^{2}}\left(Y^{*}-t\left(\lambda_{2}\right)\right)^{\prime}\left(Y^{*}-t\left(\lambda_{2}\right)\right)\right]  \tag{16}\\
&=E_{w_{1}}\left[\begin{array}{l}
-\frac{n}{2} \ln 2 \pi \sigma_{2}^{2}-\frac{1}{2 \sigma_{2}^{2}}\left(Y^{*}-t\left(\lambda_{1}\right)\right)^{\prime}\left(Y^{*}-t\left(\lambda_{1}\right)\right) \\
+\left(t\left(\lambda_{1}\right)-t\left(\lambda_{2}\right)\right)^{\prime}\left(t\left(\lambda_{1}\right)-t\left(\lambda_{2}\right)\right)
\end{array}\right]  \tag{17}\\
&=-\frac{n}{2} \ln 2 \pi \sigma_{2}^{2}-\frac{1}{2 \sigma_{2}^{2}}\left[n \sigma_{1}^{2}+\left(t\left(\lambda_{1}\right)-t\left(\lambda_{2}\right)\right)^{\prime}\left(t\left(\lambda_{1}\right)-t\left(\lambda_{2}\right)\right)\right] . \tag{18}
\end{align*}
$$

## UDOMBOSO ET AL.

Expand $D(\mathbf{W}, \mathbf{W})$ as

$$
\begin{equation*}
D(\mathbf{W}, \mathbf{W})=-\frac{n}{2} \ln 2 \pi \sigma_{\mathbf{W}}^{2}-\frac{1}{2 \sigma_{\mathbf{W}}^{2}}\left[n \sigma_{\mathbf{w}}^{2}+(t(\hat{\lambda})-t(\lambda))^{\prime}(t(\hat{\lambda})-t(\lambda))\right] \tag{19}
\end{equation*}
$$

Expanding $t(\hat{\lambda})$ in order one at $\hat{\lambda}=\lambda$,

$$
\begin{equation*}
t(\hat{\lambda}) \cong t(\lambda)+\frac{\partial t}{\partial \hat{\lambda}}(\hat{\lambda}-\lambda) \tag{20}
\end{equation*}
$$

This results in

$$
\begin{align*}
& D(\mathbf{W}, \mathbf{W}) \cong-\frac{1}{2}\left\{\begin{array}{l}
n \ln 2 \pi \sigma_{\mathbf{W}}^{2}+ \\
\frac{1}{\sigma_{\mathbf{W}}^{2}}\left[\begin{array}{l}
2 \sigma_{\mathbf{W}}^{2}+\left[t(\lambda)+\frac{\partial t}{\partial \hat{\lambda}}(\hat{\lambda}-\lambda)-t(\lambda)\right]^{\prime} \\
{\left[t(\lambda)+\frac{\partial t}{\partial \hat{\lambda}}(\hat{\lambda}-\lambda)-t(\lambda)\right]}
\end{array}\right]
\end{array}\right\}  \tag{21}\\
& =-\frac{1}{2}\left\{n \ln 2 \pi \sigma_{\mathrm{w}}^{2}+\frac{1}{\sigma_{\mathrm{w}}^{2}}\left[2 \sigma_{\mathrm{w}}^{2}+(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right]\right\} \tag{22}
\end{align*}
$$

Similarly,

$$
\begin{align*}
D(\mathbf{W}, \mathbf{W}) & =-\frac{n}{2} \ln 2 \pi \sigma_{\mathbf{W}}^{2}-\frac{1}{2 \sigma_{\mathbf{W}}^{2}}\left[n \sigma_{\mathbf{W}}^{2}+(t(\hat{\lambda})-t(\hat{\lambda}))^{\prime}(t(\hat{\lambda})-t(\hat{\lambda}))\right]  \tag{23}\\
& =-\frac{1}{2}\left\{n \ln 2 \pi \sigma_{\mathbf{W}}^{2}+n\right\} \tag{24}
\end{align*}
$$

Thus, the second term of (11) becomes

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

$$
\begin{align*}
& D(\mathbf{W}, \mathbf{W})-D(\mathbf{W}, \mathbf{W}) \cong-\frac{1}{2}\left\{\begin{array}{l}
n \ln 2 \pi \sigma_{\mathbf{W}}^{2}+ \\
\frac{1}{2 \sigma_{\mathbf{W}}^{2}}\left[\begin{array}{l}
\left.n \sigma_{\mathbf{W}}^{2}+(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\right]-n \ln 2 \pi \sigma_{\mathbf{W}}^{2} \\
{\left[\frac{\partial t}{\partial \hat{\lambda}}\right]}
\end{array}\right\}, ~
\end{array}\right.  \tag{25}\\
& =-\frac{1}{2}\left\{\begin{array}{l}
n \ln 2 \pi \sigma_{\mathrm{W}}^{2}+n \frac{\sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}^{2}}+ \\
\frac{1}{\sigma_{\mathrm{W}}^{2}}\left[(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right]-n \ln 2 \pi \sigma_{\mathrm{W}}^{2}-n
\end{array}\right\}  \tag{26}\\
& =-\frac{1}{2}\left\{\begin{array}{l}
n\left[\ln 2 \pi \sigma_{\mathrm{W}}^{2}-n \ln 2 \pi \sigma_{\mathrm{W}}^{2}+\frac{\sigma_{\mathrm{w}}^{2}}{\sigma_{\mathrm{W}}^{2}}\right] \\
+\frac{1}{\sigma_{\mathrm{W}}^{2}}\left[(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right]-n
\end{array}\right\}  \tag{27}\\
& =-\frac{1}{2}\left\{\begin{array}{l}
n\left[\ln \frac{2 \pi \sigma_{\mathrm{W}}^{2}}{2 \pi \sigma_{\mathrm{W}}^{2}}+\frac{\sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}^{2}}\right]+ \\
\frac{1}{\sigma_{\mathrm{W}}^{2}}\left[(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right]-n
\end{array}\right\}  \tag{28}\\
& =-\frac{1}{2}\left\{\begin{array}{l}
n\left[\ln \frac{\sigma_{\mathbf{W}}^{2}}{\sigma_{\mathbf{W}}}+\frac{\sigma_{\mathbf{w}}^{2}}{\sigma_{\mathbf{W}}^{2}}\right]+ \\
\frac{1}{\sigma_{\mathbf{W}}^{2}}\left[(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right]-n
\end{array}\right\} \tag{29}
\end{align*}
$$

The distribution of

## UDOMBOSO ET AL.

$$
\frac{2 \sigma_{\mathrm{w}}^{2}}{\sigma_{\mathrm{w}}^{2}} \sim \chi_{n-p}
$$

and

$$
\frac{1}{\sigma_{\mathrm{w}}^{2}}\left[(\hat{\lambda}-\lambda)\left[\frac{\partial t}{\partial \hat{\lambda}}\right]^{\prime}\left[\frac{\partial t}{\partial \hat{\lambda}}\right](\hat{\lambda}-\lambda)\right] \sim \chi_{p}
$$

Therefore,

$$
\begin{align*}
D(\mathbf{W}, \mathbf{W})-D(\mathbf{W}, \mathbf{W}) & =-\frac{1}{2}\left\{n \ln \frac{\sigma_{\mathbf{W}}^{2}}{\sigma_{\mathbf{W}}}+(n-p)+p-n\right\}  \tag{30}\\
& =-\frac{1}{2}\left\{n \ln \frac{\sigma_{\mathbf{W}}^{2}}{\sigma_{\mathbf{W}}^{2}}\right\} \tag{31}
\end{align*}
$$

Taking expectation, the above becomes

$$
\begin{equation*}
E[D(\mathbf{W}, \mathbf{W})-D(\mathbf{W}, \mathbf{W})] \cong-\frac{1}{2} E\left\{n \ln \frac{\sigma_{\mathbf{W}}^{2}}{\sigma_{\mathbf{W}}}\right\} \tag{32}
\end{equation*}
$$

Bickel and Doksum (1977) noted that by taking a second order expansion of $\ln \chi_{d f}$ about $d f$ and evaluating the expectation of the result, the following relation ensues,

$$
\begin{equation*}
E\left[\ln \chi_{d f}\right]=\ln d f-\frac{1}{d f}+o\left[\frac{1}{(d f)^{2}}\right] \tag{33}
\end{equation*}
$$

where $d f$ is degrees of freedom. Write

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

$$
\begin{equation*}
E\left[n \ln \frac{\sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}^{2}}\right]=n E\left[\ln \frac{n \sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}^{2}}\right]-n \ln n \tag{34}
\end{equation*}
$$

By Bickel and Doksum (1977) relation, and according to Cavanaugh (1997, 1999),

$$
\begin{equation*}
E\left[n \ln \frac{\sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}}\right]=n\left\{\ln (n-p)-\frac{1}{n-p}+o\left[\frac{1}{(n-p)^{2}}\right]\right\}-n \ln n \tag{35}
\end{equation*}
$$

The first-order expansion of $\ln (n-p)$ is

$$
\begin{equation*}
\ln (n-p)=\ln n-\frac{p}{n}+o\left(\frac{p}{n}\right)^{2} \tag{36}
\end{equation*}
$$

Thus,

$$
\begin{align*}
E\left[n \ln \frac{\sigma_{\mathrm{W}}^{2}}{\sigma_{\mathrm{W}}}\right]=n\left\{\operatorname{lnn}-\frac{p}{n}\right. & \left.+o\left(\frac{p}{n}\right)^{2}-\frac{1}{n-p}+o\left[\frac{1}{(n-p)^{2}}\right]\right\}-n \ln n  \tag{37}\\
& \cong-\left\{p+\frac{n}{n-p}\right\}  \tag{38}\\
& =-\left\{\frac{n p-p^{2}+n}{n-p}\right\} \tag{39}
\end{align*}
$$

Putting this result back in (32),

$$
\begin{gather*}
E[D(\mathbf{W}, \mathbf{W})-D(\mathbf{W}, \mathbf{W})] \cong-\frac{1}{2}\left\{-\left[\frac{n p-p^{2}+n}{n-p}\right]\right\}  \tag{40}\\
=\frac{n p-p^{2}+n}{2(n-p)} \tag{41}
\end{gather*}
$$

Thus, the alternative NIC becomes

## UDOMBOSO ET AL.

$$
\begin{equation*}
\mathrm{ANIC}=\mathrm{NIC}+\frac{n p-p^{2}+n}{2(n-p)} \tag{42}
\end{equation*}
$$

which is a correction for the biased NIC.

## Results

## Illustrative Examples:

The following illustrations demonstrate the power of the adjusted network information criterion in accounting for sample size. Anders (1996) proposed a statistical neural network model given as

$$
\begin{equation*}
\boldsymbol{y}=f(\mathbf{X}, \mathbf{w})+\boldsymbol{u} \tag{43}
\end{equation*}
$$

where $\boldsymbol{y}$ is the dependent variable, $\mathbf{X}=\left(\boldsymbol{x}_{\mathbf{0}} \equiv \mathbf{1}, \boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{\mathbf{I}}\right)$ is a vector of independent variables, $\mathbf{w}=(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the network weight: ' $\boldsymbol{\alpha}$ ' is the weight of the input unit, ' $\boldsymbol{\beta}$ ' is the weight of the hidden unit, and ' $\gamma$ ' is the weight of the output unit, and $\boldsymbol{u}_{i}$ is the stochastic term that is normally distributed (that is, $\left.\boldsymbol{u}_{i} \sim N\left(0, \sigma^{2} I_{n}\right)\right)$.
$f(\mathbf{X}, \mathbf{w})$ is the artificial neural network function, expressed as

$$
\begin{equation*}
f(\mathbf{X}, \mathbf{w})=\alpha \mathbf{X}+\sum_{h=1}^{H} \boldsymbol{\beta}_{\boldsymbol{h}} g\left(\sum_{i=0}^{I} \gamma_{h i} \boldsymbol{x}_{i}\right), \tag{44}
\end{equation*}
$$

where $g($.$) is the transfer function.$
The proposed convoluted form of the artificial neural network function used in this study is

$$
\begin{equation*}
f(\mathbf{X}, \mathbf{w})=\boldsymbol{\alpha} \mathbf{X}+\sum_{h=1}^{H} \boldsymbol{\beta}_{\boldsymbol{h}}\left[g_{1}\left(\sum_{i=0}^{I} \gamma_{h i} \boldsymbol{x}_{i}\right) g_{2}\left(\sum_{i=0}^{I} \gamma_{h i} x_{i}\right)\right], \tag{45}
\end{equation*}
$$

and thus, the form of the statistical neural network model proposed is

$$
\begin{equation*}
\boldsymbol{y}=\boldsymbol{\alpha} \mathbf{X}+\sum_{h=1}^{H} \boldsymbol{\beta}_{\boldsymbol{h}}\left[g_{1}\left(\sum_{i=0}^{I} \gamma_{h i} \boldsymbol{x}_{\boldsymbol{i}}\right) g_{2}\left(\sum_{i=0}^{I} \gamma_{h i} x_{i}\right)\right]+\boldsymbol{u}_{\boldsymbol{i}} \boldsymbol{u}_{\boldsymbol{j}}, \tag{46}
\end{equation*}
$$

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

where $\boldsymbol{y}$ is the dependent variable, $\mathbf{X}=\left(\boldsymbol{x}_{\mathbf{0}} \equiv \mathbf{1}, \boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{\mathbf{I}}\right)$ is a vector of independent variables, $\mathbf{w}=(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the network weight: ' $\boldsymbol{\alpha}$ ' is the weight of the input unit, ' $\boldsymbol{\beta}$ ' is the weight of the hidden unit, and ' $\gamma$ ' is the weight of the output unit, $\boldsymbol{u}_{i}$ and $\boldsymbol{u}_{j}$ are the stochastic terms that are normally distributed (that is, $\boldsymbol{u}_{i}, \boldsymbol{u}_{j} \sim N\left(0, \sigma^{2} I_{n}\right)$ ), and $g_{1}($.$) and g_{2}($.$) are the transfer functions.$

The choice of the transfer functions used was based on preliminary investigations of the fifteen (15) transfer functions which uses hidden neurons that included $2,5,10,50$, and 100 at 1000 iterations. Best performances came from Hyperbolic Tangent transfer function (TANH), Hyperbolic Tangent Sigmoid transfer function (TANSIG), and Symmetric Saturating Linear transfer function (SATLINS), respectively. Similarly, further investigation was conducted on the choice of convolution, and it was found out that best performance was obtained in the convolution of the Symmetric Saturating Linear transfer function and the Hyperbolic Tangent transfer function (SATLINS_TANH), followed by the convolution of the Symmetric Saturating Linear transfer function and the Hyperbolic Tangent Sigmoid transfer function (SATLINS_TANSIG). The data used for the analyses used in this research were split into two -2 and 3 . The hidden neurons used include $2,5,10,20,40,60,80$, and 100 , while the sample sizes include $10,20,40,60,80,100,125,150,175,200,250,300$, and 400.

Based on two (2) variables, it is shown in Table 1 that the values of NIC across samples, while Table 2 shows the values of ANIC across the samples. It is shown in Table 3 that the sample points at which the values of NIC and ANIC are low in each heterogeneous models in comparison to the root (homogeneous) models.

Table 1. Model Selections across Samples based on NIC (2 Variables)

|  | $n=$ | 10 | 20 | 40 | 60 | 80 | 100 | 125 | 150 | 175 | 200 | 250 | 300 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| SATLINS | 0.0038 | 0.0026 | 0.0239 | 0.0021 | 0.0002 | 0.0007 | 0.0013 | 0.0011 | 0.0044 | 0.0039 | 0.0012 | 0.0031 | 0.0068 |
| TANH | 0.0054 | 0.0217 | 0.0016 | 0.0006 | 0.0113 | 0.0003 | 0.0005 | 0.0021 | 0.0023 | 0.0021 | 0.0017 | 0.0029 | 0.0045 |
| TANSIG | 0.0031 | 0.0120 | 0.0017 | 0.0047 | 0.0023 | 0.0003 | 0.0113 | 0.0011 | 0.0038 | 0.0024 | 0.0017 | 0.0052 | 0.0044 |
| SATLINS_TANH | 0.0066 | 0.0227 | 0.0028 | 0.0008 | 0.0110 | 0.0001 | 0.0007 | 0.0004 | 0.0011 | 0.0024 | 0.0024 | 0.0023 | 0.0037 |
| SATLINS_TANSIG | 0.0049 | 0.0125 | 0.0056 | 0.0010 | 0.0013 | 0.0003 | 0.0018 | 0.0019 | 0.0050 | 0.0039 | 0.0007 | 0.0041 | 0.0043 |

## UDOMBOSO ET AL.

Table 2. Model Selections across Samples based on ANIC (2 Variables)

|  | ANIC |  |  |  |  |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| $n=$ | 10 | 20 | 40 | 60 | 80 | 100 | 125 | 150 | 175 | 200 | 250 | 300 | 400 |
| SATLINS | 1.6217 | 1.5581 | 1.5500 | 1.5154 | 1.5130 | 1.5107 | 1.5091 | 1.5069 | 1.5048 | 1.5046 | 1.5043 | 1.5051 | 1.5080 |
| TANH | 1.6073 | 1.5261 | 1.5224 | 1.5154 | 1.5015 | 1.5095 | 1.5083 | 1.5078 | 1.5064 | 1.5046 | 1.5045 | 1.5178 | 1.5248 |
| TANSIG | 1.5627 | 1.5185 | 1.5199 | 1.5093 | 1.5099 | 1.5193 | 1.5164 | 1.5080 | 1.5063 | 1.5050 | 1.5076 | 1.5056 | 1.6102 |
| SATLINS_TANH | 1.6025 | 1.5245 | 1.5215 | 1.5149 | 1.5012 | 1.5091 | 1.5080 | 1.5059 | 1.5071 | 1.5039 | 1.5201 | 1.5252 | 1.5119 |
| SATLINS_TANSIG | 1.5257 | 1.5260 | 1.5151 | 1.5120 | 1.5089 | 1.5074 | 1.5062 | 1.5039 | 1.5066 | 1.5056 | 1.5047 | 1.5961 | 1.5706 |

Table 3. Sample points at which NIC and ANIC are low in each heterogeneous model in comparison to the root models (2 Variables)

| Model | Sample Size $n$ |  |
| ---: | ---: | ---: |
| NIC | ANIC |  |
| SATLINS_TANH | $100,150,175,400$ | $10,20,40,60,80,100,125,150,200$ |
| SATLINS_TANSIG | $100,250,400$ | $10,40,80,100,125,150$ |



Figure 1. Graph of NIC based on Sample Sizes (2 Variables)

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION



Figure 2. Graph of ANIC based on Sample Sizes (2 Variables)

Correspondingly based on two (2) variables, Figure 1 is the graph of NIC across samples, while Figure 2 is the graph of ANIC across samples. The models in ANIC are almost parallel between sample number 10 and 150 inclusive.

Similarly, based on three (3) variables, Table 4 shows the values of NIC across samples, while Table 5 shows the values of ANIC across the samples. Table 6 shows the sample points at which the values of NIC and ANIC are low in each heterogeneous models in comparison to the root (homogeneous) models.

Table 4. Model Selections across Samples based on NIC (3 Variables)

|  | $n=$ | 10 | 20 | 40 | 60 | 80 | 100 | 125 | 150 | 175 | 200 | 250 | 300 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | 400 |  |  |  |  |  |  |  |  |  |  |  |  |
| SATLINS | 0.4682 | 0.0306 | 0.0196 | 0.0363 | 0.0210 | 0.0561 | 0.0090 | 0.0166 | 0.0154 | 0.0139 | 0.0203 | 0.0230 | 0.0436 |
| TANH | 0.3184 | 1.0532 | 0.0301 | 0.0350 | 0.0197 | 0.0158 | 0.0141 | 0.0228 | 0.0154 | 0.0213 | 0.0195 | 0.0225 | 0.0736 |
| TANSIG | 0.3115 | 0.1102 | 0.0216 | 0.0537 | 0.0160 | 0.0189 | 0.0149 | 0.0213 | 0.0173 | 0.0254 | 0.0165 | 0.0206 | 0.0489 |
| SATLINS_TANH | 0.3540 | 0.0274 | 0.0245 | 0.0159 | 0.0193 | 0.0137 | 0.0159 | 0.0471 | 0.0159 | 0.0192 | 0.0112 | 0.0179 | 0.0462 |

## UDOMBOSO ET AL.

Table 5. Model Selections across Samples based on ANIC (3 Variables)

|  | ANIC |  |  |  |  |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| n= | 10 | 20 | 40 | 60 | 80 | 100 | 125 | 150 | 175 | 200 | 250 | 300 | 400 |
| SATLINS | 2.1172 | 2.1083 | 2.0572 | 2.0269 | 2.0405 | 2.0684 | 2.0209 | 2.0230 | 2.0215 | 2.0186 | 2.0227 | 2.0229 | 2.0349 |
| TANH | 2.4044 | 3.1075 | 2.0372 | 2.0144 | 2.0388 | 2.0276 | 2.0142 | 2.0199 | 2.0177 | 2.0238 | 2.0203 | 2.0109 | 2.0039 |
| TANSIG | 2.0076 | 2.1847 | 2.0383 | 2.0748 | 2.0338 | 2.0344 | 2.0248 | 2.0159 | 2.0216 | 2.0234 | 2.0156 | 2.0145 | 2.0223 |
| SATLINS_TANH | 2.2510 | 2.0752 | 2.0464 | 2.0383 | 2.0349 | 2.0261 | 2.0243 | 1.9935 | 2.0207 | 2.0258 | 2.0116 | 2.0170 | 1.9995 |
| SATLINS_TANSIG | 2.1847 | 2.1356 | 2.0093 | 2.0413 | 2.0368 | 2.0312 | 2.0248 | 2.0223 | 2.0168 | 2.0140 | 2.0086 | 2.0192 | 1.8820 |

Table 6. Sample points at which NIC and ANIC are low in each heterogeneous model in comparison to the root models (3 Variables)

| Model | Sample Size $n$ |  |
| ---: | ---: | ---: |
| NIC | ANIC |  |
| SATLINS_TANH | $20,60,80,100,250,300$ | $20,100,150,250,400$ |
| SATLINS_TANSIG | $60,250,300$ | $40,175,200,250,400$ |



Figure 3. Graph of NIC based on Sample Sizes (3 Variables)

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION



Figure 4. Graph of ANIC based on Sample Sizes (3 Variables)

Correspondingly based on three (3) variables, Figure 3 is the graph of NIC across samples, while Figure 4 is the graph of ANIC across samples. The models in ANIC became almost parallel from around sample number 20 and 40 up till sample number 400.

A test shows significant difference between the homogeneous and heterogeneous models ( $p<0.05$ ). Rates of selection for the heterogeneous models are respectively $72.9 \%$, and $72.1 \%$ using NIC, against $66.9 \%$, $55.9 \%$ and $65.1 \%$ respectively for the homogeneous models, while with ANIC the heterogeneous models have rates of selection respectively as $66.9 \%$ and $66.8 \%$, against $66.7 \%$, $66.2 \%$, and 66.6 for the respective homogeneous models. The results of the ANIC demonstrate the high precision of SNN models at large samples.

## UDOMBOSO ET AL.



Figure 5. Overall Rates of Efficiency and Selection of the SNN Models: SATLINS_TANH, SATLINS_TANSIG, SATLINS, TANH, TANSIG

## Conclusion

An ANIC criterion was derived, based on Kullback's symmetric divergence, for model selection in some Statistical Neural Network models. The analyses show that on a general note, the ANIC improves model selection in more sample sizes than does the NIC. Because neural network is a data-driven model, then more attention should be paid to the sample size when determining the model to be selected.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov \& F. Csaki, (Eds.), Second International Symposium Information Theory (pp. 267-281). Budapest: Akademia Kiado.

## ADJUSTED NETWORK INFORMATION FOR SNN MODEL SELECTION

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723. doi: 10.1109/TAC.1974.1100705

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Mathematical Statistics, 30(1), 9-14.

Amemiya, T. (1980). Selection of regressors. International Economic Review, 21(2), 331-354.

Anders, U. (1996). Statistical model building for neural networks. $\sigma^{\text {th }}$ International AFIR Colloquium. Nürnberg, Germany.

Bickel, P. J., \& Doksum, K. A. (1977). Mathematical statistics: Basic ideas and selected topics. New Jersey, NJ: Prentice-Hall.

Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. Statistics \& Probability Letters, 33(2), 201-208. doi: 10.1016/S0167-7152(96)00128-9

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. Statistics \& Probability Letters, 42(4), 333-343. doi: 10.1016/S0167-7152(98)00200-4

Hafidi, B., \& Mkhadri, A. (2006). A corrected Akaike criterion based on Kullback's symmetric divergence: applications in time series, multiple and multivariate regression. Computational Statistics \& Data Analysis, 50(6), 1524-1550. doi: 10.1016/j.csda.2005.01.007

Hannan, E., \& Quinn, B. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B (Methodological), 41(2), 190-195.

Hurvich, C. M., Shumway, R., \& Tsai, C.-L. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika, 77(4), 709-719. doi: 10.1093/biomet/77.4.709

Hurvich, C. M., \& Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297-307. doi: 10.1093/biomet/76.2.297

Hurvich, C. M., \& Tsai, C.-L. (1993). Corrected Akaike information criterion for vector autoregressive model selection. Journal of Time Series Analysis, 14(3), 271-279. doi: 10.1111/j.1467-9892.1993.tb00144.x

Kullback, S. (1968). Information theory and statistics. New York, NY: Dover.

## UDOMBOSO ET AL.

Murata, N., Yoshizawa, S., \& Amari, S. (1994). Network information criterion determining the number of hidden units for artificial neural network models. IEEE Transactions on Neural Networks, 5(6), 865-872. doi: 10.1109/72.329683

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464. doi: 10.1214/aos/1176344136

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics - Theory and Methods, 7(1), 13-26. doi: 10.1080/03610927808827599

# Limited Failure Censored Life Test Sampling Plan in Burr Type X Distribution 

R. R. L. Kantam<br>Acharya Nagarjuna University Andhra Pradesh, India

M. S. Ravikumar<br>Konaseema Institute of Medical Sciences \& Research Foundation<br>Andhra Pradesh, India

The Burr type X distribution is considered as a life time random variable of a product whose lots are to be decided for acceptance or otherwise on the basis of sample lifetimes drawn from the lot. The sample is divided into various groups in order to develop a group sampling plan in such a way that the life testing experiment is terminated as soon as the first failure in each group is observed. The acceptance criterion based on the theory of order statistics is proposed and is shown to be more economical than a criterion proposed in the earlier similar works.

Keywords: Single sampling, lot acceptance, group sampling plan, truncated life tests, reliability test plans, order statistics

## Introduction

Acceptance sampling is concerned with inspection and decision making regarding products. Life tests are experiments carried out on sample products in order to assess the life time of an item (time to its failure or the time it stops working satisfactorily). A common practice in life test is to terminate the test at a prefixed time and record the number of failures that occurred during that time period or when a prefixed number of failures is realised. The former termination is generally called truncated life tests/time censored life test and the latter is called a failure censored life test. If the quality of a product is measured through the life time, sampling plans to determine acceptability of a product with respect to life time are called Reliability Sampling Plans.

In life test sampling plans a common constraint is the duration of total time spent on testing. Sampling plans based on time truncated life tests would address this constraint to some extent. When the life time random variable is assumed to

[^25]
## KANTAM \& RAVIKUMAR

follow a specific continuous probability distribution, sampling plans are developed by various researchers covering a wide spectrum of probability models. Epstein (1954) was one of the foremost works about acceptance sampling plans based on truncated life tests with the exponential distribution as the probability model. Other researchers in this direction are as follows: Goode and Kao (1961) worked with the Weibull model which includes the exponential distribution as a particular case. Gupta and Groll (1961) and Gupta (1962) considered the gamma and log-normal distributions, respectively. More recently, the studies of Kantam, Rosaiah, and Srinivasa Rao (2001), Baklizi (2003), Baklizi and El-Masri (2004), Rosaiah and Kantam (2005), Balakrishanan, Lieva and López (2007), Aslam and Kantam (2008), Srinivasa Rao, Ghitany, and Kantam (2009), Rosaiah, Kantam, and Srinivasa Rao (2009), Srinivasa Rao and Kantam (2010), Lio, Tsai, and Wu (2009), Lio, Tsai, and Wu (2010), Lu (2011), Kantam, Sriram, and Suhasini (2012), Srinivasa Rao, Kantam, Rosaiah, and Pratapa Reddy (2012), Srinivasa Rao and Kantam (2013), Kantam and Sriram (2013), Subba Rao, Prasad, and Kantam (2013), Kantam, Sriram, and Suhasini (2013), Rosaiah, Kantam, Rama Krishnan, and Siva Kumar (2014), Subba Rao, Naga Durgamamba, and Kantam (2014) and the references therein, are related to construction of acceptance sampling plans based on truncated life tests with different probability models. In all these works, given the termination time of a life test, the construction of the sampling plan consists of determining the minimum number of sample items that are to be life-tested and the acceptance number beyond which the observed failures out of the life-tested items of the sample lead to rejection of the submitted lot, conditioned on pre specified producer's and consumer's risks.

However, if a failure censored life test is under consideration, one has to wait till a pre specified number of failures out of the sample items that are being tested is realised. Sometimes the life of product might be quite long possibly resulting in even a failure censored life-testing plan to be long time consuming. Johnson (1964) proposed a sampling plan in which the experimenter can decide to group the test units into several groups and then conduct the life-tests on all the groups simultaneously until the first failure in each group is realised. Based on the recorded first failure time in each group if a decision process about the acceptance/rejection of submitted lot is developed the procedure may be named as Limited Failure Censored Life Test Sampling Plan (LFCLTSP). Balasooriya (1995) developed such a sampling plan for the two parameter exponential distribution though the specific name is not given as LFCLTSP. Wu and Tsai (2000), Wu, Tsai, and Ouyang (2001), Jun, Balamurali, and Lee (2006) have proposed LFCLTSP when the underlying lifetime random variable follows

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Weibull distribution, with respective distinct approaches in working out the parameters of the sampling plan. The scheme of life testing and termination process of LFCLTSP is named by some researchers as 'Sudden death testing' (for example Pascual \& Meeker, 1998; Jun et al., 2006). 'Limited failure censored life tests' is the name proposed by Wu et al. (2001). Our suggested name is Limited Failure Censored Life Test Sampling Plan (LFCLTSP). Thus, the purpose of this study is to develop LFCLTSPs for one of the models of Burr (1942) - Burr type X distribution on lines of Jun et al. (2006). A new criterion is also suggested that is more economical.

## Construction of LFCLTSP (Jun et al. 2006)

The purpose of proposing LFCLTSP is to reduce testing time. The total number of products to be tested, say $N$ is divided into groups of equal size according to the number of available experimental testers. Thus there are $n$ items in each group and a total of $m$ groups may be considered for this grouping so that $N=m \times n$. The items in each group are tested identically and simultaneously on different testers. The first group of items is run until the first failure occurs. At this point the surviving items are suspended and removed from testing. An equal set of new items numbering $n$ is next tested until the first failure. This process is repeated until one failure is generated from each of the $m$ groups. In the end, $m$ failures are observed while $(n-1) m$ items are suspended. Wu et al. (2001) named this testing process as "limited failure censored life test". The sample information so obtained can be utilized for deciding upon the acceptance of the lot from which the original sample of $N$ is put for testing. According to the characteristics of testers a group size $n$ is usually specified but the total number of groups $m$ should be determined. For that a variable sampling plan is proposed by Jun et al. (2006) with the following assumptions/specifications

- The life time $X$ follows a Weibull distribution with a known shape parameter ( $k$ ).
- $\quad$ There is a lower specification limit $(L)$ regarding the life time.
- $\quad p_{0}$ is a desirable lot quality level (proportion of non conformities) at the pre specified producer's risk $\alpha$.
- $\quad p_{1}\left(>p_{0}\right)$ is an undesirable lot quality level (proportion of non conformities) at the pre specified consumer's risk $\beta$.


## KANTAM \& RAVIKUMAR

## Sampling Plan

The cumulative distribution function (cdf) of the base line distribution (Weibull) is given by

$$
\begin{equation*}
F(x)=1-\exp \left\{-x^{k}\right\} \tag{1}
\end{equation*}
$$

The fraction non-conforming or unreliability is expressed by

$$
\begin{equation*}
p=\operatorname{Pr}\{X<L\}=F(L) \tag{2}
\end{equation*}
$$

If $p$ is given, the corresponding $L$ is obtained from

$$
\begin{equation*}
w=L^{k}=-\ln (1-p) \tag{3}
\end{equation*}
$$

The proposed sampling plan of Jun et al. (2006) is as follows:
(i) Draw a random sample of size $N=m \times n$ and allocate $n$ items to each of the $m$ groups.
(ii) Observe $Y_{i}$ the time to the first failure in the $i^{\text {th }} \operatorname{group}(i=1,2, \ldots, m)$.
(iii) Calculate the quantity $V=\sum_{i=1}^{m} Y_{i}^{k}$.
(iv) Accept the lot if $V \geq c L^{k}$ and reject the lot otherwise ( $c$ may be called acceptability constant - a concept similar to the acceptance number in time truncated reliability test plans).

The number of groups $m$ and the acceptability constant $c$ are called the parameters of the sampling plan and will be determined by the following procedure:

Since $Y_{i}$ is the first order statistic in a sample of size $n$ from Weibull distribution with shape parameter $k$ its cdf is given by

$$
\begin{equation*}
\operatorname{Pr}\left(Y_{i} \leq y\right)=1-\exp \left(-n y^{k}\right) \tag{4}
\end{equation*}
$$

which is the cdf of a Weibull distribution with shape parameter $k$ and scale parameter $n^{1 / k}$. Therefore the variables $Y_{i}^{k}$ follow i.i.d exponential with scale parameter $n$ and as such $V=\sum_{i=1}^{m} Y_{i}^{k}$ follows a gamma distribution with shape

## LFCLTSP IN BURR TYPE X DISTRIBUTION

parameter $m$ and scale parameter $n$. Thus the quantity $2 n V$ follows a chi-square distribution with $2 m$ degrees of freedom so that the probability of acceptance of the lot for a lot quality level $p$ is given by

$$
\begin{equation*}
P_{a}(p)=\operatorname{Pr}\left\{V \geq c L^{k} \mid p\right\}=\operatorname{Pr}\left\{2 n V \geq 2 n c L^{k} \mid p\right\}=1-G_{2 m}(2 n c w) \tag{5}
\end{equation*}
$$

where $w$ is the solution of equation (3) and $G_{l}$ is the cdf of a chisquare variate with $l$ degrees of freedom. As in Fertig and Mann (1980), the probability of acceptance should be at least $(1-\alpha)$ at the desirable/acceptable lot quality level $p_{0}$ where $\alpha$ is producer's risk. Similarly, the probability of acceptance should not be more than $\beta$ at the undesirable/tolerance lot quality level $p_{1}$, where $\beta$ is consumer's risk. These two remarks lead to the following two inequalities

$$
\begin{gather*}
1-G_{2 m}\left(2 n c w_{0}\right) \geq 1-\alpha  \tag{6}\\
1-G_{2 m}\left(2 n c w_{1}\right) \leq \beta, \tag{7}
\end{gather*}
$$

If $\chi_{q, l}^{2}$ denotes the percentile point of tail probability $q$ in the chi-square distribution with $l$ degrees of freedom then, from (6), (7),

$$
\begin{align*}
2 n c w_{0} & \leq \chi_{1-\alpha, 2 m}^{2}  \tag{8}\\
2 n c w_{1} & \geq \chi_{\beta, 2 m}^{2} \tag{9}
\end{align*}
$$

which jointly lead to

$$
\begin{equation*}
\frac{w_{0}}{w_{1}} \leq \frac{\chi_{1-\alpha, 2 m}^{2}}{\chi_{\beta, 2 m}^{2}} \tag{10}
\end{equation*}
$$

Therefore, $m$ can be obtained by the smallest integer satisfying (10). The acceptability constant $c$ can be obtained from the equality case in either of the expressions (8), (9). It can be noticed that the number of groups $m$ is determined independently of the group size $n$ and also of the shape parameter $k$. Jun et al.

## KANTAM \& RAVIKUMAR

(2006) have evaluated $m, n c$ for $\alpha=0.05$ and $\beta=0.1$ at selected combinations of $p_{0}, p_{1}$. The corresponding table is reproduced below:

Table 1. Design parameters of sampling plans ( $\alpha=0.05, \beta=0.1$ )

|  | $p_{0}$ | $p_{1}$ | $g$ |
| :---: | ---: | ---: | ---: |
|  | 0.002 | 18.7 | 12201.0 |
|  | 0.004 | 5.1 | 2025.6 |
|  | 0.005 | 3.9 | 1308.0 |
|  | 0.010 | 2.1 | 408.5 |
|  | 0.050 | 1.0 | 44.4 |
|  | 0.100 | 0.8 | 18.6 |
|  | 0.010 | 0.010 | 18.6 |
| 2417.6 |  |  |  |
|  | 0.020 | 7.7 | 757.7 |
|  | 0.025 | 5.1 | 399.3 |
|  | 0.05 | 3.9 | 256.5 |
|  | 0.25 | 2.1 | 79.0 |
|  | 0.02 | 0.9 | 7.6 |
|  | 0.04 | 18.5 | 1195.3 |
|  | 0.05 | 5.0 | 196.0 |
|  | 0.10 | 3.8 | 125.7 |
|  | 0.15 | 2.1 | 37.8 |
|  | 0.3 | 1.6 | 20.0 |
|  | 0.1 | 1.1 | 7.0 |
|  | 0.2 | 17.4 | 217.7 |
|  | 0.05 | 0.3 | 4.6 |
|  | 0.5 | 2.5 | 20.7 |
|  | 0.2 | 1.8 | 14.3 |
|  | 0.4 | 16.1 | 9.1 |
|  | 0.5 | 4.0 | 13.9 |
|  | 3.0 | 7.7 |  |
|  |  |  |  |

For the sake of convenience in presentation, this procedure of Jun et al. (2006) is called Method-I and adopts the same for Burr type X distribution to construct LFCLTSP below.

## LFCLTSP for Burr type X distributed Lifetimes: Method-I

Let the life time of a product be given by Burr type X distribution with shape parameter $k$ so that cdf is given by

$$
\begin{equation*}
F(x)=\left(1-e^{-x^{2}}\right)^{k} \tag{11}
\end{equation*}
$$

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Let $L$ denote the $p^{\text {th }}$ quantile of a Burr type X variate.

$$
\begin{equation*}
\text { i.e., } F(L)=p \tag{12}
\end{equation*}
$$

If $p$ is given, the corresponding $L$ is obtained from

$$
\begin{equation*}
w=L=\sqrt{-\ln \left(1-p^{1 / k}\right)} \tag{13}
\end{equation*}
$$

Product with life time less than $L$ is considered nonconforming. Suppose the producer and the consumer have an agreement that lots with nonconforming fraction less than or equal to $p_{0}$ are presumed to be good and have to be accepted with probability of at least $1-\alpha$. Here $\alpha$ is called producer's risk. Furthermore suppose that lots with non conforming fraction greater than $p_{1}\left(>p_{0}\right)$ are not acceptable to the consumer and should be rejected with a probability of at least $1-\beta$. Here $\beta$ is called consumer's risk.

If a random sample of $N$ items grouped into $m$ groups of size $n$ each is put to test, an LFCLTSP on lines of Jun et al. (2006) can be constructed with the following decision process.

- $\quad$ Observe $Y_{i}$ the time to the first failure in the $i^{\text {th }} \operatorname{group}(i=1,2, \ldots, m)$.
- Calculate the quantity $V=\sum_{i=1}^{m} Y_{i}$.
- $\quad$ Accept the lot if $V \geq c L$ and reject the lot otherwise ( $c$ may be called acceptability constant - a concept similar to the acceptance number in time truncated reliability test plans).

In order to get the plan parameters $m$ and $c$, the percentiles of the sampling distribution of $V$ are needed, which is the sum of $m$ i.i.d observations on the first order statistic in a random sample of size $n$ modelled by Burr type $X$ distribution with shape parameter $k$. In view of the mathematical structure of the Burr type X model the sampling distribution of $V$ cannot be analytically tractable. Hence, consider the empirical sampling distribution of V for various known values of the shape parameter $k$ and tabulated the percentiles of $V$ for $k=1.5(0.5) 3 ; m=2(1) 10$; $n=5,10$ in Tables 2 through 5.

## KANTAM \& RAVIKUMAR

Table 2. Percentiles of $V=\sum_{i=1}^{m} Y_{i}$ at $k=1.5$

| m | $\begin{aligned} & \hline p \\ & n \\ & \hline \end{aligned}$ | 0.99865 | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.00135 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 5 | 2.177892 | 2.004307 | 1.895265 | 1.754279 | 1.641112 | 1.511261 | 0.726597 | 0.628004 | 0.545044 | 0.454419 | 0.400317 | 0.308623 |
|  | 10 | 1.575851 | 1.49449 | 1.445218 | 1.351926 | 1.27116 | 1.170157 | 0.569494 | 0.490655 | 0.429326 | 0.363279 | 0.321958 | 0.276647 |
| 3 | 5 | 2.831163 | 2.706339 | 2.58504 | 2.437557 | 2.313499 | 2.17402 | 1.198332 | 1.071406 | 0.970376 | 0.87162 | 0.801438 | 0.695547 |
|  | 10 | 2.197636 | 2.064846 | 1.995025 | 1.881046 | 1.78807 | 1.684527 | 0.935104 | 0.839882 | 0.758188 | 0.671293 | 0.62377 | 0.559463 |
| 4 | 5 | 3.647301 | 3.449295 | 3.310793 | 3.12908 | 2.97969 | 2.80371 | 1.67707 | 1.5263 | 1.389551 | 1.26126 | 1.149709 | 1.052957 |
|  | 10 | 2.826993 | 2.664092 | 2.570817 | 2.413215 | 2.301425 | 2.173745 | 1.315808 | 1.197378 | 1.098862 | 0.997235 | 0.92406 | 0.81006 |
| 5 | 5 | 4.350084 | 4.118747 | 3.969957 | 3.7743 | 3.612841 | 3.416829 | 2.170072 | 2.013894 | 1.878884 | 1.720819 | 1.618063 | 1.444002 |
|  | 10 | 12.35316 | 10.94002 | 10.13745 | 8.904167 | 7.859382 | 6.615174 | 1.765038 | 1.607898 | 1.474932 | 1.27609 | 1.109361 | 0.734325 |
| 6 | 5 | 5.034357 | 4.809434 | 4.648151 | 4.403913 | 4.239143 | 4.02923 | 2.640285 | 2.47937 | 2.336891 | 2.162175 | 2.042361 | 1.836678 |
|  | 10 | 3.875745 | 3.709052 | 3.602054 | 3.436095 | 3.28752 | 3.127581 | 2.081978 | 1.947834 | 1.82679 | 1.704452 | 1.610753 | 1.434844 |
| 7 | 5 | 5.753342 | 5.502226 | 5.357075 | 5.07924 | 4.8843 | 4.659395 | 3.158127 | 2.937147 | 2.796224 | 2.619996 | 2.491437 | 2.274716 |
|  | 10 | 4.474 | 4.197021 | 4.080527 | 3.92798 | 3.770507 | 3.610251 | 2.466902 | 2.325082 | 2.18972 | 2.048043 | 1.969056 | 1.812718 |
| 8 | 5 | 6.395463 | 6.164418 | 5.99307 | 5.723568 | 5.509403 | 5.256372 | 3.66206 | 3.466886 | 3.297804 | 3.083538 | 2.943184 | 2.722184 |
|  | 10 | 4.958382 | 4.753225 | 4.612685 | 4.430195 | 4.271718 | 4.088844 | 2.869967 | 2.709541 | 2.575134 | 2.42246 | 2.313287 | 2.101565 |
| 9 | 5 | 7.094235 | 6.772925 | 6.588319 | 6.343406 | 6.142674 | 5.868219 | 4.172239 | 3.952084 | 3.755656 | 3.52943 | 3.368455 | 3.118947 |
|  | 10 | 5.431611 | 5.218219 | 5.081496 | 4.902999 | 4.740855 | 4.549297 | 3.260287 | 3.085957 | 2.930402 | 2.757371 | 2.63884 | 2.414776 |
| 10 | 5 | 7.802554 | 7.440521 | 7.246377 | 7.000981 | 6.743447 | 6.488565 | 4.678347 | 4.460186 | 4.265542 | 4.0395 | 3.91871 | 3.691152 |
|  | 10 | 5.922311 | 5.721521 | 5.599249 | 5.377788 | 5.224957 | 5.021406 | 3.664299 | 3.467201 | 3.329653 | 3.15404 | 3.027771 | 2.857264 |

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 3. Percentiles of $V=\sum_{i=1}^{m} Y_{i}$ at $k=2$

| m | $\begin{gathered} p \\ n \\ n \end{gathered}$ | 0.99865 | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.00135 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 5 | 2.334626 | 2.197253 | 2.107378 | 2.00437 | 1.890244 | 1.755955 | 0.951999 | 0.85061 | 0.770884 | 0.6691 | 0.598262 | 0.491655 |
|  | 10 | 1.837277 | 1.740968 | 1.683879 | 1.590693 | 1.515541 | 1.422791 | 0.789355 | 0.707058 | 0.634459 | 0.559604 | 0.508588 | 0.425463 |
| 3 | 5 | 3.22835 | 3.041119 | 2.953291 | 2.801527 | 2.669792 | 2.521244 | 1.532702 | 1.404577 | 1.304342 | 1.169315 | 1.108815 | 0.998243 |
|  | 10 | 2.586898 | 2.426711 | 2.357428 | 2.246726 | 2.15384 | 2.038838 | 1.267581 | 1.165772 | 1.08047 | 0.985238 | 0.921347 | 0.830126 |
| 4 | 5 | 4.09782 | 3.860049 | 3.758922 | 3.59698 | 3.446504 | 3.272358 | 2.134587 | 1.979105 | 1.851842 | 1.715081 | 1.636566 | 1.496763 |
|  | 10 | 3.25933 | 3.113185 | 3.03314 | 2.910781 | 2.797988 | 2.658206 | 1.760559 | 1.64645 | 1.540285 | 1.430059 | 1.356785 | 1.208283 |
| 5 | 5 | 4.910748 | 4.690049 | 4.535921 | 4.355348 | 4.197799 | 4.015986 | 2.767222 | 2.597823 | 2.436065 | 2.270212 | 2.149996 | 1.928493 |
|  | 10 | 3.926676 | 3.772913 | 3.676726 | 3.525628 | 3.403761 | 3.260695 | 2.254587 | 2.118493 | 1.990416 | 1.882084 | 1.787045 | 1.662601 |
| 6 | 5 | 5.68631 | 5.456727 | 5.323453 | 5.119146 | 4.942976 | 4.743159 | 3.364193 | 3.188782 | 3.028525 | 2.861073 | 2.750664 | 2.51819 |
|  | 10 | 4.67039 | 4.4496 | 4.342586 | 4.163724 | 4.019068 | 3.85217 | 2.755476 | 2.61692 | 2.470129 | 2.328519 | 2.238464 | 2.092608 |
| 7 | 5 | 6.450999 | 6.221178 | 6.119936 | 5.888118 | 5.703747 | 5.487192 | 3.98455 | 3.782213 | 3.615684 | 3.414431 | 3.27489 | 3.007482 |
|  | 10 | 5.301514 | 5.121319 | 4.957179 | 4.764079 | 4.625306 | 4.454871 | 3.270242 | 3.11377 | 2.977862 | 2.836983 | 2.716608 | 2.564213 |
| 8 | 5 | 7.257875 | 7.028246 | 6.920731 | 6.647952 | 6.453962 | 6.211597 | 4.620773 | 4.419366 | 4.24421 | 4.02282 | 3.883549 | 3.6345 |
|  | 10 | 5.884038 | 5.674714 | 5.559945 | 5.374511 | 5.222974 | 5.03431 | 3.783271 | 3.616054 | 3.472336 | 3.30749 | 3.207673 | 3.001297 |
| 9 | 5 | 8.09619 | 7.859907 | 7.663112 | 7.404836 | 7.185362 | 6.929384 | 5.234979 | 4.991544 | 4.793696 | 4.601039 | 4.480408 | 4.229538 |
|  | 10 | 6.522471 | 6.308458 | 6.216781 | 6.005329 | 5.845744 | 5.63685 | 4.29928 | 4.11053 | 3.946653 | 3.751242 | 3.63625 | 3.409376 |
| 10 | 5 | 9.004663 | 8.570265 | 8.405551 | 8.140922 | 7.901065 | 7.642368 | 5.878805 | 5.634311 | 5.412461 | 5.177099 | 5.016886 | 4.7515 |
|  | 10 | 7.219127 | 6.952429 | 6.825122 | 6.602637 | 6.428021 | 6.218982 | 4.822058 | 4.635115 | 4.473158 | 4.275834 | 4.119699 | 3.913186 |

## KANTAM \& RAVIKUMAR

Table 4. Percentiles of $V=\sum_{i=1}^{m} Y_{i}$ at $k=2.5$

| m | $\begin{aligned} & p \\ & n \\ & n \end{aligned}$ | 0.99865 | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.00135 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 5 | 2.468259 | 2.346736 | 2.268899 | 2.159342 | 2.055733 | 1.933521 | 1.140584 | 1.035646 | 0.945613 | 0.842214 | 0.761873 | 0.664364 |
|  | 10 | 2.049563 | 1.941881 | 1.874171 | 1.777782 | 1.69742 | 1.610381 | 0.98222 | 0.892383 | 0.825022 | 0.747394 | 0.686008 | 0.581007 |
| 3 | 5 | 3.477266 | 3.304669 | 3.206421 | 3.074597 | 2.947724 | 2.798189 | 1.828369 | 1.705361 | 1.589362 | 1.444766 | 1.362453 | 1.234796 |
|  | 10 | 2.848432 | 2.708178 | 2.635159 | 2.524335 | 2.429885 | 2.322525 | 1.53788 | 1.434265 | 1.346467 | 1.240148 | 1.178894 | 1.081697 |
| 4 | 5 | 4.34472 | 4.233665 | 4.130113 | 3.943907 | 3.803379 | 3.636949 | 2.516907 | 2.364628 | 2.225065 | 2.067376 | 1.964329 | 1.693457 |
|  | 10 | 3.63085 | 3.47622 | 3.37126 | 3.252098 | 3.147937 | 3.019605 | 2.132412 | 2.012767 | 1.904756 | 1.774562 | 1.705587 | 1.610282 |
| 5 | 5 | 5.302715 | 5.14715 | 5.013002 | 4.815496 | 4.644212 | 4.463374 | 3.226286 | 3.05078 | 2.913382 | 2.749309 | 2.649468 | 2.465423 |
|  | 10 | 4.419297 | 4.238808 | 4.137307 | 3.99215 | 3.860642 | 3.724371 | 2.719253 | 2.578875 | 2.445154 | 2.309705 | 2.210798 | 2.077999 |
| 6 | 5 | 6.310054 | 6.046179 | 5.863146 | 5.671537 | 5.510671 | 5.306801 | 3.914177 | 3.734768 | 3.566162 | 3.40839 | 3.280709 | 3.105496 |
|  | 10 | 5.149388 | 4.954964 | 4.874275 | 4.727925 | 4.583576 | 4.420357 | 3.322092 | 3.177341 | 3.050563 | 2.894142 | 2.764584 | 2.559992 |
| 7 | 5 | 7.205234 | 6.949778 | 6.813811 | 6.554688 | 6.356546 | 6.131491 | 4.652032 | 4.439628 | 4.251104 | 4.045369 | 3.886589 | 3.644996 |
|  | 10 | 5.871197 | 5.708455 | 5.592766 | 5.429955 | 5.273217 | 5.110513 | 3.931232 | 3.779749 | 3.646526 | 3.489141 | 3.375452 | 3.204921 |
| 8 | 5 | 7.984277 | 7.757269 | 7.597745 | 7.367686 | 7.159602 | 6.938196 | 5.367689 | 5.141553 | 4.930601 | 4.724564 | 4.605638 | 4.358607 |
|  | 10 | 6.67911 | 6.430736 | 6.31225 | 6.143098 | 5.980323 | 5.790149 | 4.524136 | 4.356781 | 4.203633 | 3.990518 | 3.874942 | 3.713165 |
| 9 | 5 | 8.985175 | 8.649383 | 8.484864 | 8.252281 | 8.020122 | 7.76861 | 6.08215 | 5.845327 | 5.649002 | 5.409941 | 5.242056 | 4.959961 |
|  | 10 | 7.448108 | 7.168401 | 7.019431 | 6.843368 | 6.664836 | 6.467801 | 5.140985 | 4.959848 | 4.799899 | 4.579462 | 4.474995 | 4.237564 |
| 10 |  | 9.800283 | 9.510817 | 9.340881 | 9.026834 | 8.817583 | 8.572455 | 6.806433 | 6.554773 | 6.34551 | 6.116703 | 5.947915 | 5.690806 |
|  | 10 | 8.165481 | 7.897884 | 7.751006 | 7.536716 | 7.355737 | 7.151375 | 5.737844 | 5.531289 | 5.360084 | 5.169283 | 5.035914 | 4.815527 |

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 5. Percentiles of $V=\sum_{i=1}^{m} Y_{i}$ at $k=3$

| m | $\begin{aligned} & p \\ & n \\ & n \end{aligned}$ | 0.99865 | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.00135 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 2 | 5 | 2.608441 | 2.473904 | 2.411283 | 2.292199 | 2.192362 | 2.077583 | 1.297579 | 1.197155 | 1.109135 | 1.004346 | 0.939656 | 0.829669 |
|  | 10 | 2.165301 | 2.075306 | 2.019814 | 1.919237 | 1.84604 | 1.757886 | 1.134309 | 1.045123 | 0.968032 | 0.883864 | 0.833268 | 0.742663 |
| 3 | 5 | 3.683211 | 3.521215 | 3.433714 | 3.297114 | 3.164566 | 3.024332 | 2.054279 | 1.929827 | 1.817847 | 1.703655 | 1.628594 | 1.463765 |
|  | 10 | 3.070413 | 2.948811 | 2.859479 | 2.751485 | 2.647107 | 2.544672 | 1.77378 | 1.669841 | 1.579924 | 1.480296 | 1.418421 | 1.302424 |
| 4 | 5 | 4.664032 | 4.523223 | 4.389786 | 4.229283 | 4.108187 | 3.939568 | 2.82075 | 2.675108 | 2.546479 | 2.380471 | 2.26397 | 2.127763 |
|  | 10 | 3.873799 | 3.754457 | 3.677123 | 3.566902 | 3.458809 | 3.32237 | 2.444186 | 2.319325 | 2.200637 | 2.067358 | 2.005182 | 1.859126 |
| 5 | 5 | 5.748489 | 5.546665 | 5.401009 | 5.195914 | 5.022025 | 4.837675 | 3.599143 | 3.430413 | 3.281878 | 3.119621 | 3.007574 | 2.832034 |
|  | 10 | 4.792635 | 4.623547 | 4.509994 | 4.37538 | 4.249666 | 4.103716 | 3.108066 | 2.963333 | 2.845844 | 2.706644 | 2.627772 | 2.482635 |
| 6 | 5 | 6.69915 | 6.489924 | 6.349406 | 6.138978 | 5.953869 | 5.751055 | 4.376599 | 4.196276 | 4.034022 | 3.858248 | 3.724574 | 3.472215 |
|  | 10 | 5.630339 | 5.43406 | 5.32593 | 5.184179 | 5.043455 | 4.883834 | 3.791391 | 3.629649 | 3.505016 | 3.354572 | 3.259626 | 3.033245 |
| 7 | 5 | 7.676058 | 7.437145 | 7.273444 | 7.065268 | 6.877302 | 6.656454 | 5.167565 | 4.962574 | 4.780826 | 4.583857 | 4.456209 | 4.271837 |
|  | 10 | 6.379116 | 6.22337 | 6.107936 | 5.950745 | 5.80989 | 5.641027 | 4.463186 | 4.289654 | 4.148583 | 3.983681 | 3.881626 | 3.707637 |
| 8 |  | 8.597227 | 8.333983 | 8.168087 | 7.933648 | 7.749028 | 7.53491 | 5.966976 | 5.74813 | 5.543404 | 5.322675 | 5.162271 | 4.849312 |
|  | 10 | 7.229972 | 7.069915 | 6.958532 | 6.73594 | 6.591474 | 6.399465 | 5.142293 | 4.971053 | 4.82937 | 4.649567 | 4.553311 | 4.325541 |
| 9 | 5 | 9.592283 | 9.262802 | 9.095726 | 8.845289 | 8.653274 | 8.412515 | 6.770802 | 6.530221 | 6.326654 | 6.088762 | 5.95204 | 5.680317 |
|  | 10 | 7.980247 | 7.808114 | 7.680098 | 7.476421 | 7.340753 | 7.151641 | 5.819063 | 5.618981 | 5.458145 | 5.287944 | 5.141031 | 4.962647 |
| 10 | 5 | 10.65452 | 10.25902 | 10.07409 | 9.798465 | 9.57316 | 9.320885 | 7.579234 | 7.356013 | 7.111197 | 6.872162 | 6.685894 | 6.374154 |
|  | 10 | 8.888584 | 8.627152 | 8.485921 | 8.290417 | 8.122677 | 7.916383 | 6.519429 | 6.308877 | 6.138436 | 5.943742 | 5.811467 | 5.586315 |

## KANTAM \& RAVIKUMAR

If $G($.$) stands for the cdf of the random variable V$, the percentiles in Tables 2 through 5 are the values of $G^{-1}(p)$. If $G_{k}^{-1}(q)$ stands for the $q^{\text {th }}$ percentile of $V$ with the shape parameter $k$ the following inequalities are parallel to the expressions (6) through (10).

$$
\begin{gather*}
G_{k}\left(n c w_{0}\right) \leq \alpha  \tag{14}\\
G_{k}\left(n c w_{1}\right) \geq 1-\beta  \tag{15}\\
n c w_{0} \leq G_{k}^{-1}(1-\alpha)  \tag{16}\\
n c w_{1} \geq G_{k}^{-1}(\beta) \tag{17}
\end{gather*}
$$

which jointly lead to

$$
\begin{equation*}
\frac{w_{0}}{w_{1}} \leq \frac{G_{k}^{-1}(1-\alpha)}{G_{k}^{-1}(\beta)} \tag{18}
\end{equation*}
$$

Therefore, $m$ can be obtained by the smallest integer satisfying (18). The acceptability constant $c$ can be obtained from the equality case in either of the expressions (16), (17). We have tabulated the values of $m$ and $c$ determined for the same combinations of $p_{0}, p_{1}$ as chosen by Jun et al. (2006) and are presented in Tables 6 through 9 for $k=1.5(0.5) 3$.

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 6. Design parameters of LFCLTSP ( $\alpha=0.05, \beta=0.1, k=1.5$ )

| $p_{0}$ | $p_{1}$ |  | c |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  | $m$ | $n=5$ | $n=10$ |
| 0.001 | 0.002 | ----- | ----- | ----- |
|  | 0.004 | 7 | 29.29783 | 23.19252 |
|  | 0.005 | 5 | 20.08845 | 16.03866 |
|  | 0.010 | 3 | 10.6872 | 8.377761 |
|  | 0.050 | 2 | 6.264294 | 4.894248 |
|  | 0.100 | 2 | 6.264294 | 4.894248 |
| 0.005 | 0.010 | ---- | ----- | ----- |
|  | 0.015 | 10 | 25.891 | 20.12681 |
|  | 0.020 | 6 | 14.39253 | 11.30701 |
|  | 0.025 | 5 | 11.69048 | 9.333712 |
|  | 0.05 | 2 | 3.64551 | 2.848211 |
|  | 0.25 | 2 | 3.64551 | 2.848211 |
| 0.01 | 0.02 | ----- | ----- | -- |
|  | 0.04 | 6 | 11.37282 | 8.934676 |
|  |  | 5 |  |  |
|  | 0.10 | 2 | 2.880642 | 2.250625 |
|  | 0.15 | 2 | 2.880642 | 2.250625 |
|  | 0.3 | 2 | 2.880642 | 2.250625 |
| 0.05 | 0.1 | ---- | ----- | ----- |
|  | 0.2 | 5 | 5.273139 | 4.210088 |
|  | 0.25 | 4 | 3.996433 | 3.13519 |
|  | 0.3 | 3 | 2.805348 | 2.19913 |
|  | 0.5 | 2 | 1.644353 | 1.284721 |
| 0.1 | 0.2 | 19 | 18.30972 | ---- |
|  | 0.4 | 4 | 3.098573 | 2.430822 |
|  | 0.5 | 3 | 2.175083 | 1.705062 |

## KANTAM \& RAVIKUMAR

Table 7. Design parameters of LFCLTSP ( $\alpha=0.05, \beta=0.1, k=2$ )

| $p_{0}$ | $p_{1}$ |  | c |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  | m | $n=5$ | $n=10$ |
| 0.001 | 0.002 | ----- | ----- | ----- |
|  | 0.004 | 8 | 24.65361 | 20.1723 |
|  | 0.005 | 6 | 17.78875 | 14.59859 |
|  | 0.010 | 3 | 7.835488 | 6.503305 |
|  | 0.050 | 2 | 4.745161 | 3.944351 |
|  | 0.100 | 2 | 4.745161 | 3.944351 |
| 0.005 | 0.010 | ----- | ----- | ----- |
|  | 0.015 | 11 | 23.02939 | ----- |
|  | 0.020 | 7 | 13.96657 | 11.49821 |
|  | 0.025 | 5 | 9.592978 | 7.822957 |
|  | 0.05 | 3 | 5.18668 | 4.304845 |
|  | 0.25 | 2 | 3.141047 | 2.610952 |
| 0.01 | 0.02 | ----- | ----- | ----- |
|  | 0.04 | 7 | 11.65218 | 9.592849 |
|  | 0.05 | 5 | 8.003328 | 6.526616 |
|  | 0.10 | 3 | 4.327196 | 3.59149 |
|  | 0.15 | 2 | 2.620545 | 2.178292 |
|  | 0.3 | 2 | 2.620545 | 2.178292 |
| 0.05 | 0.1 | ----- | ----- | ---- |
|  | 0.2 | 5 | 5.163768 | 4.21099 |
|  | 0.25 | 4 | 3.933925 | 3.272697 |
|  | 0.3 | 3 | 2.791919 | 2.317239 |
|  | 0.5 | 2 | 1.690782 | 1.40544 |
| 0.1 | 0.2 | 19 | 18.27998 | -- |
|  | 0.4 | 4 | 3.209982 | 2.670436 |
|  | 0.5 | 3 | 2.278134 | 1.890807 |

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 8. Design parameters of LFCLTSP ( $\alpha=0.05, \beta=0.1, k=2.5$ )

| $p_{0}$ | $p_{1}$ | m | $c$ $n=5$ | $n=5$ |
| :---: | :---: | :---: | :---: | :---: |
| 0.001 | 0.002 | ----- | ----- | ----- |
|  | 0.004 | 8 | 22.8966 | 19.42811 |
|  | 0.005 | 6 | 17.39037 | 14.80557 |
|  | 0.010 | 3 | 6.680032 | 5.618128 |
|  | 0.050 | 2 | 4.056706 | 3.495534 |
|  | 0.100 | 2 | 4.056706 | 3.495534 |
| 0.005 | 0.010 | ----- | ----- | ----- |
|  | 0.015 | 11 | 24.39888 | ----- |
|  | 0.020 | 7 | 14.37326 | 12.17942 |
|  | 0.025 | 5 | 10.44058 | 8.882286 |
|  | 0.05 | 3 | 4.767352 | 4.009501 |
|  | 0.25 | 2 | 2.895158 | 2.494665 |
| 0.01 | 0.02 | ----- | ----- | --- |
|  | 0.04 | 7 | 10.68762 | 9.099074 |
|  | 0.05 | 5 | 7.344211 | 6.208183 |
|  | 0.10 | 3 | 4.105354 | 3.452738 |
|  | 0.15 | 2 | 2.493134 | 2.148254 |
|  | 0.3 | 2 | 2.493134 | 2.148254 |
| 0.05 | 0.1 | --- | ----- | ----- |
|  | 0.2 | 5 | 6.232243 | 5.302059 |
|  | 0.25 | 4 | 3.945877 | 3.358724 |
|  | 0.3 | 3 | 2.845752 | 2.393372 |
|  | 0.5 | 2 | 1.728192 | 1.489128 |
| 0.1 | 0.2 | 19 | 19.32857 | ----- |
|  | 0.4 | 4 | 3.318712 | 2.824882 |
|  | 0.5 | 3 | 2.393443 | 2.012965 |

## KANTAM \& RAVIKUMAR

Table 9. Design parameters of LFCLTSP ( $\alpha=0.05, \beta=0.1, k=3$ )

| $p_{0}$ | $p_{1}$ | m $c$ |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  |  | m | $n=5$ | $n=5$ |
| 0.001 | 0.002 | ----- | ----- | ----- |
|  | 0.004 | 9 | 20.11819 | 17.31086 |
|  | 0.005 | 7 | 15.28861 | 13.21549 |
|  | 0.010 | 4 | 8.241426 | 7.145336 |
|  | 0.050 | 2 | 3.688174 | 3.219797 |
|  | 0.100 | 2 | 3.688174 | 3.219797 |
| 0.005 | 0.010 | ----- | ----- | ---- |
|  | 0.015 | 14 | 24.21504 | ----- |
|  | 0.020 | 8 | 13.2736 | 11.47917 |
|  | 0.025 | 6 | 9.690052 | 8.381595 |
|  | 0.05 | 3 | 4.456362 | 3.856001 |
|  | 0.25 | 2 | 2.764474 | 2.413401 |
| 0.01 | 0.02 | ----- | ----- | ----- |
|  | 0.04 | 8 | 11.6694 | 10.09184 |
|  | 0.05 | 6 | 8.518946 | 7.368625 |
|  | 0.10 | 3 | 3.917781 | 3.389978 |
|  | 0.15 | 2 | 2.430369 | 2.121726 |
|  | 0.3 | 2 | 2.430369 | 2.121726 |
| 0.05 | 0.1 | ----- | ----- | ----- |
|  | 0.2 | 6 | 6.19041 | 5.354513 |
|  | 0.25 | 4 | 3.94636 | 3.421503 |
|  | 0.3 | 3 | 2.84691 | 2.463375 |
|  | 0.5 | 2 | 1.766061 | 1.541781 |
| 0.1 | 0.2 | 20 | 19.35645 | ----- |
|  | 0.4 | 4 | 3.386708 | 2.936284 |
|  | 0.5 | 3 | 2.443176 | 2.114032 |

It may be noted that $m$ is solved as integer values only and $m, c$ depend on the shape parameter $k$ of the Burr type X distribution.

## LFCLTSP for Burr type X distributed Lifetimes: Method-II

The statistic $V=\sum_{i=1}^{m} Y_{i}$ introduced for the decision process of the sampling plan seems to have been considered as the total test time to get the limited failure censored sample $-Y_{1}, Y_{2}, \ldots, Y_{m}$ which are $m$ first order statistics in $m$ independent random samples of size n each. If $Z$ denotes the maximum of $Y_{1}, Y_{2}, \ldots, Y_{m}$ it may also be viewed as the total test time/experimental time as opined by Kantam and Srinivasa Rao (2004). Hence, larger realized value of $Z$ can be considered as an indication that the products in the submitted lot have longer life prompting one to

## LFCLTSP IN BURR TYPE X DISTRIBUTION

consider the lot as a good lot for acceptability. In other words " $\mathrm{Z}>c L$ " can be taken as a criterion of acceptance of the lot. Thus, for Method-II the following decision rule is proposed:
(i) Draw a random sample of size $N=m \times n$ and allocate $n$ items to each of the $m$ groups.
(ii) Observe $Y_{i}$ the time to the first failure in the $i^{\text {th }} \operatorname{group}(I=1,2, \ldots, m)$.
(iii) Identify the quantity $Z=\operatorname{Max}\left(Y_{1}, Y_{2}, Y_{3}, \ldots, Y_{m}\right)$.
(iv) Accept the lot if $\mathrm{Z} \geq c L$ and reject the lot otherwise ( $c$ may be called acceptability constant - a concept similar to the acceptance number in time truncated reliability test plans).

Using the theory of order statistics, the cdf of $Z$ may be obtained in a closed form as long as the cdf of the base line distribution is in a closed form. Hence, the percentiles of $Z$ can be used to get the design parameters $m, c$ analytically. For the focal distribution, Burr type X distribution with shape parameter $k$, the following is the analytical procedure of calculating design parameters of LFCLTSP by Method-II.

The cdf of Burr type X with shape parameter $k$ is

$$
\begin{equation*}
F(x)=\left(1-e^{-x^{2}}\right)^{k} \tag{19}
\end{equation*}
$$

Let $X_{1}, X_{2}, X_{3}, \ldots, X_{n}$ be a random sample of size $n$ from (19) The cdf of least of $X_{1}, X_{2}, X_{3}, \ldots, X_{n}$ is given by

$$
\begin{equation*}
F_{(1)}(x)=1-[1-F(x)]^{n} \tag{20}
\end{equation*}
$$

That is,

$$
\begin{equation*}
F_{(1)}(x)=1-\left[1-\left(1-e^{-x^{2}}\right)^{k}\right]^{n} \tag{21}
\end{equation*}
$$

$Y_{1}, Y_{2}, Y_{3}, \ldots, Y_{m}$ of the limited failure censored test are now a random sample of size $m$ from $F_{(1)}(x)$. Hence, the cdf of $Z$ - the largest of $Y_{1}, Y_{2}, Y_{3}, \ldots, Y_{m}$ is given by

## KANTAM \& RAVIKUMAR

$$
\begin{gather*}
G_{(m)}(z)=\left[F_{1}(z)\right]^{m}  \tag{22}\\
\text { i.e., } G_{(m)}(z)=\left[1-\left[1-\left(1-e^{-z^{2}}\right)^{k}\right]^{n}\right]^{m} \tag{23}
\end{gather*}
$$

As a corollary if $k=1$ then RHS of (23) becomes

$$
\begin{equation*}
G_{(m)}(z)=\left(1-e^{-n z^{2}}\right)^{m}, \tag{24}
\end{equation*}
$$

which correspondents to the cdf of $Z$ when the base line distribution is the well known Rayleigh distribution which in turn is a special case of Weibull distribution. The design parameters $m$ and $c$ of LFCLTSP are obtained with the help of percentiles of $G_{(m)}(z)$ given in (23). If $\alpha$ and $\beta$ are respectively the producer's and consumer's risks for desirable/acceptable lot quality level $p_{0}$, undesirable/lot tolerance quality level $p_{1}$ then $m$ and $c$ are the solutions of the following two inequalities.

$$
\begin{gather*}
G_{m}\left(c w_{0}\right) \leq \alpha  \tag{25}\\
G_{m}\left(c w_{1}\right) \geq 1-\beta \tag{26}
\end{gather*}
$$

where $w_{0}$ and $w_{1}$ are as defined above.
The inequalities (25), (26) respectively imply

$$
\begin{gather*}
c w_{0} \leq G_{m}^{-1}(1-\alpha)  \tag{27}\\
c w_{1} \geq G_{m}^{-1}(\beta) \tag{28}
\end{gather*}
$$

which jointly lead to

$$
\begin{equation*}
\frac{w_{0}}{w_{1}} \leq \frac{G_{m}^{-1}(1-\alpha)}{G_{m}^{-1}(\beta)} \tag{29}
\end{equation*}
$$

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Therefore, $m$ can be obtained by the smallest integer satisfying (29). The acceptability constant $c$ can be obtained from the equality case in either of the expressions (27), (28). The values of $m$ and $c$ were analytically determined for the same combinations of $p_{0}, p_{1}$ as chosen by Jun et al. (2006) and are presented in Tables 10 through 13 for $k=1.5(0.5) 3$ along with the values of the design parameters of LFCLTSP of Method-I also for the sake of comparison. The values of $m$ obtained for Method-II can be seen to be consistently smaller than or equal to those of Method-I, thus indicating less number of items to be put to life test in Method-II and hence giving a preference to Method-II over Method-I.

Table 10. Design parameters of LFCLTSP of Methods -I and II at $k=1.5, \alpha=0.05$ and $\beta=0.1$


## KANTAM \& RAVIKUMAR

Table 11. Design parameters of LFCLTSP of Methods-I and II at $k=2, \alpha=0.05$ and $\beta=0.1$


## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 12. Design parameters of LFCLTSP of Methods-I and II at $k=2.5, \alpha=0.05$ and $\beta=0.1$

| $p_{0}$ | $p_{1}$ | $m$ |  |  |  | c |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $n=5$ |  | $n=10$ |  | $n=5$ |  | $n=10$ |  |
|  |  | I | 1 | 1 | II | I | II | I | II |
| 0.001 | 0.002 | - |  | ----- | 5 | - | 3.226744 | -- | 2.606892 |
|  | 0.004 | 8 | 3 | 8 | 2 | 22.8966 | 2.698795 | 19.42811 | 1.995706 |
|  | 0.005 | 6 |  | 6 | 2 | 17.39037 | 2.340102 | 14.80557 | 1.995706 |
|  | 0.010 | 3 |  | 3 | 2 | 6.680032 | 2.340102 | 5.618128 | 1.995706 |
|  | 0.050 | 2 |  | 2 | 2 | 4.056706 | 2.340102 | 3.495534 | 1.995706 |
|  | 0.100 | 2 | 2 | 2 | 2 | 4.056706 | 2.340102 | 3.495534 | 1.995706 |
| 0.005 | 0.010 | ----- | 6 | ----- | 5 | ----- | 2.302837 | ----- | 1.860466 |
|  | 0.015 | 11 | 4 | 11 | 3 | 24.39888 | 2.091057 | ----- | 1.632736 |
|  | 0.020 | 7 |  | 7 | 2 | 14.37326 | 1.926055 | 12.17942 | 1.42428 |
|  | 0.025 | 5 |  | 5 | 2 | 10.44058 | 1.670065 | 8.882286 | 1.42428 |
|  | 0.05 | 3 |  | 3 | 2 | 4.767352 | 1.670065 | 4.009501 | 1.42428 |
|  | 0.25 | 2 |  | 2 | 2 | 2.895158 | 1.670065 | 2.494665 | 1.42428 |
| 0.01 | 0.02 | ---- | 6 | --- | 5 | -- | 1.983063 | ----- | 1.60212 |
|  | 0.04 | 7 | 3 | 7 | 2 | 10.68762 | 1.658601 | 9.099074 | 1.226503 |
|  | 0.05 | 5 |  | 5 | 2 | 7.344211 | 1.438159 | 6.208183 | 1.226503 |
|  | 0.10 | 3 |  | 3 | 2 | 4.105354 | 1.438159 | 3.452738 | 1.226503 |
|  | 0.15 | 2 |  | 2 | 2 | 2.493134 | 1.438159 | 2.148254 | 1.226503 |
|  | 0.3 | 2 |  | 2 | 2 | 2.493134 | 1.438159 | 2.148254 | 1.226503 |
| 0.05 | 0.1 | -- |  | -- | 4 | ---- | 1.374621 | ----- | 1.05379 |
|  | 0.2 | 5 |  | 5 | 2 | 6.232243 | 0.996904 | 5.302059 | 0.850188 |
|  | 0.25 | 4 |  | 4 | 2 | 3.945877 | 0.996904 | 3.358724 | 0.850188 |
|  | 0.3 | 3 |  | 3 | 2 | 2.845752 | 0.996904 | 2.393372 | 0.850188 |
|  | 0.5 | 2 |  | 2 | 2 | 1.728192 | 0.996904 | 1.489128 | 0.850188 |
| 0.1 | 0.2 | 19 |  | 19 | 4 | 19.32857 | 1.10974 | - | 0.886299 |
|  | 0.4 | 4 |  | 4 | 2 | 3.318712 | 0.838454 | 2.824882 | 0.715058 |
|  | 0.5 | 3 | 2 | 3 | 2 | 2.393443 | 0.838454 | 2.012965 | 0.715058 |

## KANTAM \& RAVIKUMAR

Table 13. Design parameters of LFCLTSP of Methods-I and II at $k=3, \alpha=0.05$ and $\beta=0.1$

| $p_{0}$ | $p_{1}$ | $m$ |  |  |  |  | c |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $n=5$ |  |  | $n=10$ |  | $n=5$ |  | $n=10$ |  |
|  |  | I | 1 |  | , | II | 1 | II | 1 | II |
| 0.001 | 0.002 | ----- | 6 |  | ----- | 5 | ----- | 2.769706 | ----- | 2.290677 |
|  | 0.004 | 9 | 2 |  | 9 | 2 | 20.11819 | 2.082676 | 17.31086 | 1.811738 |
|  | 0.005 | 7 | 2 |  | 7 | 2 | 15.28861 | 2.082676 | 13.21549 | 1.811738 |
|  | 0.010 | 4 | 2 |  | 4 | 2 | 8.241426 | 2.082676 | 7.145336 | 1.811738 |
|  | 0.050 | 2 | 2 |  | 2 | 2 | 3.688174 | 2.082676 | 3.219797 | 1.811738 |
|  | 0.100 | 2 | 2 |  | 2 | 2 | 3.688174 | 2.082676 | 3.219797 | 1.811738 |
| 0.005 | 0.010 | ----- | 6 |  | ----- | 5 | ----- | 2.076035 | ----- | 1.716979 |
|  | 0.015 | 14 | 3 |  | 14 | 3 | 24.21504 | 1.770461 | ----- | 1.530374 |
|  | 0.020 | 8 | 2 |  | 8 | 2 | 13.2736 | 1.561071 | 11.47917 | 1.357989 |
|  | $0.025$ | 6 | 2 |  | 6 | 2 | 9.690052 | 1.561071 | 8.381595 | 1.357989 |
|  | 0.05 | 3 | 2 |  | 3 | 2 | 4.456362 | 1.561071 | 3.856001 | 1.357989 |
|  | 0.25 | 2 |  |  | 2 | 2 | 2.764474 | 1.561071 | 2.413401 | 1.357989 |
| 0.01 | 0.02 | -- | 6 |  | --- | 5 | ---- | 1.825133 | ----- | 1.509471 |
|  | 0.04 | 8 | 2 |  | 8 | 2 | 11.6694 | 1.372406 | 10.09184 | 1.193868 |
|  | 0.05 | 6 | 2 |  | 6 | 2 | 8.518946 | 1.372406 | 7.368625 | 1.193868 |
|  | 0.10 | 3 | 2 |  | 3 | 2 | 3.917781 | 1.372406 | 3.389978 | 1.193868 |
|  | 0.15 | 2 | 2 |  | 2 | 2 | 2.430369 | 1.372406 | 2.121726 | 1.193868 |
|  | 0.3 | 2 | 2 |  | 2 | 2 | 2.430369 | 1.372406 | 2.121726 | 1.193868 |
| 0.05 | 0.1 | ----- | 5 |  | ----- | 4 | ----- | 1.278518 | ----- | 1.047212 |
|  | 0.2 | 6 | 2 |  | 6 | 2 | 6.19041 | 0.997278 | 5.354513 | 0.86754 |
|  | 0.25 | 4 | 2 |  | 4 | 2 | 3.94636 | 0.997278 | 3.421503 | 0.86754 |
|  | 0.3 | 3 |  |  | 3 | 2 | 2.84691 | 0.997278 | 2.463375 | 0.86754 |
|  | 0.5 | 2 | 2 |  | 2 | 2 | 1.766061 | 0.997278 | 1.541781 | 0.86754 |
| 0.1 | 0.2 | 20 | 5 |  | 20 | 4 | 19.35645 | 1.097205 | ----- | 0.898702 |
|  | 0.4 | 4 | 2 |  | 4 | 2 | 3.386708 | 0.855849 | 2.936284 | 0.74451 |
|  | 0.5 | 3 | 2 |  | 3 | 2 | 2.443176 | 0.855849 | 2.114032 | 0.74451 |

When $k=1$ Burr type X is a Rayleigh distribution which is a Weibull distribution with shape parameter $=2$. Jun et al. (2006) observed that their LFCLTSP for Weibull distribution is invariant of its shape parameter. As matter of comparison, design parameters of LFCLTSP of Method-II were computed for Burr type X at $k=1$ also, so that these become the parameters of LFCLTSP for Weibull distribution with shape 2. These are given Table 14.

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Table 14. Design parameters of LFCLTSP of Method-II at $k=1, \alpha=0.05$ and $\beta=0.1$

|  | $p_{0}$ | $m$ |  | $c$ |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
|  | $p_{1}$ | $n=5$ | $n=5$ | $n=5$ | $n=10$ |
|  | 0.002 | 5 | 4 | 12.6215 | 7.999934 |
| 0.001 | 0.004 | 2 | 2 | 7.112941 | 5.029608 |
|  | 0.005 | 2 | 2 | 7.112941 | 5.029608 |
|  | 0.010 | 2 | 2 | 7.112941 | 5.029608 |
|  | 0.050 | 2 | 2 | 7.112941 | 5.029608 |
|  | 0.100 | 2 | 2 | 7.112941 | 5.029608 |
|  | 0.010 | 5 | 4 | 5.63885 | 3.574094 |
|  | 0.015 | 3 | 2 | 4.281842 | 2.247055 |
|  | 0.020 | 2 | 2 | 3.177816 | 2.247055 |
|  | 0.025 | 2 | 2 | 3.177816 | 2.247055 |
|  | 0.05 | 2 | 2 | 3.177816 | 2.247055 |
|  | 0.25 | 2 | 2 | 3.177816 | 2.247055 |
|  | 0.02 | 5 | 4 | 3.982257 | 2.524089 |
|  | 0.04 | 2 | 2 | 2.244231 | 1.586911 |
|  | 0.05 | 2 | 2 | 2.244231 | 1.586911 |
|  | 0.10 | 2 | 2 | 2.244231 | 1.586911 |
|  | 0.15 | 2 | 2 | 2.244231 | 1.586911 |
|  | 0.3 | 2 | 2 | 2.244231 | 1.586911 |
|  | 0.1 | 5 | 4 | 1.762744 | 1.117287 |
|  | 0.2 | 2 | 2 | 0.993408 | 0.702445 |
|  | 0.25 | 2 | 2 | 0.993408 | 0.702445 |
|  | 0.3 | 2 | 2 | 0.993408 | 0.702445 |
|  | 0.5 | 2 | 2 | 0.993408 | 0.702445 |
|  | 0.2 | 5 | 3 | 1.229931 | 0.660398 |
|  | 0.4 | 2 | 0.933944 | 0.490122 |  |
|  | 0.5 | 2 | 0.693137 | 0.490122 |  |

Comparison of Tables 1 and 14 also indicate that Method-II is preferable to Method-I in constructing LFCLTSP for Rayleigh distributed life times.

## Illustration

The quality assurance in a bearing manufacturing process states that $p_{0}=0.01$, $p_{1}=0.04, \alpha=0.05, \beta=0.1$ the number of test positions (size of each group, $n=10$. For this information Table -2.1 of Jun et al. (2006) suggests $m=5, c=196$. Accordingly a random sample of size $N=50$ items are put to test in five groups with 10 items in each group. The observed first failure times in the five groups are $Y_{1}=120, Y_{2}=200, Y_{3}=185, Y_{4}=55, Y_{5}=265$. Assuming that the life times follow Weibull distribution with shape parameter 2 and a lower

## KANTAM \& RAVIKUMAR

specification of $L=100$ they have calculated $V=\sum_{i=1}^{5} Y_{i}^{2}=161875$ and the acceptability constant $c L^{2}=196000$ since $V<c L^{2}$ they decided the submitted lot to be rejected.

Adopting the same information to Burr type X distribution we take the shape parameter of Burr type X namely $k=1$. Then it becomes the Rayleigh distribution which is also a Weibull with shape parameter 2 . For the sake of comparison with the sampling plan of Jun et al. (2006), at the above $p_{0}, p_{1}, \alpha, \beta$, $n=10$, we get from Table 14 as $m=2$, and acceptability constant $c=1.586911$ then $c L=158.6911$. $Z=$ the maximum of $55,120=120$. Since $Z<c L$. i.e., $120<158.691$, the lot is to be rejected.

From this example, the approach reached the decision of rejecting the lot by conducting limited failure censored life test for only two groups of 10 items each, whereas that of Jun et al. (2006) required the experiment to be conducted for 5 groups of 10 items each resulting in higher cost of experimentation and larger number of destructions. In that way, the Method-II is preferable to the Method-I proposed by Jun et al. (2006). Moreover, it may be recalled that $V, Z$ are defined as

$$
\begin{gathered}
V=\sum_{i=1}^{m} Y_{i} \\
Z=\operatorname{Max}\left(Y_{1}, Y_{2}, \ldots, Y_{m}\right) .
\end{gathered}
$$

If $c$ is the acceptability constant and $L$ is the lower specification, $Z>c L \Rightarrow V>c L$. That is acceptance by Method-II implies acceptance by Method-I, so that as far as acceptance decision is considered Method-II gives a stronger conclusion implying the same decision by Method-I.

## References

Aslam, M., \& Kantam, R. R. L. (2008). Economic reliability acceptance sampling based on truncated life tests in the Birnbaum-Saunders distribution. Pakistan Journal of Statistics, 24(4), 269-276.

Baklizi, A. (2003). Acceptance sampling based on truncated Life tests in the Pareto distribution of the second kind. Advances and Applications in Statistics, 3(1), 33-48.

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Baklizi, A., \& El Masri, A. E. Q. (2004). Acceptance sampling based on truncated life tests in the Birnbaum Saunders model. Risk Analysis, 24(6), 1453-1457. doi: 10.1111/j.0272-4332.2004.00541.x

Balakrishnan, N., Leiva, V., \& López, J. (2007). Acceptance sampling plans from truncated life tests based on the generalised Birnbaum-Saunders distribution. Communications in Statistics - Simulation and Computation, 36(3), 643-656. doi: 10.1080/03610910701207819

Balasooriya, U. (1995). Failure-censored reliability sampling plans for the exponential distribution. Journal of Statistical Computation and Simulation, 52(4), 337-349. doi: 10.1080/00949659508811684

Burr, I. W. (1942). Cumulative frequency functions. The Annals of Mathematical Statistics, 13(2), 215-232. doi: 10.1214/aoms/1177731607

Epstein, B. (1954). Truncated life tests in the exponential case. Annals of Mathematical Statistics, 25(3), 555-564. doi: 10.1214/aoms/1177728723

Fertig, K. W., \& Mann, N. R. (1980). Life-test sampling plans for twoparameter Weibull populations. Technometrics, 22(2), 165-177.
doi: 10.1080/00401706.1980.10486131
Goode, H. P., \& Kao, J. H. K. (1961). Sampling plans based on the Weibull distribution. Proceedings of Seventh National Symposium on Reliability and Quality Control (pp. 24-40), Philadelphia, PA: Institute of Radio Engineers.

Gupta, S. S. (1962). Life test sampling plans for normal and lognormal distribution. Technometrics, 4(2), 151-175.
doi: 10.1080/00401706.1962.10490002
Gupta, S. S., \& Groll, P. A. (1961). Gamma distribution in acceptance sampling based on life tests. Journal of the American Statistical Association, 56(296), 942-970. doi: 10.1080/01621459.1961.10482137

Johnson, L. G. (1964). Theory and technique of variation research. Amsterdam: Elsevier.

Jun, C.-H., Balamurali, S., \& Lee, S.-H. (2006). Variables sampling plans for Weibull distributed lifetimes under sudden death testing. IEEE Transactions Reliability, 55(1), 53-58. doi: 10.1109/TR.2005.863802

Kantam, R. R. L., Rosaiah, K., \& Srinivasa Rao, G. (2001). Acceptance sampling based on life tests: Log-logistic model. Journal of Applied Statistics, 28(1), 121-128. doi: 10.1080/02664760120011644

## KANTAM \& RAVIKUMAR

Kantam, R. R. L., \& Srinivasa Rao, G. (2004). A note on savings in experimental time under type II censoring. Economic Quality Control, 19(1), 91-95. doi: 10.1515/EQC.2004.91

Kantam, R. R. L., \& Sriram, B. (2013). Economic reliability test plans based on Rayleigh distribution. Journal of Statistics, 20, 88-101.

Kantam, R. R. L., Sriram, B., \& Suhasini, A. (2012). Reliability test plans: Exponentiated exponential distribution. Mathematical Sciences International Research Journal, 1(3), 1022-1032.

Kantam, R. R. L., Sriram, B., \& Suhasini, A. (2013). Reliability test plans based on log-logistic distribution. International Journal of Research in Commerce, IT \& Management, 3(9), 49-57.

Lio, Y. L., Tsai, T.-R. \& Wu, S.-J. (2009). Acceptance sampling plans from truncated life tests based on the Birnbaum - Saunders distribution for percentiles. Communications in Statistics - Simulation and Computation, 39(1), 119-136. doi: 10.1080/03610910903350508

Lio, Y. L., Tsai, T.-R. \& Wu, S.-J. (2010). Acceptance sampling plans from truncated life tests based on the Burr type XII percentiles. Journal of the Chinese Institute of Industrial Engineers, 27(4), 270-280. doi: 10.1080/10170661003791029

Lu, W. (2011). Acceptance sampling plans based on truncated life tests for Maxwell distribution. Pakistan Journal of Statistics, 27(2), 159-170.

Pascual, F. G., \& Meeker, W. Q. (1998). The modified sudden death test: Planning life tests with a limited number of test positions. Journal of Testing and Evaluation, 26(5), 434-443. doi: 10.1520/jte12692j

Rosaiah, K., \& Kantam, R. R. L. (2005). Acceptance sampling based on the inverse Rayleigh distribution. Economic Quality Control, 20(2), 277-286.

Rosaiah, K., Kantam, R. R. L., Rama Krishna, V., \& Siva Kumar, D. C. U. (2014). An economic reliability test plan for type - I generalized half logistic distribution. Journal of Chemical, Biological and Physical Sciences, 4(2), 14861493.

Rosaiah, K., Kantam, R. R. L., \& Srinivasa Rao, B. (2009). Reliability test plan for half logistic distribution. Calcutta Statistical Association Bulletin, 61(241/244), 183-196.

Srinivasa Rao, G., Ghitany, M. E., \& Kantam, R. R. L. (2009). MarshallOlkin extended Lomax distribution: An economic reliability test plan. International Journal of Applied Mathematics, 22(1), 139-148.

## LFCLTSP IN BURR TYPE X DISTRIBUTION

Srinivasa Rao, G., \& Kantam, R. R. L. (2010). Acceptance sampling plans from truncated life tests based on the log-logistic distributions for percentiles. Economic Quality Control, 25(2), 153-167. doi: 10.1515/eqc.2010.008

Srinivasa Rao, G., \& Kantam, R. R. L. (2012). Two-stage estimation in loglogistic model. Pakistan Journal of Statistics, 28(3), 331-339.

Srinivasa Rao, B., \& Kantam, R. R. L. (2013). Acceptance sampling plans for percentiles of half logistic distribution. International Journal of Reliability, Quality and Safety Engineering, 20(5), 1-13. doi: 10.1142/S0218539313500162

Srinivasa Rao, G., Kantam, R. R. L., Rosaiah, K., \& Pratapa Reddy, J. (2012). Acceptance sampling plans for percentiles based on the inverse Rayleigh distribution. Electronic Journal of Applied Statistical Analysis, 5(2), 164-177. doi: 10.1285/i20705948v5n2p164

Subba Rao, R., Naga Durgamamba, A., \& Kantam, R. R. L. (2014). Acceptance sampling plans: Size biased Lomax model. Universal Journal of Applied Mathematics, 2(4), 176-183. doi: 10.13189/ujam.2014.020403

Subba Rao, R., Prasad, G., \& Kantam, R. R. L. (2013). Acceptance sampling on life tests: Exponentiated Pareto model. Engineering Sciences International Research Journal, 1 (1), 152-158.

Wu, J.-W., \& Tsai, W.-L. (2000). Failure-censored sampling plan for the Weibull distribution. Information and Management Sciences, 11(2), 13-25.

Wu, J.-W., Tsai, T.-R., \& Ouyang, L.-Y. (2001). Limited failure-censored life test for the Weibull distribution. IEEE Transactions on Reliability, 50(1), 107111.doi:10.1109/24.935024

# On Generalizing Cumulative Ordered Regression Models 

Robert W. Walker<br>Willamette University<br>Salem, OR


#### Abstract

We examine models that relax proportionality in cumulative ordered regression models. Something fundamental arising from ordered variables and stochastic ordering implies a partitioning. Efforts to relax proportionality also relax the ability to collapse an inherently multidimensional problem to a partitioning of the (unidimensional) real line. It is surprising and unfortunate to find that deviations from proportionality are sufficient to generate internal contradictions; undecidable propositions must exist by relaxing proportional odds without other relevant and significant changes in the underlying model. We prove a single theorem linking continuous support and partitions of a latent space to show that for these two characteristics to be simultaneously satisfied, the model must be the proportional-odds model. Conditioning on the adjacency that is closely related to the partitioning is fruitful, but at this point we join the class of continuation-ratio models. Alternatively, Anderson's (1984) stereotype model is quite general and nests ordered and unordered choice models, but again we have left the domain of cumulative models. Adopting multidimensional cumulative models or imposing covariate-specific thresholds are the only certain methods for avoiding these troubles in the cumulative framework. It is generically impossible to generalize the cumulative class of ordered regression models in ways consistent with the spirit of generalized cumulative regression models. Monte Carlo studies also demonstrate the general principles.


Keywords: Proportional odds models, partial proportional odds models, Monte Carlo simulation

## Introduction

Generalizations of common cumulative models for ordered phenomena are considered. The parallelism inherent to cumulative models such as the (proportional odds) ordered logit/probit model (McKelvey \& Zavoina, 1975; McCullagh, 1980) is seen as limiting and workers in numerous statistical literatures have worked on generalizations. ${ }^{1}$ According to Google Scholar, 407 papers cited

[^26]Peterson and Harrell's (1990) presentation of a partial proportional-odds model and 478 cited the work by Brant (1990) on testing proportional odds as of March 6, 2015; the generalization appears in McCullagh and Nelder (1989) and Agresti (2002). Test statistics have been proposed for testing this specification as null hypothesis against a more general specification (Brant, 1990) and software for Stata (Williams, 2006) and Yee's (2010) library for R (R Development Core Team, 2009) allow these models to be estimated.

Following Peterson and Harrell (1990), Cox (1995), and numerous others in medical statistics, social science has used these models for ordered scales related to social policy and racial attitudes (Branton \& Jones, 2005); Fullerton (2009) presents a sociological analysis of income quartiles; Gannon (2009) examines self-reported disability status using this generalization. A spate of articles in the Journal of Modern Applied Statistical Methods (e.g., O’Connell \& Liu, 2011; Liu \& Koirala, 2012) developed diagnostics for the model and deployed the model for educational outcomes. The model received some attention in the field of health economics (Lindeboom \& van Doorslaer, 2004). Generalized threshold models (Maddala, 1983; Terza, 1985) are similar and the general argument applies to the class of location-scale models. To our surprise, there is no obvious way to generalize the model while retaining two basic assumptions that motivate cumulative models. It is cumulative in both probabilistic foundations and in name.

Given the widespread attention to generalizations of the cumulative model, it is surprising to find that the model only sensibly exists in the presence of proportional odds or the parallel equivalent. Partitioning a unidimensional latent space uniquely with functions of covariates is constrained by the requirement that everything match at the boundaries of any two adjacent partitions with cumulative probabilities. Defining a model and finding conditions where optima exist can be quite different from a model with sensible statistical microfoundations; these two ideas diverge when generalizing the cumulative ordered regression model. When models are employed for describing and estimating physical, social, or biological processes, internal contradictions pose significant difficulties because it is not clear how we return the parameters to their substantive context in a way that is consistent with the assumptions that facilitate estimation.

## The Argument

Anderson (1984) distinguished ordered variables that are grouped continuous ordered groupings of an unobserved continuous outcome - from assessments judgments or grades somehow combining (possibly) multiple inputs. Ordered

ROBERT W. WALKER

responses can represent a coarsened latent variable, as in income quartiles, Likert scales (in many cases), feeling thermometers, and grades given to things ranging from diamonds to dairy products. Although the continuous variable cannot be observed, the groupings - partitions - may be observed as an ordered outcome. The ordered outcome can be inverted to a partition of the latent scale and differences on this scale are often of substantive interest; many seemingly ordered phenomena do not obviously present for finer measurement but the latent measure remains substantively interesting. Nevertheless, an ordered variable can be an assessed or judged outcome.

A variable of the second type is generated by an assessor who possesses an indeterminate amount of information before producing his judgment of the grade of the ordered variable. For example, Anderson and Philips (1981) refer to the "extent of pain relief after treatment": worse, same, slight improvement, marked improvement or complete relief. In principle, there is a single, unobservable, continuous variable related to this ordered scale, but in practice, the doctor making the assessment will use several pieces of information in making his judgment on the observed category. For example, he might use severity of pain, kind of pain, consistency in time and degree of disability. We will refer to these variables of the second type as "assessed" ordered categorical variables and argue that, in general, a different approach to modelling regression relationships is appropriate for the two types. (Anderson, 1984, p. 2)

Anderson's argument suggested that the presence of multiple inputs requires a model that need not assume an underlying order but instead allows order to emerge (or not) as a special case of a more general model. The arguments underlying the stereotype model of Anderson (1984) are precisely focused on dimensionality (the number of latent dimensions), ordering (and whether or not it obtains), and distinguishability (do covariates distinguish categories?) with a model that can assess each in a null hypothetical framework. The model derives from category probabilities rather than a cumulative scale. Though, the outcome variable itself, $y$, can also be represented by sets of ratios of cumulative probabilities with some assumed distribution, the statistical principles that are engaged require a sensible probability model and the cumulative framework becomes quite limiting. At some point, the cumulative approach requires a well-defined cumulative distribution; this is deeply constraining and leads to an internal contradiction in "generalizations" of cumulative ordered regression models.

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

Consider a $J>2$ category ordinal variable $y_{i}$ observed on units $i \in N$. The canonical distribution for such outcomes is the multinomial. Frequently, analysts employ the notion of a latent variable, $y_{i}^{*}$, crossing thresholds defined by the $J+1$ vector $\boldsymbol{\tau}$, with $\tau_{0}=-\infty$ and $\tau_{J}=\infty$. A key component to cumulative models is a partition linking an observation rule and a latent unidimensional continuous variable. We define such a rule as Assumption 1.

Assumption 1: Mutually exclusive and exhaustive partition:

$$
\begin{equation*}
y_{i}=j i f \text { if and only if } \tau_{j}>y_{i}^{*} \geq \tau_{j-1} \tag{1}
\end{equation*}
$$

for

$$
j=1,2, \ldots, J, \quad i \in N, \quad \sum_{j \in J} \operatorname{pr}(y=j)=1
$$

Equivalently, $\tau$ could be viewed as a function so that $\tau: y^{*} \rightarrow y$. It is many-to-one, but it is special because it is an ordered partition. The ordering can be inverted to imply a unique set of inequality relations that must apply to $y^{*}$. Though we have yet to define $y^{*}$, we will place some structure on randomness.

Assumption 2: $\quad \epsilon_{i}$ are independent and identically distributed with probability density function f and cumulative distribution F such that $\operatorname{supp}(\epsilon)=\mathbb{R}$ and $\mu(\epsilon)=1$ (measure one).

The errors will ultimately give a distribution to the random variable of interest; the random variable will inherit the distribution of $\epsilon$ conditional on a true model consisting of covariates. Of a driving force in the statistical logic, the latent variable, as a function of covariates $X$, requires structure. This leads to Definition 1.

## Definition 1: Proportional Odds: Linearity in latent variables.

$$
\begin{equation*}
y_{i}^{*}=\mathbf{X}_{i} \boldsymbol{\beta}+\epsilon_{i} \tag{2}
\end{equation*}
$$

where $\mathbf{X}_{i}$ is a (row) vector of centered covariates for $i, \boldsymbol{\beta}$ is a (column) vector of parameters of interest with $\left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)-1$ existing. ${ }^{2}$ This is the source of our notion of "parallelism". The latent variable is a linear function of covariates and parameters

## ROBERT W. WALKER

and this yields a set of parallel planes. When the covariates are bounded, the randomness in the latent variable is then inherited from $\epsilon .^{3}$ We will assume independence between the covariates and random errors.

The argument will apply equally to commonly used F with continuous support (normal, logistic, cloglog (Gumbel), Cauchy) on $\mathbb{R}$. Substituting (2) into (1) yields (for all $j \in J$ ), for a general cumulative distribution function F determined by the assumed distribution of $\epsilon_{i}$,

$$
\begin{equation*}
\mathcal{L}_{\mathrm{F}}=\sum_{i=1}^{N} \sum_{j=1}^{J} \operatorname{Ind}\left[y_{i}=j\right] \times \ln \left[\mathrm{F}\left(\tau_{j}-y_{j}^{*}\right)-\mathrm{F}\left(\tau_{j-1}-y_{i}^{*}\right)\right] \tag{3}
\end{equation*}
$$

One appealing feature of this proportional odds model is the ease of use. It is intuitively pleasing to link ordered categories to some underlying continuum that is determined by covariates and to allow the marginal effects to be well defined across all outcomes. In effect, we have a linear regression for ordered outcomes that does not impose an (likely fallacious) interval-scale interpretation. It is also intuitively restrictive.

The "partial-proportional odds" model (Brant, 1990; McCullagh \& Nelder, 1989; Peterson \& Harrell, 1990) employs subscripted $\boldsymbol{\beta}$ with the implicit idea that some (or all) regressors have varying impacts that depend on the comparison, as in Definition 2.

## Definition 2: Partial/Non Proportional Odds: Conditionally (on j) linear in

 the latent variable.$$
\begin{equation*}
y_{i}^{*}=\mathbf{X}_{i} \boldsymbol{\beta}_{j}+\epsilon_{i} \tag{4}
\end{equation*}
$$

The proportional-odds assumption is that $\mathrm{F}=\Lambda$ or that $\epsilon_{i}$ are i.i.d. logistic and that $\boldsymbol{\beta}_{1}=\boldsymbol{\beta}_{2}=\ldots=\boldsymbol{\beta}_{J-1}$; this model can be estimated by maximizing (3) with (2) under general conditions owing to properties illuminated by Pratt (1981) or by employing generalized linear models with conditional mean function as implied by (2). Insert (4) into (3) to yield

$$
\begin{equation*}
\mathcal{L}_{\mathrm{F}}=\sum_{i=1}^{N} \sum_{j=1}^{J} \operatorname{Ind}\left[y_{i}=j\right] \times \ln \left[\mathrm{F}\left(\tau_{j}-\mathbf{X}_{i} \boldsymbol{\beta}_{j}\right)-\mathrm{F}\left(\tau_{j-1}-\mathbf{X}_{i} \boldsymbol{\beta}_{j-1}\right)\right] \tag{5}
\end{equation*}
$$

It is straightforward to identify the parameters as deviations $\left(\beta_{j}=\beta_{1}+\delta_{j}\right)$ from a base parameter with a simple view toward whether or not proportional odds obtains in Wald statistics, score tests, or approximating the likelihood ratio. ${ }^{4}$ At first glance, (5) is a very useful generalization because the underlying linear structure of the proportional odds type model seems excessively limiting even with creative functional forms for the covariates. Generalizing the model maintains the significant intuitive appeal of the cumulative model for parameters that it is sensible to believe map linearly onto the latent scale with the flexibility of altering relationships in a way that uniquely leverages the adjacency of ordered data. There is obvious gain to the exercise that is quite appealing by retaining the simplicity of unidimensionality; at what cost? The unidimensional cumulative foundation, if the model is an appropriate partition of the latent space, requires that this hold for each $i \in N$.

In the literature on partial proportional odds models, much has been made of conditions for sensible estimates. Estimating the model is distinct from requiring predicted responses to be nonnegative. Conditions must hold on $\boldsymbol{\beta}_{j}$ and $\boldsymbol{\tau}$ for estimates to exist (the parameters and thresholds are jointly bounded) and these conditions are weaker than those required for nonnegative category probabilities. ${ }^{5}$ The latter is the usual criterion for assessing the model. Unfortunately, the set of models we can estimate is itself a proper subset of models that contradict their own probabilistic foundations. Put simply, models may be estimated with nonnegative probabilities for each ordered category that have no well-behaved latent variable satisfying Assumptions 1 and 2 with Definition 2; this is the central demonstration of Theorem 1. Research has remained focused on testing (Brant, 1990), estimating (Peterson \& Harrell, 1990), and generalizing (Maddala, 1983; Terza, 1985; Cox, 1995; Williams, 2006) ordered regression models using Definition 2 for which no such generalization exists.

Theorem 1: $\quad$ Assumption 1 and Assumption 2 avoid internal contradiction if and only if observational equivalence holds between Definition 1 and Definition 2 ((2) and (4)).

## Proof:

1. Suppose that Assumption 2 holds and (2) $\neq$ (4). (4) generically requires $\exists j$ : $\beta_{j} \neq \beta_{j-1}$ and, perhaps more importantly that $y_{j}^{*} \neq y_{j-1}^{*}$. Assumption 2, recalling parameters to scale, allows us to write,

## ROBERT W. WALKER

$\beta_{j}=\beta_{j} / \sigma$ which then scales $\mathrm{y}^{*}=\mathrm{y}^{*} / \sigma$. With the assumption that $\mathbf{X}$ is centered, there must be an $x$ to which one of the following two conditions apply because no column of $x$ can fail to have support on both sides of the center without contradicting full rank.

Case (a): $\quad$ Suppose $\beta_{j}>\beta_{j-1}$ in (4). $\sigma$ is fixed under Assumption $2 .{ }^{6}$ Consider $y_{j}^{*}=\tau_{j}$ which is justified by the support of $\epsilon$ on $\mathbb{R}$. Under the supposition, $y_{j}^{*}=\tau_{j}>y_{j-1}^{*}$ (because $\beta_{j}>\beta_{j-1} \Rightarrow \beta_{j} x>\beta_{j-1} x$ ) in (4). $y$ is undefined; invoking Assumption 1 yields $y_{j}^{*}=j+1$ while $y_{j-1}^{*}=j$.

Case (b): $\quad$ Suppose $\beta_{j-1}>\beta_{j}$. Consider $y_{j-1}^{*}=\tau_{j}$ which is justified by the support of $\epsilon$ on $\mathbb{R}$. Under the supposition, $y_{j-1}^{*}=\tau_{j}>y_{j}^{*}$ (because $\beta_{j-1}>\beta_{j} \Rightarrow \beta_{j-1} x>\beta_{j} x$ ) in (4). $y$ is undefined; invoking Assumption 1 yields $y_{j-1}^{*}=j+1$ while $y_{j}^{*}=j$.
2. Suppose that Assumption 1 holds and $(2) \neq(4)$. Assumption 1 allows us to write the probability that $y=j$ sums to one, the logic will follow the above. That all of the observations sum to one will contradict continuous support. To show this, generically write

$$
\begin{align*}
& \sum_{j \in J} \operatorname{Pr}\left(y_{i}=j\right)=\int_{-\infty-y_{j}^{*}}^{\tau_{1}-y_{i}^{*}} \mathrm{f}(\epsilon) d \epsilon+\int_{\tau_{1}-y_{i}^{*}}^{\tau_{2}-y_{i}^{*}} \mathrm{f}(\epsilon) d \epsilon+\ldots+\int_{\tau_{J-1}-y_{i}^{*}}^{\infty-y_{i}^{*}} \mathrm{f}(\epsilon) d \epsilon  \tag{7}\\
& \int_{-\infty}^{\infty} \mathrm{f}(\epsilon) d \epsilon=1 \tag{8}
\end{align*}
$$

Under the proportional odds model, all is fine and Assumption 2 is satisfied. We have integrated the real line satisfying the restrictions on $f$ and $F$. Now let us examine (4).

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

$$
\sum_{j \in J} \operatorname{Pr}\left(y_{i}=j\right)=\int_{\tau_{0}-y_{i 0}^{*}}^{\tau_{1}-y_{i 1}^{*}} \mathrm{f}(\epsilon) d \epsilon+\int_{\tau_{1}-y_{i 2}^{*}}^{\tau_{2}-y_{i 2}^{*}} \mathrm{f}(\epsilon) d \epsilon+\ldots+\int_{\tau_{J-1}-y_{i, J-1}^{*}}^{\tau_{J}-y_{i, J-1}^{*}} \mathrm{f}(\epsilon) d \epsilon
$$

The boundaries of the integrals were moved by assuming (4) but leaving $\tau$ fixed. To view this more cleanly, expand the integral about the fact that $y_{j}^{*} \neq y_{j-1}$. The two cases from before will appear parenthetically.

$$
\begin{align*}
& \sum_{j \in J} \operatorname{Pr}\left(y_{i}=j\right)=\int_{-\infty-y_{i 1}^{*}}^{\tau_{1}-y_{i 1}^{*}} \mathrm{f}(\epsilon) d \epsilon \overbrace{\left(+\int_{\tau_{1}-y_{i 1}^{*}}^{\tau_{2}-y_{i 2}^{*}} \mathrm{f}(\epsilon) d \epsilon\right)}^{y_{i \sim}^{*}>y_{i 1}^{*}} \overbrace{\left[-\int_{\tau_{1}-y_{i 2}}^{\tau_{1}-y_{12}^{*}} \mathrm{f}(\epsilon) d \epsilon\right]}^{y_{i 1}^{*}<y_{i 1}^{*}} \\
& +\int_{\tau_{1}-y_{i 2}^{*}}^{\tau_{2}-y_{i 2}^{*}} \mathrm{f}(\epsilon) d \epsilon \underbrace{\left(+\int_{\tau_{2}-y_{i 2}^{*}}^{\tau_{2}-y_{i 3}^{*}} \mathrm{f}(\epsilon) d \epsilon\right)}_{y_{i 3}^{*}>y_{i 2}^{*}} \underbrace{\left[-\int_{\tau_{2}-y_{i 3}^{*}}^{\tau_{2}-y_{i 2}^{*}} \mathrm{f}(\epsilon) d \epsilon\right]}_{y_{i 3}^{*} y_{y_{i 2}^{*}}^{*}}  \tag{9}\\
& +\ldots+\int_{\tau_{J-1}-y_{i, J-1}^{*}}^{\infty-y_{i,-1}^{*}} \mathrm{f}(\epsilon) d \epsilon \\
& \neq 1
\end{align*}
$$

Generically, the parenthetical (or bracketed) elements will be non-zero unless $(2)=(4)$. Moreover, these regions, given support on $\mathbb{R}$, are not countable and the probability that the two or multiple terms offset is a set of measure zero. Up to this set of measure zero, $\sum_{j \in J} \operatorname{Pr}\left(y_{j}=j\right) \neq 1$ contradicting Assumption 2.

## Discussion

The proportional-odds/parallel model is the $J-1$ dimensional solution that uniquely collapses to a marginal distribution. Efforts to make the model more realistic, such as the structure defined in Definition 2, ultimately make it less realistic in the sense that its properties cannot be studied under its assumptions. The reason is that the assumptions are internally contradictory when combined with Definition 2 . The models become internally contradictory of their own probability formulations when they deviate from the proportional odds model. The underlying latent variable is a strict order under the proportional-odds assumption and deviations can violate this ordering. These deviations from this underlying ordering wreak havoc on the probabilistic foundations.

## ROBERT W. WALKER

We will replicate the Monte Carlo simulation evidence in Peterson and Harrell (1990, Table 6, Design 4) to provide a context. When the odds-ratio formulation common is considered to the ordered logistic regression model, write

$$
\begin{equation*}
\frac{\operatorname{Pr}\left(y_{i} \leq j \mid \mathbf{X}_{i}\right)}{\operatorname{Pr}\left(y_{i}>j \mid \mathbf{X}_{i}\right)}=\exp \left(-\mathbf{X}_{i} \boldsymbol{\beta}_{j}\right), \quad j=1,2, \ldots, J-1 \tag{10}
\end{equation*}
$$

Consider the experiment reported as Design 4 in Peterson and Harrell (1990, p. 216) that defines a four-category $y . \mathbf{X}$ is a set of five completely crossed binary predictors ( $2^{5}$ ) of ten observations each $(N=320)$. Peterson and Harrell (1990) set $\boldsymbol{\beta}=0.5$ with the exception of $\boldsymbol{\beta}_{25}=1 . \boldsymbol{\alpha}$ are constants (or inverse cutpoints such that $\tau$ in (1)) are set to $\boldsymbol{\alpha}=\{0.405,-0.847,-2.2\}$. The key to their strategy is in independent multinomial sampling.

Begin with ratios of categories specified along some cumulative scale but curiously no appeal to a random variable. $y$ will ultimately result from creating cumulative probabilities and comparing them with model estimates. Peterson and Harrell (1990, p. 208) define

$$
\begin{equation*}
C_{i j}=\operatorname{Pr}\left(y \geq j \mid \mathbf{X}_{i}\right)=\frac{1}{1+\exp \left(-\alpha_{j}-\mathbf{X}_{i} \boldsymbol{\beta}_{j}\right)} \tag{11}
\end{equation*}
$$

Because this defines the cumulative distribution function of a logistically distributed random variable, work backward to examine the distribution(s?) of this logistic random variable. As in Peterson and Harrell (1990, Design 4), suppose $x_{1}=x_{2}=x_{3}=x_{4}=0, x_{5}=1$; this implies $\mathbf{X}_{i} \boldsymbol{\beta}=\{0.5,1,0.5\}$. Figure 1 illustrates a part of the difficulty.

The top panel of Figure 1 provides the cumulative probability plots obtained from all thirty-two possible combinations of our five binary predictors arranged along the $x$-axis ordered lexicographically first by $\sum_{i} x_{i}$ and then by $i$. The $y$-axis presents the cumulative probabilities (3-purple, 2-orange, 1 -blue, 0 -red) according the partial proportional-odds model. As the $x$-axis increases, the probability of higher categories increases. The unfilled circles represent predictions from the first and third equations (which happen to be equal) and the filled circles represent predictions from the second equation. The unfilled circles define the cumulative probabilities for the lowest (blue) and highest (purple) categories while the (orange) filled circles define outcomes in the interior categories. As expected from the

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

parameters given before, recalling that these parameters have log odds-ratio interpretations, the highest category becomes quite common.


Figure 1. Peterson and Harrell (1990): Table 6, Design 4: The top panel plots cumulative probabilities derived from the partial proportional-odds cumulative logit color coded as in the legend; Filled circles represent probabilities derived from $j=2$; Open circles represent probabilities derived from $j=\{1,3\}$; The solid lines capture the cumulative probabilities as they enter the "partial proportional-odds" likelihood; The bottom panel displays logistic densities for $\mathbf{X} \boldsymbol{\beta}_{j}=0.5$ above zero and for $\mathbf{X} \boldsymbol{\beta}_{j}=1$ below zero; The non-hatched areas represent areas such that partitioning fails

That the probability of category two is shrinking is a product of the nonproportional-odds and the oft-noted issue of negative probabilities is a
necessary consequence of this shrinkage. That they do not cross in the observed data is taken as a signal of the underlying validity of the model (and estimates) when the very fact that such lines can cross with a cumulative probability model contradict the foundations of cumulative probabilities.

The scale has been assumed fixed by implication, $y \sim \lambda\left(\mathbf{X}_{i} \boldsymbol{\beta}_{j}, \sigma=1\right)$, where $\lambda$ is a logistic distribution characterized by location $\mathbf{X}_{i} \boldsymbol{\beta}_{j}$ and scale equal to one. ${ }^{7}$ The density depends on the outcome because $j$ enters the conditional expectation. The way to resolve this is to set $(2)=(4)$. If it is assumed that the cutpoints between categories for this logistic random variable to be constants and that the model is true, then a logistically distributed random variable with continuous support on $\mathbb{R}$ arises and Theorem 1 applies. This only works when (2) = (4).

Peterson and Harrell (1990) instead use the sequence of logit cumulative odds. Define a multinomial random variable using $C_{i j}$ as a partition of the unit interval $(0,1)$ for input probabilities. This is the equivalent of invoking Assumption 1. The problem is that drawing cumulative probabilities in a uniform fashion over the unit interval and inverting them to the logit-scale, given that the logit is a one-to-one transformation, implies continuous support on $\mathbb{R}$. Under Theorem 1, this cannot be valid unless it is done under the proportional variant. On a superficial level, the approach resolves an inconsistency such that simulation succeeds with probability one.

In the process, avoid defining a random variable excepting $y$ and take a cumulative probability over an undefined logistic random variable. Invoke a logistic random variable to estimate $C_{i j}$ alongside $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{j \text {. }}$. Order only enters to the extent that the multinomial distribution is drawn as a partition of the logistic distribution. But here is where the problem emerges. A uniform random variable, call it $u$, gives the hypothetical cumulative probability. Taking $C_{i j}$ as given can generate y according to which interval $u$ happens to fall into for each $i$. Theorem 1 dictates a generic problem with this strategy; either the logistic distribution does not have continuous support or it does not generically integrate to one. In either case, sidestepping the specification of the random variable also allows us to sidestep the uncomfortable realization that the random variable we invoke does not and cannot have the properties that we have assumed. This is illustrated in the bottom panel of Figure 1.

The bottom panel of Figure 1 displays an example of the implied logistic densities from Peterson and Harrell (1990), Design 4. The hatched areas represent portions of the density that satisfy partitioning while the blank areas showcase the area of partitioning failure. How does partitioning fail? In general, if one equation
produces some outcome $\tilde{y}$ and the other equation produces $\hat{y}$ and $\tilde{y} \neq \hat{y}$, we can say the partition failed because the resulting value of $y$ is not unique (or does not exist).

The problem is very similar to the issues of completeness and coherence in the econometric study of simultaneous equations with limited outcomes (Heckman, 1978; Gourieroux, Laffont, \& Monfort, 1980; Dagenais, 1999; Tamer, 2003; Lewbel, 2007). ${ }^{8}$ Depending on the sign of differences in adjacent regression coefficients, the offending regions are characterized in (9). This is of consequence because the size of the regions in which the model is internally contradictory is increasing in the differences (the size of deviations from parallelism) and the boundaries depend on $\mathbf{X}_{i}$. Deviations from proportional odds (or an equivalent parallelism of planes in $\mathbb{R}^{p}$ where $p$ is the column rank of $\mathbf{X}$ ) are sufficient to break the most basic of assumptions about sources of randomness and notions of order, both quite sensible. Larger deviations from parallelism increase the measure of contradictions.

The generic fact that $y^{*}$ becomes multidimensional under deviations from parallelism or proportionality causes the problem. Under deviations from proportional odds with a well behaved cumulative distribution, no such constraint exists that is not a jointly identified function of thresholds for each unique $x$ and the parameters of interest $\boldsymbol{\beta}_{j \text {. Thought of as a function, } \boldsymbol{\tau} \text { must match at the boundaries }}$ of the cumulative distribution for $\mathbf{y}^{*}(\operatorname{and} \boldsymbol{\epsilon})$ to have continuous support. For this to work, $\tau_{j}$ cannot be invariant to $i$ unless $\mathbf{X}$ is also invariant to $i$. Of course, if $\mathbf{X}$ is globally invariant to $i, \mathbf{X}$ is a constant. When $\mathbf{X}$ is a variable, the trouble reemerges. Before presenting a Monte Carlo simulation, two related issues are mentioned.

There are a variety of ways to test parallelism of the regression slopes. The previous demonstration begs the question of what alternatives such tests embody. If the model does not exist except at the null hypothesis, a rejection of the null seems entirely uninformative because it offers no insights into the nature of the problem. These tests do not obviously lead to some more general class of models in which parallelism is a restriction. With this in mind, turn to an analysis of solutions to the more general problem in models that are not internally contradictory.

In the nonrecursive simultaneous equations setting, Dagenais (1999) restricts the support of $\boldsymbol{\epsilon}$. Although this is technically correct and logically sound, it seems hard to intuitively justify for most substantive applications and impossible to verify in practice. Another obvious solution emerges from the ideas of Maddala (1983) and Terza (1985). The generalized threshold model parameterizes the thresholds

ROBERT W. WALKER

$\left(\tau_{j}\right)$, instead of allowing regression coefficients to vary on the basis of the comparisons $\left(\beta_{j} \neq \beta_{j-1}\right)$. To be precise, write

$$
\begin{equation*}
\tau_{i j}=\mathbf{X}_{i} \boldsymbol{\psi}_{j} \tag{12}
\end{equation*}
$$

and expand the vector $\mathbf{X}_{i}$ to include a constant. As it happens, the model is an isomorphism to the previous case (4) and all the same results apply. Suppose instead, construct a model based on the varying thresholds where the variation in the thresholds is specific to the row rank of $\mathbf{X}$. In other words, maintain the aforementioned parallelism, but allow the thresholds to be specific to observed covariates. Of course, with continuous covariates, this is not at all helpful, but with discrete regressors and large samples, such a model can be estimated and all of the relevant thresholds can also be estimated so long as each J is observed for each unique row of $\mathbf{X}$. When there are no such observations, the problems of Chamberlain's (1980) fixed effects estimator when outcomes do not vary arises.

A brief R (R Development Core Team, 2009) simulation example showcases the severity of the problem (Appendix A provides a logistic example). ${ }^{9}$ The simulation is constructed with a single binary regressor and a uniform regressor on $[-1,1]$. Set $\tau=\{-0.5,0.5\}, \beta_{11}=0.05, \beta_{12}=0.1, \beta_{2}=1$ for 1000 observations and perturbed the latent variable with standard logistic, normal, Gumbel (cloglog), and Cauchy errors before applying Assumption 1 to yield results. It is important to note that with 1000 observations and relatively small effect sizes (as these are), under the proportional odds logistic regression, roughly $6.25 \%$ of 10,000 iterations reject the hypothesis that $\beta=0$ when $\beta=0.10$. The effects are so small there is almost no power. Even under these minute deviations, answers fail to exist. Turning to the evidence reported in Figure 2, the number of failures in invoking partitioning is bounded below by zero and bounded above by just over $3 \%$. The graphic makes clear that a non-zero fraction of outcomes are undefined (in all 10000 Monte Carlo trials) as reported in Figure 2. Given Theorem 1, it comes as no surprise that all are susceptible; the model contains an internal contradiction unless it is the parallel version.


Figure 2: Undefined outcomes from partial-proportional odds: 10,000 Monte Carlo trials

## Conclusion

Modeling multidimensionality is a useful endeavor and it is not prohibitive. The difficulties are in conceptualizing the substance of such dimensions in applications and linking them together to obtain a stochastic order. It is important that our efforts remain true to the underlying probability structures that generate the data. Models that cannot be inverted cannot be studied in any meaningful way. To the extent that models are meant to capture the processes that generated them, generalizing
nonparallelism in a cumulative framework under Assumptions 1 and 2 is impossible and the probability of contradiction goes to one as sample sizes become infinite. The parallel version of the model is exceptionally useful for many problems, but generalizations of the model must carefully handle the restrictions imposed by their cumulative foundations.

Although the main demonstration is a negative one, hope is not lost. A wellstudied and widely known class of ordered regression models can accommodate non-parallel effects and retain some cumulative foundations. Form odds-ratios for the sequential (Fahrmeir \& Tutz, 1994) or continuation-ratio logit (Agresti, 2002) as

$$
\begin{equation*}
\frac{\operatorname{Pr}\left(y_{i}=j \mid \mathbf{X}_{i}, y_{i} \geq j\right)}{\operatorname{Pr}\left(y_{i}>j \mid \mathbf{X}_{i}, y_{i} \geq j\right)}=\exp \left(-\mathbf{X}_{i} \boldsymbol{\beta}_{j}\right), \quad j=1,2, \ldots, J-1 \tag{13}
\end{equation*}
$$

The solution has two parts. First, condition on the observed data and this resolves the incompleteness of the generalized cumulative regression model. Second, the models are mixtures of category and cumulative probabilities and, more importantly, are inherently multidimensional in the non-parallel case. However, such models conform to the more basic intuition that each unique linear function captured by $\boldsymbol{\beta}_{j}$ must yield a unique dimension. The generalized cumulative regression model is a multidimensional model collapsed to a single dimension. The collapsing works if and only if the model is the proportional odds model.

## Acknowledgements

The author thanks Alan Agresti, William H. Greene, Andrew Martin and Peter McCullagh, Christopher Winship, and participants in the Statistics Seminar of the Department of Mathematics at Washington University for commentary and suggestions. An equivalent and simpler geometric presentation can be constructed using equiprobability contours, as Peter McCullagh suggested, though the linkages to the coherence/completeness problem are less obvious (though they can also be represented geometrically). This approach seems more direct.

## References

Agresti, A. (2002). Categorical data analysis (2nd ed.). New York, NY: John Wiley and Sons.

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

Anderson, J. A. (1984). Regression and ordered categorical variables. Journal of the Royal Statistical Society. Series B (Methodological), 46(1), 1-30. Retrieved from http://www.jstor.org/stable/2345457

Anderson, J. A., \& Philips, P. R. (1981). Regression, discrimination, and measurement models for ordered categorical variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 30(1), 22-31. doi: 10.2307/2346654

Boes, S., \& Winkelmann, R. (2006). Ordered response models. Allgemeines Statistisches Archiv, 90(1), 167-181. doi: 10.1007/s10182-006-0228-y

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics, 46(4), 1171-1178. doi: 10.2307/2532457

Branton, R. P., \& Jones, B. S. (2005). Reexamining racial attitudes: The conditional relationship between diversity and socioeconomic environment. American Journal of Political Science, 49(2), 359-372. doi: 10.1111/j.00925853.2005.00128.x

Chamberlain, G. (1980). Analysis of covariance with qualitative data. The Review of Economic Studies, 47(1), 225-238. doi: 10.2307/2297110

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. Statistics in Medicine, 14(11), 1191-1203. doi: 10.1002/sim. 4780141105

Dagenais, M. (1999). A simultaneous probit model. Cahiers Économiques de Bruxelles, 163(3), 325-346.

Fahrmeir, L., \& Tutz, G. (1994). Multivariate statistical modelling based on generalized linear models. New York, NY: Springer.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. Sociological Methods \& Research, 38(2), 306-347. doi: 10.1177/0049124109346162

Gannon, B. (2009). The influence of economic incentives on reported disability status. Health Economics, 18(7), 743-759. doi: 10.1002/hec. 1399

Gourieroux, C., Laffont, J. J., \& Monfort, A. (1980). Coherency conditions in simultaneous linear equation models with endogenous switching regimes. Econometrica, 48(3), 675-696.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equations system. Econometrica, 46(4), 931-959.

ROBERT W. WALKER

Lewbel, A. (2007). Coherency and completeness of structural models containing a dummy endogenous variable. International Economic Review, 48(4), 1379-1392. doi: 10.1111/j.1468-2354.2007.00466.x

Lindeboom, M., \& van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health. Journal of Health Economics, 23(6), 1083-1099. doi: 10.1016/j.jhealeco.2004.01.002

Liu, X., \& Koirala, H. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. Journal of Modern Applied Statistical Methods, 11(1), 242-254. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol11/iss1/21/

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Beverly Hills, CA: Sage Publications.

Maddala, G. S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge, UK: Cambridge University Press.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society. Series B (Methodological), 42(2), 109-142. Retrieved from http://www.jstor.org/stable/2984952

McCullagh, P. J. (2005). The proportional odds model. In P. Armitage \& T. Colton (Eds.), Encyclopedia of biostatistics (2nd ed.). New York, NY: Wiley.

McCullagh, P., \& Nelder, J. A. (1989). Generalized linear models (2nd ed.). Boca Raton, FL: Chapman \& Hall/CRC.

McKelvey, R. D., \& Zavoina, Z. (1975). A statistical model for the analysis of ordinal level dependent variables. The Journal of Mathematical Sociology, 4(1), 103-120. doi: 10.1080/0022250X.1975.9989847

O'Connell, A. A., \& Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. Journal of Modern Applied Statistical Methods, $10(1), 139-175$. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol10/iss1/15/

Peterson, B., \& Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. Journal of the Royal Statistical Society. Series C (Applied Statistics), 39(2), 205-217. doi: 10.2307/2347760

Pratt, J. W. (1981). Concavity of the log likelihood. Journal of the American Statistical Association, 76(373), 103-106. doi: 10.1080/01621459.1981.10477613

R Development Core Team. (2009). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for Statistical Computing.

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

Tamer, E. (2003). Incomplete simultaneous discrete response models with multiple equilibria. Review of Economic Studies, 70(1):147-165. doi:

### 10.1111/1467-937X. 00240

Terza, J. V. (1985). Ordinal probit: A generalization. Communications in Statistics - Theory and Methods, 14(1), 1-11. doi: 10.1080/03610928508828893

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. The Stata Journal, 6(1), 58-82.

Winship, C., \& Mare, R. D. (1984). Regression models with ordinal variables. American Sociological Review, 49(4), 512-525. Retrieved from http://www.jstor.org/stable/2095465

Yee, T. W. (2010). The VGAM package for categorical data analysis. Journal of Statistical Software, 32(10), 1-34. doi: 10.18637/jss.v032.i10

## Endnotes

1. To our knowledge, the first suggestion of such a model is contained in Winship and Mare (1984, p. 519). Long (1997, ch. 6) calls this "parallel regressions". It is latent parallelism. Parallelism holds in the latent variable representation though not in probabilities under asymmetry.
2. We could only require full column rank and finite $x$.
3. In the class of models we consider, parameters are generally estimated to scale. The standard deviation of this error is the most commonly used method of scaling.
4. Boes and Winkelmann (2006) show that such a model is similar, in likelihood terms, to what is known as the generalized threshold model of Maddala (1983) and Terza (1985).
5. These conditions are elaborated by McCullagh (2005). $\tau_{j}-\tau_{j-1}>\mathbf{X}_{i}\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{j-1}\right)$
6. But see the class of location-scale models. Cox (1995) discussed generalizations of the location-scale model that nest, as special cases, the partial-proportional odds model of Peterson and Harrell (1990). These results generalize to that case because the scale parameters cannot collapse to zero and the measure of the set of contradictions, though possibly shrinking, similarly does not collapse to zero.

## ROBERT W. WALKER

7. The standard logistic distribution has variance equal to $\frac{\pi^{2}}{3} s^{2}$, where $s$ is a scaling parameter. If we set $s=\frac{\sqrt{3}}{\pi}$, we can make the variance one.
8. Coherence, in simultaneous equations with limited outcomes, refers to nonexistence of solutions. Completeness refers to multiplicity. These problems often arise in the analysis of simultaneous move, discrete action game theory and are tantamount to lack of existence, in the coherence case, and lack of uniqueness, in the completeness case, of equilibrium.
9. Peterson and Harrell (1990) were able to undertake the Monte Carlo simulations that they report because the parameters, as they set them, do not cross and they rely on probabilities fed to the canonical multinomial distribution rather than simulating latent quantities. Were they to have done the latter, they would have realized this.

## ON GENERALIZING CUMULATIVE ORDERED REGRESSION MODELS

## Appendix A. An R Example

```
> x1 <- sample(c(0, 1), size = 1000, replace = TRUE)
> x2 <- runif(1000, -1, 1)
> eps <- rlogis(1000)
> y.star1 <- 0.05 * x1 + x2 + eps
> y.star2 <- 0.1 * x1 + x2 + eps
> y1 <- (y.star1 < -0.5 & y.star2 < -0.5)
> y2 <- (y.star1 > -0.5 & y.star2 > -0.5 & y.star1 < 0.5 & y.star2 <
+ 0.5)
> y3 <- (y.star1 > 0.5 & y.star2 > 0.5)
> y<- y1 + 2 * y2 + 3 * y3
> table(y)
y
    0}12
    14 392 216 378
> bad.result <- data.frame(y.star1, y.star2, y1, y2, y3, y)
> bad.result[y == 0, ]
            y.star1 y.star2 y1 y2 y3 y
40 -0.5114175 -0.4614175 FALSE FALSE FALSE 0
152 0.4553046 0.5053046 FALSE FALSE FALSE 0
163 0.4736140 0.5236140 FALSE FALSE FALSE 0
333-0.5033633-0.4533633 FALSE FALSE FALSE 0
4 1 7 0 . 4 5 1 9 1 7 3 ~ 0 . 5 0 1 9 1 7 3 ~ F A L S E ~ F A L S E ~ F A L S E ~ 0 ~
4 4 9 0 . 4 5 0 7 8 0 7 ~ 0 . 5 0 0 7 8 0 7 ~ F A L S E ~ F A L S E ~ F A L S E ~ 0 ~
464 0.4668629 0.5168629 FALSE FALSE FALSE 0
468 0.4720030 0.5220030 FALSE FALSE FALSE 0
663 0.4846675 0.5346675 FALSE FALSE FALSE 0
676 0.4669751 0.5169751 FALSE FALSE FALSE 0
677 0.4820676 0.5320676 FALSE FALSE FALSE 0
833-0.5296321 -0.4796321 FALSE FALSE FALSE 0
834-0.5144424 -0.4644424 FALSE FALSE FALSE 0
880 0.4776592 0.5276592 FALSE FALSE FALSE 0
```


# Estimation of Parameters of Misclassified Size Biased Borel Distribution 

B. S. Trivedi<br>Ahmedabad University<br>Navrangpura, Ahmedabad, India

M. N. Patel<br>Gujarat University<br>Ahmedabad, India


#### Abstract

A misclassified size-biased Borel Distribution (MSBBD), where some of the observations corresponding to $x=c+1$ are wrongly reported as $x=c$ with probability $\alpha$, is defined. Various estimation methods like the method of maximum likelihood (ML), method of moments, and the Bayes estimation for the parameters of the MSBB distribution are used. The performance of the estimators are studied using simulated bias and simulated risk. Simulation studies are carried out for different values of the parameters and sample size.


Keywords: Borel distribution, misclassification, size-biased, method of moments, maximum likelihood, Bayes estimation

## Introduction

The Borel distribution is a discrete probability distribution, arising in contexts including branching processes and queueing theory. If the number of offspring that an organism has is Poisson-distributed, and if the average number of offspring of each organism is no bigger than 1, then the descendants of each individual will ultimately become extinct. The number of descendants that an individual ultimately has in that situation is a random variable distributed according to a Borel distribution.

Borel (1942) defined a one parameter Borel distribution as

$$
\begin{equation*}
\mathrm{P}(X=x)=\mathrm{p}(x ; \theta)=\frac{(1+x)^{x-1}}{x!} \theta^{x} \mathrm{e}^{-(1+x) \theta} ; \quad 0<\theta<1, x=1,2,3, \ldots \tag{1}
\end{equation*}
$$

[^27]
## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

This distribution describes a distribution of the number of customers served before a queue vanishes under condition of a single queue with random arrival times (at constant rate) of customers and a constant time occupied in serving each customer.

Gupta (1974) defined the Modified Power Series Distribution (MPSD) with probability function given by

$$
\begin{equation*}
\mathrm{P}_{1}(X=x)=\mathrm{a}(x) \frac{(g(\theta))^{x}}{\mathrm{f}(\theta)}, \quad x \in \mathrm{~T} \tag{2}
\end{equation*}
$$

where $\mathrm{a}(x)>0$, T is a subset of the set of non-negative integers, $\mathrm{g}(\theta)$ and $\mathrm{f}(\theta)$ are positive, finite, and differentiable, and $\theta$ is the parameter.

Hassan and Ahmad (2009) showed the Borel distribution is a particular case of modified power series distribution (MPSD) with

$$
\begin{equation*}
\mathrm{a}(x)=\frac{(1+x)^{x-1}}{x!}, \quad \mathrm{g}(\theta)=\theta \mathrm{e}^{-\theta}, \quad \mathrm{f}(\theta)=\mathrm{e}^{\theta} \tag{3}
\end{equation*}
$$

in (2).
The Borel-Tanner distribution generalizes the Borel distribution. Let $k$ be a positive integer. If $x_{1}, x_{2}, \ldots, x_{k}$ are independent and each has Borel distribution with parameter $\theta$, then their sum $w=x_{1}+x_{2}+\ldots+x_{k}$ is said to have the Borel-Tanner distribution with parameters $\theta$ and $k$. This gives the distribution of the total number of individuals in a Poisson-Galton-Watson process starting with $k$ individuals in the first generation, or of the time taken for an M/D/1 queue to empty starting with $k$ jobs in the queue. The case $k=1$ is simply the Borel distribution above.

Here, the M/D/1 queue represents the queue length in a system having a single server, where arrivals are determined by a Poisson process and job service times are fixed (deterministic). An extension of this model with more than one server is the M/D/c queue.

## Size-Biased Borel Distribution

Size-biased distributions are a special case of the more general form known as weighted distributions. Weighted distributions have numerous applications in forestry and ecology.

Size-biased distributions were first introduced by Fisher (1934) to model ascertainment bias; weighted distributions were later formalized in a unifying

## TRIVEDI \& PATEL

theory by Rao (1965). Such distributions arise naturally in practice when observations from a sample are recorded with unequal probability, such as from probability proportional to size (PPS) designs. In short, if the random variable $X$ has distribution $\mathrm{f}(x ; \theta)$, with unknown parameter $\theta$, then the corresponding weighted distribution is of the form

$$
\begin{equation*}
\mathrm{f}^{w}(x ; \theta)=\frac{\mathrm{w}(x) \mathrm{f}(x ; \theta)}{\mathrm{E}\{\mathrm{w}(x)\}} \tag{4}
\end{equation*}
$$

where $\mathrm{w}(x)$ is a non-negative weight function such that $\mathrm{E}\{\mathrm{w}(x)\}$ exists.
The size-biased Borel distribution is also derived from the size-biased MPSD as it is a particular case of the MPSD. A size-biased MPSD is obtained by taking the weight of MPSD (2) as $x$, given by

$$
\begin{align*}
\mathrm{P}(X=x) & =\frac{b_{x}(\mathrm{~g}(\theta))^{x}}{\mu(\theta) \mathrm{f}(\theta)} \\
& =\frac{b_{x}(\mathrm{~g}(\theta))^{x}}{\mathrm{f}^{*}(\theta)} \tag{5}
\end{align*}
$$

where $b_{x}=x \mathrm{a}(x)$ and $\mathrm{f}^{*}(\theta)=\mu(\theta) \mathrm{f}(\theta)$.
Now, by taking

$$
\begin{equation*}
b_{x}=x \mathrm{a}(x)=\frac{(1+x)^{x-1}}{(x-1)!}, \mu(\theta)=\frac{\theta}{1-\theta}, \mathrm{g}(\theta)=\theta \mathrm{e}^{-\theta}, \mathrm{f}(\theta)=\mathrm{e}^{\theta}, \quad 0<\theta<1 \tag{6}
\end{equation*}
$$

a size-biased Borel distribution is obtained with p.m.f. given by

$$
\begin{equation*}
\mathrm{P}(X=x)=\frac{(1+x)^{x-1}}{(x-1)!} \theta^{x-1}(1-\theta) \mathrm{e}^{-\theta(1+x)}, \quad x=1,2,3, \ldots \tag{7}
\end{equation*}
$$

## Misclassified Size-Biased Borel Distribution

A dependent variable which is a discrete response causes the estimated coefficients to be inconsistent in a probit or logit model when misclassification is present. By

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

'misclassification' we mean that the response is reported or recorded in the wrong category; for example, a variable is recorded as a one when it should have the value zero. This mistake might easily happen in an interview setting where the respondent misunderstands the question or the interviewer simply checks the wrong box. Other data sources where the researcher suspects measurement error, such as historical data, certainly exist as well. It will be shown that, when a dependent variable is misclassified in a probit or logit setting, the resulting coefficients are biased and inconsistent.

Assume that some of the values $(c+1)$ are erroneously reported as $c$, and let the probabilities of these observation be $\alpha$. Then the resulting distribution of the size-biased random variable $X$ is called the misclassified size-biased distribution. Trivedi and Patel (2013) have considered misclassified size-biased generalized negative binomial distributions and parameter estimation. The misclassified sizebiased Borel distribution can be obtained as

$$
\begin{align*}
p_{x} & =\mathrm{P}(X=x) \\
& = \begin{cases}\left(\theta \mathrm{e}^{-\theta}\right)^{c}\left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}, & x=c \\
(1-\alpha)(c+2)^{c}\left(\theta \mathrm{e}^{-\theta}\right)^{c+1}\left\{c!\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}, & x=c+1 \\
\frac{(1+i)^{i-1}\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{(i-1)!}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}, & x \in \mathrm{~S}\end{cases} \tag{8}
\end{align*}
$$

where $S$ is the set of non-negative integers excluding integers $c$ and $c+1,0 \leq \alpha \leq 1$, $0<\theta<1$, and $x=1,2,3, \ldots$. The mean and variance of this distribution are obtained from the moments of misclassified size-biased MPSD given by Hassan and Ahmad (2009) as

$$
\begin{equation*}
\text { Mean }=\mu_{1}^{\prime}=\frac{1}{(1-\theta)^{2}}+\frac{\theta}{(1-\theta)}-\alpha(1-\theta) b_{c+1} \mathrm{~g}^{c} \mathrm{e}^{-2 \theta} \tag{9}
\end{equation*}
$$

$$
\begin{align*}
\text { Variance } & =\mu_{2} \\
& =\frac{\left(3 \theta-\theta^{2}\right)}{(1-\theta)^{4}}+\left[\begin{array}{l}
\alpha\left\{\begin{array}{c}
2+2 \theta(1-\theta)-(2 c+1)(1-\theta)^{2} \\
-\alpha b_{c+1} \mathrm{~g}^{c}(1-\theta)^{3} \mathrm{e}^{-2 \theta}
\end{array}\right\} \\
\times\left\{b_{c+1} \mathrm{~g}^{c}(1-\theta)^{3} \mathrm{e}^{-2 \theta}\right\}
\end{array}\right] \tag{10}
\end{align*}
$$

## Method of Maximum Likelihood Estimation

Let $x_{1}, x_{2}, \ldots, x_{k}$ be the probable values of the random variable $X$ in a random sample of misclassified size-biased Borel distribution and $n_{k}$ denote the number of observations corresponding to the value $x_{k}$ in the sample (where $k>0$ ). Thus the likelihood function $L$ is given by

$$
\begin{align*}
\mathrm{L} & \propto \prod_{i=1}^{k} P_{i}^{n_{i}} \\
= & P_{c}^{n_{c}} P_{c+1}^{n_{c+1}} \prod_{i \neq c, c+1} P_{i}^{n_{i}} \\
= & {\left[\left(\theta \mathrm{e}^{-\theta}\right)^{c}\left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right]^{n_{c}} }  \tag{11}\\
& \times\left[(1-\alpha)(c+2)^{c}\left(\theta \mathrm{e}^{-\theta}\right)^{c+1}\left\{c!\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right]^{n_{c+1}} \\
& \times \prod_{i \neq c, c+1}\left[\frac{(1+i)^{i-1}\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{(i-1)!}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right]^{n_{i}}
\end{align*}
$$

where

$$
\sum_{i=1}^{k} n_{i}=N
$$

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

$$
\begin{align*}
\ln \mathrm{L}= & n_{c} \ln \left[\left(\theta \mathrm{e}^{-\theta}\right)^{c}\left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right] \\
& +n_{c+1} \ln \left[(1-\alpha)(c+2)^{c}\left(\theta \mathrm{e}^{-\theta}\right)^{c+1}\left\{c!\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right] \\
& +\sum_{i \neq c, c+1} n_{i} \ln \left[\frac{(1+i)^{i-1}\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{(i-1)!}\left\{\frac{\theta \mathrm{e}^{\theta}}{(1-\theta)}\right\}^{-1}\right] \\
= & n_{c} c \ln \theta-n_{c} c \theta+n_{c} \ln \left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}-n_{c} \ln \theta  \tag{12}\\
& -n_{c} \theta+n_{c} \ln (1-\theta)+n_{c+1} \ln (1-\alpha)+n_{c+1} c \ln (c+2) \\
& +n_{c+1}(c+1) \ln \theta-n_{c+1}(c+1) \theta-n_{c+1} \ln c!-n_{c+1} \ln \theta-n_{c+1} \theta \\
& +n_{c+1} \ln (1-\theta)+\sum_{i \neq c, c+1} n_{i}\left\{\begin{array}{r}
-i \theta-\ln (i-1)!-\theta+\ln (i+1)+(i-1) \ln \theta \\
-i \theta)
\end{array}\right\}
\end{align*}
$$

Let the derivative of $\ln \mathrm{L}$ with respect to $\alpha$ and $\theta$ be zero. The solutions of $\frac{\partial \ln 1}{\partial \alpha}=0$ and $\frac{\partial \ln 1}{\partial \theta}=0$ gives us the ML estimators of $\alpha$ and $\theta:$

$$
\begin{align*}
& \frac{\partial \ln 1}{\partial \alpha}=n_{c} \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}^{-1}+n_{c+1} \frac{(-1)}{(1-\alpha)}  \tag{13}\\
& \frac{\partial \ln 1}{\partial \theta}=\left(\frac{1}{\theta}-1\right) \sum_{i=1}^{k} i n_{i}+n_{c} \frac{\alpha \frac{(2+c)^{c}}{c!}\left(\mathrm{e}^{-\theta}-\theta \mathrm{e}^{-\theta}\right)}{\left\{\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right\}}-\left(\frac{1}{\theta}+1\right) \sum_{i=1}^{k} n_{i}  \tag{14}\\
&-\frac{1}{1-\theta} \sum_{i=1}^{k} n_{i}
\end{align*}
$$

## TRIVEDI \& PATEL

Equating $\frac{\partial \ln 1}{\partial \alpha}$ and $\frac{\partial \ln 1}{\partial \theta}$ to zero, we get

$$
\begin{gather*}
\alpha=\frac{n_{c}(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)-n_{c+1} c(1+c)^{c-1}}{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\left(n_{c}+n_{c+1}\right)}  \tag{15}\\
\left(\frac{1-\theta}{\theta}\right) \sum_{i=1}^{k} i_{i}+n_{c}\left\{\frac{c(1+c)^{c-1}}{\alpha(2+c)^{c} \mathrm{e}^{-\theta}(1-\theta)}+\frac{\theta}{(1-\theta)}\right\}^{-1}  \tag{16}\\
\quad-\left(\frac{1+\theta}{\theta}\right) \sum_{i=1}^{k} n_{i}-\left(\frac{1}{1-\theta}\right) \sum_{i=1}^{k} n_{i}=0
\end{gather*}
$$

In the equation (16), substituting $\alpha$ from the equation (15), we get an equation consisting only parameter $\theta$, say $\mathrm{g}(\theta)=0$. By solving this equation for $\theta$ using any iterative method, we get the solution, known as the MLE of $\theta$. Using this MLE of $\theta$ in (15), we get the MLE of $\alpha$.

## Asymptotic Variance-Covariance Matrix of ML Estimators

The second order derivatives with respect to $\alpha$ and $\theta$ of the likelihood function L are obtained as below:

$$
\begin{gather*}
\frac{\partial^{2} \ln \mathrm{~L}}{\partial \alpha^{2}}=-\frac{n_{c}(2+c)^{2 c}\left(\theta \mathrm{e}^{-\theta}\right)^{2}}{\left\{c(1+c)^{c-1}+\alpha(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\right\}^{2}}-\frac{n_{c+1}}{(1-\alpha)^{2}}  \tag{17}\\
\frac{\partial^{2} \ln \mathrm{~L}}{\partial \theta \partial \alpha}=\frac{n_{c}(2+c)^{c} \mathrm{e}^{-\theta}(1-\theta)\left\{c(1+c)^{c-1}\right\}}{\left\{c(1+c)^{c-1}+\alpha(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\right\}^{2}} \tag{18}
\end{gather*}
$$

$$
\begin{align*}
\frac{\partial^{2} \ln \mathrm{~L}}{\partial \theta^{2}}=\left(\frac{-1}{\theta^{2}}\right) & \sum_{i=1}^{k} \operatorname{in}_{i}+\left(\frac{1-2 \theta}{\theta^{2}(1-\theta)^{2}}\right) \sum_{i=1}^{k} n_{i} \\
& -\frac{n_{c} \alpha(2+c)^{c} \mathrm{e}^{-\theta}\left\{(2-\theta)\left[c(1+c)^{c-1}+\alpha(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\right]\right\}}{\left\{c(1+c)^{c-1}+\alpha(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\right\}^{2}}  \tag{19}\\
& +\frac{n_{c} \alpha(2+c)^{c} \mathrm{e}^{-\theta}\left\{\mathrm{e}^{-\theta}(1-\theta)^{2} \alpha(2+c)^{c}\right\}}{\left\{c(1+c)^{c-1}+\alpha(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)\right\}^{2}}
\end{align*}
$$

Using the above equations, the asymptotic variance covariance matrix $\Sigma$ of MLE is obtained from the inverse of the Fisher information matrix

$$
\mathrm{J}(\theta, \alpha)=\left[\begin{array}{ll}
-\mathrm{E}\left(\frac{\partial^{2} \ln \mathrm{~L}}{\partial \theta^{2}}\right) & -\mathrm{E}\left(\frac{\partial^{2} \ln \mathrm{~L}}{\partial \theta \partial \alpha}\right)  \tag{20}\\
-\mathrm{E}\left(\frac{\partial^{2} \ln \mathrm{~L}}{\partial \alpha \partial \alpha}\right) & -\mathrm{E}\left(\frac{\partial^{2} \ln \mathrm{~L}}{\partial \alpha^{2}}\right)
\end{array}\right]
$$

That is

$$
\boldsymbol{\Sigma}=\left[\begin{array}{cc}
\mathrm{v}(\theta) & \operatorname{cov}(\theta, \alpha)  \tag{21}\\
\operatorname{cov}(\theta, \alpha) & \mathrm{v}(\alpha)
\end{array}\right], \quad \mathrm{SE}(\hat{\theta})=\sqrt{\mathrm{v}(\theta)}, \mathrm{SE}(\hat{\alpha})=\sqrt{\mathrm{v}(\hat{\alpha})}
$$

## Method of Moments

The mean and variance of the misclassified size-biased Borel distribution are

$$
\begin{equation*}
\text { Mean }=\mu_{1}^{\prime}=\frac{1}{(1-\theta)^{2}}+\frac{\theta}{(1-\theta)}-\alpha(1-\theta) b_{c+1} \mathrm{~g}^{c} \mathrm{e}^{-2 \theta} \tag{22}
\end{equation*}
$$

$$
\begin{align*}
\text { Variance } & =\mu_{2} \\
& =\frac{\left(3 \theta-\theta^{2}\right)}{(1-\theta)^{4}}+\left\{\begin{array}{c}
\alpha\left[\begin{array}{c}
2+2 \theta(1-\theta)-(2 c+1)(1-\theta)^{2} \\
-\alpha b_{c+1} \mathrm{~g}^{c}(1-\theta)^{3} \mathrm{e}^{-2 \theta}
\end{array}\right] \\
\times\left[b_{c+1} \mathrm{~g}^{c}(1-\theta)^{-1} \mathrm{e}^{-2 \theta}\right]
\end{array}\right] \tag{23}
\end{align*}
$$

The recurrence relation of row moments of the misclassified size-biased Borel distribution is

$$
\begin{equation*}
\mu_{r+1}^{\prime}=\frac{\mathrm{g}}{\mathrm{~g}^{\prime}} \frac{\partial \mu_{r}^{\prime}}{\partial \theta}+\alpha \frac{\left(\mu_{r}^{\prime}-c^{r}\right) \mathrm{g}^{c+1} b_{c+1}}{\mathrm{f} \mu}+\mu_{1}^{\prime} \mu_{r}^{\prime} \tag{24}
\end{equation*}
$$

where $\mathrm{g}(\theta), \mathrm{f}(\theta), \mu(\theta)$, and $b_{x}$ are as per (6). By taking different values of $r$, different row moments are obtained. Taking $r=1$ will obtain the second row moments of the misclassified size-biased Borel distribution.

$$
\begin{equation*}
\mu_{2}^{\prime}=\frac{\mathrm{g}}{\mathrm{~g}^{\prime}} \frac{\partial \mu_{1}^{\prime}}{\partial \theta}+\alpha \frac{\left(\mu_{1}^{\prime}-c^{1}\right) \mathrm{g}^{c+1} b_{c+1}}{\mathrm{f} \mu}+\left(\mu_{1}^{\prime}\right)^{2} \tag{25}
\end{equation*}
$$

Solving (22) and (25) for $\alpha$ and $\theta$ yields moment estimators of $\alpha$ and $\theta$.
The explicit form cannot be obtained for the moment estimators but, by the method of iteration, the solution for the equations may be obtained.

## Asymptotic Variance-Covariance Matrix of Moment Estimators

Denote $\mu_{1}^{\prime}$ by $\mathrm{H}_{1}(\theta, \alpha)$ and $\mu_{2}^{\prime}$ by $\mathrm{H}_{2}(\theta, \alpha)$, i.e.

$$
\begin{equation*}
\mathrm{H}_{1}(\theta, \alpha)=\frac{1}{(1-\theta)^{2}}+\frac{\theta}{(1-\theta)}-\alpha(1-\theta) b_{c+1} \mathrm{~g}^{c} \mathrm{e}^{-2 \theta} \tag{26}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{H}_{2}(\theta, \alpha)=\frac{\mathrm{g}}{\mathrm{~g}^{\prime}} \frac{\partial \mu_{1}^{\prime}}{\partial \theta}+\alpha \frac{\left(\mu_{1}^{\prime}-c^{1}\right) \mathrm{g}^{c+1} b_{c+1}}{\mathrm{f} \mu}+\left(\mu_{1}^{\prime}\right)^{2} \tag{27}
\end{equation*}
$$

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

Then, the asymptotic variance-covariance matrix of moment estimators $\tilde{\theta}$ and $\tilde{\alpha}$ are given by

$$
\begin{align*}
\mathbf{V} & =\mathbf{A}^{-1} \boldsymbol{\Sigma}\left(\mathbf{A}^{-1}\right)^{\prime} \\
& =\left[\begin{array}{cc}
\mathrm{v}(\tilde{\theta}) & \operatorname{cov}(\tilde{\theta}, \tilde{\alpha}) \\
\operatorname{cov}(\tilde{\theta}, \tilde{\alpha}) & \mathrm{v}(\tilde{\alpha})
\end{array}\right] \tag{28}
\end{align*}
$$

where the matrix $\mathbf{A}$ is

$$
\mathbf{A}=\left[\begin{array}{ll}
\frac{\partial \mathrm{H}_{1}}{\partial \theta} & \frac{\partial \mathrm{H}_{1}}{\partial \alpha}  \tag{29}\\
\frac{\partial \mathrm{H}_{2}}{\partial \theta} & \frac{\partial \mathrm{H}_{2}}{\partial \alpha}
\end{array}\right]=\left[\begin{array}{ll}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{array}\right]
$$

and

$$
\boldsymbol{\Sigma}=\left[\begin{array}{cc}
\mathrm{v}\left(\mathrm{~m}_{1}^{\prime}\right) & \operatorname{cov}\left(\mathrm{m}_{1}^{\prime}, \mathrm{m}_{2}^{\prime}\right)  \tag{30}\\
\operatorname{cov}\left(\mathrm{m}_{1}^{\prime}, \mathrm{m}_{2}^{\prime}\right) & \mathrm{v}\left(\mathrm{~m}_{2}^{\prime}\right)
\end{array}\right]=\left[\begin{array}{cc}
\sigma_{11} & \sigma_{12} \\
\sigma_{21} & \sigma_{22}
\end{array}\right]
$$

where

$$
\begin{aligned}
& \mathrm{V}\left(\mathrm{~m}_{r}^{\prime}\right)=\frac{\mu_{2 r}^{\prime}-\left(\mu_{r}^{\prime}\right)^{2}}{n}, \quad r=1,2 \\
& \operatorname{COV}\left(\mathrm{~m}_{r}^{\prime}, \mathrm{m}_{s}^{\prime}\right)=\frac{\mu_{r+s}^{\prime}-\mu_{r}^{\prime} \mu_{s}^{\prime}}{n}, \quad r \neq s=1,2
\end{aligned}
$$

and $\mathrm{m}_{r}^{\prime}$ is the $r^{\text {th }}$ sample raw moment of the MSBPL distribution, i.e.

$$
\mathrm{m}_{r}^{\prime}=\frac{1}{n} \sum_{i=1}^{n} x_{i}^{r}
$$

## TRIVEDI \& PATEL

## Bayes Estimation

The ML method, as well as other classical approaches, is based only on the empirical information provided by the data. However, when there is some technical knowledge on the parameters of the distribution available, a Bayes procedure seems to be an attractive inferential method. The Bayes procedure is based on a posterior density, say $\pi(\alpha, \theta \mid x)$, which is proportional to the product of the likelihood function $\mathrm{L}(\alpha, \theta \mid x)$ with a prior joint density, say $\mathrm{g}(\alpha, \theta)$, representing the uncertainty on the parameters values. Assume before the observations were made knowledge about the parameters $\alpha$ and $\theta$ was vague. Consequently, the non-informative vague prior $\pi_{1}(\alpha)=\mathrm{g}_{1}(\alpha)=1$ is applicable to a good approximation.

The non-informative priors of $\alpha$ and $\theta$ are

$$
\begin{align*}
& \pi_{1}(\alpha)=g_{1}(\alpha)=1  \tag{31}\\
& \pi_{2}(\theta)=g_{2}(\theta)=1 \tag{32}
\end{align*}
$$

Hence, the joint prior of $\theta$ and $\alpha$ is given by

$$
\begin{align*}
& \mathrm{g}(\theta, \alpha)=\mathrm{g}_{1}(\alpha) \mathrm{g}_{2}(\theta)  \tag{33}\\
& \mathrm{g}(\alpha, \theta)=\pi_{1}(\alpha) \pi_{2}(\theta)=1
\end{align*}
$$

If L is the likelihood function indexed by a continuous parameter $\boldsymbol{\Theta}=(\theta, \alpha)$ with prior density $\mathrm{g}(\theta, \alpha)$, then the posterior density for $\boldsymbol{\Theta}$ is given by

$$
\begin{align*}
& \pi(\boldsymbol{\Theta} \mid x)=\frac{\mathrm{L}(\boldsymbol{\Theta}) \mathrm{g}(\alpha, \theta)}{\int_{0}^{1} \int_{0}^{1} \mathrm{~L}(\boldsymbol{\Theta}) \mathrm{g}(\alpha, \theta) d \alpha d \theta} \\
& =\frac{\left[\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\right]\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\int_{0}^{1}\left[\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\right] \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}\left(\theta \mathrm{e}^{-\theta}\right)^{j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \int_{0}^{1} \alpha^{j}(1-\alpha)^{n_{c+1}} d \alpha d \theta} \\
& =\frac{\left[\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\right]\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\int_{0}^{1}\left[\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\right] \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}\left(\theta \mathrm{e}^{-\theta}\right)^{j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} d \theta} \\
& =\frac{\left[\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\right]\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \int_{0}^{1} \theta \sum_{i=1}^{k}\left[n_{i}(i-1)+j\right](1-\theta) \sum_{i=1}^{k} n_{i} e^{-\theta}\left[\sum_{i=1}^{k} n_{i}(i+1)+j\right] d \theta} \tag{34}
\end{align*}
$$

## TRIVEDI \& PATEL

Using the result given by Gradshteĭn and Ryzhik (2007, p. 347),

$$
\begin{equation*}
\pi(\boldsymbol{\Theta} \mid x)=\frac{\prod_{i=1}^{k}\left[\frac{\left(\theta \mathrm{e}^{-\theta}\right)^{i}}{\theta \mathrm{e}^{\theta} / 1-\theta}\right]^{n_{i}}\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta \mathrm{e}^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1, \mu+\gamma+2 ;-\eta)} \tag{35}
\end{equation*}
$$

where

$$
\begin{aligned}
\Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)=1 & +\frac{(\gamma+1)}{(\mu+\gamma+2)}\left(\frac{-\eta}{1!}\right)+\frac{(\gamma+1)(\gamma+2)}{(\mu+\gamma+2)(m+\gamma+3)}\left(\frac{-\eta^{2}}{2!}\right)+\frac{(\gamma+1)(\gamma+2)(\gamma+3)}{(\mu+\gamma+2)(m+\gamma+3)(m+\gamma+4)}\left(\frac{-\eta^{3}}{3!}\right) \\
& +\frac{(\gamma+1)(\gamma+2)(\gamma+3)(\gamma+4)}{(\mu+\gamma+2)(m+\gamma+3)(m+\gamma+4)(m+\gamma+5)}\left(\frac{-\eta^{4}}{4!}\right)+\ldots
\end{aligned}
$$

where

$$
\begin{equation*}
\gamma=\sum_{i=1}^{k} n_{i}(i-1)+j=\mathrm{N}(\bar{x}-1)+j, \quad \mu=\sum_{i=1}^{k} n_{i}=N, \quad \eta=-\left[\left(\sum_{i=1}^{k} n_{i}(i+1)\right)+j\right]=\mathrm{N}(\bar{x}+1)+j \tag{36}
\end{equation*}
$$

From (35), the marginal posterior of $\alpha$ will be

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

$$
\begin{align*}
\pi(\alpha \mid x) & =\int_{0}^{1} \pi(\alpha, \theta \mid x) d \theta \\
& =\int_{0}^{1} \frac{\prod_{i=1}^{k}\left[\left(\theta e^{-\theta}\right)^{i} / \frac{\theta e^{\theta}}{(1-\theta)}\right]^{n_{i}}\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta e^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} d \theta \\
& =\frac{(1-\alpha)^{n_{c+1}} \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{\alpha^{j}(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} \tag{37}
\end{align*}
$$

From (37), the Bayes estimate of $\alpha$ is given by

$$
\begin{aligned}
\hat{\alpha}_{B S} & =\int_{0}^{1} \alpha \pi(\alpha \mid x) d \alpha \\
& =\int_{0}^{1} \alpha \frac{(1-\alpha)^{n_{c+1}} \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{\alpha^{j}(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)}{\sum_{j=0}^{n_{c}}\binom{n}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} d \alpha
\end{aligned}
$$

## TRIVEDI \& PATEL

$$
\begin{equation*}
=\frac{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+2\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} \tag{38}
\end{equation*}
$$

Similarly, from (35), the marginal posterior of $\theta$ will be

$$
\begin{align*}
\pi(\theta \mid x) & =\int_{0}^{1} \pi(\alpha, \theta \mid x) d \alpha \\
& =\int_{0}^{1} \frac{\prod_{i=1}^{k}\left[\left(\theta e^{-\theta}\right)^{i} / \frac{\theta e^{\theta}}{(1-\theta)}\right]^{n_{i}}\left[\frac{(1+c)^{c-1}}{(c-1)!}+\alpha \frac{(2+c)^{c}\left(\theta e^{-\theta}\right)}{c!}\right]^{n_{c}}(1-\alpha)^{n_{c+1}}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} d \alpha \\
& =\frac{\theta^{\sum_{i=1}^{k} n_{i}(i-1)+j}(1-\theta)^{\sum_{i=1}^{k} e^{-\theta\left(\left[\sum_{l=1}^{k} n_{i(i+1)}\right)+j\right]}} \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} \tag{39}
\end{align*}
$$

From (39), the Bayes estimate of $\theta$ is given by

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

$$
\begin{align*}
& \hat{\theta}_{B S}=\int_{0}^{1} \theta \pi(\theta \mid x) d \theta \\
& =\int_{0}^{1} \theta \frac{\theta^{\sum_{k=1}^{k} n_{c}(i-1)+j}(1-\theta)^{\sum_{k=1}^{k} n_{i}} \mathrm{e}^{-\theta\left[\left(\sum_{k=1}^{k} n_{c}(i+1)\right)+j\right.} \sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)}}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j}} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta) \quad d \theta \\
& =\frac{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+, j+1\right)} \beta_{(\mu+1, \gamma+2)} \Phi(\gamma+2 ; \mu+\gamma+3 ;-\eta)}{\sum_{j=0}^{n_{c}}\binom{n_{c}}{j} \frac{(2+c)^{c j}}{(c!)^{j}}\left[\frac{(1+c)^{c-1}}{(c-1)!}\right]^{n_{c}-j} \beta_{\left(n_{c+1}+1, j+1\right)} \beta_{(\mu+1, \gamma+1)} \Phi(\gamma+1 ; \mu+\gamma+2 ;-\eta)} \tag{40}
\end{align*}
$$

where $\gamma, \mu$, and $\eta$ are as given in (36) above.

## Simulation Study

One thousand random samples, each of size $n$, were generated by using Monte Carlo simulation with different choices of sample size $n, \theta, \alpha$, and value of $c=1$ from the misclassified size-biased Borel distribution defined in equation (8). Using these different values of sample size $n, \theta$, and $\alpha$, we calculated the simulated risk (SR) and simulated bias of estimators $\alpha$ and $\theta$ by the method of MLE, method of moments, and Bayes estimation. The simulated results are shown in Tables 1 and 2. The $\operatorname{SR}$ is defined as

$$
\mathrm{SR}=\sqrt{\frac{\sum_{i=1}^{1000}\left(\hat{\theta}_{1}-\theta\right)^{2}}{1000}}
$$

## Conclusion

A comparison was made between different methods of estimation for the parameters of the misclassified size-biased Borel distribution. From Table 1 and 2, it was found that the method of maximum likelihood estimator works better compared to the moment estimator and the Bayes estimator on the basis of SR. As sample size increases, SR of both parameters of all three methods decreases. For fixed misclassification error $\alpha$, as $\theta$ increases, the SR of $\alpha$ and $\theta$ decreases in the case of maximum likelihood estimation, moment estimation method, and Bayes estimation. For fixed values of $\theta$ and sample size $n$, as $\alpha$ increases, there is not much difference in the SR of $\alpha$ as well as $\theta$. At the same time, if these values were compared in context of sample size, observe that, for a fixed value of $\theta$ and as $\alpha$ increases, the SR of $\alpha$ and $\theta$ decreases in most of the cases with the increase in sample size. As sample size increases, the bias in $\alpha$ and $\theta$ decreases in the case of all the three methods.

## References

Borel, E. (1942). Sur l'emploi du theoreme de Bernoulli pour faciliter le calcul d'un infinite de coefficients. Application au probleme de l' attente a un guichet. Comptes Rendus, Acedemic des Sciences, Paris, Series A, 214, 452-456.

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

Fisher, R. A. (1934). The effects of methods of ascertainments upon the estimation of frequencies. Annals of Eugenics, 6(1), 13-25. doi: 10.1111/j.14691809.1934.tb02105.x

Gradshteĭn, I. S., \& Ryzhik, I. M. (2007). Tables of integrals, series and products (7th ed.). Oxford: Academic Press.

Gupta, R. C. (1974). Modified power series distribution and some of its applications. Sankhyā: The Indian Journal of Statistics, Series B, 36(3), 288-298. Available from http://www.jstor.org/stable/25051912

Hassan, A., \& Ahmad, P. B. (2009). Misclassification in size-biased modified power series distribution and its applications. Journal of the Korean Society for Industrial and Applied Mathematics South Korea, 13(1), 55-72. Retrieved from http://mathnet.or.kr/mathnet/thesis_file/jksiam-v13n1p055.pdf

Rao, C. R. (1965). On discrete distributions arising out of ascertainment. In G. P. Patil (Ed.), Classical and contagious discrete distributions (pp. 320-332). Calcutta, India: Pergamon Press and Statistical Publishing Society.

Trivedi, B. S., \& Patel, M. N. (2013). Estimation in misclassified size-biased generalized negative binomial distribution. Mathematics and Statistics, 1(2), 7485. doi: $10.13189 / \mathrm{ms} .2013 .010210$

## TRIVEDI \& PATEL

## Appendix A

Table 1. Simulated risk of ML, moment, and Bayes estimators for different values of $\alpha, \theta$, and sample size $n$

| $\theta$ | $\boldsymbol{\alpha}$ | $n$ | ML |  | Moment |  | Bayes |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $\mathbf{S R}(\theta)$ | SR(a) | SR( $\boldsymbol{\theta}$ ) | SR(a) | SR( $\theta$ ) | SR(a) |
| 0.03 | 0.12 | 20 | 0.070621 | 0.687617 | 0.070009 | 0.730376 | 0.428511 | 0.731791 |
|  |  | 50 | 0.017214 | 0.662598 | 0.070000 | 0.716240 | 0.366045 | 0.722317 |
|  |  | 90 | 0.028486 | 0.576466 | 0.070000 | 0.661494 | 0.342946 | 0.729741 |
|  | 0.15 | 20 | 0.086849 | 0.637623 | 0.090139 | 0.712136 | 0.428910 | 0.774369 |
|  |  | 50 | 0.018903 | 0.615088 | 0.070000 | 0.695215 | 0.366122 | 0.695103 |
|  |  | 90 | 0.016796 | 0.386507 | 0.070000 | 0.649579 | 0.343211 | 0.675911 |
|  | 0.20 | 20 | 0.072757 | 0.600000 | 0.075005 | 0.681954 | 0.428803 | 0.683406 |
|  |  | 50 | 0.022814 | 0.489319 | 0.070000 | 0.668798 | 0.365955 | 0.653836 |
|  |  | 90 | 0.022814 | 0.489319 | 0.070000 | 0.668798 | 0.365955 | 0.653836 |
| 0.06 | 0.12 | 20 | 0.040157 | 0.409082 | 0.042393 | 0.606705 | 0.408054 | 0.659958 |
|  |  | 50 | 0.012628 | 0.391981 | 0.040017 | 0.591911 | 0.349603 | 0.628596 |
|  |  | 90 | 0.015325 | 0.280505 | 0.040019 | 0.524187 | 0.327791 | 0.610602 |
|  | 0.15 | 20 | 0.034921 | 0.525708 | 0.042064 | 0.564374 | 0.407451 | 0.595577 |
|  |  | 50 | 0.032482 | 0.237705 | 0.040083 | 0.559160 | 0.348794 | 0.565870 |
|  |  | 90 | 0.030247 | 0.194459 | 0.040000 | 0.508564 | 0.327905 | 0.567689 |
|  | 0.20 | 20 | 0.041125 | 0.453903 | 0.041515 | 0.533885 | 0.408379 | 0.554755 |
|  |  | 50 | 0.031410 | 0.319943 | 0.040203 | 0.521217 | 0.350368 | 0.546684 |
|  |  | 90 | 0.029152 | 0.212999 | 0.040016 | 0.476619 | 0.328233 | 0.531593 |
| 0.09 | 0.12 | 20 | 0.031714 | 0.386623 | 0.034880 | 0.413639 | 0.392743 | 0.556575 |
|  |  | 50 | 0.028941 | 0.338622 | 0.029383 | 0.376251 | 0.338557 | 0.558982 |
|  |  | 90 | 0.003557 | 0.010874 | 0.012466 | 0.336422 | 0.320139 | 0.556492 |
|  | 0.15 | 20 | 0.040798 | 0.301392 | 0.043413 | 0.409123 | 0.392115 | 0.556858 |
|  |  | 50 | 0.023699 | 0.105444 | 0.025690 | 0.347586 | 0.339796 | 0.539688 |
|  |  | 90 | 0.020707 | 0.086850 | 0.021824 | 0.321968 | 0.319821 | 0.520310 |
|  | 0.20 | 20 | 0.032107 | 0.361397 | 0.032177 | 0.415253 | 0.391115 | 0.504882 |
|  |  | 50 | 0.021050 | 0.240808 | 0.024129 | 0.348901 | 0.339039 | 0.499720 |
|  |  | 90 | 0.014885 | 0.214971 | 0.021792 | 0.326637 | 0.319959 | 0.454457 |

## MISCLASSIFIED SIZE-BIASED BOREL DISTRIBUTION

Table 2. Simulated Bias of ML, Moment and Bayes estimators for different values of $\alpha, \theta$, and sample size $n$

| $\theta$ | $\alpha$ | $n$ | ML |  | Moment |  | Bayes |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $\operatorname{Bias}(\theta)$ | Bias( $\alpha$ ) | $\operatorname{Bias}(\theta)$ | Bias( $\alpha$ ) | $\operatorname{Bias}(\theta)$ | Bias( $\alpha$ ) |
| 0.03 | 0.12 | 20 | 0.070474 | 0.691185 | 0.070010 | 0.699012 | 0.428432 | 0.696161 |
|  |  | 50 | 0.026334 | 0.571479 | 0.070000 | 0.637222 | 0.365951 | 0.619558 |
|  |  | 90 | 0.016826 | 0.555057 | 0.070000 | 0.558062 | 0.342872 | 0.528030 |
|  | 0.15 | 20 | 0.084899 | 0.613982 | 0.070120 | 0.694661 | 0.428820 | 0.668716 |
|  |  | 50 | 0.017858 | 0.506054 | 0.070000 | 0.684802 | 0.366023 | 0.641862 |
|  |  | 90 | 0.012158 | 0.374220 | 0.070000 | 0.647191 | 0.343131 | 0.604162 |
|  | 0.20 | 20 | 0.072757 | 0.357243 | 0.070005 | 0.688818 | 0.428718 | 0.667923 |
|  |  | 50 | 0.020000 | 0.348958 | 0.070000 | 0.662306 | 0.365868 | 0.659111 |
|  |  | 90 | 0.002144 | 0.292983 | 0.070000 | 0.622817 | 0.343073 | 0.657824 |
| 0.06 | 0.12 | 20 | 0.046324 | 0.193095 | 0.041704 | 0.575521 | 0.407688 | 0.649596 |
|  |  | 50 | 0.042542 | 0.146035 | 0.040018 | 0.550282 | 0.349271 | 0.623053 |
|  |  | 90 | 0.035325 | 0.080505 | 0.040017 | 0.545392 | 0.327534 | 0.622236 |
|  | 0.15 | 20 | 0.059598 | 0.334418 | 0.041511 | 0.557115 | 0.407108 | 0.685204 |
|  |  | 50 | 0.051860 | 0.290584 | 0.040073 | 0.502591 | 0.348482 | 0.600330 |
|  |  | 90 | 0.015826 | 0.263941 | 0.039999 | 0.482067 | 0.327645 | 0.600231 |
|  | 0.20 | 20 | 0.058381 | 0.366643 | 0.041210 | 0.422684 | 0.408050 | 0.583953 |
|  |  | 50 | 0.043795 | 0.205674 | 0.040177 | 0.377713 | 0.349991 | 0.569713 |
|  |  | 90 | 0.039152 | 0.202999 | 0.040012 | 0.351268 | 0.327979 | 0.568386 |
| 0.09 | 0.12 | 20 | 0.024845 | 0.190314 | 0.018532 | 0.223233 | 0.391976 | 0.542166 |
|  |  | 50 | 0.005821 | 0.282052 | 0.013171 | 0.233392 | 0.337880 | 0.551094 |
|  |  | 90 | 0.003557 | 0.010874 | 0.011603 | 0.210933 | 0.319659 | 0.552079 |
|  | 0.15 | 20 | 0.040859 | 0.167709 | 0.017899 | 0.196935 | 0.391373 | 0.552278 |
|  |  | 50 | 0.021317 | 0.008981 | 0.013323 | 0.191992 | 0.339088 | 0.538764 |
|  |  | 90 | 0.020707 | 0.008685 | 0.011186 | 0.191486 | 0.319373 | 0.535741 |
|  | 0.20 | 20 | 0.025665 | 0.115874 | 0.016674 | 0.193710 | 0.390411 | 0.499345 |
|  |  | 50 | 0.019843 | 0.071383 | 0.012421 | 0.183469 | 0.338378 | 0.491407 |
|  |  | 90 | 0.015508 | 0.021350 | 0.011075 | 0.175713 | 0.319515 | 0.490788 |

# Estimating the Parameter of Exponential Distribution under Type II Censoring From Fuzzy Data 

Iman Makhdoom<br>Payame Noor University Tehran, Iran

Parviz Nasiri<br>Payame Noor University<br>Tehran, Iran

Abbas Pak<br>Shahrekord University<br>Shahrekord, Iran

The problem of estimating the parameter of Exponential distribution on the basis of type II censoring scheme is considered when the available data are in the form of fuzzy numbers. The Bayes estimate of the unknown parameter is obtained by using the approximation forms of Lindley (1980) and Tierney and Kadane (1986) under the assumption of gamma prior. The highest posterior density (HPD) estimate of the parameter of interest is found. A Monte Carlo simulation is used to compare the performances of the different methods. A real data set is investigated to illustrate the applicability of the proposed methods.

Keywords: Type II censoring, fuzzy lifetime data, exponential distribution, Bayesian estimation

## Introduction

In life testing and reliability studies, the experimenter may not always obtain complete information on failure times for all experimental units. Data obtained from such experiments are called censored data. One of the most common censoring scheme is Type II (failure) censoring, where the life testing experiment will be terminated upon the $r^{\text {th }}$ ( $r$ is pre-fixed) failure. This scheme is often adopted for toxicology experiments and life testing applications by engineers as it has been proven to save time and money. Several authors have addressed inferential issues based on Type II censored samples; for example, Ng , Kundu, and Balakrishnan (2006) discussed point and interval estimation for the two parameter Birnbaum-Saunders distribution base on Type II censored samples. Balakrishnan and Han (2008) considered inference for a simple step-stress model

[^28]
## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

from exponential distribution under Type II censoring. Iliopoulos and Balakrishnan (2011) studied likelihood inference for Laplace distribution based on Type II censored samples. Dey and Kuo (1991) obtained a new class of empirical Bayes estimator for exponential distribution parameter from Type II censored data. Singh and Kumar (2007) considered Bayesian estimation of the exponential parameter under a multiply Type II Censoring scheme. Kundu and Raqab (2012) addressed Bayesian inference for Weibull distribution under Type II censoring scheme.

The above research results are based on precise lifetime data. However, in real situations, some collected data might be imprecise quantities. For instance, the lifetime of a battery may be reported as: 'about $1000 h$ ', 'approximately $1400 h$ ', 'almost between $1000 h$ and $1200 h$ ', 'essentially less than $1200 h$ ', and so on. The lack of precision of such data can be described using fuzzy sets. The classical statistical estimation methods are not appropriate to deal with such imprecise cases. Therefore, the conventional procedures used for estimating the parameter of Exponential distribution will have to be adapted to the new situation.

In recent years, several researchers considered applying the fuzzy sets to estimation theory. Gertner and Zhu (1996) considered Bayesian estimation in forest surveys when samples or prior information are fuzzy. Huang, Zuo, and Sun (2006) proposed a new method to determine the membership function of the estimates of the parameters and the reliability function of multiparameter lifetime distributions. Coppi, Gil, and Kiers (1991) presented some applications of fuzzy techniques in statistical analysis. Akbari and Rezaei (2007) proposed a new method for uniformly minimum variance unbiased fuzzy point estimation. Pak, Parham, and Saraj $(2013,2014)$ conducted a series of studies to develop the inferential procedures for the lifetime distributions on the basis of fuzzy numbers. However, there are no reports on estimating the parameter of Exponential parameter from Type II fuzzy censored data. Hence, the purpose of this study is to consider Bayesian estimation of the parameter of Exponential distribution under Type II censoring scheme when the lifetime observations are reported in the form of fuzzy numbers.

Below are the main definitions of fuzzy sets and some of the formula:

Definition 1: Let $X$ be a universe set. A fuzzy set $\tilde{A}$ in $X$ is defined by a membership function $\mu_{\tilde{A}}(x) \rightarrow[0,1]$, where $\mu_{\tilde{A}}(x), \forall x \in X$, indicates the degree of $x$ in $A$.

## MAKHDOOM ET AL.

Definition 2: A fuzzy subset $\tilde{A}$ of the universe set $X$ is normal iff $\sup _{x \in X} \mu_{\tilde{A}}(x)=1$, where $X$ is the universe set.
Definition 3: A fuzzy subset $\tilde{A}$ of universe set $X$ is convex iff $\mu_{\tilde{A}}(\lambda x+(1-\lambda) y) \geq \min \left\{\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y)\right\}, \forall x, y \in X, \forall \lambda \in[0,1]$.
Definition 4: A fuzzy set $\tilde{x}$ is a fuzzy number iff $\tilde{x}$ be normal and convex on $X$.

In all of fuzzy types of presentation, $L R$-type fuzzy numbers are most used as in linguistic, decision making, knowledge representation, medical diagnosis, control systems, databases. Therefore, we shall focus on the set of $L R$-type fuzzy numbers.

Suppose that $L: \mathbb{R}^{+} \rightarrow[0,1]$ and $R: \mathbb{R}^{+} \rightarrow[0,1]$ be two continuous functions with the following properties:

1) $\quad L(-x)=L(x), R(-x)=R(x)$.
2) $L(0)=1, R(0)=1$.
3) $\quad L$ and $R$ be decreasing in $[0, \infty)$.
4) $\quad \lim _{x \rightarrow \infty} L(x)=0, \lim _{x \rightarrow \infty} R(x)=0$

Definition 5: A fuzzy number $\tilde{x}$ is said to be an $L R$-type fuzzy number if

$$
\mu_{\tilde{x}}(x)= \begin{cases}L\left(\frac{m-x}{\alpha}\right) & x \leq m \\ R\left(\frac{x-m}{\beta}\right) & x \geq m\end{cases}
$$

where $m$ characterizes the mean value of $\tilde{x}$, while $\alpha$ and $\beta$ are the left and the right coefficient of fuzziness, respectively. Symbolically, the $L R$-type fuzzy number is denoted by $\tilde{x}=(\alpha, m, \beta)$.

## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

## Data, likelihood and parameter estimation

Consider a generalization of the likelihood function based on Type II censoring when the lifetime observations are reported in the form of $L R$-type fuzzy numbers. The Bayes estimate of the unknown parameter will then be obtained using suitable conjugate prior of the unknown parameter, and the highest posterior density estimation will be discussed.

## Fuzzy lifetime data and the likelihood function

Suppose that $n$ independent units are placed on a life test with the corresponding lifetimes $X_{1}, \ldots, X_{n}$. It is assumed that these variables are independent and identically distributed as Exponential $E(\lambda)$, with probability density function (pdf)

$$
\begin{equation*}
f(x ; \lambda)=\lambda \exp (-\lambda x), x>o, \lambda>0 . \tag{1}
\end{equation*}
$$

Prior to the experiment, a number $r<n$ is determined and the experiment is terminated after the $r^{\text {th }}$ failure. Now consider the problem where under the Type II censoring scheme, failure times are not observed precisely and only partial information about them are available in the form of fuzzy numbers $\tilde{x}_{i}=\left(\alpha_{i}, m_{i}, \beta_{i}\right), I=1, \ldots, r$, with the corresponding membership functions $\mu_{\tilde{x}_{1}}\left(x_{1}\right), \ldots, \mu_{\tilde{x}_{r}}\left(x_{r}\right)$. Let the maximum value of the means of these fuzzy numbers to be $m_{(\mathrm{r})}$. The lifetime of $n-r$ surviving units, which are removed from the test after the $m^{\text {th }}$ failure, can be encoded as fuzzy numbers $\tilde{x}_{r+1}, \ldots, \tilde{x}_{n}$ with the membership functions

$$
\mu_{\tilde{x}_{j}}(x)=\left\{\begin{array}{ll}
0 & x \leq m_{(r)} \\
1 & x>m_{(r)}
\end{array}, \quad j=r+1, \ldots, n .\right.
$$

The fuzzy data $\tilde{\mathbf{x}}=\left(\tilde{x}_{1}, \ldots, \tilde{x}_{n}\right)$ is thus the vector of observed lifetimes. Then, by using Zadeh's definition of the probability of a fuzzy event (Zadeh, 1968), the corresponding observed-data likelihood function can be obtained as

$$
\begin{equation*}
\ell(\tilde{\mathbf{x}} ; \lambda)=\lambda^{r} \exp \left[-(n-r) \lambda m_{(r)}\right] \prod_{i=1}^{r} \int \exp (-\lambda x) \mu_{\bar{x}_{i}}(x) d x . \tag{2}
\end{equation*}
$$

## MAKHDOOM ET AL.

## Bayesian estimation

In recent decades, the Bayes viewpoint, as a powerful and valid alternative to traditional statistical perspectives, has received frequent attention for statistical inference. Consider the Bayesian estimation of the unknown parameter $\lambda$. As conjugate prior for $\lambda$, we take the $\operatorname{Gamma}(a, b)$ density with pdf given by

$$
\begin{equation*}
\pi(\lambda)=\frac{b^{a}}{\Gamma(a)} \lambda^{a-1} \exp (-\lambda b), \lambda>0 \tag{3}
\end{equation*}
$$

where $a>0$ and $b>0$. Based on this prior, the posterior density function of $\lambda$ given the data can be written as follows:

$$
\begin{equation*}
\pi(\lambda \mid \tilde{\mathbf{x}})=\frac{\lambda^{r+a-1} \exp \left[-\lambda(n-r) m_{(r)}+b\right] \prod_{i=1}^{r} \int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}{\int_{0}^{\infty} \lambda^{r+a-1} \exp \left[-\lambda(n-r) m_{(r)}+b\right] \prod_{i=1}^{r}\left(\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x\right) d \lambda} \tag{4}
\end{equation*}
$$

Then, under a squared error loss function, the Bayes estimate of any function of $\lambda$, say $g(\lambda)$, is

$$
\begin{equation*}
E(h(\lambda) \mid \tilde{\mathbf{x}})=\frac{\int_{0}^{\infty} h(\lambda) \lambda^{r+a-1} \exp \left[-\lambda(n-r) m_{(r)}+b\right] \prod_{i=1}^{r}\left(\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x\right) d \lambda}{\int_{0}^{\infty} \lambda^{r+a-1} \exp \left[-\lambda(n-r) m_{(r)}+b\right] \prod_{i=1}^{r}\left(\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x\right) d \lambda} \tag{5}
\end{equation*}
$$

Note that (5) can not be obtained analytically; therefore, adopt two approximations-Lindley's approximation and Tierney and Kadane's approximation for computing the Bayes estimate.

## Lindley's approximation

Setting $F(\lambda)=\ln \pi(\lambda)+\ln \ell(\tilde{\mathbf{x}} ; \lambda) \equiv \vartheta(\lambda)+L(\lambda),(5)$ can be rewritten as

## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

$$
\begin{equation*}
E(h(\lambda) \mid \tilde{\mathbf{x}})=\frac{\int_{0}^{\infty} h(\lambda) e^{F(\lambda)} d \lambda}{\int_{0}^{\infty} e^{F(\lambda)} d \lambda} . \tag{6}
\end{equation*}
$$

Then, by using Lindley's approximation (see Lindley, 1980), the ratio of the two integrals in (6) can be obtained as

$$
\begin{equation*}
h(\lambda)+\frac{1}{2} h_{11} \delta_{11}+\vartheta_{1} h_{1} \delta_{11}+\frac{1}{2} F_{3} \delta_{11}^{2} h_{1}, \tag{7}
\end{equation*}
$$

where

$$
\begin{gathered}
h_{1}=\frac{d h(\lambda)}{d \lambda}, h_{11}=\frac{d^{2} h(\lambda)}{d \lambda^{2}}, \vartheta_{1}=\frac{d \vartheta(\lambda)}{d \lambda} \\
F_{3}=\frac{\partial^{3} F(\lambda)}{\partial \lambda^{3}}, \delta_{11}=\left[-\frac{\partial^{2} F(\lambda)}{\partial \lambda^{2}}\right]^{-1} .
\end{gathered}
$$

Evaluating all the expressions in (7) at the maximum likelihood estimate (MLE) of $\lambda$ produces the approximation $\hat{h}_{B}$ to (6). In this case,

$$
F(\lambda)=r \log \lambda+\sum_{i=1}^{r} \log \int \exp (-\lambda x) \mu_{x_{i}}(x) d x-(n-r) \lambda m_{(r)} .
$$

The MLE of $\lambda$, say $\hat{\lambda}$, is the solution of the equation

$$
\frac{\partial F(\lambda)}{\partial \lambda}=\frac{r}{\lambda}-\sum_{i=1}^{r} \frac{\int x \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}{\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}-(n-r) m_{(r)}=0
$$

Now, to apply Lindley's form in (7), first obtain

$$
\delta_{11}=-\frac{r}{\hat{\lambda}^{2}}+\sum_{i=1}^{r}\left\{\frac{\int x^{2} \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x}-\left[\frac{\int x \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x}\right]^{2}\right\}
$$

## MAKHDOOM ET AL.

$$
\begin{aligned}
& F_{3}=\frac{2 r}{\hat{\lambda}^{3}}-\sum_{i=1}^{r}\left\{\begin{array}{l}
\frac{\int x^{3} \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x} \\
-\frac{\left\lfloor x^{2} \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x\right\rfloor\left[x \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x\right]}{\left[\int \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x\right]}
\end{array}\right\} \\
& +2 \sum_{i=1}^{r}\left\{\frac{\int x \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{\tilde{x}_{i}}(x) d x} \times\binom{\frac{\int x^{2} \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{x_{i}}(x) d x}}{-\left[\frac{\int x \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}{\int \exp (-\hat{\lambda} x) \mu_{\hat{x}_{i}}(x) d x}\right]}\right\}
\end{aligned}
$$

The approximate Bayes of $\lambda$, say $\hat{\lambda}_{B}$, for the squared error loss function is the posterior mean of $h(\lambda)=\lambda$, which is by (7) as follows.

$$
\begin{equation*}
\hat{\lambda}_{B}=\hat{\lambda}+\left[\frac{a-1}{\hat{\lambda}}-b\right] \delta_{11}+\frac{1}{2} F_{3} \delta_{11}^{2} . \tag{8}
\end{equation*}
$$

## Tierney and Kadane's approximation

Setting $W(\lambda)=L(\lambda) / n$ and $W^{*}(\lambda)=[\ln h(\lambda)+L(\lambda)] / n$, the expression in (6) can be re-expressed as

$$
\begin{equation*}
E(h(\hat{\lambda}) \mid \tilde{\mathbf{x}})=\frac{\int_{0}^{\infty} h(\lambda) e^{n W^{*}(\lambda)} d \lambda}{\int_{0}^{\infty} e^{n W(\lambda)} d \lambda} . \tag{9}
\end{equation*}
$$

Following Tierney and Kadane (1986), (9) can be approximated as the following form:

## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

$$
\begin{equation*}
\hat{g}_{B T}(\lambda)=\left(\frac{\left|\phi^{*}\right|}{|\phi|}\right)^{\frac{1}{2}} \exp \left\{n\left[W^{*}\left(\tilde{\lambda}^{*}\right)-W(\tilde{\lambda})\right]\right\}, \tag{10}
\end{equation*}
$$

where $\tilde{\lambda}^{*}$ and $\tilde{\lambda}$ maximize $W^{*}(\lambda)$ and $W(\lambda)$, respectively, and $\phi^{*}$ and $\phi$ are minus the inverse of the second derivatives of $W^{*}(\lambda)$ and $W(\lambda)$ at $\tilde{\lambda}^{*}$ and $\tilde{\lambda}$, respectively.

In this case,

$$
W(\lambda)=\frac{1}{n}\left\{\begin{array}{l}
k+(r+a-1) \log \lambda-\lambda\left[b+(n-r) m_{(r)}\right]  \tag{11}\\
+\sum_{i=1}^{r} \log \int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x
\end{array}\right\},
$$

where $k$ is a constant, and

$$
\begin{equation*}
W^{*}(\lambda)=H(\lambda)+\frac{1}{n} \ln \lambda . \tag{12}
\end{equation*}
$$

Substituting for (11) and (12) in (10), the Bayes estimate $\hat{\lambda}_{B T}$ of a function $h(\lambda)=\lambda$ under squared error loss can then be obtained straightforwardly.

## HPD estimation

The highest posterior density (HPD) estimation is another popular method used by the Bayesian perspective. This method is based on the maximum likelihood principle; hence, it leads to the mode of the posterior density. The HPD estimate, $\hat{\lambda}_{H}$, of $\lambda$ is obtained by solving the equation $\frac{\partial \pi(\lambda \mid \tilde{\mathbf{x}})}{\partial \lambda}=0$ where

$$
\begin{equation*}
\frac{\partial \pi(\lambda \mid \tilde{\mathbf{x}})}{\partial \lambda}=\frac{r+a-1}{\lambda}-\left[b+(n-r) m_{(r)}\right]-\sum_{i=1}^{r} \frac{\int x \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}{\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x} . \tag{13}
\end{equation*}
$$

However, the solution cannot be obtained explicitly. In the following, Theorem 1 discusses the existence and uniqueness of the HPD estimate of $\lambda$.

## MAKHDOOM ET AL.

Theorem 1. Let $g(\lambda)$ denote the function on the right-hand side of the expression in (13). Then the root of the equation $g(\lambda)=0$ exists and is unique.

Proof. From (13) it is easily seen that $\lim _{\lambda \rightarrow \infty} g(\lambda)=\infty$. Also, note that $g(\lambda)<\frac{r+a-1}{\lambda}, \forall \lambda \in(0, \infty)$, and consequently

$$
\lim _{\lambda \rightarrow \infty} g(\lambda)<\lim _{\lambda \rightarrow \infty} \frac{r+a-1}{\lambda}=0 \quad \forall \lambda \in(0, \infty)
$$

Therefore, the equation $g(\lambda)=0$ has at least one root in $(0, \infty)$. To prove that the root is unique, we consider the first derivative of $g, g(' \lambda)$, given by

$$
g(' \lambda)=-\frac{r+a-1}{\lambda^{2}}+\sum_{i=1}^{r} \frac{\partial^{2}}{\partial \lambda^{2}} \log \int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x
$$

Let $u(\lambda)=\exp (-\lambda x)$ and $v_{i}(\lambda)=\int \exp (-\lambda x) \mu_{\hat{x}_{i}}(x) d x$. Then $g$ (' $\left.\lambda\right)$ can be written as

$$
g(' \lambda)=-\frac{r+a-1}{\lambda^{2}}+\sum_{i=1}^{r} \frac{\partial^{2}}{\partial \lambda^{2}} \log v_{i}(\lambda)
$$

It is clearly that $u(\lambda)$ is a log-concave function of $\lambda$, and by the Prekopa-Leindler inequality (see Gardner, 2002) $v_{i}(\lambda), i=1, \ldots, m$, are also log-concave in $\lambda$. It follows that $g$ is a strictly decreasing function w.r.t. $\lambda$ and hence the equation $g(\lambda)=0$ has exactly one solution.

Because there is no closed form of the solution to the equation (13), an iterative numerical search such as Newton-Raphson method can be used to obtain the HPD estimate of $\lambda$. The second-order derivative form required for proceeding with the Newton-Raphson method, is obtained as follows.

$$
\frac{\partial^{2} \pi(\lambda \mid \tilde{\mathbf{x}})}{\partial \lambda^{2}}=-\frac{(r+a-1)}{\lambda^{2}}+\sum_{i=1}^{r}\left\{\frac{\int x^{2} \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}{\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}-\left(\frac{\int x \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}{\int \exp (-\lambda x) \mu_{\tilde{x}_{i}}(x) d x}\right)^{2}\right\}
$$



Figure 1. Fuzzy information system used to encode the simulated data

## Numerical Study

A Monte Carlo simulation study and one example are presented to illustrate the methods of inference developed in this paper. First, for fixed $\theta=1$ and different choices of $n$ and $r$, generated Type II censored samples were generated, say $\mathbf{x}=\left(x_{1}, \ldots, x_{r}\right)$, from the exponential distribution using the method proposed by Aggarwala and Balakrishnan (1998). Each realization of $\mathbf{x}$ was fuzzified using the fuzzy information system (see Pak et al., 2014) shown in Figure 1, corresponding to the membership functions

$$
\mu_{\tilde{x}_{1}}(x)=\left\{\begin{array}{ll}
1 & x \leq 0.25 \\
\frac{0.5-x}{0.25} & 0.25 \leq x \leq 0.5, \\
0 & \text { otherwise },
\end{array} \quad \mu_{\tilde{x}_{2}}(x)= \begin{cases}\frac{x-0.25}{0.25} & 0.25 \leq x \leq 0.5 \\
\frac{0.75-x}{0.25} & 0.5 \leq x \leq 0.75 \\
0 & \text { otherwise }\end{cases}\right.
$$

## MAKHDOOM ET AL.

$$
\begin{aligned}
& \mu_{\tilde{x}_{3}}(x)=\left\{\begin{array}{ll}
\frac{x-0.5}{0.25} & 0.5 \leq x \leq 0.75, \\
\frac{1-x}{0.25} & 0.75 \leq x \leq 1, \\
0 & \text { otherwise },
\end{array} \quad \mu_{\tilde{x}_{4}}(x)= \begin{cases}\frac{x-0.75}{0.25} & 0.75 \leq x \leq 1, \\
\frac{1.25-x}{0.25} & 1 \leq x \leq 1.25, \\
0 & \text { otherwise },\end{cases} \right. \\
& \mu_{\tilde{x}_{5}}(x)=\left\{\begin{array}{ll}
\frac{x-1}{0.25} & 1 \leq x \leq 1.25, \\
\frac{1.5-x}{0.25} & 1.25 \leq x \leq 1.5, \\
0 & \text { otherwise },
\end{array} \quad \mu_{\tilde{x}_{\tilde{x}_{6}}}(x)= \begin{cases}\frac{x-1.25}{0.25} & 1.25 \leq x \leq 1.5, \\
\frac{1.75-x}{0.25} & 1.5 \leq x \leq 1.75, \\
0 & \text { otherwise },\end{cases} \right. \\
& \mu_{\tilde{x}_{7}}(x)=\left\{\begin{array}{ll}
\frac{x-1.5}{0.25} & 1.5 \leq x \leq 1.75, \\
\frac{2-x}{0.25} & 1.75 \leq x \leq 2, \\
0 & \text { otherwise },
\end{array} \quad \mu_{\tilde{x}_{\tilde{x}_{8}}}(x)= \begin{cases}x-1.75 & 1.75 \leq x \leq 2, \\
1 & x \geq 2, \\
0 & \text { otherwise } .\end{cases} \right.
\end{aligned}
$$

Then, the approximate Bayes estimates (via Lindley approximation or Tierney and Kadane approximation) and the HPD estimates of $\lambda$ for the fuzzy sample were computed under the assumption that $\lambda$ has $\operatorname{Gamma}(a, b)$ prior, including the noninformative gamma prior, i.e. $a=b=0$, and informative gamma prior, i.e. $a=b=2$. The average values and mean squared errors of the estimates, computed based on 1000 replication, are presented in Tables 1 and 2.

In viewing the tables, using Lindley approximation or Tierney and Kadane approximation for the computation of Bayes estimates gave similar estimation results. The performance of HPD estimates are better than the Bayes estimates in terms of MSE. Also, the approximate Bayes estimates based on informative prior are uniformly better than that of non-informative prior. In all the cases, it was observed that as the effective sample size $m$ increases the performances in terms of MSE become better.

## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

Table 1. Average values (AV) and mean squared errors (MSE) of the Bayes and HPD estimates of $\lambda$ based on non-informative prior $(a=b=0)$ and for different sample sizes.

| $n$ | $r$ |  | $\hat{\lambda}_{B}$ |  | $\hat{\lambda}_{B T}$ |  |  |  | $\hat{\lambda}_{H}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: | :---: |
|  |  | AV | MSE | AV | MSE | AV | MSE |  |  |
| 20 | 10 | 1.081 | 0.175 | 1.082 | 0.175 | 1.061 | 0.147 |  |  |
| 20 | 12 | 1.078 | 0.156 | 1.078 | 0.155 | 1.057 | 0.131 |  |  |
| 20 | 15 | 1.055 | 0.124 | 1.030 | 0.123 | 1.045 | 0.115 |  |  |
| 30 | 15 | 1.092 | 0.104 | 1.091 | 0.104 | 1.075 | 0.085 |  |  |
| 30 | 20 | 1.065 | 0.096 | 1.065 | 0.096 | 1.052 | 0.066 |  |  |
| 30 | 25 | 1.040 | 0.071 | 1.041 | 0.071 | 1.028 | 0.048 |  |  |
| 50 | 20 | 1.051 | 0.098 | 1.050 | 0.098 | 1.040 | 0.073 |  |  |
| 50 | 25 | 1.034 | 0.055 | 1.034 | 0.054 | 1.026 | 0.037 |  |  |
| 50 | 35 | 1.021 | 0.037 | 1.021 | 0.037 | 1.018 | 0.029 |  |  |

Table 2. Average values (AV) and mean squared errors (MSE) of the Bayes and HPD estimates of $\lambda$ based on informative prior $(a=b=2)$ and for different sample sizes.

| $n$ | $r$ |  |  |  | $\hat{\lambda}_{B}$ | $\hat{\lambda}_{B T}$ |  |  | $\hat{\lambda}_{H}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: | :---: |
|  |  | AV | MSE | AV | MSE | AV | MSE |  |  |
| 20 | 10 | 1.069 | 0.151 | 1.068 | 0.152 | 1.047 | 0.129 |  |  |
| 20 | 12 | 1.059 | 0.133 | 1.059 | 0.132 | 1.036 | 0.117 |  |  |
| 20 | 15 | 1.038 | 0.105 | 1.038 | 0.105 | 1.030 | 0.092 |  |  |
| 30 | 15 | 1.077 | 0.081 | 1.076 | 0.080 | 1.056 | 0.070 |  |  |
| 30 | 20 | 1.051 | 0.067 | 1.051 | 0.067 | 1.041 | 0.051 |  |  |
| 30 | 25 | 1.024 | 0.052 | 1.024 | 0.053 | 1.017 | 0.033 |  |  |
| 50 | 20 | 1.040 | 0.079 | 1.041 | 0.078 | 1.028 | 0.056 |  |  |
| 50 | 25 | 1.019 | 0.041 | 1.018 | 0.041 | 1.015 | 0.025 |  |  |
| 50 | 35 | 1.012 | 0.020 | 1.012 | 0.020 | 1.007 | 0.014 |  |  |

## Application example

To demonstrate the application of the proposed methods to real data, consider the following life-testing experiment in which $n=22$ identical valves are placed on test. The unknown lifetime $x_{i}$ of valve $i$ may be regarded as a realization of a random variable $X_{i}$, induced by random sampling from a total population of valves, which is distributed as Exponential by an unknown parameter of $\lambda$. A tested valve may be considered as failed, or -strictly speaking- as nonconforming, when at least one value of its parameters falls beyond specification limits. In practice, however, there isn't the possibility to measure all parameters and are not able to define precisely the moment of a failure. So, the observed failure times (in $100 h$ ) are reported in the form of lower and upper bounds, as well as a point estimate which are as follows.

## MAKHDOOM ET AL.

## Data Set:

(20.68, 25.86, 29.73), (22.72, 28.41, 32.67), (24.61, 30.77, 35.38), (26.43, 33.04, 37.99), (28.15, 35.19, 40.46), (30.29, 37.87, 43.55), (34.32, 42.91, 49.34), (35.51, 44.39, 51.04), (37.80, 47.25, 54.33),
(41.16, 51.45, 59.16), (42.52, 53.16, 61.13), (43.97, 54.97, 63.21),
(44.31, 55.39, 63.69), (46.75, 58.44, 67.20), (47.69, 59.62, 68.56),
(48.09, 60.12, 69.13), (52.27, 65.34, 75.14), (53.65, 67.07, 77.13),
(60.72, 75.91, 87.29), (63.45, 79.32, 91.21), (65.69, 82.12, 94.43), (73.48, 91.86, 105.63).

Each triple is modeled by a triangular fuzzy number $\tilde{x}_{i}$, and is interpreted as a possibility distribution related to an unknown value $x_{i}$, itself a realization of a random variable $X_{i}$. Randomness arises from the selection of objects from the total population of batteries. In contrast, fuzziness arises from the limited ability of the observer to describe the moment of a failure using numbers, which is not influenced by random factors. Consider Type II censored samples of size $r=12,15,20$ from the above data and compute the estimate of $\lambda$ using the Bayes and HPD procedures under the assumption of non-informative and informative priors. All the results are summarized in Table 3.

Table 3. Bayes and HPD estimates for application example.

| $r$ |  | $a=b=0$ |  |  | $a=b=2$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\lambda}_{B}$ | $\hat{\lambda}_{B T}$ | $\hat{\lambda}_{H}$ | $\hat{\lambda}_{B}$ | $\hat{\lambda}_{B T}$ | $\hat{\lambda}_{H}$ |  |
| 12 | 0.0118 | 0.0117 | 0.0107 | 0.0136 | 0.0135 | 0.0126 |  |
| 15 | 0.0141 | 0.0140 | 0.0131 | 0.0158 | 0.0159 | 0.0152 |  |
| 20 | 0.0163 | 0.0162 | 0.0154 | 0.0181 | 0.0181 | 0.0172 |  |

## Conclusion

Statistical analysis of exponential distribution under Type II censoring is based on precise lifetime data. Precisely reported lifetimes are common when data comes from specially designed life tests. In such a case a failure should be precisely defined, and all tested items should be continuously monitored. However, in real situations these test requirements might not be fulfilled. In these cases, it is sometimes impossible to obtain exact observations of lifetime. The obtained

## ESTIMATING THE PARAMETER OF EXPONENTIAL DISTRIBUTION

lifetime data may be imprecise most of the time. Therefore, a suitable statistical methodology is needed to handle these data as well.

The Bayesian inference for the exponential distribution parameter under Type II censoring was addressed when the lifetime observations are fuzzy numbers. Based on the results of the simulation study, the HPD procedure produces the estimates with smaller MSE than the Bayes estimates. Using the informative prior for computing the approximate Bayes estimates provides an improvement in the estimates in terms of MSE.

## References

Aggarwala, R., \& Balakrishnan, N. (1998). Some properties of progressively censored order statistics from arbitrary and uniform distributions with applications to inference and simulation. Journal of Statistical Planning and Inference, 70(1), 35-49. doi: 10.1016/S0378-3758(97)00173-0

Akbari, M. G., \& Rezaei, A. (2007). An uniformly minimum variance unbiased point estimator using fuzzy observations. Austrian Journal of Statistics, 36(4), 307-317.

Balakrishnan, N., \& Han, D. (2008). Exact inference for a simple step-stress model with competing risks for failure from exponential distribution under TypeII censoring. Journal of Statistical Planning and Inference, 138(12), 4172-4186. doi: 10.1016/j.jspi.2008.03.036

Coppi, R., Gil, M. A., \& Kiers, H. A. L. (2006). The fuzzy approach to statistical analysis. Computational Statistics \& Data Analysis, 51(1), 1-14. doi: 10.1016/j.csda.2006.05.012

Dey, D. K., \& Kuo, L. (1991). A new empirical Bayes estimator with type II censored data. Computational Statistics \& Data Analysis, 12(3), 271-279. doi: 10.1016/0167-9473(91)90111-E

Gardner, R. J. (2002). The Brunn-Minkowski inequality. Bulletin of the American Mathematical Society, 39(3), 355-405. doi: 10.1090/S0273-0979-02-00941-2

Gertner, G. Z., \& Zhu, H. (1996). Bayesian estimation in forest surveys when samples or prior information are fuzzy. Fuzzy Sets and Systems, 77(3), 277-290. doi: 10.1016/0165-0114(95)00088-7

## MAKHDOOM ET AL.

Huang, H.-Z., Zuo, M. J., \& Sun, Z.-Q. (2006). Bayesian reliability analysis for fuzzy lifetime data. Fuzzy Sets and Systems, 157(12), 1674-1686. doi: 10.1016/j.fss.2005.11.009

Iliopoulos, G., \& Balakrishnan, N. (2011). Exact likelihood inference for Laplace distribution based on Type-II censored samples. Journal of Statistical Planning and Inference, 141(3), 1224-1239. doi: 10.1016/j.jspi.2010.09.024

Kundu, D., \& Raqab, M. Z. (2012). Bayesian inference and prediction of order statistics for a Type-II censored Weibull distribution. Journal of Statistical Planning and Inference, 142(1), 41-47. doi: 10.1016/j.jspi.2011.06.019

Lindley, D. V. (1980) Approximate Bayesian methods. Trabajos de Estadistica Y de Investigacion Operativa, 31(1), 223-237. doi: 10.1007/BF02888353

Ng, H. K. T., Kundu, D., \& Balakrishnan N. (2006). Point and interval estimation for the two-parameter Birnbaum-Saunders distribution based on TypeII censored samples. Computational Statistics \& Data Analysis, 50(11), 3222-3242. doi: 10.1016/j.csda.2005.06.002

Pak, A., Parham, G. H., \& Saraj, M. (2013). Inference for the Weibull distribution based on fuzzy data. Revista Colombiana de Estadística, 36(2), 339358.

Pak, A., Parham, G. H., \& Saraj, M. (2014). Inference for the Rayleigh distribution based on progressive Type-II fuzzy censored data. Journal of Modern Applied Statistical Methods, 13(1), 287-304. Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss1/19

Tierney, L., \& Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81(393), 82-86. doi: 10.1080/01621459.1986.10478240

Singh, U., \& Kumar, A. (2007). Bayesian estimation of the exponential parameter under a multiply Type II censoring scheme. Austrian Journal of Statistics, 36(3), 227-238.

Zadeh, L. A. (1968). Probability measures of fuzzy events. Journal of Mathematical Analysis and Applications, 10, 421-427. doi: 10.1016/0022-247X(68)90078-4

# The Application of Legendre Multiwavelet Functions in Image Compression 

Elham Hashemizadeh<br>Islamic Azad University,<br>Karaj, Iran

Sohrab Rahbar<br>Islamic Azad University<br>Karaj, Iran

Legendre multiwavelets are introduced. These functions can be designed in such a way that the properties of orthogonality, polynomial approximation, and symmetry hold at the same time. In this way, they can be effectively deployed in image compression.

Keywords: Image compression, Legendre multiwavelet, multiwavelets, preprocessing, wavelets

## Introduction

In recent years, the wavelets theory has played a significant role in signal processing, especially in image processing. These wavelets are mainly scalar, where there is only one scaling function. However, multiwavelets are based on more than one scaling function. Such growing interests in multiwavelets mainly stem from the following facts: (i) multiwavelets can simultaneously possess orthogonality, symmetry, and a high order of approximation for a given support of the scaling functions (this is not possible for any real valued scaler wavelets); and (ii) multiwavelets have produced promising results in the areas of image compression.

A multiwavelet system can provide perfect reconstruction while preserving length (orthogonality), good performance at the boundaries (via linear-phase symmetry), and a high order of approximation (vanishing moments). Thus, multiwavelets offer the possibility of superior performance for image processing applications, compared with scalar wavelets. In this paper, we use linear Legendre multiwavelets in image compression, and show its usefulness through actual examples.

[^29]
## HASHEMIZADEH \& RAHBAR

## Multiwavelets

As with scalar wavelets, multiwavelets are based on the concept of multiresolution analysis (MRA). The only difference is the number of scaling functions used to generate those functions. The standard multiresolution has one scaling function $\phi(t)$, which satisfies the following properties (Aboufadel \& Schilicker, 1999; Martin \& Bell, 2001, Burrus \& Gopinath, 1998; Daubechies, 1992).

- The translates $\phi(t-k)$ are linearly independent and produce a basis for the subspace $\phi_{1}(t), \ldots, \phi_{N}(t)$.
- The dilates $\phi\left(2^{j} t-k\right)$ generate subspaces $V_{j}, j \in Z$, such that $\ldots \subset V_{-1} \subset V_{0} \subset V_{1} \subset \ldots \subset V_{j} \subset \ldots \overline{\smile_{j=-\infty}^{\infty} V_{j}}=L^{2}(R), \cap_{j=-\infty}^{\infty} V_{j}=\{0\}$.
- To obtain the subspace $V_{1}$, it is sufficient to add the family of functions $\psi(t-k), W_{0}$, to $V_{0}$, i.e., $V_{1}=V_{0} \oplus W_{0}$.

In multiwavelets, the notion of MRA is used in the same way except that the basis for $V_{0}$ is generated by the translates of $N$ different scaling functions $\phi_{1}(t-k)$, $\phi_{2}(t-k), \ldots, \phi_{N}(t-k)$. As in the scalar case, the vector $\Phi(t)=\left[\phi_{1}(t), \ldots, \phi_{N}(t)\right]^{T}$, satisfies the matrix dilation equation

$$
\Phi(t)=\sum_{k} C[k] \Phi(2 t-k)
$$

where $C[k]$ is an $N$ by $N$ matrix of coefficients. There are also $N$ wavelets $\psi_{1}(t), \ldots, \psi_{N}(t)$ satisfying the matrix wavelet equation

$$
\Psi(t)=\sum_{k} D[k] \Phi(2 t-k)
$$

where $\Psi(t)=\left[\psi_{1}(t), \ldots, \psi_{N}(t)\right]^{T}$ is a vector, and $D[k]$ is an $N$ by $N$ matrix.
The scaling functions $\phi_{1}(t), \ldots, \phi_{N}(t)$ are in $V_{0}$ whose basis is $\left\{\sqrt{2} \phi_{i}(2 t-k)\right.$ : $1 \leq i \leq N, k \in Z\}$. Thus, the scaling function and the multiwavelet functions have to satisfy the two-scale dilation equations

$$
\Phi(t)=\sqrt{2} \sum_{k} H_{k} \Phi(2 t-k) \text { and } \Psi(t)=\sqrt{2} \sum_{k} G_{k} \Phi(2 t-k),
$$

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

where $G_{k}$ and $H_{k} \in l^{2}(2)^{N \times N}$ are $N \times N$ matrices of filter coefficients. Moreover, the set of scaling functions $\quad\left\{\phi_{i, j, k}(t): 1 \leq i \leq N, k \in Z\right\} \quad$ with $\phi_{i j, k}(t)=2^{-j / 2} \phi_{i}\left(t / 2^{j}-k\right)$ is a Riesz basis for $V_{i}$. We also write

$$
\left[\varphi_{1, j, k}(t), \ldots, \varphi_{N, j, k}(t)\right]^{T}=\Phi_{j, k}(t) .
$$

In the case of orthonormal multiscaling functions, $\left\{\phi_{i}(t): 1 \leq i \leq N, k \in Z\right\}$ is not just a Riesz basis, but is orthonormal, i.e.,

$$
\langle\Phi(t), \Phi(t-k)\rangle=\int \Phi(t) \Phi^{T}(t-k) d t=I_{N} \delta_{k, 0}, k \in Z
$$

where $I_{N}$ is the $N \times N$ identity matrix. This implies that

$$
\sum_{k} G_{k} G_{2 l+k}^{T}=I_{r} \delta_{0, l} l \in Z,
$$

In scalar wavelets, this means that the sum of squares of low-pass filter coefficients equals unity, and the filter is orthogonal to its even translates.

## Linear Legendre Multiwavelets

A pair of linear Legendre scaling functions $\phi_{1}(x)$ and $\phi_{2}(x)$ on [ 0,1$]$ are introduced and depicted in Figure 1.

$$
\phi_{1}(x)=1, \text { and } \phi_{2}(x)=\sqrt{3}(2 x-1) .
$$

## HASHEMIZADEH \& RAHBAR



Figure 1. A pair of linear Legendre scaling functions. The left is $\phi_{1}(x)$ and the right is $\phi_{2}(x)$.

The integer translates of $\phi_{1}(x)$ together with the ones of $\phi_{2}(x)$ span a subspace $V_{0}$. Furthermore, the translates of $1 / 2$ scaled version of $\phi_{1}(x)$ and $\phi_{2}(x)$ that span $V_{1} \supset V_{0}$ are given by

$$
\left\{\begin{array}{cc}
\varphi_{1}(2 x-k)=1 & \text { for } \frac{k}{2} \leq x<\frac{k+1}{2} \\
\varphi_{2}(2 x-k)=\sqrt{3}(4 x-2 k-1) & \text { for } \frac{k}{2} \leq x<\frac{k+1}{2}
\end{array}\right.
$$

The two-scale relations for linear Legendre scaling functions should express each of $\phi_{1}(x)$ and $\phi_{2}(x)$ in terms of the four scaling functions $\phi_{i}(2 x)$ and $\phi_{i}(2 x-1)$, $i=1,2$ (Udea \& Lodha, 1995). We propose these relations to be

$$
\left[\begin{array}{l}
\phi_{1}(x)  \tag{1}\\
\phi_{2}(x)
\end{array}\right]=\left[\begin{array}{cccc}
1 & 0 & 1 & 0 \\
-\frac{\sqrt{3}}{2} & \frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2}
\end{array}\right]=\left[\begin{array}{c}
\phi_{1}(2 x) \\
\phi_{2}(2 x) \\
\phi_{1}(2 x-1) \\
\phi_{2}(2 x-1)
\end{array}\right]
$$

Let $\psi_{1}(x)$ and $\psi_{2}(x)$ be wavelets on [0,1] defined based on the linear Legendre scaling functions. By using (1), $\psi_{1}(x)$ and $\psi_{2}(x)$ can be expressed in terms of $\phi_{i}(2 x)$ and $\phi_{i}(2 x-1), i=1,2$. The conditions of being orthonormal and vanishing moments help reduce the two-scale relation for linear Legendre multiwavelets as follows.

$$
\left[\begin{array}{l}
\psi_{1}(x)  \tag{2}\\
\psi_{2}(x)
\end{array}\right]=\left[\begin{array}{cccc}
0 & -1 & 0 & 1 \\
\frac{1}{2} & \frac{\sqrt{3}}{2} & -\frac{1}{2} & \frac{\sqrt{3}}{2}
\end{array}\right]\left[\begin{array}{c}
\phi_{1}(2 x) \\
\phi_{2}(2 x) \\
\phi_{1}(2 x-1) \\
\phi_{2}(2 x-1)
\end{array}\right]
$$

Then, the explicit formulae for $\psi_{1}(x)$ and $\psi_{2}(x)$ are

$$
\psi_{1}(x)= \pm\left\{\begin{array}{ll}
-\sqrt{3}(4 x-1), & 0 \leq x<\frac{1}{2} \\
+\sqrt{3}(4 x-3), & \frac{1}{2} \leq x<1
\end{array}, \text { and } \psi_{2}(x)= \pm\left\{\begin{array}{ll}
6 x-1, & 0 \leq x<\frac{1}{2} \\
6 x-5, & \frac{1}{2} \leq x<1
\end{array} .\right.\right.
$$

## Decomposition relations

The matrix equation

$$
\left[\begin{array}{l}
\phi_{1}(x)  \tag{3}\\
\phi_{2}(x) \\
\psi_{1}(x) \\
\psi_{2}(x)
\end{array}\right]=\left[\begin{array}{cccc}
1 & 0 & 1 & 0 \\
-\frac{\sqrt{3}}{2} & \frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} \\
0 & -1 & 0 & 1 \\
\frac{1}{2} & \frac{\sqrt{3}}{2} & -\frac{1}{2} & \frac{\sqrt{3}}{2}
\end{array}\right]\left[\begin{array}{c}
\phi_{1}(2 x) \\
\phi_{2}(2 x) \\
\phi_{1}(2 x-1) \\
\phi_{2}(2 x-1)
\end{array}\right]
$$

which combines (1) and (2) is called the reconstruction relation. Note that $\phi_{1}(x), \phi_{2}(x) \in V_{0}$ and $\psi_{1}(x), \psi_{2}(x) \in W_{0}$. This four bases are expressed in terms of $\phi_{1}(2 x), \phi_{2}(2 x), \phi_{1}(2 x-1)$ and $\phi_{2}(2 x-1)$ in $V_{1}$ subspace.

The decomposition relation is simply defined as the inverse of the reconstruction relation defined in (3). The square matrix in (3) is orthogonal with constant magnitude. This is because $\left\{\phi_{1}(x), \phi_{2}(x), \psi_{1}(x), \psi_{2}(x)\right\}$ and $\left\{\phi_{1}(2 x), \phi_{2}(2 x), \phi_{1}(2 x-1), \phi_{2}(2 x-1)\right\}$. Thanks to the property $M^{-1}=M^{T}$ of any orthogonal matrix, the decomposition relation would be

## HASHEMIZADEH \& RAHBAR

$$
\left[\begin{array}{c}
\varphi_{1}(2 x) \\
\varphi_{2}(2 x) \\
\varphi_{2}(2 x-1) \\
\varphi_{2}(2 x-1)
\end{array}\right]=\frac{1}{2}\left[\begin{array}{cccc}
1 & -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \\
0 & \frac{1}{2} & -1 & \frac{\sqrt{3}}{2} \\
1 & \frac{\sqrt{3}}{2} & 0 & -\frac{1}{2} \\
0 & \frac{1}{2} & 1 & \frac{\sqrt{3}}{2}
\end{array}\right]\left[\begin{array}{c}
\varphi_{1}(x) \\
\varphi_{2}(x) \\
\psi_{1}(x) \\
\psi_{2}(x)
\end{array}\right]
$$

## Preprocessing and sampling

Preprocessing aims to convert the given scalar input signal of length $N$ to a sequence of two-dimensional vectors $v_{0, k}$. This is used in the analysis algorithm as shown in Figure 2. Here, $N$ is assumed to be a multiple of a power of 2, and therefore it even. The signal $\left\{X_{k}\right\}$ is taken to be a function $x($.$) which is observed$ at integer time points. If the preprocessing produces $N$ two-dimensional vectors, it is said to be an oversampling scheme. If it produces $N / 2$ two-dimensional vectors, the result is a critical sampling. After the wavelet reconstruction step (synthesis), e.g., inverse DMWT, a postfilter is applied. Evidently, prefiltering, wavelet transform, inverse transform, and postfiltering should exactly recover the input signal.


Figure 2. A multiwavelet filter bank which is iterated once.

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

## Repeated Row Preprocessing: Oversampling Scheme

The obvious way to get two input rows from a given signal is to repeat the signal. Two identical rows go into the multifilter bank. It then begins oversampling of the data by a factor of two. Although, the usefulness of oversampling has been proven, it requires more calculations than critical sampling. Furthermore, in data compression applications, one intends to remove redundancy, not to increase it. In the case of one-dimensional signals the repeated row scheme is convenient to be implemented (Strela \& Walden, 1998).

## Matrix (approximation) Preprocessing: Critical Sampling

Let $x(t)$ belongs to $V_{0}$ and is generated by the translates of linear Legendre multiscaling functions so that $x(t)=\sum_{k}\left[v_{0, k}^{(0)} \varphi_{1}(t-k)+v_{0, k}^{(1)} \varphi_{2}(t-k)\right]$.

Suppose that the input sequence samples are made at the half-integers, i.e., $X_{2 k}=x(k)$, and $X_{2 k+1}=x\left(k+\frac{1}{2}\right)$.

From Fig. 1 for the linear Legendre multiscaling function, the only nonzero function values at the integers and half-integers are $\phi_{1}(0)=1, \phi_{2}\left(\frac{1}{2}\right)=1$, and $\phi_{2}(0)=-\sqrt{3}$.

Thus,

$$
X_{2 k}=v_{0, k}^{(0)}-\sqrt{3} v_{0, k}^{(1)} \text {, and } X_{2 k+1}=v_{0, k}^{(0)} .
$$

Hence,

$$
\begin{equation*}
v_{0, k}^{(0)}=X_{2 k+1}, \tag{4}
\end{equation*}
$$

and

$$
\begin{equation*}
v_{0, k}^{(1)}=\frac{1}{\sqrt{3}} X_{2 k+1}-\frac{1}{\sqrt{3}} X_{2 k} \tag{5}
\end{equation*}
$$

Equations (4) and (5) give the required vector $v_{0, k}$ for $k=0, \ldots, N / 2-1$.We can write it in matrix form as follows.

## HASHEMIZADEH \& RAHBAR

$$
\left[\begin{array}{c}
v_{0, k}^{(0)} \\
v_{0, k}^{(1)}
\end{array}\right]=\left[\begin{array}{cc}
0 & 1 \\
-\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}
\end{array}\right]\left[\begin{array}{c}
X_{2 k} \\
X_{2 k+1}
\end{array}\right]
$$

In fact, approximation prefiltering is a special case of matrix prefiltering,

$$
v_{0, k}=\sum_{m=0}^{M} P_{m}\left[\begin{array}{c}
X_{2(m+k)} \\
X_{2(m+k)+1}
\end{array}\right] .
$$

where $P_{0}, P_{1}, \ldots, P_{M}$ are $2 \times 2$ matrices. Thus, for linear Legendre multiwavelets, we have

$$
P_{0}=\left[\begin{array}{cc}
0 & 1 \\
-\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}
\end{array}\right], \text { and } P_{i}=0_{2}, i=1, \ldots, M
$$

where $0_{2}$ is a $2 \times 2$ matrix of zeros.
If prefilters are $N$-dimensional, the matrix preprocessing is represented by $\rho X=V_{0}$ where now $\rho$ is $N \times N, X$ is $N \times 1$, and $V_{0}$ is $N \times 1$. Then, we have

$$
\left[\begin{array}{ccccccc}
P_{0} & P_{1} & \cdots & P_{M} & 0_{2} & \cdots & \cdots \\
0_{2} & P_{0} & P_{1} & \cdots & P_{M} & 0_{2} & \cdots \\
& & & & & & \ddots
\end{array}\right]\left[\begin{array}{c}
X_{0} \\
X_{1} \\
X_{2} \\
X_{3} \\
\vdots
\end{array}\right]=\left[\begin{array}{c}
v_{0,0}^{(0)} \\
v_{0,0}^{(1)} \\
v_{0,1}^{(0)} \\
v_{0,1}^{(1)} \\
\vdots
\end{array}\right],
$$

where $P_{m}$ 's are $2 \times 2$ matrices (Strela, Heller, Strang, Topiwala, \& Heil, 1999; Strela, 1996; Kim \& Li, 2003).

## Image Compression Using Legendre Wavelets

The notions stated above were applied to compress a given image. In doing so, critical sampling was deployed on a well-known photo of Barbara. Initially

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

preprocess all the rows (see Strela \& Walden, 1998), which results in the picture shown in Figure 3.


Figure 3. The result of row preprocessing on Barbara's photo.

Then, preprocess all columns. The amount of data remains unchanged, since we use critical sampling.


Figure 4. Result of column preprocessing on the photo in Figure 3.

At the next step, perform the 2D-wavelet cascade. For linear Legendre multiwavelets, the multiwavelet filter bank is obtained as follow. From (3), we have

$$
H_{1}=\left[\begin{array}{cc}
1 & 0 \\
-\frac{\sqrt{3}}{2} & \frac{1}{2}
\end{array}\right], H_{2}=\left[\begin{array}{cc}
1 & 0 \\
\frac{\sqrt{3}}{2} & \frac{1}{2}
\end{array}\right], G_{1}=\left[\begin{array}{cc}
0 & -1 \\
\frac{1}{2} & \frac{\sqrt{3}}{2}
\end{array}\right], \text { and } G_{2}=\left[\begin{array}{cc}
0 & 1 \\
-\frac{1}{2} & \frac{\sqrt{3}}{2}
\end{array}\right],
$$

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

where $H_{1}, H_{2}, G_{1}$, and $G_{2}$ are matrix filters. Using the ideas of matrix block multiplication (Van Fleet, 2000), the above filters can be combined into the matrix below.

$$
\left[\begin{array}{cccccccccc}
\ddots & \ddots & \ddots & & & & & & & \\
& 0_{2} & H_{1} & H_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & \\
\cdots & 0_{2} & 0_{2} & 0_{2} & H_{1} & H_{2} & 0_{2} & 0_{2} & 0_{2} & \cdots \\
& 0_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & H_{1} & H_{2} & 0_{2} & \\
\hline & 0_{2} & G_{1} & G_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & \\
\cdots & 0_{2} & 0_{2} & 0_{2} & G_{1} & G_{2} & 0_{2} & 0_{2} & 0_{2} & \cdots \\
& 0_{2} & 0_{2} & 0_{2} & 0_{2} & 0_{2} & G_{1} & G_{2} & 0_{2} & \\
& & & & & & & \ddots & \ddots & \ddots
\end{array}\right],
$$

where $0_{2}$ is a $2 \times 2$ zero matrix. The two-channel matrix filter bank operates on two input data streams, filters them into four output data streams, and then, downsamples each of the results by a factor of two. After one step of the 2D multiwavelet cascade algorithm for linear Legendre with approximation preprocessing, the photo of Figure 5 appears.

## HASHEMIZADEH \& RAHBAR



Figure 5. The result of one-step 2D multiwavelet cascade algorithm with linear Legendre wavelets and approximation preprocessing.

The pattern of image subbands is shown in Figure 6 (Sudhakar \& Jayaraman, 2005). By comparing Figure 5 with Figure 6, it is seen that almost all information is in the block of $G_{1} G_{1}$. This reflects the fact that, for linear Legendre, one scaling function is symmetric while the other is anti-symmetric.

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

| $H_{2} G_{1}$ | $H_{2} G_{2}$ | $H_{2} H_{l}$ | $H_{2} H_{2}$ |
| :--- | :--- | :--- | :--- |
| $H_{l} G_{1}$ | $H_{l} G_{2}$ | $H_{l} H_{l}$ | $H_{l} H_{2}$ |
| $G_{2} G_{1}$ | $G_{2} G_{2}$ | $G_{2} H_{l}$ | $G_{2} H_{2}$ |
| $G_{1} G_{1}$ | $G_{1} G_{2}$ | $G_{1} H_{l}$ | $G_{1} H_{2}$ |

Figure 6. The pattern of image subbands.

Figure 7 compares the cumulative energy in the original image with the one resulted from one-step Legendre multiwavelet transform.


Figure 7. The cumulative energy of the original image and the one produced by the onestep Legendre multiwavelet transform.

The next step is to quantize the components of the Legendre multiwavelet transform. Any values of the transform that are small in modulus will be

## HASHEMIZADEH \& RAHBAR

converted to 0 . The modified transform is then be converted to Huffman codes and the resulting bit stream is encoded. The produced file is markedly smaller than the original one.


Figure 8. Some implementations of linear Legendre multiwavelet compression on Barbara's image.

## MULTIWAVELET FUNCTIONS IN IMAGE COMPRESSION

## Conclusion

The technique of using Legendre multiwavelets in image compression, as deployed in this paper, shows great advantages compared to the wavelet-based and even other multiwavelet-based methods. This is due to the fact that in a linear Legendre multiwavelet one scaling function is symmetric while the other one is anti-symmetric. In this way, it concentrates well on the image and excludes a tiny amount of details. Linear Legendre multiwavelets were used, but, the higher order Legendre multiwavelets can be employed (Rahbar, 2004).

## References

Aboufadel, E. \& Schilicker, S. (1999). Discovering wavelets. New York, NY: Wiley.

Barbara [Digital Image]. Retrieved from https://homepages.cae.wisc.edu/~ece533/images/barbara.bmp

Burrus, C. S., Gopinath, R. A., \& Guo, H. (1998). Introduction to wavelets and wavelet transforms: A primer. Upper Saddle River, NJ: Prentice Hall.

Daubechies, I. (1992). Ten lectures on wavelets. Philadelphia, PA: SIAM.
Kim, W., \& Li, C.-C. (2003). On preconditioning multiwavelet system for image compression. International Journal of Wavelets, Multiresolution and Information Processing, l(1), 51-74. doi: 10.1142/S0219691303000049

Martin, M. B. \& Bell, A. E. (2001). New image compression techniques using multiwavelets and multiwavelet packets. IEEE Transactions on Image Processing, 10(4), 500-510. doi: 10.1109/83.913585

Rahbar, S. (2004). Solving Fredholm Integral Equation Using Legendre Wavelet Functions. WSEAS Transactions on Mathematics, 3(3), 591-594. doi: 10.1.1.552.2048

Strela, V. (1996). Multiwavelets: Theory and applications (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Strela, V., Heller, P. N., Strang, G., Topiwala, P., \& Heil, C. (1999).The application of multiwavelet filterbanks to image processing. IEEE Transactions on Image Processing, 8(4), 548-563. doi: 10.1109/83.753742

Strela, V. \& Walden, A. (1998). Signal and image denoising via wavelet thresholding: Orthogonal and biorthogonal, scalar and multiple wavelet

## HASHEMIZADEH \& RAHBAR

transforms (Technical Report TR-98-01). London: Imperial College Statistics Section.

Sudhakar, R \& Jayaraman, S. (2005). Fingerprint compression using multiwavelets. International Journal of Signal Processing, 2(2), 78-87.

Udea, M \& Lodha, S. (1995). Wavelets: An elementary introduction and examples (Technical Report). Santa Cruz, CA: University of California at Santa Cruz.

Van Fleet. P. J. (2000). An elementary introduction to wavelets and their applications. Unpublished manuscript.

# Bayesian Inference for Median of the Lognormal Distribution 

K. Aruna Rao<br>SDM Degree College (Autonomous)<br>Ujire, India

Juliet G. D'Cunha<br>Mangalore University<br>Mangalagangorthri, India

Lognormal distribution has many applications. The past research papers concentrated on the estimation of the mean of this distribution. This paper develops credible interval for the median of the lognormal distribution. The estimated coverage probability and average length of the credible interval is compared with the confidence interval using Monte Carlo simulation.

Keywords: Lognormal distribution, credible interval, coverage probability, confidence interval, Monte Carlo simulation

## Introduction

The lognormal distribution is widely used in the analysis of rainfall (Ananthakrishnan, \& Soman, 1989), survival analysis (Kalbfleisch \& Prentice, 2002; Lawless, 2003) and in the analysis of stock market data (D'Cunha \& Rao, 2014a). Length biased lognormal distribution is used in the analysis of data from oil field exploration studies (Ratnaparkhi \& Naik-Nimbalkar, 2012; see reference therein). In the analysis of stock market data, although lognormal distribution is not directly used, analysis is carried out using log transformation which in turn implies that the underlying distribution is lognormal.

The lognormal distribution belongs to log location scale family. The salient feature of the log location scale family is that the coefficient of variation (CV) of the distribution depends only on the scale parameter and not on the location parameter. When the log location scale family of distribution is obtained through the symmetric location scale family, the median of the distribution is $\exp$ (location parameter). In this case the median is invariant under distributional transformation

[^30]
## RAO \& D'CUNHA

of the data. The estimation of mean, median and variance of the lognormal distribution has a long history.

Zellner (1971) proposed several estimators of the mean and median of the lognormal distribution and obtained minimum mean square error estimator of these parameters in some class of estimators. Bayes estimators of these parameters were also considered. Subsequently, Padgett and Wei (1977) developed Bayes estimator of reliability function for the lognormal distribution. They used two types of priors namely normal prior for mean and gamma prior for the inverse of the scale parameter; the other prior is the vague prior of Jeffrey (see Ghosh, Delampady, \& Samanta, 2006).

This was extended by Padgett and Johnson (1983), where they obtained lower bounds on reliability function of the two parameter lognormal distribution. Sarabia, Castillo, Gómez-Déniz, and Vázquez-Polo (2005) proposed a class of bivariate conjugate priors for $\mu$ and $\sigma$ of the lognormal distribution using the conditional specification. Several procedures were also suggested for the estimation of the hyperparameters. Harvey and van der Merwe (2012) compared the Bayesian credible interval for the means and variances of lognormal distribution.

The performance of the credible interval is compared with credible/confidence interval suggested by Zhou and Tu (2000) and Krishnamoorthy and Mathew (2003). In the last section the authors discuss about bivariate lognormal distribution and obtain Bayesian confidence intervals for the difference between two correlated lognormal means and for the ratio of lognormal variances. The conclusion was the Bayes credible interval has shorter length compared to the length of the other intervals.

D'Cunha and Rao (2014a) developed Bayesian credible interval for the CV of the lognormal distribution and compared it with the confidence interval obtained by the maximum likelihood estimator. They showed that, under mild regularity conditions, Bayes estimator for the mean of the lognormal distribution exists. Thus, research after Zellner (1971) did not focus on median of the lognormal distribution.

The lognormal distribution is positively skewed and for skewed distributions, median is a better estimator rather than the mean, which is affected by extreme values. In medical studies, median survival time is often reported than the mean survival time. This motivates deriving Bayes credible interval for the median of the lognormal distribution. Under absolute error loss function, the Bayes estimator is the minimum average risk estimator.

## BAYESIAN INFERENCE FOR MEDIAN LOGNORMAL DISTRIBUTION

Among the Bayesian significance tests, for testing a sharp null hypotheses namely the test based on credible interval, the Bayes factor and the full Bayesian significance test (FBST), credible interval is the simplest to compute and straight forward decision can be taken as in the case of significance tests. Thus, the purpose of this study is to compare the performance the Bayes credible interval with the confidence interval obtained from the maximum likelihood estimator (MLE).

## Bayes Estimator for the Median of the Lognormal Distribution

Let $\mu$ and $\sigma$ denote the $\log$ location and scale parameter of the lognormal distribution. Given a random sample of size $n, x_{1}, \ldots, x_{n}$ from this distribution, let $Z_{i}=\log X_{i}, i=1, \ldots, n$, where $Z$ follows normal distribution with parameters $\mu$ and $\sigma^{2}$ and maximum likelihood estimator of $\mu$ and $\sigma^{2}$ are $\bar{Z}$ and $S_{z}^{2}$ respectively, where $\bar{Z}=\frac{\sum_{i=1}^{n} Z_{i}}{n}$ and $S_{z}^{2}=\frac{\sum_{i=1}^{n}\left(Z_{i}-\bar{Z}\right)^{2}}{n}$. Using invariance property of maximum likelihood estimator (Kale (1999)), the maximum likelihood estimate (MLE) of the median of the lognormal distribution namely $\theta=e^{\mu}$ is given by $\hat{\theta}=e^{\hat{\mu}}=e^{\bar{Z}}$. The asymptotic variance of $\hat{\theta}$ can be obtained using delta method and is given by

$$
\begin{equation*}
\operatorname{var}(\hat{\theta})=e^{2 \mu} \frac{\sigma^{2}}{n}+o\left(n^{-1}\right) \tag{1}
\end{equation*}
$$

In the above expression $\frac{n-1}{n} \approx 1$. The $100(1-\alpha) \%$ asymptotic confidence interval for $\hat{\theta}$ is given by $\hat{\theta} \pm Z_{\alpha / 2} S . E .(\hat{\theta})$, where $Z_{\alpha / 2}$ refers to upper $\alpha / 2$ th percentile value of the standard normal distribution and S.E. $(\hat{\theta})$ refers to estimated standard error $(\hat{\theta})$. The estimate of $\mu$ and $\sigma^{2}$ is obtained by substituting the value of $\hat{\mu}$ and $\hat{\sigma}^{2}$ in the expression for variance of $\hat{\theta}$.

Four objective priors are considered: the uniform prior $\pi(\mu, \sigma)=1$, right invariant prior $\pi(\mu, \sigma)=1 / \sigma$, left invariant Jeffreys prior given by $\pi(\mu, \sigma)=1 / \sigma^{2}$ and Jeffreys rule prior $\pi(\mu, \sigma)=1 / \sigma^{3}$. For a discussion of these priors see Berger (1985) and Ghosh et al. (2006).

## RAO \& D'CUNHA

The choice of the right invariant prior stems from the fact that $Z$ follows normal distribution and the right invariant prior used in this paper is the one that is suggested for location scale family (Ghosh et al., 2006). The advantage of objective Bayesian analysis is that the prediction remains the same irrespective of the decision maker. The procedure can be applied universally given the past data. The posterior density $\pi\left(\mu, \sigma \mid z_{1}, \ldots, z_{n}\right)$ for the uniform, right invariant, left invariant Jeffreys priors and Jeffreys rule prior are given by the following expressions,

$$
\begin{align*}
& \pi\left(\mu, \sigma \mid z_{1}, \ldots, \mathrm{z}_{n}\right)=\frac{1}{\sqrt{2 \pi} \frac{\sigma}{\sqrt{n}}} e^{-\frac{1}{2}(\bar{z}-\mu)^{2} \sigma^{2} / n} \frac{\left(\frac{(n-1) s_{z^{2}}}{2}\right)^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)}  \tag{2}\\
& \left(\frac{1}{\sigma^{2}}\right)^{\left(\frac{n+1}{2}\right)-1} e^{-\frac{1}{2}\left(\frac{n-1}{\sigma^{2}}\right) s_{z}^{2}} \text { (using uniform prior) } \\
& \pi\left(\mu, \sigma \mid z_{1}, \ldots, \mathrm{z}_{n}\right)=\frac{1}{\sqrt{2 \pi} \frac{\sigma}{\sqrt{n}}} e^{-\frac{1(\bar{z}-\mu)^{2}}{\sigma^{2} / n}} \frac{\left(\frac{(n-1) s_{z^{2}}}{2}\right)^{\frac{n+2}{2}}}{\Gamma\left(\frac{n+2}{2}\right)}  \tag{3}\\
& \left(\frac{1}{\sigma^{2}}\right)^{\left(\frac{n+2}{2}\right)-1} e^{-\frac{1}{2}\left(\frac{n-1}{\sigma^{2}}\right) s_{z}^{2}} \text { (using right invariate prior) } \\
& \left(\pi\left(\mu, \sigma \mid z_{1}, \ldots, \mathrm{z}_{n}\right)=\frac{1}{\sqrt{2 \pi} \frac{\sigma}{\sqrt{n}}} e^{-\frac{1(\bar{z}-\mu)^{2}}{\sigma^{2} / n}} \frac{\left(\frac{(n-1) s_{z^{2}}}{2}\right)^{\frac{n+3}{2}}}{\Gamma\left(\frac{n+3}{2}\right)}\right.  \tag{4}\\
& \left(\frac{1}{\sigma^{2}}\right)^{\left(\frac{n+3}{2}\right)-1} e^{-\frac{1}{2}\left(\frac{n-1}{\sigma^{2}}\right) s_{\frac{2}{2}}} \text { (using left invariate prior) }
\end{align*}
$$

## BAYESIAN INFERENCE FOR MEDIAN LOGNORMAL DISTRIBUTION

$$
\begin{align*}
& \pi\left(\mu, \sigma \mid z_{1}, \ldots, \mathrm{z}_{n}\right)= \frac{1}{\sqrt{2 \pi} \frac{\sigma}{\sqrt{n}}} e^{-\frac{1(\bar{z}-\mu)^{2}}{\sigma^{2} / n}} \frac{\left(\frac{(n-1) s_{z^{2}}}{2}\right)^{\frac{n+4}{2}}}{\Gamma\left(\frac{n+4}{2}\right)}  \tag{5}\\
&\left(\frac{1}{\sigma^{2}}\right)^{\left(\frac{n+4}{2}\right)-1} e^{-\frac{1}{2}\left(\frac{n-1}{\sigma^{2}}\right) s_{z}^{2}} \text { (using Jeffreys rule prior) }
\end{align*}
$$

Although the independent prior is used for $\mu$ and $\sigma$, the posterior density has a bivariate correlated distribution. The Bayes estimator of $\theta$ is $E\left(\theta \mid z_{1}, \ldots, z_{n}\right)$, where expectation is taken with respect to the posterior density of $\mu$ and $\sigma$.

## Monte Carlo Simulation

Observations of size $n$ are generated from normal distribution with mean $\mu$ and variance $\sigma^{2}$. The Bayes estimator of the median of the lognormal distribution is $E\left(e^{\mu}\right)$, where expectation is taken with respect to the posterior density of $\pi(\mu, \sigma \mid$ data ). This expectation has no closed form solution and Monte Carlo integration is carried out using importance sampling approach. Since the posterior density is the product of $\eta=1 / \sigma^{2}$ which is gamma and the conditional density of $\mu \mid \sigma^{2}$ which is normal, an observation is generated for $\eta$ from gamma density and using this value of $\eta$, an observation is generated for $\mu$ from normal density. This constitutes a pair of observations $(\eta, \mu)$ from the bivariate posterior density.

Using 10,000 simulations the Bayes estimator for the median of the lognormal distribution is given by $E\left(e^{\mu} \mid z_{1}, \ldots, z_{n}\right)=\frac{1}{M} \sum_{i=1}^{M}\left(e^{\mu_{i}}\right)$, where $M$ denotes the number of paired samples generated from the posterior distribution. In this paper we have used $M=10,000$. The equitailed credible interval has the limit $(1-\alpha / 2)^{\text {th }}$ and $\alpha / 2^{\text {th }}$ upper percentile value of the posterior distribution of $e^{\mu}$. For each sample the confidence interval is given by $\hat{\theta} \pm Z_{\alpha / 2} S . E .(\hat{\theta})$, where $\hat{\theta}=e^{\hat{\mu}}$ and $Z_{\alpha / 2}$ refers to upper $\alpha / 2$ th percentile value of the standard normal distribution and S.E. $(\hat{\theta})$ refers to estimated standard error $(\hat{\theta})$. Using 1000 simulations, the proportion of times the true median lies inside the credible/confidence interval gives the estimated $100(1-\alpha) \%$ credible/confidence level and is referred as coverage probability for brevity. For the investigation the

## RAO \& D'CUNHA

value of $\mu$ is fixed at $\log (1000)$ and the CV value used for the investigation ranges from $0.1,0.3,0.5,0.7,1,1.5,2$ and 2.5. It may be recalled that $\mathrm{CV}=\left(e^{\sigma^{2}}-1\right)^{1 / 2}$.

The sample sizes considered are $n=10,20,40,60,80,100,150$ and 200. The total number of configurations are 256 ( 8 sample size $\times 8 \mathrm{CV}$ values $\times 4$ priors). The average time required for the computation of credible interval is approximately 30 minutes for each sample size and CV.

## Results

Presented in Table 1 are the number of times coverage probability is maintained by the credible/confidence interval for 8 combinations of CV across sample sizes. We say that a credible/confidence interval maintains credible/confidence level of $(1-\alpha)=0.95$ if the coverage probability is in the interval of 0.940 to 0.960 . The confidence interval maintains the level for a sample size of $n \geq 60$ and the credible interval maintains the level for a sample size $n \geq 80$. The table also presents the average length of the credible/confidence interval. The average length is computed whenever the credible/confidence level is maintained.

Table 1. Coverage probability of the credible and confidence interval for the Median across sample sizes for 8 combinations of specified values of CV

| $n$ | Bayes Procedure (Equitailed) |  |  |  |  |  |  |  | MLE(Equitailed) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Numb pro |  |  |  | Average Length |  |  |  | Number of times coverage probability | Average |
|  | U | R | L | JR | U | R | L | JR | is maintained | length |
| 10 | 0 | 0 | 0 | 0 | * | * | * | * | 0 | * |
| 20 | 3 | 1 | 2 | 0 | 604.93 | 612.17 | * | * | 1 | 86.03 |
| 40 | 4 | 3 | 3 | 1 | 544.27 | 449.52 | 279.71 | 275.94 | 3 | 492.29 |
| 60 | 5 | 7 | 3 | 3 | 433.82 | 489.59 | 485.34 | 481.19 | 7 | 436.37 |
| 80 | 6 | 6 | 6 | 5 | 395.89 | 393.52 | 391.06 | 440.87 | 6 | 296.97 |
| 100 | 7 | 7 | 8 | 6 | 273.59 | 252.46 | 251.19 | 272.75 | 7 | 217.95 |
| 150 | 6 | 5 | 6 | 5 | 305.89 | 304.94 | 244.34 | 273.68 | 5 | 254.19 |
| 200 | 7 | 5 | 7 | 7 | 185.89 | 185.49 | 202.39 | 184.57 | 5 | 197.64 |
| overall | 38 | 34 | 35 | 27 | 392.04 | 383.96 | 309.00 | 321.5 | 34 | 283.06 |

*Note. Whenever coverage probability is not maintained average length has not been calculated. U-Uniform prior, R-Right invariant prior, L-Left invariant prior, JR-Jeffreys rule prior.

Presented in Table 2 are the average length of the credible interval and the confidence interval for all the sample sizes. The table has been constructed such that the average length of the credible interval is computed over the confidence interval where the nominal confidence level is maintained for each of the prior.

## BAYESIAN INFERENCE FOR MEDIAN LOGNORMAL DISTRIBUTION

The average length of the credible interval is shorter compared to confidence interval.

Table 2. Average length of the credible/confidence interval for all sample size for each prior.

| Prior | Number of times coverage probability maintained out of 64 configurations | Number of times credible interval has shorter length | Number of times confidence interval has shorter length | Average length of the credible interval | Average length of the confidence interval |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Uniform | 38 | 18 | 20 | 196.77 | 197.32 |
| Right invariant | 35 | 27 | 8 | 243.23 | 244.64 |
| Left invariant | 30 | 30 | 0 | 292.39 | 294.58 |
| Jeffreys Rule | 27 | 27 | 0 | 304.47 | 308.38 |

An investigation was also carried out to find out the influence of the variability in the data regarding coverage probability and length of the credible/confidence interval. Presented in Table 3 are the coverage probability and length of the credible/confidence interval for sample size $n=100$. The length of the confidence/credible interval increases with the increase in value of CV, upto the value of $\mathrm{CV}=2$, then there is a decrease in the length for the value of $C V=2.5$, specific reason for this type of behavior is not known. The length of the credible interval for most of the values of CV is marginally shorter than the confidence interval.

Table 3. Length of the confidence/credible interval for various values of CV when sample size $=100$.

| Sample <br> size | Conf/cred <br> interval <br> based on | 0.1 | 0.3 | 0.5 | 0.7 | 1 | 1.5 | 2 | 2.5 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | MLE | 39.08 | 115.16 | 184.99 | 247.49 | 326.88 | 427.06 | 500.25 | 184.99 |
|  |  | $(0.952)$ | $(0.952)$ | $(0.942)$ | $(0.946)$ | $(0.945)$ | $(0.944)$ | $(0.937)$ | $(0.942)$ |
|  | Uniform | 38.95 | 114.79 | 184.49 | 247.29 | 327.26 | 428.90 | 503.76 | 184.49 |
|  |  | $(0.951)$ | $(0.962)$ | $(0.945)$ | $(0.941)$ | $(0.951)$ | $(0.951)$ | $(0.941)$ | $(0.945)$ |
| 100 | Right | 38.74 | 114.26 | 183.71 | 246.09 | 325.42 | 426.47 | 501.28 | 183.71 |
|  |  | $(0.951)$ | $(0.960)$ | $(0.944)$ | $(0.941)$ | $(0.951)$ | $(0.951)$ | $(0.941)$ | $(0.944)$ |
|  | Left | 38.57 | 113.71 | 182.75 | 244.72 | 324.01 | 424.61 | 498.39 | 182.75 |
|  | $(0.946)$ | $(0.959)$ | $(0.944)$ | $(0.940)$ | $(0.946)$ | $(0.946)$ | $(0.940)$ | $(0.944)$ |  |
|  | Jeffreys | 38.38 | 113.10 | 181.75 | 243.72 | 322.40 | 422.49 | 496.37 | 181.75 |
|  | Rule | $(0.942)$ | $(0.958)$ | $(0.939)$ | $(0.940)$ | $(0.942)$ | $(0.942)$ | $(0.940)$ | $(0.930)$ |

## RAO \& D'CUNHA

## Illustrative Example

To illustrate the use of Bayes procedure for the median of lognormal distribution, consider the data on 31 consecutive daily Carbon Monoxide measurements (parts per million - ppm) taken by an oil refinery northeast of San Francisco and nine measurements on the same stack taken by the Bay Area Air Quality Management District (BAAQMD). The data are available from http://lib.stat.emu.edu/DASL/. The sample is carbon monoxide measurements taken by the Bay Area Air Quality Management District (BAAQMD. The hypothesis of interest is

$$
H_{0}: \text { Median Carbon Monoxide measurement of BAAQMD }=58.81 \mathrm{ppm}
$$

The value 58.81 ppm is the estimated median value obtained by using the data on 31 consecutive daily Carbon Monoxide measurements taken by oil refinery to the northeast of San Francisco. The results are summarized in Table 4. Notice that the estimated median value $=58.81 \mathrm{ppm}$ does not lie inside any of the credible/confidence interval, thus we conclude that median carbon monoxide measurement of BAAQMD $\neq 58.81 \mathrm{ppm}$.

Table 4. Credible/confidence interval and length of the credible/confidence interval for 4 priors under Bayes and Maximum Likelihood estimation for BAAQMD data.

| Procedure | Prior | Estimate | Credible/confidence <br> interval | Length of the <br> Credible/confidence interval |
| ---: | ---: | ---: | ---: | ---: |
|  | Uniform | 20.27 | $(10.60,36.03)$ | 25.43 |
| Bayes | Right | 20.23 | $(10.97,34.15)$ | 23.18 |
|  | Left | 20.12 | $(11.40,33.29)$ | 21.90 |
|  | Jeffreys Rule | 20.08 | $(11.09,33.12)$ | 22.03 |
| MLE | - | 19.36 | $(7.03,31.69)$ | 24.65 |

## Conclusion

The performance of the Bayes credible interval was investigated for the median of the lognormal distribution. It has many applications and most of the previous papers propose credible intervals for the mean of the lognormal distribution. The median of the lognormal distribution depends only on the log location parameter and should be the right choice as the measure of location rather than the mean. Lognormal distribution is right skewed and mean of the distribution is a function of $\log$ location and $\log$ scale parameter. Thus, the mean is very much influenced by the variability in the data when the underlined distribution is lognormal.

## BAYESIAN INFERENCE FOR MEDIAN LOGNORMAL DISTRIBUTION

Confidence interval is computationally simple. The present investigation revealed that Bayes credible interval has shorter average length compared to the confidence interval and is recommended.

Note: A program in the MATLAB software version 7.0 for computation of credible interval is available.

## References

Ananthakrishnan, R., \& Soman, M. K. (1989). Statistical distribution of daily rainfall and its association with the co-efficient of variation of rainfall series. International Journal of Climatology, 9(5), 485-500. doi:10.1002/joc. 3370090504

Berger, J. O. (1985). Statistical decision theory and Bayesian analysis (2 ${ }^{\text {nd }}$ ed.). New York, NY: Springer-Verlag.

D’Cunha, J. G., \& Rao, K. A. (2014a). Bayesian inference for volatility of stock prices. Journal of Modern Applied Statistical Methods, 13(2), 493-505. Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss2/29

D’Cunha, J. G., \& Rao, K. A. (2014b). Bayesian inference for mean of the lognormal distribution. International Journal of Scientific and Research Publications, 4(10), 195-203.

Ghosh, J. K., Delampady, M., \& Samanta, T. (2007). An introduction to Bayesian analysis: Theory and methods. USA: Springer Science \& Business Media.

Harvey, J., \& van der Merwe A. J. (2012). Bayesian confidence intervals for means and variances of lognormal and bivariate lognormal distributions. Journal of Statistical Planning and Inference, 142(6), 1294-1309.
doi: 10.1016/j.jspi.2011.12.006
Kalbfleisch, J. D., \& Prentice, R. L. (2002). The statistical analysis of failure time data (2 ${ }^{\text {nd }}$ ed.). New York, NY: John Wiley \& Sons.

Kale, B. K. (1999). A first course on parametric inference (2nd ed.). Narosa.
Krishnamoorthy, K., \& Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized P -values and generalized confidence intervals. Journal of Statistical Planning and Inference, 115(1), 103-121. doi: 10.1016/S0378-3758(02)00153-2

Lawless, J. F. (2003). Statistical models and methods for lifetime data (2 ${ }^{\text {nd }}$ ed.). New York, NY: John Wiley \& Sons.

## RAO \& D'CUNHA

Padgett, W. J., \& Johnson, M. P. (1983). Some Bayesian lower bounds on reliability in the lognormal distribution. Canadian Journal of Statistics, 11 (2), 137-147. doi: 10.2307/3314981

Padgett, W. J., \& Wei, L. J. (1977). Bayes estimation of reliability for the two-parameter lognormal distribution. Communications in Statistics-Theory and Methods, 6(5), 443-457. doi: 10.1080/03610927708827505

Ratnaparkhi, M. V., \& Naik-Nimbalkar, U. V. (2012). The length-biased lognormal distribution and its application in the analysis of data from oil field exploration studies. Journal of Modern Applied Statistical Methods, 11(1), 255260. Available at: http://digitalcommons.wayne.edu/jmasm/vol11/iss1/22

Sarabia, J. M., Castillo, E., Gómez-Déniz, E., \& Vázquez-Polo, F. J. (2005).
A class of conjugate priors for log-normal claims based on conditional specification. The Journal of Risk and Insurance, 72(3), 479-495. doi: 10.1111/j.1539-6975.2005.00133.x

Zellner, A. (1971). Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression. Journal of the American Statistical Association, 66(334), 327-330. doi: 10.1080/01621459.1971.10482263

Zhou, X.-H., Tu, W. (2000). Interval estimation for the ratio in means of log-normally distributed medical costs with zero values. Computational Statistics and Data Analysis, 35(2), 201-210. doi: 10.1016/S0167-9473(00)00009-8

# Developing Bayesian-based Confidence Bounds for Non-identically Distributed Observations using the Lyapunov Condition 

Garry M. Jacyna<br>The MITRE Corporation

Scott L. Rosen<br>The MITRE Corporation

The purpose of this paper is to establish a direct method for assessing the confidence in the detection and identification probabilities for segmented observations that are not identically distributed across assigned segments within a region. This paper arrives at easily computable confidence intervals by showing through mathematical analysis that:
I. The probability of successful detection within each test segment can be characterized by a Beta distribution;
II. The distribution of a weighted sum of independent but non-identically distributed sample means is asymptotically Normally distributed by the Lyapunov variant of the Central Limit Theorem, i.e., the approximation improves as the number of samples increases;
III. Given that the distribution of the sample means convergences to a Normal distribution, the confidence intervals about the observed sample means for both the detection and identification probabilities can be determined in closed form for multiple target types.

The motivation for this approach is the need to determine the exceedance probabilities to support a Systems Acceptance Test based on collected data.

Keywords: Bayesian inference, analysis of designed experiments, beta distribution, Lyapunov condition

## Background

A System Acceptance Test (SAT) requiring confirmatory data analysis (Box, Hunter, \& Hunter, 2005) driven by apriori and politically deducted hypotheses is needed to assess the impact of a specific acquisition on two key system level performance parameters for a particular region: probability of detection ( Pd ) and probability of identification (Pid). The difficulty with this assessment is that

## JACYNA \& ROSEN

within this region of interest, only a finite portion is covered by each sensor's area-of-regard (AOR). In addition, there are an infinite number of threat compositions and avenues of ingress and egress (i.e., routes) that are possible throughout the area. Only a small sampling of operationally valid (traversable by the threat) routes across the region is executed and these are used to characterize the performance measures across the entire area. See Figure 1 for an illustrative view of this concept.


-     - Sample Route
-     - "Operationally Valid" Route
- Sensor Area of Regard
$\square$ - Project Area

Figure 1. A high level illustration depicting the relationship among a project area and the sample routes, operational valid routes, and sensor areas-of-regards within it.

The overwhelming size of the test area introduces an additional test constraint. This is addressed by approximating the route samples with segmentlevel performance observations by considering a segment to be a contiguous subset of a given route. Moreover, a trial in this system acceptance test is defined as being a test observation made on a single segment. The subset of segments chosen for the test fall within a given sensor's AOR and belong to an operationally valid route as shown in Figure 2.

A difficult analysis problem arises when attempting to compute a system level estimate of performance involving an associated confidence bound and exceedance probability from segment-level observations made on small sample routes. This is because each route segment has a different underlining probability distribution that is a function of the different target/system/environmental factors present at the time of observation. Computing confidence intervals for Pd and Pid for individual segments is straightforward, but determining a single overall

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

confidence interval for the sample set as a whole is not trivial since these independent random variables are drawn from different underlying detection and identification probability distributions. In this case it is not immediately apparent that the Central Limit Theorem (CLT) applies (Karr, 1993, p. 190-192).


Figure 2. An illustration depicting the relationship among a segment, an "operationally valid" route, and a sensor AOR.

It is shown mathematically that the Lyapunov variant (Karr, 1993, p. 190192) of the Central Limit Theorem (CLT) can be used to establish the Normality of the weighted sample means for Pd and Pid generated from this systems acceptance test. Furthermore, it is shown that a sample size of at least 1 trial for 10 unique segments is sufficient for approximating the resulting mean Pd and Pid observations by a Gaussian or Normal distribution.

Following this Normality result, the corresponding confidence intervals are then generated from the sample detection and identification proportions obtained in the test. Example computations are used to illustrate their implementation. A confidence interval calculator is then discussed for generating hypothetical confidence interval and exceedance probability values based upon inputted sample sizes for each individual segment and projected sample means for Pd and Pid. This calculator was then in turn used for shaping the experimental design for the test.

## JACYNA \& ROSEN

## Introduction to the Analysis Problem

The work presented below describes, demonstrates, and justifies mathematically an approach for computing the confidence intervals associated with system level Pd and Pid observations. A few problem assumptions are necessary. These assumptions are as follows:

## Problem Assumptions:

A. A detection trial pertains to the traversal of a single item of interest across an entire segment.
B. There is a binary outcome for an identification trial; the detection is successfully or correctly identified or the detection is unsuccessfully or incorrectly identified.
C. The sample probability of detection obtained from test constitutes the number of successful detections divided by the number of detection trials.
D. The sample probability of identification obtained from test constitutes the number of successful identifications divided by the number of identification trials
E. A single success probability $p$ can characterize the probability of successful detection along a whole segment
F. A single success probability $p$ cannot characterize the probability of successful identification across a whole segment, but can characterize the probability of successful identification for an individual identification trial within a segment.
G. The success probabilities for detection and successful identification are uniformly distributed between 0 and 1 across the sample set of segments.

## Outline of the Approach

The probability density function (pdf) for the system Pd sample mean is derived as a function of segment-level observations from test. The analysis shows, through a mathematical proof and supporting Monte Carlo computations, that the distribution can be approximated as a Normal distribution.

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

In addition, through the application of a mixture distribution (across the identification trials within a segment), it is shown that the Normality results for Pd will also apply to Pid. Recommended confidence intervals then follow for Pd and Pid that are supported by this Normality result. This result is also valid in the case of a single target type or multiple target types as explained in the section focusing on Pid. A case study follows to illustrate how the reader can apply these confidence intervals. Below is a summary of the concepts that are presented and verified through mathematical analysis:
i) Through Bayes theorem, the probability of successful detection on each segment can be characterized by a Beta distribution.
a. The weighted system detection probability is a convolution of Beta distributions and is referred to as an Augmented Beta Distribution.
b. The distribution of the system sample mean is equivalent to the weighted system Pd distribution and, therefore, can also be characterized by the derived Augmented Beta Distribution.
ii) The Augmented Beta distribution is shown through a mathematical proof to be approximately normally-distributed by the Lyapunov variant of the Central Limit Theorem.
a. The Lyapunov Central Limit Theorem specifies certain conditions that are sufficient to establish that the sum or average of a large number of independent observations is normally-distributed even if the observations are generated from different underlying probability distributions.
b. The Lyapunov conditions hold when the threat arrival weights are uniformly distributed or when the selection of segments is equally likely.
c. The Lyapunov conditions hold when the threat arrival weights are greater than zero for all but a finite number of segments.
d. It is illustrated through an empirical computational study that the system sample mean rapidly converges to a Normal distribution within a 30 segment Test design alternative.

## JACYNA \& ROSEN

iii) Based on the fact that the system weighted sample means are approximately Normally distributed, the confidence about the observed sample means for Pd is:

$$
\begin{aligned}
C & =\left(2 \pi \sigma_{\ell}^{2}\right)^{-1 / 2} \int_{p^{*}}^{1} \exp \left\{-\left(s-m_{\ell}\right)^{2} / 2 \sigma_{\ell}^{2}\right\} d s \\
& =F\left[\left(1-m_{\ell}\right) / \sigma_{\ell}\right]-F\left[\left(p^{*}-m_{\ell}\right) / \sigma_{\ell}\right]
\end{aligned}
$$

where: $p^{*}$ is the exceedance probability or specified acceptable value for $\mathrm{Pd}, F(x)$ is the Standard Normal Distribution:

$$
F(x)=(2 \pi)^{-1 / 2} \int_{-\infty}^{x} \exp \left(-s^{2} / 2\right) d s
$$

$m_{\ell}$ and $\sigma_{\ell}^{2}$ are the sample mean and variance, respectively:

$$
m_{\ell}=\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right), \quad \sigma_{\ell}^{2}=\sum_{i=1}^{M} w_{i}^{2} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)}
$$

and $N_{i}$ and $n_{i}$ are the total number of detection attempts and actual detections/identifications (see iv) observed in test, respectively, for each segment $i$, where $M$ denotes the total number of segments.
iv) The Normality results for the weighted system sample mean also hold for the system Pid sample mean when the probability of successful identification on a segment is considered to be a mixture distribution as shown in the section focusing on Pid. The confidence interval can then be computed analogously as above. Furthermore, the Normality results also hold for multiple target types if the probability of detection is similarly considered to be a mixture distribution.

## The Probability Density Function for Pd

This section develops the probability density function (pdf) for the weighted detection probability ( $\mathrm{Pd)}$ across all segments assuming that the measurements relative to each segment are independent but not identically distributed. Each

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

segment $i$ has an associated test outcome $S_{i}=\left\{n_{i}, N_{i}\right\}$ which is a record of the number of detection successes $n_{i}$ out of $N_{i}$ possible trials. It is assumed that the outcome of each segment is statistically independent of any other segment outcome and that the unknown success probability for each segment is $p_{i}$. This implies that each segment $n_{i}$ is binomially-distributed with known number of trials $N_{i}$ and unknown probability $p_{i}$. The pdfs associated with each $p_{i}$ are shown to be the well-known Beta distribution for an uninformative prior (i.e., a prior pdf that is uniformly distributed in the interval [0, 1]). Following that, this result is generalized for the weighted probability function $\sum_{i} w_{i} p_{i}$ for the system routes/segments when the segment outcomes $S_{i}$ are independent but not identically distributed and the known arrival weights for each segment are given by $w_{i}$. A confidence interval is also determined for the probability that the weighted mean probability is greater than or equal to some exceedance probability $p^{*}$. The appropriateness of the Gaussian approximation to this general problem as the number of components in the weighted mean (i.e., number of segments) becomes large is then shown. Following that, an illustration of the approach is given to show that convergence to a Gaussian distribution is reached within the number of segments and trials allocated for test. Then a discussion of how the Normality result can be extend to Pid is provided.

## The Beta Distribution for Segment Pd

Recognizing that the probability of detection $p_{i}$ for a specific segment $i$ is an unknown parameter, it is desirable to quantify this parameter with its own probability distribution. Now, determine the pdf associated with the probability of a successful detection $p_{i}$ for a generic segment based on a fixed number of trials $N_{i}$ and the number of successes from these trials $n_{i}$. From Bayes theorem (Bernardo \& Smith, 2000, p. 241-255), this conditional $\operatorname{pdf} f\left(p_{i} \mid n_{i}: N_{i}\right)$ is:

$$
\begin{equation*}
f_{p}\left(p_{i} \mid n_{i}: N_{i}\right)=\kappa B\left(n_{i}, N_{i} ; p_{i}\right) f_{p}\left(p_{i}\right), \tag{1}
\end{equation*}
$$

where the corresponding likelihood function $B\left(n_{i}, N_{i} ; p_{i}\right)=\binom{N_{i}}{n_{i}} p_{i}^{n_{i}}\left(1-p_{i}\right)^{N_{i}-n_{i}}$ is the Binomial distribution, the prior distribution $f_{p}\left(p_{i}\right)=1$ if $0 \leq p_{i} \leq 1$; otherwise, $f_{p}\left(p_{i}\right)=0$ (using a uniformed prior assumption), $\kappa$ is a proportionality constant and $n_{i}$ denotes the number of successful detections out of $N_{i}$ trials. It

## JACYNA \& ROSEN

follows that $\kappa^{-1}=\int_{0}^{1} B\left(n_{i}, N_{i} ; p_{i}\right) f_{p}\left(p_{i}\right) d p_{i}$. Using the integral identity (Abramowitz \& Stegun, 1965, p.258):

$$
\begin{equation*}
\beta(n+1, N-n+1)=\int_{0}^{1} p^{n}(1-p)^{N-n} d p=n!(N-n)!/(N+1)! \tag{2}
\end{equation*}
$$

where $\beta(x, y)=\int_{0}^{1} p^{x-1}(1-p)^{y-1}$ is the Beta function, (1) can be rewritten as:

$$
\begin{equation*}
f_{p}\left(p_{i} \mid n_{i} ; N_{i}\right)=p_{i}^{n_{i}}\left(1-p_{i}\right)^{N_{i}-n_{i}} / \beta\left(n_{i}+1, N_{i}-n_{i}+1\right) \tag{3}
\end{equation*}
$$

This density is referred to as the Beta Distribution.
A confidence interval is defined by $C$ and $p^{*} ; C$ is interpreted as the probability that the true value of the unknown parameter $p$ lies between the exceedance probability $p^{*}$ and 1 . In particular:

$$
\begin{equation*}
C=\operatorname{Prob}\left(p_{i} \geq p^{*} \mid n_{i}, N_{i}\right)=\int_{p^{*}}^{1} p_{i}^{n_{i}}\left(1-p_{i}\right)^{N_{i}-n_{i}} d p_{i} / \beta\left(n_{i}+1, N_{i}-n_{i}+1\right) . \tag{4}
\end{equation*}
$$

## The Augmented Beta Distribution Result for a System Pd Sample Mean

Suppose there are $M$ segments with $M$ associated test outcomes $S_{1}, S_{2}, \ldots, S_{M}$, where, as before, each set $S_{i}$ records the number of trials, $N_{i}$, and the number of successful detections, $n_{i}$. It is assumed that the test results $n_{i}$ are independent but not necessarily identically distributed. The unknown detection probability parameters associated with the $M$ segments are labeled as $p_{1}, p_{2}, \ldots p_{M}$. The detection probability of the regional system should be represented by a weighted average of the segment detection probabilities, in which the segment weights $w_{i}$ are computed from the relative proportion of threat traffic through the region expected to occur in segment $i ; 0 \leq w_{i} \leq 1$ and $\sum_{i} w_{i}=1$. Therefore, the weighted average is a convex combination of the segment statistics $\ell=\sum_{i} w_{i} p_{1}$ and $0 \leq \ell \leq 1$. The value $\ell$ is a system-wide metric of detection performance.

To understand the regional system sample mean, examine the following joint pdf for the posterior detection probabilities across $M$ segments. Using vector notation to express segment detection probability parameters, segment sample sizes, and number of segment detections as $\mathbf{p}=\left[p_{1}, p_{2}, \ldots, p_{M}\right]$, $\mathbf{N}=\left[N_{1}, N_{2}, \ldots, N_{M}\right]$ and $\mathbf{n}=\left[n_{1}, n_{2}, \ldots, n_{M}\right]$, respectively, this joint posterior

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

probability is $f_{p}(\mathbf{p} \mid \mathbf{n} ; \mathbf{N})$. Because the sets of measurements are statistically independent and each segment measurement set $S_{i}$ is strictly a function of the probability parameter $p_{i}$ we can write:

$$
\begin{equation*}
f_{\mathbf{p}}(\mathbf{p} \mid \mathbf{n}, \mathbf{N})=\prod_{i=1}^{M} f_{p}\left(p_{i} \mid n_{i}, N_{i}\right)=\prod_{i=1}^{M} \beta^{-1}\left(n_{i}+1, N_{i}-n_{i}-1\right) p_{i}^{n_{i}}\left(1-p_{i}\right)^{N_{i}-n_{i}} \tag{5}
\end{equation*}
$$

For clarity in the derivations below, the weighted estimate $\ell$ is notated as $\ell=\sum_{i} y_{i}$, where $y_{i}=w_{i} p_{i}$. Here $y_{i}$ is a one-to-one transformation of $p_{i}$ such that $p_{i}=y_{i} / w_{i}$. Since each $p_{i}$ is Beta distributed, the distribution of $y_{i}$ is also a Beta distribution:

$$
\begin{equation*}
f_{y}\left(y_{i} \mid n_{i} ; N_{i}\right)=w_{i}^{-1} f_{p}\left(p_{i} / w_{i} \mid n_{i} ; N_{i}\right), \tag{6}
\end{equation*}
$$

where $f_{p}\left(p_{i} / w_{i} \mid n_{i} ; N_{i}\right)$ is given by (3) upon substituting $p_{i} / w_{i}$ for $p_{i}$. In addition, since the set of conditional estimates are statistically independent by virtue of (5), write:

$$
\begin{equation*}
f_{\ell}\left(\ell \mid n_{1}, N_{1} ; \ldots ; n_{M} N_{M}\right)=f_{y 1}\left(\ell \mid n_{1}, N_{1}\right) \otimes \cdots \otimes f_{y M}\left(\ell \mid n_{M}, N_{M}\right), \tag{7}
\end{equation*}
$$

where $\otimes$ denotes a convolution operation, i.e., $f(y) \otimes g(y)=\int f(y-x) g(x) d x$. Finally, the associated confidence $C$, analogous to (4), for the weighted estimate $\ell$ of regional system Pd can be expressed as:

$$
\begin{equation*}
C=\operatorname{Prob}\left(\ell \geq p^{*} \mid n_{1}, N_{1} ; \ldots, n_{M}, N_{M}\right)=\int_{p^{*}}^{1} f_{\ell}\left(\ell \mid n_{1}, N_{1} ; \ldots, n_{M}, N_{M}\right) d \ell . \tag{8}
\end{equation*}
$$

The probability density function $f_{\ell}\left(\ell \mid n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)$ is referred to as the Augmented Beta Distribution. Note that this is not a Beta distribution. The integral does not have a tractable closed form solution, but could be evaluated for specific parameter values through numerical methods.

The Augmented Beta Distribution $f_{\ell}\left(\ell \mid n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)$ is also appropriate when multiple target types are present within Test. This occurs when the weighted estimate $\ell$ for regional system Pd consists of a Beta distributed success probability $p_{i}$ for each target type within each segment and $f_{\ell}\left(\ell \mid n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)$ results from a convolution across segment success

## JACYNA \& ROSEN

probabilities $p_{i}$, which are themselves a convolution of target type success probabilities within the segment.

## The Gaussian Approximation of System Pd

Returning to the weighted mean for system Pd (i.e., $\ell$ ), assume that $\ell$ is approximately Gaussian distributed for a sufficiently large number of segments $M$. Because $\quad \ell=\sum_{i} w_{i} p_{i}$, the corresponding mean $m_{\ell}\left(n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)=\sum_{i} w_{i} \mathrm{E}\left(p_{i}\right)$, where $\mathrm{E}($.$) is the expectation operator.$ However, since $p_{i}$ is Beta distributed, $\mathrm{E}\left(p_{i}\right)=\left(n_{i}+1\right) /\left(N_{i}+2\right)$ and the mean $m_{\ell}$ can be written as (Abramowitz \& Stegun, 1965, p. 930):

$$
\begin{equation*}
m_{\ell}\left(n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)=\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right) . \tag{9}
\end{equation*}
$$

Similarly, the associated variance is:

$$
\begin{equation*}
\sigma_{\ell}^{2}=\sum_{i=1}^{M} w_{i}^{2} \operatorname{Var}\left(p_{i}\right) \tag{10}
\end{equation*}
$$

where $\operatorname{Var}\left(p_{i}\right)=\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right) /\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)$ (Abramowitz \& Stegun, 1965, p. 930). Thus:

$$
\begin{equation*}
\sigma_{\ell}^{2}=\operatorname{Var}\left(\ell \mid n_{1}, N_{1} ; \ldots ; n_{M}, N_{M}\right)=\sum_{i=1}^{M} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)} \tag{11}
\end{equation*}
$$

Substituting these expressions for the mean and variance into the Standard Normal Distribution, allows us to compute an approximate ( $1-p^{*}$ ) confidence interval for $\ell$ as:

$$
\begin{align*}
C & =\left(2 \pi \sigma_{\ell}^{2}\right)^{-1 / 2} \int_{p^{*}}^{1} \exp \left\{-\left(s-m_{\ell}\right)^{2} / 2 \sigma_{\ell}^{2}\right\} d s  \tag{12}\\
& =F\left[\left(1-m_{\ell}\right) / \sigma_{\ell}\right]-F\left[\left(p^{*}-m_{\ell}\right) / \sigma_{\ell}\right],
\end{align*}
$$

where $F(x)=(2 \pi)^{-1 / 2} \int_{-\infty}^{x} \exp \left(-s^{2} / 2\right) d s$ is the Standard Normal Distribution.

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

The above assumption that the Augmented Beta Distribution converges to a Gaussian distribution for a large number of aggregate trials $\sum_{i} N_{i}$ holds when the following three conditions are satisfied:
I) The random variables $y_{i}$ are independent and have finite mean $\mu_{i}$ and variance $\sigma_{i}^{2}$.
II) A raw moment greater than $2+\delta$ is finite, i.e., $\mathrm{E}\left(\left|y_{i}\right|^{2+\delta}\right)$ is bounded for some $\delta>0$ and for every $1 \leq i \leq M$.
III) $\quad \lim _{M \rightarrow \infty} \frac{1}{S_{M}^{2+\delta}} \sum_{i=1}^{M} \mathrm{E}\left(\left|y_{i}-\mu_{i}\right|^{2+\delta}\right)=0$ for every $1 \leq i \leq M$ and for some $\delta>0$ (known as Lyapunov's Condition), where $S_{M}^{2+\delta}=\left(\sum_{i=1}^{M} \sigma_{i}^{2}\right)^{(2+\delta) / 2}$.

These three conditions describe the Lyapunov variant of the Central Limit Theorem. Proving that these conditions are satisfied for this problem will establish that the regional system sample mean is approximately Gaussian or normally-distributed. The details of this proof are given below. Recall by the previous definition that $y_{i}=w_{i} p_{i}$.

## Proof:

I. It is easily verified that both $\mu_{i}$ and $\sigma_{i}^{2}$ are bounded from (9) and (10).
II.
a. $\mathrm{E}\left(\left|y_{i}\right|^{2+\delta}\right)$ being bounded implies that there exists a real number $R \leq \infty$ such that $\mathrm{E}\left(\left|y_{i}\right|^{2+\delta}\right) \leq R$ for all $y_{i}=w_{i} p_{i}$
b. Letting $\delta=2$, it can be shown that $\mathrm{E}\left(\left|y_{i}\right|^{4}\right)$ is bounded, based on the use of the recursion relation $\mathrm{E}\left(y_{i}^{k}\right)=\left(n_{i}+k\right) \mathrm{E}\left(y_{i}^{k-1}\right) /\left(N_{i}+k+1\right)$ and the fact that $\mathrm{E}\left(y_{i}^{2}\right)$ is bounded (Johnson, Miller, \& Freund, 1995, p. 586).
III.
a. It can be shown that $\mathrm{E}\left(\left|y_{i}-\mu_{i}\right|^{4}\right) \leq w_{i}^{4}$, since $0 \leq \mathrm{E}\left(\left|p_{i}-\mathrm{E}\left(p_{i}\right)\right|^{4}\right) \leq 1$, which follows from the fact that:

## JACYNA \& ROSEN

$$
\begin{aligned}
\text { i. } & & y_{i}=w_{i} p_{i} . \\
\text { ii. } & & 0 \leq p_{i} \leq 1 . \\
\text { iii. } & & 0 \leq \mathrm{E}\left(p_{i}\right) \leq 1 .
\end{aligned}
$$

b. Additionally, because $S_{M}^{2}=\sum_{i=1}^{M} \sigma_{i}^{2}$, as previously defined in condition III, conclude $S_{M}^{2} \geq \sigma_{0}^{2} \sum_{i=1}^{M} w_{i}^{2}$, when $\sigma_{0}^{2}=\min _{i}\left[\operatorname{Var}\left(p_{i}\right)\right]>0$, using the fact that there always exists a smallest non-zero $\operatorname{Var}\left(p_{i}\right)$ by virtue of (10).
c. Combining the results of the first two steps, i.e., $0 \leq \mathrm{E}\left(\left|p_{i}-\mathrm{E}\left(p_{i}\right)\right|^{4}\right) \leq 1$ and $S_{M}^{2} \geq \sigma_{0}^{2} \sum_{i=1}^{M} w_{i}^{2}$, the expression $S_{M}^{2} \sum_{i=1}^{M} \mathrm{E}\left(\left|p_{i}-\mathrm{E}\left(p_{i}\right)\right|^{4}\right)$ is bounded by:

$$
\begin{equation*}
0 \leq \frac{1}{S_{M}^{4}} \sum_{i=i}^{M} E\left(\left|y_{i}-\mu_{i}\right|^{4}\right) \leq \frac{1}{\sigma_{0}^{4}} \frac{\sum_{i=1}^{M} w_{i}^{4}}{\left(\sum_{i=1}^{M} w_{i}^{2}\right)^{2}} . \tag{13}
\end{equation*}
$$

This bound implies that if $\lim _{M \rightarrow \infty} \frac{\sum_{i=1}^{M} w_{i}^{4}}{\left(\sum_{i=1}^{M} w_{i}^{2}\right)^{2}}=0$, then the following limit $\lim _{M \rightarrow \infty} \frac{1}{S_{M}^{4}} \sum_{i=1}^{M} E\left(\left|y_{i}-\mu_{i}\right|^{4}\right)=0$, which allows us to conclude that condition III is satisfied.
d. Without loss of generality, assume that the weights are bounded above and below by $w_{U}$ and $w_{L}$, respectively, such that $0 \leq w_{L} \leq w_{i} \leq w_{U} \leq 1$ Even if there did exist a finite number of zero weights, the remaining weights could be reindexed so that $w_{L}>0$. Now, it follows that $\sum_{i=1}^{M} w_{i}^{4} \leq M w_{U}^{4}$ and $\quad \sum_{i=1}^{M} w_{i}^{2} \geq M w_{L}^{2}$. This implies that $\sum_{i=1}^{M} w_{i}^{4} /\left(\sum_{i=1}^{M} w_{i}^{2}\right)^{4} \leq\left(w_{U} / w_{L}\right)^{4} / M$. From (13), conclude that the Lyapunov condition is satisfied since
$\lim _{M \rightarrow \infty} \frac{1}{s_{M}^{4}} \sum_{i=1}^{M} E\left(\left|y_{i}-\mu_{i}\right|^{4}\right)=0$ and condition III is satisfied when there are an infinite number of nonzero but bounded weights.
e. For the case where all the weights are uniform, $w_{i}=1 / M$; $1 \leq i \leq M$, it follows from the results in (d) that condition III is again satisfied.

By the Lyapunov variant of the Central Limit Theorem, conditions I-III being true imply that the sum or average of $y_{i}$ is Gaussian for large $M$ and that the regional system sample means for Pd are Gaussian. However, for the highly unlikely cases where all but a small number of arrival weights $w_{i}$ are zero, then the right-hand side of (13) may no longer be zero in the limit of large $M$ resulting in the Gaussian approximation becoming invalid. Intuitively though, it can be reasoned that this case is not possible, because a segment of an operationally valid route cannot have a zero probability of being traversed by an item-of-interest. This would be especially true for a segment selected for test.

## Example of the Augmented Beta and its Gaussian Approximation

As an illustration, consider the following hypothetical example where a weighted detection probability estimate is constructed from ten independent segments. The number of trials per segment for this example is $5,10,8,6,9,7,4,5,8$, and 9 , respectively. Note, however, that the number of trials per segment has no impact on the convergence to Normality. The corresponding number of successful detections declared is $4,8,7,5,7,5,2,4,6$, and 8 , respectively. In addition, the probability associated with choosing a given segment is $0.2,0.1,0.3,0.05,0.1$, $0.025,0.15,0.025,0.025$, and 0.025 , respectively. Using (7), the Augmented Beta Distribution can be computed and is of the form illustrated in Figure 3. The convolutions are approximated discretely for a step size $r=0.0001$ so that the integrated density in the interval $[0,1]$ is nearly unity. The distribution is shown to be both uni-modal and approximately symmetric about its mean value ( $\sim 0.7$ ). In addition, the distribution can be well-approximated by a Gaussian distributionit passes both the Kolmogorov-Smirnov and Chi-Square Goodness-of-Fit tests.

This illustration suggests that the Augmented Beta Distribution converges to the Gaussian distribution using a batch of 10 segments. Therefore, it is reasonable to infer that the regional system sample mean, which is going to be obtained from a batch of $N$ segments with $N>30$, will rapidly converge to the Gaussian

## JACYNA \& ROSEN

distribution. The proof provided above that as $M$ gets large, the regional system sample mean becomes Gaussian is borne out in this illustration which suggests that this convergence begins to occur when $M=10$.


Figure 3. The Augmented Beta Distribution corresponding to 10 independent segments, where the number of hypothetical trials per segment is $N=[5,10,8,6,9,7,4,5,8,9]$, the number of detections is $n=[4,8,7,5,7,5,2,4,6,8]$ and the associated probability of choosing each segment is $q=[0.2,0.1,0.3,0.05,0.1,0.025,0.15,0.025,0.025$, 0.025].

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION



Figure 4. The confidence $C$ for the Augmented Beta Distribution associated with the weighted detection probabilities for 10 independent segments, where the number of hypothetical trials per segment is $N=[5,10,8,6,9,7,4,5,8,9]$, the number of detections is $n=[4,8,7,5,7,5,2,4,6,8]$ and the associated probability of choosing each trial is $q=[0.2,0.1,0.3,0.05,0.1,0.025,0.15,0.025,0.025,0.025]$.

Figure 4 is a plot of the weighted detection probability confidence $C$ as a function of the exceedance probability $p^{*}$ for both the Augmented Beta Distribution and its corresponding Gaussian approximation. It is apparent that both distribution functions are nearly identical. This is true when the aggregate number of trials $\sum_{i} N_{i}$ is sufficiently large. For a confidence $C=0.90$, the exceedance probability is approximately 0.65 for either the exact or Gaussian approximation.

## Extension to Pid

The previous analysis for Pd presented above assumed that each segment $i$ has a single, but unknown success probability $p_{i}$ representative of the entire segment.

## JACYNA \& ROSEN

Consider the case where a specific segment $i$ consists of multiple distinct and unknown success probabilities $p_{i, j}$, where $j=1,2, \ldots, J$. This situation is relevant to the identification problem for a given segment, where a single success probability $p$ is not representative of the probability of successful or correct identification across a whole segment because multiple distinct success probabilities $p_{i, j}$ exist. Our reasoning is based on the argument that the window of processing for an identification trial is much smaller than for a detection trial. In particular, the window for detection extends across the entire segment. Because identification is a human-driven action comprised of multiple concurrent processing tasks, the trial window cannot possibility extend across the whole segment. Therefore, identification attempts within a segment can occur at different locations where the system attributes can vary. This implies that the success probabilities between any two identification trials, even within a segment, cannot be assumed to be equal.

Consider how the analysis above can be extended to address Pid, where multiple distinct and unknown success probabilities $p_{i, j}$ occur within a segment. Similar to the derivation above, assume that each segment consists of $N_{i}$ trials but now $n_{i}$ corresponds to successfully declared identifications. The number of successful identifications can be characterized by $n_{i}=\sum_{n=1}^{N_{i}} \chi_{n, i}$, where $\chi_{n, i}=\{0,1\}$ is an indicator function that represents an incorrect or correct identification, respectively. It is assumed that $\chi_{n, i}$ can be drawn from a mixture distribution of the form:

$$
\chi_{n, i} \sim\left\{\begin{array}{cc}
B\left(x, 1 ; p_{1, i}\right) & \text { with probability } q_{1, i}  \tag{14}\\
B\left(x, 1 ; p_{2, i}\right) & \text { with probability } q_{2, i} \\
\vdots \\
B\left(x, 1 ; p_{j, i}\right) & \text { with probability } q_{\mathrm{j}, i}
\end{array}\right.
$$

where $B\left(x, 1, p_{j, i}\right)$ is a binomial distribution for a single trial subinterval within a segment $i$, having success probability $p_{j, i}, x=\{0,1\}$ and $q_{j, i}$ is the probability that $\chi_{n, i}$ is drawn from the distribution characterizing subinterval $j$ on segment $i$. Let $p_{i}$ be the probability associated with the random variable $\chi_{n, i}$ such that:

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

$$
\begin{equation*}
p_{i}=\operatorname{Prob}\left(\chi_{n, i}\right)=\sum_{j=1} q_{j, i} p_{j, i} \tag{15}
\end{equation*}
$$

Given that each trial subinterval is independent of any other trial (naturally resulting since the location window varies for each identification trial) and the mixture is uniform across all subintervals, $n_{i}$ is binomially-distributed with average success probability $p_{i}$. For this analysis, it is not necessary to know the probabilities $p_{j, i}$ and $q_{j, i}$. From this result, it now follows that the Augmented Beta Distribution can be used for Pid and the Lyapunov convergence proof outlined above is valid for this generalization provided that the exceedance probability is interpreted as the average system identification performance along a given segment. The exact expression for the confidence $C$ in (8) and its Gaussian approximation ( $c f$. (12)) can then be used without modification. Moreover, the mixture distribution characterization also allows for a relaxing of the assumption that the success probabilities for detection must be equal for trials within a segment. This suggests that these confidence intervals can also be used on test designs consisting of multiple target types.

The following example illustrates how a mixture distribution of uniform subintervals within a segment with varying success probabilities results in the average number of successes on the segment being binomially distributed. Suppose there are 10 independent trials $\left(N_{i}=10\right)$ along a given segment $i$, where the unknown success probabilities are $p_{j, i}=[0.9,0.7,0.4,0.8]$ and the occurrence probabilities associated with these success probabilities are $q_{j, i}=0.25$ for $j=1,2, \ldots, 4$. Figure 5 depicts the distribution function for the number of successful detection attempts resulting from a Monte-Carlo sampling of the mixture distribution as defined previously in (15) with the $p_{j, i}$ and $q_{j, i}$ values noted above.

The blue bars represent the histogram resulting from the Monte-Carlo sampling with the theoretical binomial distribution (red curve) overlaid using the average success probability defined in (15). The results show excellent agreement between the two distributions and justify the use of applying the Gaussian approximation results for Pd to Pid .

## JACYNA \& ROSEN



Figure 5. A comparison of the Monte Carlo-based distribution function and the theoretical binomial distribution for the following mixture distribution: $p_{j, i}=[0.9,0.7,0.4,0.8]$ and $q_{j, i}=0.25$ for $j=1,2, \ldots, 4$, where $N_{i}=10$.

## Confidence Interval Equations and Sample Calculations

Consider the equations for constructing the confidence interval for the system mean Pd and Pid, as verified in the analysis provided above. The method generates the probability or confidence that the regional system sample mean exceeds a given threshold and is obtained under a Normal approximation when the segment success probabilities $p_{i}$ are Beta distributed. The equations for constructing the confidence interval are given below and are followed by a numerical example depicting their implementation on hypothetical data.

Given that the distribution of regional system sample mean is approximately Normally distributed and the segment success probabilities $p_{i}$ are Beta distributed, the following equation is used to compute the probability that the system Pd and Pid mean exceeds a given threshold $p^{*}$ with some confidence $C$ :

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

$$
\begin{aligned}
C & =\left(2 \pi \sigma_{\ell}^{2}\right)^{-1 / 2} \int_{p^{*}}^{1} \exp \left\{-\left(s-m_{\ell}\right)^{2} / 2 \sigma_{\ell}^{2}\right\} d s \\
& =F\left[\left(1-m_{\ell}\right) / \sigma_{\ell}\right]-F\left[\left(p^{*}-m_{\ell}\right) / \sigma_{\ell}\right]
\end{aligned}
$$

where

$$
\begin{aligned}
F(x) & =(2 \pi)^{-1 / 2} \int_{-\infty}^{x} \exp \left(-s^{2} / 2\right) d s, \\
m_{\ell} & =\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right), \quad \sigma_{\ell}^{2}=\sum_{i=1}^{M} w_{i}^{2} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)}
\end{aligned}
$$

$N_{i}$ and $n_{i}$ are the total number of detection attempts and successfully declared detections in SAT, respectively, for each segment $i$.

The following example shows how the above equation for Pd and Pid confidence intervals can be implemented using hypothetical test results consisting of observed detections and identifications distributed across non-identical segments within a region. The hypothetical test involves 56 segments with 5 potential trials occurring on each segment for detection and identification. The reader should note here that each segment may involve any mixture of target types since the proposed methodology is valid under any target type configuration supported by the program's current experimental design. Table 1 summarizes the hypothetical detection and identification observations.

A ' 1 ' appearing in Table 1 denotes a successful detection or identification while a ' 0 ' represents an unsuccessful attempt. An ' $x$ ' labeled within the identification columns indicates that a trial is not counted due to an unsuccessful detection.

The number of detection trials observed during this hypothetical test is 280 with 224 detections successfully declared. This, therefore, results in 224 potential identifications. The sample mean for Pd is simply $224 / 280=0.80$ and similarly the sample mean for Pid is $202 / 224=0.90$.

## JACYNA \& ROSEN

Table 1. Hypothetical Test Results

| Segment | Detection Trial |  |  |  |  | Identification Trial |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | \#1 | \#2 | \#3 | \#4 | \#5 | \#1 | \#2 | \#3 | \#4 | \#5 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | $x$ | 1 | 1 |
| 3 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | $x$ |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 | x | 1 | 1 | 1 | 1 |
| 8 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 11 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | x |
| 13 | 0 | 0 | 1 | 1 | 1 | x | x | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | x | x |
| 15 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | x | x |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 17 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 18 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 19 | 1 | 0 | 0 | 1 | 1 | 0 | x | x | 1 | 1 |
| 20 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 21 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 22 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 | 1 | 1 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 24 | 1 | 0 | 1 | 0 | 1 | 1 | x | 1 | x | 1 |
| 25 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 26 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 |
| 27 | 0 | 1 | 0 | 1 | 1 | x | 1 | x | 1 | 1 |
| 28 | 0 | 1 | 0 | 1 | 1 | x | 1 | x | 1 | 1 |
| 29 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 | 1 |
| 30 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 31 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | x |
| 32 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 |
| 33 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 |
| 34 | 0 | 1 | 1 | 1 | 0 | x | 1 | 1 | 1 | x |
| 35 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 37 | 1 | 1 | 1 | 0 | 0 | x | 1 | 1 | x | x |
| 38 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 39 | 1 | 0 | 0 | 1 | 1 | 1 | x | $x$ | 1 | 1 |
| 40 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 41 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 42 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | x | 1 | 1 |
| 43 | 0 | 0 | 1 | 1 | 1 | x | x | 1 | 1 | 1 |
| 44 | 1 | 0 | 1 | 1 | 1 | 1 | x | 1 | 1 | 1 |
| 45 | 1 | 0 | 0 | 1 | 1 | 1 | x | x | 0 | 1 |
| 46 | 0 | 0 | 0 | 0 | 0 | x | x | x | x | x |
| 47 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | x |
| 48 | 1 | 0 | 1 | 1 | 1 | 0 | x | 1 | 0 | 0 |
| 49 | 1 | 0 | 0 | 1 | 1 | 0 | x | x | 1 | 1 |
| 50 | 0 | 1 | 1 | 0 | 0 | x | 1 | 1 | x | x |
| 51 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 52 | 0 | 1 | 1 | 1 | 0 | x | 1 | 1 | 1 | x |
| 53 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 54 | 1 | 0 | 1 | 0 | 1 | 1 | x | 1 | $x$ | 1 |
| 55 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 56 | 0 | 1 | 0 | 1 | 1 | x | 1 | x | 1 | 1 |

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

To compute the confidence on the Pd and Pid mean, first compute the first and second central moments from (9) and (11), respectively, for both Pd and Pid:

$$
m_{l}=\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right), \quad \sigma_{l}^{2}=\sum_{i=1}^{M} w_{i}^{2} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)}
$$

These calculations for the first and second central moments are shown separately in the following two sub-sections for the Pd and Pid confidence intervals. The remaining steps required to establish that the regional sample mean Pd or Pid is greater than $p^{*}$ are also provided within each subsection.

## Confidence of Pd>p*

Because $w_{i}=1 / 30$ due to a necessary assumption of uniform threat arrival weights, the following calculations can be performed for $m_{l}$ and $\sigma_{l}^{2}$, respectively:

$$
\begin{aligned}
& m_{l}=\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right)=\sum_{i=1}^{30}\left(n_{i}+1\right) / 210=0.7145 \\
& \sigma_{l}^{2}=\sum_{i=1}^{M} w_{i}^{2} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)}=0.0004
\end{aligned}
$$

Converting the above values for mean and variance into a standardized random variable $z$ for an assumed exceedance probability $p^{*}=0.7$ :

$$
z_{1}=\left(p^{*-m_{l}}\right) / \sigma_{l}=(0.7-0.7145) / \sqrt{0.0004}=-0.7245,
$$

and

$$
z_{2}=\left(1-m_{l}\right) / \sigma_{l}=(1-0.7145) / \sqrt{0.0004}=14.2751
$$

with $z$ having the Standard Normal Distribution: $F(x)=(2 \pi)^{-1 / 2} \int_{-\infty}^{x} \exp (-s / 2) d s$.

## JACYNA \& ROSEN

Then, through the use of a standard look-up table for the Standard Normal Distribution (Johnson, Miller, \& Freund, 1994, p. 586), the probability that the mean system Pd is greater than $p^{*}$ is $F(14.2751)-F(-0.7245)=0.7656$.

## Confidence of Pid $>\boldsymbol{p}^{*}$

Again, since $w_{i}=1 / 30$, the following calculations can be performed for $m_{l}$ and $\sigma_{l}^{2}$, respectively:

$$
\begin{aligned}
& m_{l}=\sum_{i=1}^{M} w_{i}\left(n_{i}+1\right) /\left(N_{i}+2\right)=\sum_{i=1}^{30}\left(n_{i}+1\right) /\left(30\left(N_{i}+2\right)\right)=0.8044 \\
& \sigma_{l}^{2}=\sum_{i=1}^{M} w_{i}^{2} \frac{\left(n_{i}+1\right)\left(N_{i}-n_{i}+1\right)}{\left(N_{i}+2\right)^{2}\left(N_{i}+3\right)}=0.0009 .
\end{aligned}
$$

Note that for Pid, the values of $N_{i}$ change for each $i$ since the number of identification trials per segment is dependent on the number of successful detections for that segment. Similarly, converting the above values for mean and variance into a standardized random variable $z$ for an assumed exceedance probability $p^{*}=0.7$ :

$$
z_{1}=\left(p^{*}-m_{l}\right) / \sigma_{l}=(0.7-0.8044) / \sqrt{0.0009}=-3.480
$$

and

$$
z_{2}=\left(1-m_{l}\right) / \sigma_{l}=(1-0.8044) / \sqrt{0.0009}=6.5200
$$

Then, through the use of a standard look-up table for the Standard Normal Distribution, the probability that the regional system mean is greater than $p^{*}$ is $F(6.5200)-F(-3.480)=0.9997$. In summary, these example calculations show that there is a $76.56 \%$ and a $99.97 \%$ statistical confidence that the true system Pd and Pid mean is above 0.7, respectively, for this hypothetical set of test observations. It is apparent from the above example that more than five trials would be beneficial if the sample mean is within 0.1 of $p^{*}$. The difference in the hypothetical test sample means of 0.80 for Pd and 0.90 for Pid versus their Bayesian posterior expected values of 0.7145 and 0.8044 , respectively, illustrate

## BAYESIAN-BASED CONFIDENCE BOUNDS LYAPUNOV CONDITION

the need for additional trials for each segment. Moreover, this example showed that sample means around 0.80 only resulted in less than $80 \%$ confidence that the true mean is above 0.7 .

## Monte Carlo Confidence Interval Projector \& Implementation

The evaluation of the various candidate test designs for this regional test involved evaluating projected confidence interval widths and exceedance probabilities. Moreover, there is a motivation to more exactly understand the relationship between the number of segments and test trials and confidence bound widths. This understanding would facilitate the decision of selecting a design with the fewest number of trials while still maintaining a strong likelihood in achieving a specified and desired confidence width. To expedite this analysis, a Monte Carlo confidence interval tool (CI Projector) is developed to automate the calculation of the confidence intervals derived above.

The CI Projector tool is coded within an Excel environment using VBA. It requires the user to input a candidate test design through specifying the number of routes, the number of segments, and the number of trials per segment by filling out the blue columns titled 'Route', '\# of segments', and 'Ni' as shown in Figure 6. Note that Ni simply refers to the number of segments on a route.

Figure 6 also illustrates the view from the model during execution. The five columns in the middle denoted 'Segment successes' are the sampled number of successes for each segment during each iteration. Sample mean and variance values for the segments, routes, and area are tallied and averaged after each iteration. This also provides the user with a subjective understanding of the amount of variability present in the confidence interval widths from run to run.

## JACYNA \& ROSEN



Figure 6. CI Projector Screenshot

The success distribution for each segment is of course dependent on the projected sample mean for Pd and Pid. The user also inputs the projected sample means for Pd and Pid in addition to the number of Monte Carlo replications to be performed as shown below in Figure 7. Sample means for Pd must be defined for each object type passing through the system, and for adverse weather. The projected Pid sample mean strictly represents an average over all of these components.

| Iterations | 500 | Object A Pd | 0.65 |  |
| :--- | :---: | :--- | :---: | :---: |
| Ped Trials | 7 | Object B Pd | 0.9 | Simulate Test |
|  |  |  |  |  |
| Veh Trials | 7 | Object C Pd | 0.65 |  |
| Pid mean | 0.7 |  |  |  |

Figure 7. CI Projector Inputs

After defining the specific test design within the tool, the number of iterations must also be assigned by the user. The default value is 500 , but the user can adjust this value to any number. Due to the fast speed at which CI Projector runs coupled with the existence of a small amount of variability within the resulting confidence interval widths, a high number of iterations is recommended and should always be used. Lastly, the user simply clicks on the 'Simulate Test' button as shown to execute the model.

Sample outputs for the test projector tool are provided in Figure 8. Provided in the upper portion of the output columns are the necessary numerical values to compute the exceedance probability for a threshold value of interest that can be easily inferred through an attached table. The user also directly obtains lower and upper bounds for $60 \%, 70 \%$, and $80 \%$ confidence intervals.

| $\begin{gathered} \text { Var Pd } \\ 8.36796 \mathrm{E}-05 \end{gathered}$ | z1 -- Pd | z2 -- Pd | $\mathrm{P}(\mathrm{Pd} \gg .7)$ |  |
| :---: | :---: | :---: | :---: | :---: |
|  | -13.12814316 | 17.48080031 | compute w/ attached table |  |
|  | Average z1-Pd | Average z2-Pd | $\mathrm{P}(\mathrm{Pd}>.7)$ |  |
|  | -11.10064147 | 16.37713576 | compute w/ attached table |  |
|  | Lower Bound | Upper Bound | Avg Lower Bound | Avg Lpper Bound |
| 60\% interval <br> 70\% interval <br> 80 \% interval | 0.812407672 | 0.82777573 | 0.804260071 | 0.821446427 |
|  | 0.810578142 | 0.82960526 | 0.802214077 | 0.823492422 |
|  | 0.808382705 | 0.831800697 | 0.799758883 | 0.825947616 |

Figure 8. Sample Output Confidence Intervals from CI Projector

This process of inputting candidate test designs, running the model, and observing resulting confidence interval widths and exceedance probabilities can be easily repeated to evaluate a wide range of test designs. The CI Projector tool provides an environment conducive to short scenario set-up time and run time.

## JACYNA \& ROSEN

Numerous designs are evaluated using this tool, giving the decision maker reasonable assurance that the desired confidence interval widths and exceedance probabilities will be achieved as a result of an efficient test design.

## Conclusion

The probability density function (pdf) for regional system sample means is derived by considering them as a weighted estimate of the successful detection probability $\sum_{i} w_{i} p_{i}$ under a number of independent but not necessarily identically-distributed Bernoulli trials. It was shown that the resulting distribution (Augmented Beta Distribution) for $\sum_{i} w_{i} p_{i}$ is a convolution of non-identical Beta Distributions (cf. (7)). From this result, the corresponding confidence that the weighted estimate exceeds a given exceedance probability can be determined exactly from the Augmented Beta Distribution expression (cf. (8)). Then, the results were extended to address the regional system Pid sample mean. The analysis for the Pid sample means was more complicated and was based on the use of mixture distributions for cases where the success probability is no longer constant across a segment. It was shown that the same Pd results apply here provided that the exceedance probability is interpreted as the average system exceedance probability across the entire segment.

It was also shown through the Lyapunov variant of the Central Limit Theorem that the Augmented Beta Distribution converges to a Gaussian distribution as the number of segments grows large. The mathematical proof supplied showed that the Lyapunov conditions are satisfied for uniform segment priors $w_{i}$. Given this result, a simpler Gaussian approximation (cf. (12)) can be used to compute the confidence for both the regional system Pd and Pid sample mean that involves a simple aggregation of the detection and identification events observed across all segments during Test. These confidence intervals can be computed for individual target types and for a regional system average of all target types.

## References

Abramowitz, M. \& Stegun, I. A. (1965a). Handbook of mathematical functions: With formulas, graphs, and mathematical tables. New York, NY: Dover.

Bernardo, J. M. \& Smith, A. F. (2000). Bayesian theory. New York, NY: Wiley. p. 241-255.

Box, G. E. P., Hunter, J. S. \& Hunter, W. G. (2005). Statistics for experimenters ( $2^{\text {nd }}$ ed.). New York, NY: John Wiley and Sons.

Johnson, N. L., Kotz, S. \& Balakrishnan, N. (1995). Continuous univariate distributions (Vol. 2). New York, NY: Wiley.

Johnson, R. A., Miller, I., \& Freund, J. E. (1994). Miller and Fruend's probability and statistics for engineers. Englewood Cliffs, NJ: Prentice Hall. p. 586.

Karr, A. F. (1993). Probability. New York, NY: Springer-Verlag. p. 190192.

# Regularized Neural Network to Identify Potential Breast Cancer: A Bayesian Approach 

Hansapani S. Rodrigo<br>University of South Florida<br>Tampa, FL

Chris P. Tsokos<br>University of South Florida<br>Tampa, FL

Taysseer Sharaf<br>University of Michigan<br>Dearborn, MI

In the current study, we have exemplified the use of Bayesian neural networks for breast cancer classification using the evidence procedure. The optimal Bayesian network has $81 \%$ overall accuracy in correctly classifying the true status of breast cancer patients, $59 \%$ sensitivity in correctly detecting the malignancy and $83 \%$ specificity in correctly detecting the non-malignancy. The area under the receiver operating characteristic curve ( 0.7940 ) shows that this is a moderate classification model.

Keywords: Multi-Layer perceptron, classification, breast cancer, Bayesian

## Introduction

Early detection of breast cancer can reduce the deadly threat to life. Including the well-known "Gail model" (Gail et al., 1989), a number of other statistical models have been proposed to assess the risk of being diagnosed with breast cancer (Claus, Risch, \& Thompson, 1993; Domchek et al., 2003; van Asperen et al., 2004). However, these models imposed some limitations in their use of risk prediction (Amir et al., 2003; Euhus, Leitch, Huth, \& Peters, 2002).

The objective of the current study is to develop a better statistical model to correctly classify the malignant breast cancer patients with their demographic factors and previous mammogram results using a multi-layer perceptron (MLP), a type of feedforward neural network. Although there exist several other models based on neural networks with the same intention, few of them have make use of the evidence approach with automatic relevance determination (ARD) prior for

[^31]
## BAYESIAN NEURAL NETWORK BREAST CANCER ID

network regularization. We have selected the optimal network based on the model evidence (or cost function) as oppose to the classical minimum square error.

In order to train MLPs, we have considered two different approaches. In the first approach, a MLP is trained in the standard setting without incorporating any prior probabilities in their weight structure, where the later approach is based on Bayesian evidence procedure and the posterior probabilities of malignancy (Hung, Shanker, \& Hu, 2002) have been obtained. These probabilities have been used as an initial measure for risk of diagnosing with incident breast cancer.

The advantage of neural networks over the other models is that, it is a selflearning model which is free of statistical assumptions. This allows neural network process to be considered as a generalization of existing statistical methodologies.

MLPs are used in a wide variety of fields including pattern recognition, cognition and decision making (Ayer et al., 2010; Floyd, Lo, Yun, Sullivan, \& Kornguth, 1994; Orr, 2001; Wu et al., 1993), where they learn by examples through training algorithms. Training can be supervised, where both inputs and their corresponding outputs are fed to the network, or can be unsupervised, where training data consist of only the inputs. During the training process, the weights and the biases of the network are continuously adjusted to minimize the error between the network's output and the target outputs (Haykin, 1999). This process leads weights and biases of the network to learn the knowledge or information about the problem.

In the Bayesian approach, the uncertainty about the weight parameters is estimated from data itself and represented by a probability distribution (Bishop, 1995). Apart from capturing the uncertainties and providing a natural interpretation on regularization techniques, Bayesian approach has some other useful aspects. Automatic relevance determination process is one of them, which can be used to identify the relative importance of different inputs. This method also allows making predictions by combining several networks (network committees) in order to obtain improved performance.

## Multi-Layer Perceptron (MLP)

MLPs are a popular class of feedforward networks which represent a multivariate non-linear function mapping between a set of input and output variables (Bishop, 1995). These networks are organized as several interconnected layers. Each layer is a collection of artificial neurons (nodes) where connections among the layers have not formed any loops, hence the name feedforward. Data have been fed

## RODRIGO ET AL.

through the input layer, and then they pass through the hidden layer, and final outcome is given by the output layer.

The complexity of a MLP is directly proportional to the number of hidden nodes. It has been shown that a network with one hidden layer accompanied by sufficient number of hidden nodes is capable of approximating any continuous function (Hornik, Stinchcombe, \& White, 1989). Therefore, we have considered a MLP with one hidden layer (Figure 1) and the final outcome is given by (1).


Figure 1. A multi-layer perceptron network (MLP)

$$
\begin{equation*}
y(x ; w)=g(a)=g\left(\sum_{j=0}^{M} w_{1 j}^{(2)} h\left(\sum_{i=0}^{d} w_{j i}^{(1)} x_{i}^{\circ}\right)\right) \tag{1}
\end{equation*}
$$

During the training process, the goal is to minimize the difference between the actual and network predictions by adjusting the weights (including biases) using some optimization algorithms. A well trained MLP is capable of making

## BAYESIAN NEURAL NETWORK BREAST CANCER ID

reasonable predictions to unseen data, which is known as generalization. This is achieved by incorporating the regularization techniques like weight decay (Bishop, 1995). Next, we discuss some theory related to MLP for a two-class classification problem.

## Two-Class Classification Problem

For a two class classification, logistic sigmoid is selected as the activation function in the output layer. This is the activation function " $g$ " in (1), and has the form of

$$
\begin{equation*}
y(x ; w)=\frac{1}{1+\exp (-a)} \tag{2}
\end{equation*}
$$

In the Bayesian context, the $y(x ; w)$ can be interpreted as the probability of membership in class $C_{1}$ given the input vector $x$. The probability of membership of class $C_{2}$ is then given by $(1-y(x ; w))$.

## MLP with Maximum Likelihood (Standard Network)

Network training (minimizing the difference between the actual and network predictions) can be done in two ways, using conventional maximum likelihood and Bayesian approaches. In maximum likelihood, a single set of most likely values for the weights are found whereas in Bayesian, a probability distribution for weights is obtained to represent the uncertainty in the weight estimation.

For a set of training data $\left\{x^{n}, t^{n}\right\}$ which are independent and identically distributed, the likelihood can be written as in (3) (Assuming the data are coming from a Bernoulli distribution). $G(D \mid w)$ is the negative logarithm of the likelihood which is defined as the cross entropy error function as given in (4).

$$
\begin{gather*}
P(D \mid w)=\prod_{n} p\left(t^{n} \mid x^{n}, w\right)=\prod_{n} y\left(x^{n}, w\right)^{t^{n}}\left(1-y\left(x^{n}, w\right)\right)^{1-t^{n}}  \tag{3}\\
G(D \mid w)=-\sum_{n}\left\{t^{n} \ln y\left(x^{n}: w\right)+\left(1-t^{n}\right) \ln \left(1-y\left(x^{n}: w\right)\right)\right\} \tag{4}
\end{gather*}
$$

Instead of maximizing the likelihood (since it is a monotonically decreasing function), it is more convenient to minimize the cross-entropy. When training the standard MLP in our analysis we have used this error function. The predictions on

## RODRIGO ET AL.

new data are made using the optimal set of weights through the maximum likelihood method.

## MLP with Bayesian Techniques

In training a MLP, weights are adjusted whenever a new data point is presented to the network. A probability distribution which contains the degree of confidence associated with each different weight can be used to represent this uncertainty. The choice of prior distribution and about the corresponding posterior distribution will be discussed shortly.

## Network Regularization and Gaussian Prior

Smooth network mapping can be obtained by introducing network regularization techniques. This will lead for better generalization. In the simplest setting we have used a weight decay regularizer $E_{w}$ of the form (5).

$$
\begin{equation*}
E_{w}=\frac{1}{2}\|w\|^{2} \tag{5}
\end{equation*}
$$

As smaller weights (i.e a smaller $E_{w}$ ) are preferred for network weights, we have generated the weights from a zero mean Gaussian prior (6) initially.

$$
\begin{equation*}
P(w)=\frac{1}{Z_{w}(\alpha)} \exp \left(-\frac{\alpha}{2}\|w\|^{2}\right)=\frac{1}{Z_{w}(\alpha)} \exp \left(-\alpha E_{w}\right) \tag{6}
\end{equation*}
$$

where $Z_{w}=\left(\frac{2 \pi}{\alpha}\right)^{w / 2}$ and, $\alpha$ is the inverse variance of the distribution which is known as the hyper-parameter of the prior distribution. As a part of Bayesian learning we can optimize the hyper-parameter $\alpha$ (evidence procedure).

## Posterior Distribution of Weights

The posterior probability distribution for weights can be determined according to the Bayes' theorem by incorporating the above prior (6) and the data likelihood (3),

$$
\begin{equation*}
P(w \mid D)=\circ \frac{1}{Z_{s}} \exp \left(-\left(G(D \mid w)+\alpha E_{w}\right)\right)=\frac{1}{Z_{s}} \exp (-S(w)) \tag{7}
\end{equation*}
$$

where $Z_{s}$ is the normalization constant and $S(w)$ is the regularized cost function. The most probable weight vector $w_{M P}$ is found by maximizing the posterior, or minimizing the regularized cost function. From the second order Taylor series expansion of $S(w)$ around its minimum $w_{M P}$, we can obtain the following approximation.

$$
\begin{equation*}
S(w) \approx S\left(w_{M P}\right)+\frac{1}{2}\left(w-w_{M P}\right)^{T} A\left(w-w_{M P}\right) \tag{8}
\end{equation*}
$$

Where $A$ denotes the Hessian matrix of the regularized cost function. This leads to the Gaussian approximation to posterior distribution as given in (9) where $Z_{s}^{*}$ is the normalization constant.

$$
\begin{equation*}
P(w \mid D)=\circ \frac{1}{z_{s}^{*}} \exp \left(-S\left(w_{M P}\right)-\frac{1}{2} \Delta w^{T} A \Delta w\right) \tag{9}
\end{equation*}
$$

Using the above posterior distribution, obtain the network predictions for the probability that a new input vector $x^{*}$ belongs to class $C_{1}$ as in (10). Although this prediction is not directly achievable, we can use marginalized predictions to obtain the results as suggested by (MacKay, 1992):

$$
\begin{equation*}
P\left(C_{1} \mid x^{*}, D\right)=\int P\left(C_{1} \mid x^{*}, w\right) P(w \mid D) d w=\int y(x, w) P(w \mid D) d w \tag{10}
\end{equation*}
$$

## The Evidence Procedure

Prior to finding the above $w_{M P}$, it is needed to find the most probable hyperparameter $\alpha_{M P}$, which maximizes the posterior probability of weights in Bayesian setting (MacKay, 1996). This $\alpha_{M P}$ is obtained using the evidence $p(D \mid \alpha)$, by integrating the product of data likelihood and the prior distribution of the weights as given in (11).

$$
\begin{equation*}
p(D \mid \alpha)=\int p(D \mid w) p(w \mid \alpha) d \alpha \tag{11}
\end{equation*}
$$

## RODRIGO ET AL.

After several modifications, the logarithm of the evidence can be represented as in (12). The first term is the negative value of the regularized cost function, and the next two terms are the Occam factors that represent the ratio of posterior volume to prior volume. A network with higher number of hidden nodes has a large prior volume and thus, has a small Occam factor. Hence, these Occam factors act to penalize complex models and the evidence represents a trade-off between the accuracy and the complexity (MacKay, 1992).

$$
\begin{equation*}
\log E v=-S(\boldsymbol{w})+\log \left(O C C_{w}\right)+\log \left(O C C_{\alpha}\right) \tag{12}
\end{equation*}
$$

Periodically re-estimate $\alpha$ according to (13), in order to get the greatest $\log$ evidence value where $\gamma$ represents the effective number of weights whose values are controlled by the data rather than by the prior. Using that $\alpha_{M P}$ we can calculate the $w_{M P}$ (Thodberg, 1996). More details regarding this can be find in (Bishop, 1995).

$$
\begin{equation*}
\alpha^{N e w}=\frac{\gamma}{2 E_{w}} \tag{13}
\end{equation*}
$$

## The Automatic Relevance Determination

In the Bayesian setting, we can associate a separate hyper-parameter to each input variable which represents the inverse variance of the prior distribution of the weights fanning out from that input (Nabney, 2002). Optimal values for these hyper-parameters are obtained using the evidence procedure. So the weights connected to irrelevant inputs are automatically set to small values and this is known as the ARD approach.

## Committees

We can form a committee of networks to improve the prediction accuracies by combining several networks with different architectures. These networks can have different numbers of hidden nodes and/or they can be trained with different random initializations.

The simplest form of a committee, which involves taking the average predictions of the outputs of the $L$ networks, is given by (14). This will improve the accuracy of the predictions over an individual network output (Nabney, 2002).

$$
\begin{equation*}
y_{\text {COM }}(x)=\frac{1}{L} \sum_{i=1}^{L} y_{i}(x) \tag{14}
\end{equation*}
$$

## Methodology

## Implementation of MLPs

## Study Population

The data for this study are taken from Breast Cancer Surveillance Consortium (Barlow et al., 2006) for the period 1996 to 2002. The participating registries have obtained annual approvals from its institutional review board.

The data sample contains the information on menopausal type, age, breast density, ethnicity (Hispanic), body mass index (BMI), age at first birth, personal or family history of breast cancer, prior breast procedures, results of the last mammogram, type of menopause and current hormone therapy for each white woman. These women were aged from 35 to 84 years, and more details are available in Table 1.

## Implementation of the Standard and Bayesian MLPs

Training and testing data sets were created by partitioning the whole data sets each with $75 \%$ and $25 \%$ of data. A random sample out of the non-malignant group in the training set is selected and merged that with the malignant group in order to obtain a balanced training set. Table 2 represents the composition of data.

Different MLPs were trained using both standard and Bayesian approaches with varying number of hidden nodes from 1 to 25 . For all of these MLPs, a logistic sigmoid activation function and scaled conjugate gradient (SCG) training algorithm were used. SCG is selected as it is a faster training algorithm compared to other algorithms (Penny \& Roberts, 1999).

The standard MLP is trained using 10 fold cross-validation method and without any weight regularization. In 10 fold cross-validation, the training set is divided into 10 distinct segments, where 9 of those are used to train the network while the remaining segment is used for validation. This process is repeated for each of the 10 possible choices of the segments which are omitted from the training process and the validation errors (cross-entropy error) are averaged over all 10 results. The best network (with the corresponding hidden nodes) in this approach is the one with the smallest average cross-entropy in the validation data set (Kline \& Berardi, 2005).

RODRIGO ET AL.

Table 1. Details of the Study Population

|  |  | Malignant (\%) |  | Not Malignant (\%) |  |  | Total |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Total | 1053 | 6.47 | 15218 | 93.53 | 16271 | 100.00 |
| 1 | Menopausal Type ( $\mathrm{X}_{1}$ ) |  |  |  |  |  |  |
|  | Premenopausal | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |
|  | Postmenopausal | 826 | 78.44 | 12336 | 81.06 | 13162 | 80.89 |
| 2 | Age Group ( $\mathrm{X}_{2}$ ) |  |  |  |  |  |  |
|  | 35-39 | 6 | 0.57 | 496 | 3.26 | 502 | 3.09 |
|  | 40-44 | 72 | 6.84 | 788 | 5.18 | 860 | 5.29 |
|  | 45-49 | 137 | 13.01 | 2355 | 15.48 | 2492 | 15.32 |
|  | 50-54 | 168 | 15.95 | 2695 | 17.71 | 2863 | 17.60 |
|  | 55-59 | 150 | 14.25 | 1872 | 12.30 | 2022 | 12.43 |
|  | 60-64 | 141 | 13.39 | 1663 | 10.93 | 1804 | 11.09 |
|  | 65-69 | 131 | 12.44 | 1533 | 10.07 | 1664 | 10.23 |
|  | 70-74 | 96 | 9.12 | 1477 | 9.71 | 1573 | 9.67 |
|  | 75-79 | 93 | 8.83 | 1343 | 8.83 | 1436 | 8.83 |
|  | 80-84 | 59 | 5.60 | 996 | 6.54 | 1055 | 6.48 |
| 3 | Breast Density ( $\mathrm{X}_{3}$ ) |  |  |  |  |  |  |
|  | Almost entirely fat | 31 | 2.94 | 2575 | 16.92 | 2606 | 16.02 |
|  | Scattered fibroglandular densities | 405 | 38.46 | 5319 | 34.95 | 5724 | 35.18 |
|  | Heterogeneously dense | 506 | 48.05 | 4993 | 32.81 | 5499 | 33.80 |
|  | Extremely dense | 111 | 10.54 | 2331 | 15.32 | 2442 | 15.01 |
| 4 | Hispanic $\left(X_{4}\right)$ |  |  |  |  |  |  |
|  | No | 1026 | 97.44 | 12476 | 81.98 | 13502 | 82.98 |
|  | Yes | 27 | 2.56 | 2742 | 18.02 | 2769 | 17.02 |
| 5 | $\mathrm{BMI}\left(X_{5}\right)$ |  |  |  |  |  |  |
|  | 10-24.99 | 432 | 41.03 | 4969 | 32.65 | 5401 | 33.19 |
|  | 25-29.99 | 326 | 30.96 | 4404 | 28.94 | 4730 | 29.07 |
|  | 30-34.99 | 181 | 17.19 | 3304 | 21.71 | 3485 | 21.42 |
|  | 35 or more | 114 | 10.83 | 2541 | 16.70 | 2655 | 16.32 |
| 6 | Age at First Birth ( $\mathrm{X}_{6}$ ) |  |  |  |  |  |  |
|  | Age<30 | 692 | 65.72 | 7654 | 50.30 | 8346 | 51.29 |
|  | Age 30 or greater | 154 | 14.62 | 3412 | 22.42 | 3566 | 21.92 |
|  | Nulliparous | 207 | 19.66 | 4152 | 27.28 | 4359 | 26.79 |
| 7 | Number of first degree relatives with breast ca | ( $\mathrm{X}_{7}$ ) |  |  |  |  |  |
|  | Zero | 763 | 72.46 | 8515 | 55.95 | 9278 | 57.02 |
|  | One | 252 | 23.93 | 5077 | 33.36 | 5329 | 32.75 |
|  | Two or more | 38 | 3.61 | 1626 | 10.68 | 1664 | 10.23 |
| 8 | Previous breast procedure ( $X_{8}$ ) |  |  |  |  |  |  |
|  | No | 716 | 68.00 | 8925 | 58.65 | 9641 | 59.25 |
|  | Yes | 337 | 32.00 | 6293 | 41.35 | 6630 | 40.75 |
| 9 | Result of last mammogram before the index m | ogram |  |  |  |  |  |
|  | Negative | 1032 | 98.01 | 13244 | 87.03 | 14276 | 87.74 |
|  | False positive | 21 | 1.99 | 1974 | 12.97 | 1995 | 12.26 |
| 10 | Surgical menopause ( $\mathrm{X}_{10}$ ) |  |  |  |  |  |  |
|  | Natural | 576 | 54.70 | 7000 | 46.00 | 7576 | 46.56 |
|  | Surgical | 250 | 23.74 | 5336 | 35.06 | 5586 | 34.33 |
|  | Unknown | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |
| 11 | Current hormone therapy ( $\mathrm{X}_{11}$ ) |  |  |  |  |  |  |
|  | No | 400 | 37.99 | 6382 | 41.94 | 6782 | 41.68 |
|  | Yes | 426 | 40.46 | 5954 | 39.12 | 6380 | 39.21 |
|  | Unknown or not menopausal | 227 | 21.56 | 2882 | 18.94 | 3109 | 19.11 |

## BAYESIAN NEURAL NETWORK BREAST CANCER ID

Table 2. Summary of the training and testing data sets

| Data set | Malignant | Non-Malignant | Total |
| ---: | ---: | ---: | ---: |
| Training set | 829 | 1658 | 2487 |
| Test set | 224 | 3030 | 3254 |
| Total | 1053 | 4688 | 5741 |

Under the Bayesian approach, four types of networks were trained with different weight regularization techniques. The first network is trained using 10 fold cross validation along with a weight regularization. The second and third types of the networks are trained using Bayesian evidence procedure, one without and the other with ARD prior. For both of the above types, 10 different networks were trained with 10 different random initializations to examine the effect of local minima on solutions, and they were taken to construct the network committees. The optimal MLP with the lowest average regularized cost function in the training data (or the highest average log evidence) is then selected and used to predict the posterior probability of malignancy by simply averaging 10 network predictions from each committee. Additionally, a same type of neural network with one hidden node was built for a comparison, which is functionally equivalent to a logistic regression model.

As the final network type, 10 different networks were trained on 10 different random samples with varying number of hidden nodes along with evidence process and ARD prior. The best MLP is selected using the minimum of the regularized cost function.

## Model Evaluation

The selected ANN models are evaluated based on their accuracy, sensitivity, specificity values and the area under the receiver operating characteristic curve (AUC) for the testing data (Bradley, 1997; Friedman \& Wyatt, 2005). The proportions of correctly identified malignant and non-malignant women from the ANN models are known as the model accuracies. The proportions of actual malignant patients who are correctly identified from the models are known as the sensitivities and the proportions of non-malignant women who are correctly identified from the models are known as the specificities.

A perfect desirable predictor would be described as $100 \%$ sensitive (i.e. predicting all people from the malignant group as malignant) and $100 \%$ specific (i.e. predicting all non-malignant people as non-malignant). However, for any test,

## RODRIGO ET AL.

there is usually a trade-off between these two measures, and this can be represented graphically by a receiver operating characteristic curve.

## Results

The summary of our six optimal network types is given in Table 3. Overall accuracy in the logistic network ( $6^{\text {th }}$ MLP in the table) is lower than all other MLPs except for the MLP trained without ARD prior. Moreover it has the second lowest sensitivity and specificity values with the highest error. However, these models are not directly comparable in terms of their errors, as they have different settings and different training samples.

Table 3. Classification summary of the different MLP

| No | MLP Type | Error(Cross Entropy/Cost) | Accuracy | Sensitivity | Specificity |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Standard MLP | 641.96(valid error | 78.43\% | 55.36\% | 80.13\% |
| 2 | MLP with weight regularization | 434.77(valid error 8.28 ) | 74.09\% | 53.57\% | 75.61\% |
| 3 | MLP with evidence, but without ARD prior | 548.63 | 72.99\% | 60.71\% | 73.89\% |
| 4 | MLP with both evidence and ARD prior | 582.28 | 74.15\% | 59.82\% | 75.21\% |
| 5 | MLP trained on different samples with evidence and ARD prior | 908.78 | 81.35\% | 59.38\% | 82.97\% |
| 6 | MLP with one hidden node (logistic) | 1123.10 | 73.11\% | 55.35\% | 74.42\% |

Out of these MLP types, the best network in terms of the highest accuracy and specificity is found to be the MLP trained using different samples along with both evidence procedure and ARD prior ( $5^{\text {th }}$ MLP). As can be seen, use of the evidence procedure and the ARD prior has always resulted in better sensitivities. However, use of weight regularization without any optimization (evidence process) does not provide any significant improvement over the standard network training process.

It can be concluded that use of weight regularization techniques along with evidence process gives better results in Bayesian classification for most of the time. Apart from that, use of ARD prior helps to identify the most contributing variables to the network. Overall, Bayesian methods are preferred over the standard method mainly because of the natural way of handling the weight

## BAYESIAN NEURAL NETWORK BREAST CANCER ID

regularization. By forming committees, we were able to reduce the network training error.

The minimum and maximum prediction accuracies from these MLPs are $73 \%$ and $81 \%$, respectively. Sensitivity values are varying from a minimum of $54 \%$ up to a maximum of $61 \%$ while specificity values are varying from $74 \%$ to 83\%.


Figure 2. The receiver operating characteristic curves and the AUC values

The AUC values of all the above MLPs are greater than $70 \%$, which implies a moderate classification model. Figure 2 represents the receiver operating characteristic curves with the corresponding AUC values. The posterior probabilities of malignancy were obtained from the best Bayesian MLP network selected.

ARD prior identifies the relevant importance of the inputs in the network. Table 4 includes the rankings of the variable based on these hyper-parameter values. Risk factors with smaller hyper-parameters are highly contributing to the model outcome. Being in the age group 75 to 79 is the most critical factor in diagnosing with malignant breast cancer. Having a prior false positive mammogram can be an indication of malignant breast cancer. In accordance with cancer literature, risk factors such as having heterogeneously or extremely dense breast densities, and having a BMI of 35 or more are significantly contributing to the model.

## RODRIGO ET AL.

Table 4. Rankings of the attributable variables based on the ARD prior

| Rank | Alpha (hyper-parameter) | Variable | Risk Group |
| :---: | :---: | :---: | :---: |
| 1 | 0.3841 | agegrp9 | Age group 75-79 |
| 2 | 0.5550 | lastmamm | Result of last mammogram before the index mammogram <br> - False positive |
| 3 | 0.6489 | density3 | Density - Heterogeneously dense |
| 4 | 0.6846 | density 4 | Density - Extremely dense |
| 5 | 0.8251 | bmi4 | 35 or more |
| 6 | 1.3072 | agegrp2 | Age group 40-44 |
| 7 | 1.3872 | agegrp7 | Age group 65-69 |
| 8 | 1.6989 | hispanic | Hispanic or not - Yes |
| 9 | 1.7403 | nrelbc2 | Number of first degree relatives with breast cancer - 2 or more |
| 10 | 1.9510 | hrtYes | Current hormone therapy - Yes |
| 11 | 2.0528 | agegrp10 | Age group 80-84 |
| 12 | 2.0826 | bmi2 | 25-29.99 |
| 13 | 2.1980 | agegrp8 | Age group 70-74 |
| 14 | 2.2112 | hrtNo | Current hormone therapy - No |
| 15 | 2.8161 | agegrp6 | Age group 65-69 |
| 16 | 2.9341 | bmi3 | 30-34.99 |
| 17 | 3.2299 | agegrp5 | Age group 55-59 |
| 18 | 3.6520 | nrelbc1 | Number of first degree relatives with breast cancer - One |
| 19 | 3.7138 | surgnatural | Surgical menopause - Natural |
| 20 | 4.2249 | agegrp4 | Age group 50-54 |
| 21 | 5.0616 | surgsurgical | Surgical menopause - Surgical |
| 22 | 5.1547 | brstproc | Previous breast procedure - Yes |
| 23 | 5.7224 | density2 | Density - Scattered fibroglandular densities |
| 24 | 7.2989 | menopaus | Postmenopausal or age>=55 |
| 25 | 10.1388 | agenulli | Age at first birth - Nulliparous |
| 26 | 10.5538 | agegrp3 | Age group - 45-49 |
| 27 | 11.4664 | agegreater30 | Age at first birth - Age 30 or greater |

## BAYESIAN NEURAL NETWORK BREAST CANCER ID

## Conclusion

A breast cancer detection model was introduced using artificial neural network theory. With the intention of having a better classification, different types of MLPs were developed. These models are trained using the standard and Bayesian techniques. The first two models were validated using 10 -fold cross validation and we have constructed committees for the other models. Finally all MLPs were tested on a new set of test data.

The advantage of Bayesian MLP is that it gives the posterior probabilities for classification which can be used as a priori risk of diagnosing with breast cancer. The evidence procedure is used for the network regularization along with ARD prior. Use of ARD prior did not make any significant difference in the accuracy of our optimal MLP. Use of committees also did not show much difference in the overall results compared to the single network predictions alone. However, this has helped to give a low variance in the predictions.

The highest accuracy which was obtained from one of the Bayesian MLP is about $81 \%$ and this is a significant improvement over the other methods which used the same set of real data in terms of the discriminative accuracy. ROC curve provides information about a model's classification efficiency. A good classification model was obtained for the third and the fifth MLP with more than $75 \%$ area under the ROC curve. The model may be further improved by considering more relevant risk factors and more recent data, such as different races because ethnicity is one of the significant risk factors that contributes to the malignancy of breast cancer (Xu, Kepner, \& Tsokos, 2011).

It was also confirmed that ANN may have an important role in improving the accuracy and consistency of medical diagnosis. The proposed approach in developing the ANN model is free of assumptions, as opposed to parametric regression and hence increases the validity of our findings.

## References

Amir, E., Evans, D. G., Shenton, A., Lalloo, F., Moran, A., Boggis, C., ... Howell, A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. Journal of Medical Genetics, 40(11), 807-814. doi: $10.1136 / \mathrm{jmg}$.40.11.807

Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E., Jr., \& Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural

## RODRIGO ET AL.

networks revisited: Discrimination and calibration. Cancer, 116(14), 3310-3321. doi: 10.1002/cncr. 25081

Barlow, W. E., White, E., Ballard-Barbash, R., Vacek, P. M., Titus-Ernstoff, L., Carney, P. A., ... Kerlikowske, K. (2006). Prospective breast cancer risk prediction model for women undergoing screening mammography. Journal of the National Cancer Institute, 98(17), 1204-1214. doi: 10.1093/jnci/djj331

Bishop, C. M. (1995). Neural networks for pattern recognition. New York, NY: Oxford University Press.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Claus, E. B., Risch, N., \& Thompson, W. D. (1993). The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. Breast Cancer Research and Treatment, 28(2), 115-120. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8173064

Domchek, S. M., Eisen, A., Calzone, K., Stopfer, J., Blackwood, A., \& Weber, B. L. (2003). Application of breast cancer risk prediction models in clinical practice. Journal of Clinical Oncology, 21(4), 593-601. doi: 10.1200/jco.2003.07.007

Euhus, D. M., Leitch, A. M., Huth, J. F., \& Peters, G. N. (2002). Limitations of the Gail Model in the specialized breast cancer risk assessment clinic. The Breast Journal, 8(1), 23-27. doi: 10.1046/j.1524-4741.2002.08005.x

Floyd, C. E., Jr., Lo, J. Y., Yun, A. J., Sullivan, D. C., \& Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. Cancer, 74(11), 2944-2948. doi: 10.1002/1097-0142(19941201)74:11<2944::AID-CNCR2820741109>3.0.CO;2-F

Friedman, C. P., \& Wyatt, J. (2005). Evaluation methods in biomedical informatics. Springer Science \& Business Media. Retrieved from https://books.google.com/books?id=o2XyIpz_I1sC\&pgis=1

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., \& Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute, 81(24), 1879-1886. doi: 10.1093/jnci/81.24.1879

Haykin, S. (1999). Neural networks: A comprehensive foundation. USA:
Prentice Hall.

## BAYESIAN NEURAL NETWORK BREAST CANCER ID

Hornik, K., Stinchcombe, M., \& White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366. doi: 10.1016/0893-6080(89)90020-8

Hung, M. S., Shanker, M., \& Hu, M. Y. (2002). Estimating breast cancer risks using neural networks. Journal of the Operational Research Society, 53(2), 222-231. doi: 10.1057/sj/jors/2601276

Kline, D. M., \& Berardi, V. L. (2005). Revisiting squared-error and crossentropy functions for training neural network classifiers. Neural Computing and Applications, 14(4), 310-318. doi: 10.1007/s00521-005-0467-y

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. Neural Computation, 4(5), 720-736. doi: 10.1162/neco.1992.4.5.720

MacKay, D. J. C. (1996). Hyperparameters: Optimize, or integrate out? In G. R. Heidbreder (Ed.), Maximum entropy and bayesian methods (pp. 43-59). Retrieved from http://link.springer.com/chapter/10.1007/978-94-015-87297_2 Inpapers3://publication/uuid/4C73D436-E70B-49D8-AA7E-031803718650

Nabney, I. T. (2002). NETLAB: Algorithms for pattern recognition. Gateshead: Springer.

Orr, R. K. (2001). Use of an artificial neural network to quantitate risk of malignancy for abnormal mammograms. Surgery, 129(4), 459-466. doi: $10.1067 / \mathrm{msy} .2001 .112069$

Penny, W. D., \& Roberts, S. J. (1999). Bayesian neural networks for classification: How useful is the evidence framework? Neural Networks, 12(6), 877-892. doi: 10.1016/S0893-6080(99)00040-4

Thodberg, H. H. (1996). A review of Bayesian neural networks with application to near infrared spectroscopy. IEEE Computational Intelligence Society, 7(1), 56-72. doi: 10.1109/72.478392

Van Asperen, C. J., Jonker, M. A., Jacobi, C. E., van Diemen-Homan, J. E. M., Bakker, E., Breuning, M. H., ... de Bock, G. H. (2004). Risk estimation for healthy women from breast cancer families: new insights and new strategies. Cancer Epidemiology, Biomarkers \& Prevention, 13(1), 87-93. doi: 10.1158/1055-9965.EPI-03-0090

Wu, Y., Giger, M. L., Doi, K., Vyborny, C. J., Schmidt, R. A., \& Metz, C. E. (1993). Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology, 187(1), 81-87. doi: 10.1148/radiology.187.1.8451441

## RODRIGO ET AL.

Xu, Y., Kepner, J., \& Tsokos, C. P. (2011). Identify attributable variables and interactions in breast cancer. Journal of Applied Sciences, 11(6), 1033-1038. doi: 10.3923/jas.2011.1033.1038

# Monte Carlo Simulation Design for Evaluating Normal-Based Control Chart Properties 

John N. Dyer<br>Georgia Southern University<br>Statesboro, GA

The advent of more complicated control charting schemes has necessitated the use of Monte Carlo simulation (MCS) methods. Unfortunately, few sources exist to study effective design and validation of MCS methods related to control charting. This paper describes the design, issues, considerations and limitations for conducting normal-based control chart MCS studies, including choice of random number generator, simulation size requirements, and accuracy/error in simulation estimation. This paper also describes two design strategies for MCS for control chart evaluations and provides the programming code. As a result, this paper hopes to establish de facto MCS schemes aimed at guiding researchers and practitioners in validation and control-chart evaluation MCS design.

Keywords: Monte Carlo simulation, statistical process control, random number generation

## Introduction

Various control charts exist using a control chart statistic based on the Normal distribution, including the Shewhart, R-Chart, Individuals, S-Chart, Cumulative Sum (CUSUM), Exponentially-Weighted Moving Average (EWMA), Combined EWMA-Shewhart (CES), and Reverse Moving Average (RMA), among others. The performance of many control charts have been investigated using various analytical and numerical methods such as integral equations, saddle-point approximations, and Monte Carlo Markov Chain (MCMC) methods. Monte Carlo simulation (MCS) has also grown in popularity due to the relative ease of programmatic design, and the ability to investigate additional important performance measures of a control chart such as the median run-length (MRL), runlength quantiles, and the cumulative distribution function (CDF). Unfortunately

[^32]
## JOHN N. DYER

there are few sources available for researchers and practitioners to study effective design and validation of MCS methods related to control charting.

In general, MCS includes a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results, then running multiple simulations to obtain the distribution of an unknown probabilistic entity. MC methods are used in physical and mathematical problems and are useful when it is difficult or impossible to obtain a closed-form expression, or infeasible to apply a deterministic algorithm. MC methods are mainly used in three distinct problems classes: optimization, numerical integration, and studying probability distributions of random variables. The use of random numbers as input is a defining feature of MCS. This is what turns a deterministic model into a stochastic model. Regardless of the application, simulations of this kind should be impartial, systematic and reproducible. Sources of error need to be controllable or at least isolatable. The basic steps of conducting a Monte Carlo simulation can be summarized as follows (Salleh, 2013).

1. Create a model with appropriate parameters and assumptions.
2. Generate random numbers as inputs to the model.
3. Run the simulation and record the results and desired outputs.
4. Analyze the results with statistical and/or advanced modeling tools.

Although there is no strict definition MCS, in the field of statistical process control (SPC) and control charting, and for the purposes of this paper, it is broadly defined as the use of a programmatic pseudo-random number generation replicating repeated sampling from an assumed underlying statistical distribution, for the purposes of numerical integration of a function of a control-charting statistic and/or studying run-length (RL) properties and performance of the control chart. As such, there are several considerations related to control-charting MCS, including the choice, series length, and precision of the random number generator (RNG), as well as the required simulation size and expected accuracy/error in simulation estimation.

The advantage of using MCS in this manner allows one to more fully investigate the RL properties and performance of a control chart, over a wider array of performance measures including the average run-length (ARL), MRL, standard error of the run-length (SRL) and the CDF of the run-length, as well as percentiles and quartiles. The CDF measures the cumulative proportion or percent of signals given by the $i^{\text {th }}$ period following the shift. It should be noted that the CDF completely characterizes the run length distribution, while the ARL is only the mean. Additionally, the MRL can be used in conjunction with the ARL and CDF
since it is a better measure of central tendency for skewed distributions such as the run length distribution. The MRL is defined as the median ( $50^{\text {th }}$ percentile) number of sampling periods until the control chart signals. Although traditional analytical and numerical methods such as integral equation and saddle-point approximations, as well as MCMC methods, provide good estimates of ARLs at specified control limits (CLs) of a control chart, the methods can be very cumbersome, mathematically complicated, and do not readily allow wider studies of RL properties and performance measures beyond the ARL, or simultaneous evaluation over a wide range of CLs. Additionally, in most cases the MCS can provide equal or better estimates than the traditional methods. One can also use MCS to validate findings based on other methods mentioned above.

Many researchers and practitioners involved in SPC and control charting design and implementation are very familiar with Microsoft Excel and use it extensively for analysis and modeling, regardless of the inherent problems known to exist in Excel and the RNGs employed in Excel (Ahrens \& Dieter, 1988; Knusel, 2002; Benneyan, Lloyd, \& Plsek, 2003). Additionally, Excel has many built-in functions as well as offering the user a Visual Basic for Applications (VBA) interface for programming in Excel. Excel can also be used as a prototype or beta MCS for initial studies prior to full implementation. As such, this paper does not have the purpose to intensely discuss the advantages, disadvantages, similarities, differences, etc., regarding analytical/numerical/MCMC methods versus MCS, nor is it the purpose to compare and contrast MCS using a myriad of possible programming languages, such as R, Visual Basic, C+, Java, FORTRAN, etc. Instead, this paper's purpose is to describe the basic validation design, issues, considerations and limitations for normal-based control chart MCS design, including choice of RNGs, RNG series length, MCS simulation size, and accuracy/error in MCS estimation, while exemplifying using Excel 2010. The design principles can be easily extended to other programming languages and in a variety of field requiring simulation. It is assumed that the reader is familiar with basics of elementary statistics, control charting and the normal distribution. For detailed introductions to these ideas, the reader is referred to (NIST/SEMATECH, 2012; Montgomery, 1996; Ryan, 2000; Wheeler \& Chambers, 1992).

## Normal Based Control Charts and Performance Criteria

SPC techniques have been used for decades to monitor and control a process, most often a manufacturing process, but are seeing increased use in fields broadly related to health care (Benneyan et al., 2003; Srinivasan, 2011), information technology

JOHN N. DYER

(Abdel-Aziz, Abdel \& Darwis, 2008), finance (Golosnoy \& Schmid, 2007; Severin \& Schmid, 1998), and business process monitoring and improvement (Jaing, Au, \& Tsui, 2007). The idea is to plot data (a control chart statistic) over time to aid in determining trends or changes in the process variability. In any process there exists a certain amount of inherent, common cause variability. This common cause variability is usually small, yet unavoidable. In contrast, variability from assignable causes is generally large, and can usually be removed from the process if detected. The primary use of control charts is to detect any assignable causes or process changes as quickly as possible, thus enabling quick action in elimination of the assignable cause.

Control charts can be used to monitor many aspects of the process, but the most common use is for monitoring the process mean and/or variance. To monitor the mean, individual observations or averages (or functions of) are plotted over time, where these plotted values are estimates of the process mean. Likewise, sample ranges or standard deviations are plotted against time as estimates of process variability.

When evaluating control chart performance, the ARL has typically been used to quantify performance of the chart. The ARL is defined as the average number of time periods until the control chart signals, and can be defined for both the incontrol (IC) and out-of-control (OC) cases. A more recent alternative performance criterion is the CDF (Dyer \& Barrett, 2000; Dyer, Conerly, Adams, \& Barrett, 2002; Dyer, Conerly, \& Adams, 2003; Dyer, Adams, \& Conerly, 2003; Lin \& Adams, 1996) and MRL. The CDF measures the cumulative proportion or percent of signals given by the ith period following the shift. It should be noted that the CDF completely characterizes the RL distribution, while the ARL is only the mean. Additionally, the MRL can be used in conjunction with the ARL and CDF since it is a better measure of central tendency for skewed distributions such as the RL distribution (Gan, 1993). The MRL is defined as the median ( $50^{\text {th }}$ percentile) number of time periods until the control chart signals.

Although there are a myriad of control charting schemes, many are based on an assumption of plotting a control chart statistic related to individual measures or the mean of a subgroup of measures against control-limits that are a function of the normal distribution. These control charts include the Shewhart (Shewhart, 1931/1980; Shewhart, 1939/1986; Roberts, 1959) and Individuals charts (with and without runs-rules) (Nelson, 1984; Western Electric Company, 1956), MultipleSampling schemes (Daudin, 1992; He, Grigoryan, \& Sigh, 2002; Irianto \& Shinozaki, 1998; Teoh \& Khoo, 2012; Torng \& Lee, 2009), the CUSUM (Page, 1954), EWMA (various schemes) (Hunter, 1986; Ryan, 2000) CES (Lucas \&

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

Saccucci, 1990), and RMA (Dyer, Adams, \& Conerly 2003) charts, among others. In many cases a measure of variability is also charted. The Shewhart and Individuals charts are straightforward in determining control-limits and control chart performance since the distribution of the IC and OC sequentially plotted statistics are assumed to follow an independent and identical Normal distribution (iidN), hence the RL properties can be studied using the Geometric distribution. It should be noted that the IC and OC processes follow different iidN distributions.

For example, assuming an IC process, using a Shewhart chart and $\pm 3 \sigma$ CLs, the probability of a signal $(p)$ is 0.0027 , that is, the probability the plotted statistic exceeds the CLs (although it is a false-alarm), corresponding to ARL $=370$. Since the count of the number of sampling periods until a false-alarm occurs (the run length) follows a Geometric distribution (Chen, 1997), the ARL, SRL and MRL are given by

$$
\begin{gather*}
\mu=\mathrm{ARL}=\frac{1}{p}  \tag{1}\\
\sigma=\mathrm{SRL}=\frac{\sqrt{1-p}}{p}  \tag{2}\\
\mathrm{Med}=\mathrm{MRL}=\frac{\ln (0.50)}{\ln (1-p)} \tag{3}
\end{gather*}
$$

Solving (1), (2), and (3) for $p=0.0027$ yields $\mathrm{ARL}=370, \mathrm{SRL}=370$, and MRL $=256$. For cases in which consecutive observations are independent, for the Geometric distribution, $\mu \approx \sigma$, that is, $\mathrm{ARL} \approx \mathrm{SRL}$. This result is not true of several common control charting methods such as the EWMA, CUSUM and RMA. Additionally, for the Geometric distribution the first quartile $\left(\mathrm{Q}_{1}\right)$ and third quartile $\left(\mathrm{Q}_{3}\right) \mathrm{RLs}$ are given by

$$
\begin{align*}
& \mathrm{Q}_{1}=\mathrm{QRL}_{1}=\frac{\ln (0.75)}{\ln (1-p)}  \tag{4}\\
& \mathrm{Q}_{3}=\mathrm{QRL}_{3}=\frac{\ln (0.25)}{\ln (1-p)} \tag{5}
\end{align*}
$$

Solving (4) and (5) following the above example yields QRL $_{1}=106$ and $\mathrm{QRL}_{2}=512$. In general, the $p^{\text {th }}$ percentile run-length value is given by

$$
\begin{equation*}
\mathrm{RL}_{p^{\mathrm{th}} p}=\frac{\ln \left(1-p^{\text {th }} \text { percentile }\right)}{\ln (1-p)} \tag{6}
\end{equation*}
$$

Now, assume a shift of $1 \sigma$ in the process mean. The probability of a signal is now $p=0.0228$. In this case, the $\mathrm{ARL} \approx 44, \mathrm{SRL} \approx 43$, $\mathrm{MRL} \approx 30, \mathrm{QRL}_{1} \approx 12$, and $\mathrm{QRL}_{3} \approx 60$. The IC and OC processes now follow different iidN distributions following the shift in the process mean of the underlying distribution; hence the RL distributions are different iid Geometric (iidG) distributions.

Although the underlying processes used for the CUSUM, EWMA and RMA control chart statistics are assumed to be iidN, the sequentially plotted control-chart statistics are not independent since they are a form of cumulative sums or averages; hence the RL distribution cannot be exactly described as above. The Shewhart chart is said to have "no memory" while the CUSUM, EWMA, RMA (and others) are said to have "memory" (Kalgonda, Koshti, \& Ashokan, 2011). Memory refers to if the control-chart statistic uses past data from a previous sampling period. It is known that CLs for control-charts with memory are much different than for the underlying iidN process. In this case of the lack of independence (memory), it can be difficult to use the traditional analytical or numerical methods to study RL properties and control chart performance, hence many MCS studies have been conducted to determine appropriate CLs as well as control chart performance measures (Dyer, Conerly, \& Adams, 2003; Fu \& Hu, 1999; Lin \& Adams, 1996).

Additionally, MCS can also be employed to more readily determine overall RLs when multiple charts are used to monitor the same process, such as simultaneous use of charts for the mean and variability. Most of the aforementioned control-charts also have very limited tabulations of IC and OC ARLs, and sparse literature regarding other important measures like the MRL, percentiles/quartiles, SRL, CDF, or RL distribution studies.

## Designing the Validation MCS for Control Chart Evaluation

One advantage of using the design exemplified in this paper is the ability to produce the RL distribution for many different sets of CLs simultaneously. For example, most MCS programs allow specification of a single set of CLs, thus calculating a single set of performance measures (ARL, MRL, SRL, CDF, etc.) from the

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

resulting simulated run-lengths. Alternatively, this design allows specification of up to 16,380 different sets of CLs, thus calculating as many different sets of summary measures. This would be an absurd case, but it reflects the capability of the design and the ability to use a single simulation run across all desired CLs.

Regardless of the programming language, software, or RNG being used, this author proposes two different designs of a MCS for control charting the mean or individual observations based on an underlying iidN process, as described in the 4 steps below. The two designs, D1 and D2 respectively, differ only in step 4 discussed below. In D1, a series of length $m$ random numbers is independently generated $n \operatorname{sim}$ times, where $n \operatorname{sim}=$ the number of simulation runs. Then, one RL is recorded for each separate simulation, resulting in an array of nsim RLs. For example, setting $m=14,000$ and $n \operatorname{sim}=10,000$, each independent series of length 14,000 is produced 10,000 times, resulting in 10,000 recorded RLs. In D2, a single very long series of size $m$ is generated only one time. The number of RLs within the m random numbers is recorded. The optimal values of $m$ and $n \operatorname{sim}$ for D1 and the optimal value of $m$ and expected number of RLs are discussed in a subsequent section.

The basic design steps can be easily modified to accommodate a myriad of various control charting schemes, such as monitoring mean/variability simultaneously, or employing two charts for the mean simultaneously, like the CES control chart, and schemes based on runs-rules and those such as double-sampling. The MCS design steps are as follows.

1. Use a RNG to generate a series of length m, of subsets of size n of pseudo-random iidN variables ( $n=1,2,3 \ldots$ ) representing the simulated values of the underlying iidN process $\left(x_{i}\right), i=1$ to $m$. The variable $n$ represents a subgroup size from which the appropriate statistic is calculated, such as the subgroup mean. So, the result will be a series of $m$ means of subset size $n$. It is recommended that each of the $m$ means be standardized to represent the Standard Normal distribution, that is, $z \sim \mathrm{~N}(\mu=0, \sigma=1)$ where $z_{i}$ is the standardized subgroup mean. Note, that although $n=1$ for the Individuals chart, $n$ can also be 1 for the CUSUM and EWMA, but is usually $\leq 10$.
2. Transform the series of calculated $z$ statistics from step 1 to a series of control-chart statistics (CCS), appropriate to the control chart to be studied, e.g., CUSUM, EWMA, RMA, etc.

## JOHN N. DYER

3. Establish upper/lower CLs (or a range of CLs) by which to compare the series of CCS from step 2.

## For Design 1:

4. For the IC process, compare each sequential CCS from step 2 to the CLs established in step 3. When the first occurrence of a CCS exceeds one of the CLs, record the RL as $R L=i$, that is, the location within the series of $m$ statistics wherein CCS exceeded CL. Stop step 4.

Following step 4, repeat steps 1 through 4 nsim times (where nsim is the number of simulations) and calculate the summary measures like ARL, MRL, SRL, and desired percentiles/quartiles. This will produce a total of $n \operatorname{sim}$ RLs for each set of estimates. For the simulated OC process, induce a step-shift in the mean of the process in step 1 , e.g., a $1 \sigma$ shift where $\sigma$ is the process standard deviation. Again, complete the 4 steps nsim times and calculate the summary measures, including the CDF.

## For Design 2:

4. For the IC process, compare each sequential CCS from step 2 to the CLs established in step 3. When an occurrence of a CCS exceeds one of the CLs, record the RL as $R L=i$, that is, the location within the series of $m$ statistics wherein CCS exceeded one of the CLs. After recording the first $R L$, re-index the series so that next value in the series is 1 , and then continue step 4 , re-indexing and recording each subsequent $R L$ until the entire series has been evaluated. In this case, the series-length $m$ is much longer than D1.

After evaluating the entire series calculate the summary measures like ARL, MRL, SRL, CDF and desired percentiles/quartiles. This will produce an unknown but predictable number of sets of RLs for each set of estimates. For the simulated OC process, induce a step-shift in the mean of the process in step 1, e.g., a $1 \sigma$ shift where $\sigma$ is the process standard deviation. Again, complete the 4 steps and calculate the summary measures, including the CDF.

## MCS Design Considerations

Because the RL distribution for the $\mathrm{N}(\mu, \sigma)$ Shewhart/Individuals control charts is well known, one should first design a validation MCS by which to compare estimated results with known results. There are several important considerations to be made when designing the MCS according to the section above, including the choice of RNGs, the RNG series length and burn-in period, the choice of CLs, the number of simulations to conduct, and the resulting accuracy or error of the MCS. Since we already know the RL distribution of the Shewhart and Individuals chart, any MCS should begin by first validating the study against what is known about the IC $z \sim \mathrm{~N}(\mu=0, \sigma=1)$ process. That is, before transforming data into the CCS for the CUSUM, EWMA and RMA (or others), one should first test the MCS design against known properties and RL distribution of a Shewhart or Individuals chart. If the validation design is adequate, then one can make better assumptions regarding the adequacy of the design when modified for other control charting schemes. Although the literature reflects results using MCS for many control chart studies, few if any provide information regarding the estimated accuracy of results (estimated error) or degree of confidence in estimates, and few adequately describe the MCS design, the RNGs used or a justification of simulation size.

Additionally, when extending the validation design to other control charts, one should first validate at least a few summary measures from the MCS against what is already known in the literature. As such, the topics discussed in the subsections should first be applied against the known Shewhart or Individuals results.

## Choosing the Random Number Generator

As stated in design step 1 in a previous section, one must use a RNG to generate a series of pseudo-random numbers from an assumed distribution. This can be done by calling an existing RNG function (as in Excel), or employing existing and validated RNG subroutines (e.g., the IMSL subroutine library) in languages such as FORTRAN, C, C++, Java, etc., or by writing one's own RNGs using known and validated algorithms. Note also, that a RNG is more appropriately called a pseudorandom number generator (PRNG), as it is an algorithm for generating a sequence of numbers that approximates the properties of random numbers that are "sufficiently random" to suit the intended use. Regardless of the RNG used, it should meet at least some of the statistical tests of randomness, and have a periodlength long enough to not repeat a value in a very long string of pseudo-randomly generated values. Additionally, a good RNG should allow one to set a starting seed, allowing replication of results, and the RNG should have a known period-length.

## JOHN N. DYER

In any event, the RNG must be implemented programmatically and requires some knowledge of programming.

Methods commonly used for the Normal distribution include the Ziggurat method (Marsaglia \& Tsang, 2000), the Box-Muller transform (Box \& Muller, 1958), the Marsaglia polar method (Marsaglia \& Bray, 1964), the Probit function method, the Abramowitz \& Stegun algorithm (Abramowitz \& Stegun, 1972), a recent algorithm by Acklam (2014), and methods known as Kinderman-Ramage (Kinderman \& Ramage, 1976), and Ahrens-Dieter (Ahrens \& Dieter, 1988). Of these, the Ziggurat algorithm is considered superior, but the Box-Muller transform is very good and is a very common implementation in many software and programming languages. Many other algorithms are available for both the normal and other distributions (Korn, Korn, \& Kroisandt, 2010; C. Roberts \& Casella, 1999).

The task is then to use a Uniform [0, 1] RNG to randomly generate values of p , then solve the desired distribution's cdf for values of $x$. Common uniform RNGs include implementations using the Wichmann-Hill (WH) method (Wichmann \& Hill, 1982), the Mersenne Twister algorithm (MT19937) (Matsumoto \& Nishimura, 1998), the Marsaglia Multiply-with-carry (MWC) method (Marsaglia, Zaman, \& Marsaglia, 1994), and other various methods in the classes of linear feedback shift register generators and linear congruential generators.

Using Excel 2010 VBA to replicate a series of iidN pseudo-random variates requires use of two built in functions; RND and NORM_INV. The RND function returns a floating-point Uniform [0,1] random real number, representing a probability $(p)$, where $0 \leq p<1$. The function has no arguments and the output depends on the initial seed, which is set as a function of the system clock, hence not replicable. Although Microsoft claims to have implemented the Wichmann-Hill generator for the RND function, there are many finding that it was implemented incorrectly, and that it does not pass the DIEHARD test (McCullough \& Heiser, 2008). Others suggest that for long periods, RND will create negative numbers, and the period-length is not exactly known. The RND function also returns the value 0 . The NORM_INV function returns the inverse of the normal cumulative distribution with specified mean and standard deviation. That is, given a value for $p$, NORM_INV $(p, \mu, \sigma)$ seeks that value $x$ such that the function $\operatorname{NORM\_ DIST}(x)=p$. The value of $p$ is the output from the RND function. The NORM_INV function is implemented using the Probit function method. It should be noted that Microsoft claims that that accuracy of the NORM_INV function depends on the accuracy of their NORM_DIST function (which uses the Abramowitz \& Stegun algorithm), and the quality of the search procedure in its ability to "home in on" the value of $x$ that
corresponds to the supplied value of $p$ (Microsoft, 2011). They further claim that the accuracy is up to 15 or 16 decimal places. Excel with VBA is a good software and programming platform, and this author has been satisfied with the results of implementing the above Excel functions as well as the MT19937 for the Uniform $[0,1]$ RNG, which passes the DIEHARD test, and using the Box-Muller transform for the Normal distribution.

## Calculating the Series Length

The series length m is the total number of $\mathrm{N}(\mu, \sigma)$ random numbers $(x)$ or subgroup averages to be generated and examined sequentially against the CLs in each simulation run. Again, the series of random numbers or subgroup averages should be standardized to represent a $z \sim \mathrm{~N}(\mu=0, \sigma=1)$ distribution. For any given desired ARL estimate, $m$ will be constant, and is chosen as a function of the upper percentiles of the related Geometric distribution.

In D1, the total series length $m$ should be long enough for at least one of the randomly generated $z$-values to exceed the CLs based on the desired in-control ARL. For example, a Shewhart chart with $\pm 3 \sigma$ CLs ( $p \approx 0.0027$ ), the ARL $\approx 370$. Based on a known Geometric RL distribution with $\mu=\mathrm{ARL}=370$, the $99^{\text {th }}$ percentile value of the RL distribution is 1,702 . So, setting $m=1,702$, it would be expected that with great probability, at least one of the $1,702 z$-values will exceed the CLs. Choosing the $99.9^{\text {th }}$ percentile results in $m=2,554$, and the $99.99^{\text {th }}$ percentile results in $m=3,406$. Of course, larger ARLs (corresponding to smaller values of $p$ ), like $\operatorname{ARL}=1,000(p=0.001)$, the series lengths for the $99^{\text {th }}, 99.9^{\text {th }}$, $99.99^{\text {th }}$ and $99.9999^{\text {th }}$ percentiles are $m=4,602, m=6,903, m=9,205$, and $m=13,808$, respectively. For iidN cases it is recommended that the $99.9999^{\text {th }}$ percentile value be chosen for $m$ to avoid the case of any simulation run failing to result in a recordable run-length. This almost certainly ensures that each series $m$ will produce a recordable RL. When designing a MCS to evaluate a range of desired CLs and corresponding ARLs, it is recommended that (7) below be used to select the series length to accommodate the largest expected ARL estimation (eARL) in the study. Note that the multiple of 14 corresponds to the $99.9999^{\text {th }}$ percentile. Alternatively, a multiple of 11 would result in the $99.999^{\text {th }}$ percentile, a multiple of 9 would result in the $99.99^{\text {th }}$ percentile, and so on.

$$
\begin{equation*}
m=\operatorname{Max}(e \mathrm{ARL}) * 14 \tag{7}
\end{equation*}
$$

For example, if a study were performed using multiple CLs corresponding to ARLs of $1000,900,800,700, \ldots, 100$, then $\operatorname{Max}(e \operatorname{ARL})=1000$ and $m=14,000$. Keep in mind that the series of length $m$ will be generated and evaluated $n$ sim times. For example, if $m=14,000$ and $n \operatorname{sim}=10,000$, there would be 10,000 RLs recorded for each set of CLs in the completed simulation.

The greater issue of determining m arises when the sequential values of the transformed plotted statistics, CCS, are not independent, like for the CUSUM, EWMA, and RMA. Previous research has shown the CLs need to be adjusted to accommodate the correlation between sequential CCS. This case requires more of a trial and error approach in determining $m$ since the expected RL can be much different than under the iidN assumption, and the distribution of the RL is no longer exactly Geometric. For example, if the underlying distribution is iidN, and $\pm 3 \sigma \mathrm{CLs}$ are chosen, the resulting ARL $\approx 370$ and recommended $m=5,180$, use $m=6,000$, run several thousand simulations and count the number of simulations for which no CCS exceeded the CLs in each run. If the count is zero then the choice of $m$ should be sufficient. If not, then increase $m$. Additionally, the control-charts based on cumulative sums or averages are known to have a much larger standard deviation than the $z \sim \mathrm{~N}(0,1)$ process, so the value $m$ will likely be longer than the $\mathrm{N}(0,1)$ process.

Alternatively, instead of presetting $m$, one could increment a variable by 1 until the first CCS exceeds the CLs, record the value of $m$ as the RL, then stop the run and start the next simulation. Although this is suitable for studying only one RL distribution at a time for a specified set of CLs and ARL, this researcher often performs the study over a very wide range of ARLs (up to 50 simultaneously), which is a feature that often makes MCS more desirable and faster than other methods. Recall, the capability of the design and the ability to use a single simulation run across many desired CLs is an advantage of the proposed design, and is not possible when incrementing instead of presetting m .

In D2, only one series $m$ is calculated, but is a much longer length than used in D1. The choice of $m$ and the expected number of resulting RLs depend directly again on the largest ARL to be estimated. Let $d R L=$ the desired number of RLs to evaluate properties related to the largest expected ARL (eARL) in the design. So, for D 2 , the series-length $m$ is given by (8):

$$
\begin{equation*}
m=\operatorname{Max}(e \mathrm{ARL}) * d R L \tag{8}
\end{equation*}
$$

For example, for $\mathrm{ARL}=1,000$ and $d R L=10,000, m=1,000 * 10,000=10,000,000$. So a series length of $m=10,000,000$ will result in about 10,000 recordable RLs.

Inversely, choosing m to accommodate the largest ARL estimate, the resulting number of RLs for lessor values of ARLs is given by (9):

$$
\begin{equation*}
d R L=\frac{m}{e \mathrm{ARL}} \tag{9}
\end{equation*}
$$

Using the example above, simultaneously evaluate RLs for known ARLs of $1,000,500,370$, and 50 . The expected number of useable RLs will be about 10,000 , $20,000,27,000$, and 200,000 , respectively. As a result, the error in estimating lower valued ARLs will be dramatically reduced. Keep in mind that the IC ARL is always larger than the OC ARL, and the larger the mean shift (б) the smaller the OC ARL. For example, for IC ARL $=1,000$, the OC ARLs for mean shifts of $\sigma=1, \sigma=2$, and $\sigma=3$ are $90.87,10.16$, and 2.59 , respectively. Hence, while $n s i m$ will remain the same, $m$ can be decreased significantly and the program will run much faster.

## Calculating the Burn in Period

When extending the validation model to evaluating control charts like the CUSUM, EWMA and RMA, or other control-charts with memory, the series of normal random numbers ( $z$ ) and the transformed control-chart statistics (CCS) should have a burn-in period of runs prior to the actual RLs to be evaluated in the OC process. This accommodates the cumulative nature of the transformed statistics required to mimic a steady-state of the IC series prior to evaluating the subsequent desired series-length. Zero-state simulations refer to RLs those that have been initialized at the target starting value of the control statistic. Steady-state simulations refer to RLs that are evaluated after the control chart statistic has reached a steady-state, meaning the process has been "in-control" long enough for the effect of the starting value to become negligible (Lucas \& Saccucci, 1990). So zero-state simulations require no burn-in period, but steady-state simulations do require a short burn-in period.

When evaluating the IC process, it is assumed that one starts the process from an IC zero-state, meaning there is no burn-in period. When evaluating the OC process it is assumed the series has reached a steady-state, implying a burn-in period of a stable IC process. There is no body of literature regarding burn-in for MCS, while there are quite a few articles regarding burn-in for MCMC methods. Although the burn-in period is equivocally stated, it is suggested that a burn-in period that is close to and less than the smallest expected MRL being evaluated should be adequate. Beyond those periods one might expect the burn-in process to

## JOHN N. DYER

drift toward a false-alarm state. Additionally, CLs for many control-charts like the EWMA are a function of asymptotic variance, so the burn-in period should be long enough for the assumption of the asymptotic variance to hold. For example, if evaluating the OC RL properties of an EWMA with parameter $\lambda$ with corresponding control-limits set to produce an IC ARL $=300$ and MRL $=208$, then the burn-in period less than 208 should be adequate.

The EWMA parameter $\lambda$ can be used to determine a burn-in period necessary for asymptotic CLs to be appropriate. The asymptotic time-varying component of the variance of the EWMA is given as $1-(1-\lambda)^{2 t}$, hence the asymptotic CLs are constant starting at burn-in period determined by solving for the period $t$ such that $1-(1-\lambda)^{2 t} \approx 1$. This is especially true when evaluating the OC case, assuming the period was stable and in-control during the previous IC periods. As such, the total series-length will be Burn-in $+m$, where evaluation starts at the first period of $m$ following the last burn-in period. One can easily run several independent simulations using various burn-in periods to determine the most appropriate period.

Additionally, the rational starting value of the burn-in period is a function of the assumed/estimated mean of the control chart IC process. For example, if evaluating the EWMA control-chart with parameter $\lambda=0.20$ and process $\mu=5$, the rational starting or target value would be $C C S_{1}=5$. Hence, $C C S_{2}=\lambda^{*} z+(1-\lambda)^{*} C C S_{1}=0.20^{*} z+0.80 * 5$, where $z$ is the randomly generated value such that $z \sim \mathrm{~N}(\mu=5, \sigma=1)$.

## Establishing Control Limits

Control charts for the mean of a $z \sim \mathrm{~N}(0,1)$ process are typically designed based on the desired IC ARL for a process, which in turn determines the control-limits, which is turn determines the OC ARL based on a specified shift in the process mean. In the $z \sim \mathrm{~N}(\mu, \sigma)$ process, using a Shewhart or Individuals chart, the upper/lower CLs are easily found for any desired IC ARL, since $\operatorname{ARL}=1 / p$, where $p$ is the probability of an IC signal (false-alarm). For example, for a desired IC ARL $=370$, $p=1 / 370=0.0027$.

Because half of 0.0027 is below the lower CL and half is above the upper CL, one can use a Normal Inverse function (e.g., Excel's function NORM_INV $(p / 2, \mu, \sigma))$ to determine the lower CLL $\approx-2.9967$; hence the upper $\mathrm{CLU} \approx+2.9967$, which closely correspond to the $\pm 3 \sigma$ CLs. Likewise, when a mean shift ( $\sigma$ ) occurs, one can determine the expected OC ARL for the shifted process $z \sim \mathrm{~N}(\mu+\sigma \sigma, \sigma) \quad$ as $\quad \mathrm{OC} \quad \mathrm{ARL}=1 /\left(p_{1}+p_{2}\right)$, where $p_{1}=\mathrm{P}\left(z<\mathrm{CL}_{\mathrm{L}}\right)$ and $p_{2}=\mathrm{P}\left(z>\mathrm{CLU}_{\mathrm{U}}\right)$, which are not equal probabilities in the OC case. One can use a

Normal Distribution function (e.g., Excel's function $\operatorname{NORM} \operatorname{DIST}(z, \mu, \sigma)$ ) to calculate $p_{1}$ and $p_{2}$ thus allowing calculation of the OC ARL.

When using MCS to evaluate RL properties for other control charts, such as the EWMA, one might want to evaluate RL properties over a range of CLs. As previously noted, one advantage of MCS is the ability to evaluate RL properties over a range of CLs simultaneously, like evaluating properties for ARLs ranging from 20 to 1,000 in steps of 5 units, like $20,25,30, \ldots, 1,000$. In this case a consideration that must be made regarding CLs is how many decimal places are required, which also helps in deciding the unit step-distance from one ARL to the next. For example, if the IC process is $z \sim \mathrm{~N}(0,1)$, then CLs $= \pm 2.99$ would include ARL values from 359 to 370 , while CLs $= \pm 3.00$ correspond to ARL values from 371 to 382 . The point being that the more CLs that are evaluated simultaneously, the longer the time it takes to run the simulations, some of which may be redundant and overlapping. Conversely, the additional time will always render some additional information. In general, smaller unit step-distances are more appropriate for ranges of smaller expected ARLs, while larger unit step-distances are more appropriate for ranges of larger expected ARLs.

## Determining Simulation Size and Error

For D1, the choice of simulation runs, nsim, depends on the accuracy of the RNG being used, the acceptable maximum error in estimation $(E)$ of ARLs, and the $(100-\alpha) \%$ degree of confidence in the estimation of the ARL. A common simulation size used in many published articles for MCS is $n \operatorname{sim}=10,000$, hence producing 10,000 RLs (Dyer et al., 2002; Dyer, Conerly, \& Adams, 2003; Dyer, Adams, \& Conerly, 2003; Lin \& Adams, 1996). The value is not arbitrary, but is based on the assumption of the large-sample Normal distribution of a proportion ( $p$ ), (often an unknown, hence assumed value of $p=0.50$ ), and a desired $95 \%$ confidence level (corresponding to $\alpha=0.05$ and $z \approx 2$ ) in estimation of $p$ with a margin of error $E=0.01$. The value $z$ is the value of the Standard Normal distribution such that $\alpha / 2$ is above $+z$ and below $-z$. In this case the formula is given as

$$
n \operatorname{sim}=p(1-p) *\left(\frac{z}{E}\right)^{2}=0.25\left(\frac{2}{0.01}\right)^{2}=10,000
$$

Because a primary goal of MCS is evaluating IC ARLs, and ARLs are a function of $p$ (probability of a false alarm), and $p$ is almost always much less than

## JOHN N. DYER

0.05 , this goal then is to have a very small error in estimation of $p$ (much less than 0.01 ), hence a small error in estimation of the ARL. But, we are estimating ARLs, which in turn provide estimates of $p$. So, if we assume a large-sample Normal distribution of the ARL, and we know for the Geometric distribution that the mean and standard deviation are approximately equal, that is, $\mu=\mathrm{ARL} \approx \sigma=\mathrm{SRL}$, then the following estimated equation ((10) below) is given based on D 1 for the number of simulations necessary to accommodate the largest ARL estimation (when estimating a wider array of ARLs simultaneously), based on a $(100-\alpha) \%$ degree of confidence and the maximum allowable error in estimation $(E)$ of the largest expected ARL ( $e$ ARL). Recall, this estimate does not account for the error in the RNG being used, which can increase the number of required simulations. The value $z$ is the same as described in the previous paragraph.

$$
\begin{equation*}
n \operatorname{sim} \approx\left(\frac{z * \operatorname{Max}(e \mathrm{ARL})}{E}\right)^{2} \tag{10}
\end{equation*}
$$

Conversely, the estimated error ( $E$ ) for any specified expected ARL is given by (11):

$$
\begin{equation*}
E \approx \frac{z * e \mathrm{ARL}}{\sqrt{n \operatorname{sim}}} \tag{11}
\end{equation*}
$$

Since a Shewhart chart with $p=0.001$ corresponds to ARL $=1,000$, using nsim $=10,000$ would suggest a maximum error in estimation of $E=20$ with $95 \%$ confidence, not accounting for any additional error in the RNG. A review of the research also reveals that many published MCS based studies evaluate only a very few selected ARLs, like 370 , 500, and 1,000 . Using the above scenario ( $n \operatorname{sim}=10,000, z=2.00$ ), for ARL $=500$ we expect $E=10$, and for ARL $=370$ we expect $E=7.4$. Unfortunately, to reduce the error further (or to increase the degree of confidence), requires a substantial increase in $n s i m$. For example, in the case above with $\mathrm{ARL}=1,000$, reducing the desired maximum error to $E=5$ requires increasing the number of simulations to $n \operatorname{sim}=160,000$.

In D 2 , only one long series m is generated, and the resulting expected number of RLs is given by (9). Hence, (10) is modified to estimate the series-length given by (12) below:

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

$$
\begin{equation*}
m \approx\left(\frac{z}{E}\right)^{2} * \operatorname{Max}(e \mathrm{ARL}) * \sqrt{d R L^{3}} \tag{12}
\end{equation*}
$$

Conversely, the estimated error $(E)$ for any specified ARL is given by (13):

$$
\begin{equation*}
E \approx z * \sqrt{\frac{e \mathrm{ARL}}{m}} * \sqrt[4]{d R L^{3}} \tag{13}
\end{equation*}
$$

In general, all RL distributions appear Geometric to some degree, but are not exactly iidG. As a result, one must be very cautious about extending Shewhart based error estimates to ARLs obtained for other control charts for which the sequential CCS are not iidN (but are correlated), and making assumptions about the RL distribution that is not exactly iidG. Again, as discussed in the beginning of this section, at least a few of the results should be validated against existing literature before making broader conclusions about RL properties and summary measures.

An advantage of MCS is that the method allows additional summaries such as the MRL, SRL, CDF, and quantities such as percentiles and quartiles. As such, one might want to place confidence limit bounds on the resulting estimated MRL and SRL, in essence allowing one to estimate the error on these values following the simulations. It is assumed that the D1 simulation size (nsim) or the D2 desired run-length $(d R L)$ is determined relative to a desired maximum error in estimation $(E)$ of the maximum ARL under study.

Regarding the MRL, and assuming a large-sample Normal approximation, the upper and lower confidence interval provides two ordered rank locations of the RLs, which in turn allow one to determine the upper and lower confidence intervals around the MRL, which is asymmetric. Let $\mathrm{MR}_{\mathrm{L}}$ and $\mathrm{MR}_{\mathrm{U}}$ be the lower and upper rank locations, respectively, in an ordered array of RLs of size nsim for any given ARL. (14) and (15) are provided to find the rank locations, where $z$ corresponds to the desired $(1-\alpha) \%$ degree of confidence.

$$
\begin{align*}
& \mathrm{MR}_{\mathrm{L}} \approx\left(\frac{n \operatorname{sim}}{2}\right)-z * \sqrt{\frac{n \operatorname{sim}}{4}}  \tag{14}\\
& \mathrm{MR}_{\mathrm{U}} \approx\left(\frac{n \operatorname{sim}}{2}\right)+z * \sqrt{\frac{n \operatorname{sim}}{4}} \tag{15}
\end{align*}
$$

For example, using $95 \%$ confidence and $n \operatorname{sim}=10,000, \mathrm{MR}_{\mathrm{L}}=4,900$, and $M R_{\mathrm{U}}=5100$, and the ordered RL values at rank locations 4,900 and 5,100 are the asymmetric confidence interval limits on the MRL. For any given nsim, the rank locations will always be the same, but larger nsim will result in corresponding RLs closer to the true MRL, that is, less error. Confidence intervals for rank locations for other percentiles and/or quartiles can be derived by adjusting each divisor in each of the two formulas above.

Additionally, one can determine the expected error in the MRL by examining the cumulative probabilities of the Geometric distribution for a specified expected ARL at the 50th percentile RL (median), and upper and lower RLs corresponding to $0.50 \pm$ desired maximum error. Additionally, for any given nsim, the expected maximum error $(E)$ for the MRL estimate with $(1-\alpha) \%$ degree of confidence is given by (16), where $e$ MRL is the expected MRL, and (11) is inflated by a factor of $\sqrt{0.50 * \pi} \approx 1.25$ (Stigler, 1973).

$$
\begin{equation*}
E \approx \frac{z * e \mathrm{ARL}}{\sqrt{n \operatorname{sim}}} * 1.25 \tag{16}
\end{equation*}
$$

Regarding the SRL, the large-sample distribution for any given ARL is Normally distributed, hence the confidence limits on the true SRL are a function of the desired $(100-\alpha) \%$ degree of confidence, $n$ sim, the estimated SRL, and the Chisquare distribution $\left(\chi^{2}\right)$ evaluated at functions of $\alpha / 2$ and $d f=n \operatorname{sim}-1$ degrees of freedom, so that

$$
\chi_{\text {Lower }}^{2}=\chi_{\left(1-\frac{\alpha}{2}, d f=n s i m-1\right)}^{2} \quad \text { and } \quad \chi_{\mathrm{Upper}}^{2}=\chi_{\left(\frac{\alpha}{2}, d f=n s i m-1\right)}^{2}
$$

Let $\operatorname{SRL}_{\mathrm{L}}$ and $\mathrm{SRL}_{\mathrm{U}}$ be the lower and upper confidence limits (respectively). (17) and (18) are provided to find the desired $(1-\alpha) \%$ degree of confidence interval limits, where SRL is the estimated SRL.

$$
\begin{align*}
& \mathrm{SRL}_{\mathrm{L}} \approx \frac{d f * \mathrm{SRL}}{\chi_{\mathrm{Lower}}^{2}}  \tag{17}\\
& \mathrm{SRL}_{\mathrm{U}} \approx \frac{d f * \mathrm{SRL}}{\chi_{\mathrm{Upper}}^{2}} \tag{18}
\end{align*}
$$

Additionally, the expected error can be determined based on $(1-\alpha) \%$ degree of confidence in the SRL estimation using (13), where $e$ SRL is the expected SRL.

$$
\begin{equation*}
E \approx \frac{e \operatorname{SRL} *\left(\chi_{\mathrm{Upper}}^{2}-d f\right)}{d f} \tag{19}
\end{equation*}
$$

## MCS Programs and Validation Design Example

Although MCS can be implemented in a variety of programming languages, several advantages of using Excel 2010 for MCS studies include the available built-in functions, the ability to write special functions, the VBA programming interface for macros, subroutines and functions, built-in analysis and modeling tools, and the ability to use the spread-sheet as a data repository. Even when MCS is performed in programs such as $\mathrm{R}, \mathrm{C}+$, and FORTRAN, the data arrays are often imported into Excel for analysis and modeling. It should be noted that the validation design examples have been exemplified in Excel 2010 using VBA, but also compared with results in the VBA implementation using MT19937 and the Box-Muller transform method adapted to VBA (Annen, 2013). As such, the VBA code shown in Appendix A reflect calling two Excel VBA functions; RND which generates the Uniform [0, 1] variables, and NORM_DIST to generate the series of $z \sim \mathrm{~N}(\mu, \sigma)$ random variables using input from the RND functions. The code sections reflecting the generation of random numbers can be modified to implement any other choice of RNGs. Additionally, there is no error handling code, so one must be careful about such issues as inputting $\sigma \leq 0$ or placing data of any kind into worksheet cells that are not directly related to specific input or output. The screen-shots provided also reflect formatting options set at the worksheet level and not in the VBA code, such as rounding and run-time output formats. All code and a working Excel 2010 workbook is available from the author by request.

The sections below discuss the setup and description of the worksheet and the underlying VBA code for running the validation MCS based on the Individuals control chart, with an assumed IC $\mathrm{N}(\mu, \sigma)$ process for both D 1 and D 2 . Recall, the Individuals chart is a Shewhart chart with subgroup size $n=1$. The assumption for the design is that, regardless of the underlying IC iidN process, the underlying data ( $x$ ) would be standardized $(z)$ such that $z \sim \mathrm{~N}(\mu=0, \sigma=1$ ) process. As such, for the OC process the distribution changes to a $z \sim \mathrm{~N}(\sigma, 1)$ process.

## JOHN N. DYER

## Program Worksheet Inputs

For Design 1: For D1 and a single IC simulation run, we wish to generate quantity $m$ of $z \sim \mathrm{~N}(\mu=0, \sigma=1)$ random variables, sequentially compare each $z$-value to the CLs determined by the desired ARLs, then record the location of the RL in the sequence when the first $z$-value in the series exceeds the CLs (for each specified set of CLs). This process is then replicated $n$ sim times. The result will be $n$ sim RLs (for each set of specified IC ARLs) for which we can calculate the summaries for estimated ARLs, MRLs, and SRLs. Additionally, one may repeat the entire simulation as many times as desired using the Num Runs input. So, the worksheet will include input cells to specify the number of runs (Num Runs), the series length $(m)$, the number of simulations ( $n$ sim), the in-control process mean ( $\mu$ ) and standard deviation $(\sigma)$, the mean shift ( 0 for the IC case, $\sigma \sigma$ for the OC case), and the desired IC ARLs, which correspond to CLs that will be calculated using the Excel function NORM_INV implemented in the VBA code. Note that the simulation design allows any values of $\mu$ and $\sigma$, but there is no need to set these values to anything other than 0 and 1, respectively.

Figure 1 is a screen-shot of the formatted Excel 2010 workbook reflecting the initial setup using Num Runs $=1, m=14,000$, $n \operatorname{sim}=10,000$, and IC ARLs of $1,000,500,370$, and 50 . The input parameters reflect a desire for a maximum Error of $E=20$ for estimating ARL $=1,000$ with $95 \%$ confidence. The desired IC ARLs are entered from the largest to the smallest. With this limited example we wish to use MCS validation to estimate the four ARLs, MRLs and SRLS of an Individuals control chart based on the control-limits corresponding to specified IC probabilities of $p=0.001$ (ARL = 1000), $p=0.002(\mathrm{ARL}=500), p=0.0027(\mathrm{ARL}=370)$, and $p=0.02$ (ARL $=50$ ). The corresponding calculated control-limits will be $\pm 3.2905$, $\pm 3.0902$, $\pm 2.9997$, and $\pm 2.3263$, respectively. Recall, we already know the properties of the Individuals based $z \sim \mathrm{~N}(0,1)$ process, so the MCS in the example is used to exemplify setup and validation of the estimated MCS results with expected results.

The input cells are as follows:

- $\quad$ B2 (Num Runs) $=1$. This value allows one to produce one or more independent simulations runs, hence allowing one to investigate results of the same simulation design over multiple runs.
- $\quad \mathrm{B} 3(m)=14,000$. This value is based on the series length calculation
- $\quad \mathrm{m} \approx \operatorname{Max}(\mathrm{ARL}) * 14=1,000 * 14 \approx 14,000$.
- B4 (nsim) $=10,000$. This value is based on a desired maximum error in estimation of $E=20$ for Max ARL $=1,000$, with $95 \%$ confidence. Error will necessarily be lower for smaller estimates.
- B5 $(\mu)=0$ (mean for the underlying IC process assuming a $z \sim \mathrm{~N}(0,1)$ process).
- $\quad \mathrm{B} 6(\sigma)=1$ (standard deviation for underlying IC process assuming a $z \sim \mathrm{~N}(0,1)$ process).
- B7 $(\mu$-Shift $)=0$ ( 0 for the IC process, or $>0$ for the OC process $z \sim \mathrm{~N}(\sigma \sigma, 1)$ process).
- F1:I1 (IC ARLs) = 1000 (F1), $500(\mathrm{G} 1), 370(\mathrm{H} 1), 50(\mathrm{I} 1)$, assuming one wants to evaluate these four ARLs.
- Run Button - A button to execute the VBA subroutine

For Design 2: Figure 2 is a screen-shot of the formatted Excel 2010 worksheet reflecting the initial setup using $m=10,000,000$ (B3) while the remainder of input cells are the same as used in the D1 example. Recall that for D2 we wish to generate one very long series of length $m$ of $z \sim \mathrm{~N}(0,1)$ random variables, sequentially compare each $z$ to the CLs determined by the desired ARLs, then record the location of each RL in the sequence when each $z$ in the series exceeds the CLs (for each specified set of CLs). The long series is based on the $d R L$, and is almost equivalent to $m^{*} n \operatorname{sim}$ in D1. The other inputs are the same as D1, with the same error and confidence level. The output will be almost the same is D1, except that we don't know in advance how many RLs will be produced for each set of specified IC ARLs, but can be estimated using (9).


Figure 1. Design 1 program input cells

## JOHN N. DYER

| 4 | A | B | c | , | E | F | G | H | 1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Inputs |  |  | Inputs | IC ARLs | 1000 | 500 | 370 | 50 |
| 2 | Num Runs = | 1 |  | Outputs | Upper CLs |  |  |  |  |
| 3 | $\mathrm{m}=$ | 10000000 |  |  | Lower CLs |  |  |  |  |
| 4 | $\mu=$ | 0 |  |  | Expected ARLs |  |  |  |  |
| 5 | $\sigma=$ | 1 |  |  | Expected MRLs |  |  |  |  |
| 6 | $\mu$-Shift = | 0 |  |  | Expected SRLs |  |  |  |  |
| 7 |  |  |  |  | Estimated ARLs |  |  |  |  |
| 8 | Outputs |  |  |  | Estimated MRLs |  |  |  |  |
| 9 | Start-time $=$ |  |  |  | Estimated SRLs |  |  |  |  |
| 10 | End-time $=$ |  |  |  | Count |  |  |  |  |
| 11 | Run-time $=$ |  |  | Run | Run-Lengths |  |  |  |  |
| 12 | Sim Run $=$ |  |  |  |  |  |  |  |  |

Figure 2. Design 2 program input cells

## Program Worksheet Outputs

For Design 1: Figure 3 is a screen-shot reflecting the D1 outputs of one individual run of $m=14,000$ and $n \operatorname{sim}=10,000$ simulations. All output cells are set and/or calculated using VBA code as shown in Appendix A. The worksheet output cells include a timer with Start-time, End-time and Run-time, an indicator of the exact simulation number being run at any given time (if multiple simulations are to be done), cells containing the upper and lower control-limits (based on the input ARLs), expected ARLs, MRL, and SRLs, columns to record the RLs of all simulations for each set of control-limits, and cells for the estimated ARLs, MRLs and SRLs for each IC ARL (or for each OC ARL in the OC process).

The output cells for this example are described as follows. See the screen-shot for actual exemplary values.

- $\quad$ B10 $=$ Start-time, B11 $=$ End-time, B12 $=$ Run-time
- B13 (Sim Run) = changing variable depending on exact simulation being run, starting with 1 and ending with Num Runs in B2. Only used if multiple simulations are being run on the same set of inputs.
- F2:I3 = upper and lower CLs based on input ARLs.
- F4:I4 $=$ expected ARLs based on the input $\mu$-Shift. The values are the same as the input ARLs for the IC process but will change for the OC process, and depend on the mean shift (set in input cell B7).
- F5:I5 = expected MRLs based on the input $\mu$-Shift. The values are the same for the IC process but will change for the OC process, and depend on the mean shift (set in input cell B7).


## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

- F6:I6 $=$ expected SRLs based on the input $\mu$-Shift. The values are the same for the IC process but will change for the OC process, and depend on the mean shift (set in input cell B7).
- F7:I7 = estimated ARLs (average of RLs).
- F8:I8 = estimated MRLs (median of RLs).
- F9:I9 = estimated SRLs (standard error of RLs).
- $\mathrm{F} 10: \mathrm{I} 10010=10,000$ recorded RLs for each input ARL.
- $\quad \mathrm{D} 11=$ a count of the runs of size $m$ out of $n \operatorname{sim}$ that did not produce a RL for the largest input ARL. If $m$ is selected appropriately and large enough then the value will be 0 .

For Design 2: Figure 4 is a screen-shot reflecting the D2 outputs of one individual run. All output cells are set and/or calculated using VBA code as shown in Appendix B. The worksheet output cells are the same as D1, with the exceptions

| 4 | A | B | C | D | E | F | G | H | 1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Inputs |  |  | Inputs | IC ARLs | 1000 | 500 | 370 | 50 |
| 2 | NumRuns $=$ | 1 |  | Outputs | Upper CLs | 3.2905 | 3.0902 | 2.9997 | 2.3263 |
| 3 | $\mathrm{m}=$ | 14000 |  |  | Lower CLs | -3.2905 | -3.0902 | -2.9997 | -2.3263 |
| 4 | nsim $=$ | 10000 |  |  | Expected ARLs | 1000 | 500 | 370 | 50 |
| 5 | $\mu=$ | 0 |  |  | Expected MRLs | 693 | 346 | 256 | 34 |
| 6 | $\sigma=$ | 1 |  |  | Expected SRLs | 1000 | 500 | 370 | 50 |
| 7 | $\mu$-Shift $=$ | 0 |  |  | Estimated ARLs | 1012 | 505 | 373 | 50 |
| 8 |  |  |  |  | Estimated MRLs | 701 | 346 | 259 | 35 |
| 9 | Outputs |  |  |  | Estimated SRLs | 1020 | 503 | 376 | 49 |
| 10 | Start-Time | 2/11/2014 7:27 |  | Blank Runs | Run-Lenghts | 558 | 406 | 406 | 35 |
| 11 | End-Time | 2/11/2014 8:16 |  | 0 |  | 331 | 331 | 190 | 190 |
| 12 | Run-Time | 49:09.2 |  | Run |  | 1470 | 1470 | 773 | 26 |
| 13 | Sim Run $=$ | 1 |  |  |  | 2202 | 2202 | 1625 | 115 |

Figure 3. Design 1 program output cells

| 4 | A | B | C | D | E | F | G | H | 1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | Inputs |  |  | Inputs | IC ARLs | 1000 | 500 | 370 | 50 |
| 2 | Num Runs = | 1 |  | Outputs | Upper CLs | 3.2905 | 3.0902 | 2.9997 | 2.3263 |
| 3 | $\mathrm{m}=$ | 10000000 |  |  | Lower CLs | -3.2905 | -3.0902 | -2.9997 | -2.3263 |
| 4 | $\mu=$ | 0 |  |  | Expected ARLs | 1000 | 500 | 370 | 50 |
| 5 | $\sigma=$ | 1 |  |  | Expected MRLs | 693 | 346 | 256 | 34 |
| 6 | $\mu$-Shift $=$ | 0 |  |  | Expected SRLs | 1000 | 500 | 370 | 50 |
| 7 |  |  |  |  | Estimated ARLs | 1007 | 505 | 370 | 50 |
| 8 | Outputs |  |  |  | Estimated MRLs | 696 | 346 | 256 | 35 |
| 9 | Start-time $=$ | 2/11/2014 7:18 |  |  | Estimated SRLs | 1006 | 509 | 371 | 49 |
| 10 | End-time = | 2/11/2014 7:23 |  |  | Count | 9821 | 19783 | 27015 | 200552 |
| 11 | Run-time $=$ | 0:04:42 |  | Run | Run-Lengths |  |  |  |  |
| 12 | Sim Run $=$ | 1 |  |  |  | 2353 | 2154 | 1007 | 23 |
| 13 |  |  |  |  |  | 581 | 199 | 396 | 156 |

Figure 4. Design 2 program output cells
that in D2 there is no need for a count of runs that didn't produce a RL value, and the D2 output includes a count of the number of RLs for each set of control-limits. For example, the single simulation resulted in $9,821 \mathrm{RLs}$ for $\mathrm{ARL}=1,000$ as displayed in cell F10. Cells G10, H10, and I10 display the counts for each of the additional ARLs. Note the much larger RL counts for the smaller ARL calculations.

## Program Overview

For Design 1: For D1, when the VBA code is executed for the IC $z \sim \mathrm{~N}(\mu, \sigma)$ process, the following steps occur programmatically. For each simulation run (Num Runs); (1) The Start-time is stored in cell B10, (2) each of the upper and lower control-limits are calculated and stored in cells starting in cell F2. These values correspond to the desired IC ARLs, based on input choices of $\mu$ and $\sigma$, and (3) each of the expected ARLs, MRLs and SRLs are calculated and stored in in cells starting in cell F4.

For each simulation (nsim); (4) a series of $m z$-values are generated and stored in an array. For each set of the control-limits; (5) each $z$-value is sequentially compared to each upper CL and lower CL, one at a time. When the first $z$-value exceeds its CLs, the RL is recorded in an array. After a RL has been recorded for each specified CL, the process terminates and continues to the next simulation. Steps 5 through 7 are continued until all simulations are completed. After all nsim simulations are complete, the RL array is copied into the worksheet.

Following the last simulation in step 5; (6) the summary estimated measures (estimated ARL, MRL and SRL) are calculated based on the RLs and stored in cells
starting in cell F7, and (7) the End-time is stored in cell B11 and Run-time is stored in cell B12.

When an OC process is simulated, the user sets the value of the shifted mean in worksheet cell B7, in terms of a shift in $\sigma$. For example, regardless of the choice of $\sigma$, entering the value " 3 " in cell B7 implies a $3 \sigma$ shift. In step 3 the expected OC ARLs, MRLS and SRLs are calculated and stored in cells starting in cell F4. Then in step 5 the $z$-values are generated as $z \sim \mathrm{~N}(\mu$-Shift, 1$)$. Note that $m$ is now much smaller and chosen to accommodate the largest OC ARL. For example, for IC ARL $=1,000$ and corresponding control-limits $\pm 3.2905$, a mean-shift of $1 \sigma$ results in an $\mathrm{OC} \mathrm{ARL} \approx 91$. Hence, $m=\operatorname{Max}(e \mathrm{ARL}) * 14=91 * 14=1,275$. Keeping $m=10,000$ will result in an expected error of $E=1.82$ with $95 \%$ confidence.

For Design 2: For D2, when the VBA code is executed for the IC $z \sim \mathrm{~N}(\mu, \sigma)$ process, the following steps occur programmatically. For each simulation run (Num Runs), the first 4 steps and steps 6 and 7 are the same as D1. For each of the control-limits; (5) each $z$-value is sequentially compared to each upper CL and lower CL, one at a time. Anytime a $z$-value exceeds its CLs, the RL is recorded in an array. After all RLs have been recorded for each specified CL, the procedure terminates and the RL array is copied into the worksheet. When an OC process is simulated, D2 is setup the same as D1. Again m is now much smaller and chosen to accommodate the largest OC ARL. Using the previous example, for IC $\mathrm{ARL}=1,000$ and corresponding control-limits $\pm 3.2905$, a mean-shift of $1 \sigma$ results in an $\mathrm{OC} \mathrm{ARL} \approx 91$. For $d R L=10,000, m=\operatorname{Max}(\mathrm{ARL}) * d R L=91 * 10,000$ $=900,000$.

## Other Considerations

It might be questioned why two MCS validation designs are exemplified, when both produce equivalent results. When the IC process is simulated for any controlcharting scheme, both designs are adequate and have equal results. But when simulating the OC case for cumulative schemes, like the CES, CUSUM, EWMA and RMA, D1 is required since the cumulative effect of the control-chart statistic depends on previous states, which must be based on the last simulated value of the previous control-chart statistic. That is, each new series of the OC process must be started using the last value of the stable IC burn-in period, which reflects the IC process prior to a process shift. Additionally, the design choice depends on computing time versus the personal computer's (PC) configuration and performance. D1 takes significantly more time to run on a PC since the series-
length $m$ is replicated and evaluated nsim times, but for most simulations the full series length $m$ is not necessary. Unfortunately, one doesn't know in advance when the first $z$ to exceed the CLs will occur, so $m$ must be long enough to ensure a recordable RL will occur in each separate $n$ sim. The full series of $z$-values for each nsim are stored in an array, and since the array size is relatively small, the program is less dependent on the PCs processor and memory to handle large arrays. The trade-off then relates to being able to use a PC with less processing capacity and memory, but requiring greater processing time.

The D 2 design is best used to evaluate the OC case when implementing control-charts with run-rules or multiple-sampling plans, wherein the OC statistic at any given time is not dependent on a previous value. D2 creates one very long array of size $m$ and indexes through it to count the RLs, and is thus significantly faster. The long array size though may create a limitation on computers with less memory. As a result, D2 might not be a feasible option on some PCs, but for those that can accommodate the large array in memory the processing time in reduced significantly. Additionally, the D2 design with an overly long series-length may still press the limits of any modern PC's ability to dimension an overly large array in memory. In either event, if using Excel 2010 to generate or store RLs, one must be careful of the longest expected RL output since the maximum number of rows is $1,048,576$. If the RL is expected to exceed this value then the RLs can be truncated or can be written to a text file.

## Design Validation and Error Analysis

For the sake of completion, a limited study was conducted as follows to validate designs and estimate error using two tests (T1 and T2). The test compares and contrasts results using T1 and T2 (shown below) to validate the MCS design for the IC $z \sim \mathrm{~N}(0,1)$ process using $m=14,000$, $n \operatorname{sim}=10,000$, ARLs of $1000,500,370$, and 50 , respectively, repeated NumRuns $=20$ times, and maximum error in estimation corresponding to $95 \%$ confidence. The aggregated results are shown in Tables 1 (ARLs), 2 (MRLs), and 3 (SRLs), and are further discussed.

- T1: Use Excel 2010 and VBA implementing built-in RND and NORM_DIST functions.
- T2: Use Excel 2010 and VBA implementing the Mersenne Twister algorithm and Box-Muller transform methods (Annen, 2013).


## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

Tables 1, 2, and 3 reflect simulated estimated ARLs, MRLs, and SRLs as specified above for tests T1 and T2. Note that the results in each table are sorted by ARL, MRL, and SRL in ascending order, respectively. Table 1 displays the observed (estimated) ARLs (versus expected ARLs) for the 20 independent simulation runs (sorted from lowest to highest). The summary statistics provided reflect the aggregated ARLs over the 20 runs. For T1, while the aggregated average ARLs do not all match the expected results, most of individual ARL estimates are very close to the expected ARLs, that is, most of the observed error is within the maximum expected error. The last row of the table reflects the percent of the 20 individual runs for each estimated ARL that are within the expected error, and all but expected ARL $=370$ resulted in $95 \%$ to $100 \%$ of runs within the expected error. $A R L=370$ also corresponds to an unexpectedly high standard deviation of RLs (4.35). The average aggregate error for expected $\operatorname{ARL}=1,000(E=3.85)$ is well within expected error (4.5), but the error for expected $\mathrm{ARL}=500(E=4.60)$ is $205 \%$ larger than expected error (2.24), the error for expected $\mathrm{ARL}=370(E=3.35)$ is $202 \%$ larger than the expected error (1.65), and the error for expected ARL $=50$ ( $E=0.30$ ) is $134 \%$ larger than the expected error ( 0.22 ). Although the percent difference is large the actual error is very small.

Regarding T2, the aggregated estimated ARLs are generally quite close to expected ARLs, and the standard deviation between the individual runs are typically close to those of T1, with the exception of estimated ARL $=1,000$ with a very large standard deviation (7.13). While each of the 20 individual runs for each estimated ARL is within the expected errors, the average aggregated errors are similar to that of T 1 , all having larger than expected errors.

Table 2 displays the same summary data as Table 1 but for the MRLs instead. The estimated MRLs are very close to expected MRLS. The results are largely consistent with those found in Table 1, but the maximum errors for T 1 and T 2 are typically somewhat larger than those for the ARLs. The standard deviations between median RLs as well as the standard deviations of errors are much larger than expected. The last row of the table reflects the percent of 20 individual runs for each MRL that are within the expected error. For T1, 100\% of the 20 runs for expected $\mathrm{MRL}=693(\mathrm{ARL}=1,000)$ and $\mathrm{MRL}=256(\mathrm{ARL}=370)$ are within expected error, while $90 \%$ of runs for expected $\mathrm{MRL}=346(\mathrm{ARL}=500)$ are within expected error, and $95 \%$ of runs for expected $\mathrm{MRL}=34$ ( $\mathrm{ARL}=50$ ) are within expected error. For $\mathrm{T} 2,90 \%$ of the 20 runs for expected MRL $=693$ are within expected error, $95 \%$ for $\mathrm{MRL}=346,100 \%$ for $\mathrm{MRL}=258$, and only $85 \%$ for $\mathrm{MRL}=34$.

## JOHN N. DYER

Table 1. ARL summaries for $S 1$

| Expected ARLs | 1000 |  | 500 |  | 370 |  | 50 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Simulation | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 |
| 1 | 994 | 989 | 489 | 494 | 361 | 365 | 49 | 49 |
| 2 | 995 | 994 | 490 | 494 | 362 | 367 | 49 | 49 |
| 3 | 996 | 995 | 491 | 495 | 362 | 367 | 49 | 49 |
| 4 | 997 | 996 | 492 | 495 | 365 | 368 | 50 | 49 |
| 5 | 998 | 996 | 492 | 496 | 367 | 368 | 50 | 50 |
| 6 | 998 | 997 | 493 | 498 | 368 | 370 | 50 | 50 |
| 7 | 998 | 998 | 495 | 498 | 368 | 370 | 50 | 50 |
| 8 | 999 | 999 | 495 | 498 | 369 | 370 | 50 | 50 |
| 9 | 1000 | 1001 | 496 | 500 | 370 | 371 | 50 | 50 |
| 10 | 1002 | 1001 | 496 | 501 | 370 | 371 | 50 | 50 |
| 11 | 1003 | 1001 | 497 | 502 | 371 | 372 | 50 | 50 |
| 12 | 1003 | 1002 | 497 | 502 | 371 | 373 | 50 | 50 |
| 13 | 1003 | 1003 | 498 | 503 | 371 | 373 | 50 | 50 |
| 14 | 1004 | 1003 | 499 | 504 | 372 | 374 | 50 | 50 |
| 15 | 1004 | 1005 | 499 | 504 | 372 | 374 | 50 | 50 |
| 16 | 1005 | 1008 | 499 | 505 | 372 | 374 | 50 | 50 |
| 17 | 1006 | 1008 | 500 | 505 | 374 | 375 | 50 | 51 |
| 18 | 1007 | 1012 | 502 | 506 | 374 | 376 | 51 | 51 |
| 19 | 1007 | 1014 | 504 | 508 | 375 | 376 | 51 | 51 |
| 20 | 1008 | 1017 | 504 | 508 | 377 | 376 | 51 | 51 |


| ARL Summary Statistics for 20 Simulation Runs |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Average | 1001 | 1002 | 496 | 501 | 370 | 372 | 50 | 50 |
| Median | 1002 | 1001 | 496 | 501 | 370 | 372 | 50 | 50 |
| Std Dev | 4.32 | 7.13 | 4.40 | 4.58 | 4.35 | 3.30 | 1.00 | 0.64 |
| Minimum | 994 | 989 | 489 | 494 | 361 | 365 | 49 | 49 |
| Maximum | 1008 | 1017 | 504 | 508 | 377 | 376 | 51 | 51 |


| ARL Error Summary Statistics for 20 Simulation Runs |  |  |  |  |  |  |  |
| ---: | :---: | :---: | :---: | :---: | :---: | ---: | ---: |
| Expected Error1 | 20.00 | 10.00 | 7.40 | 1.65 | 0.00 |  |  |
| Expected Error2 | 4.47 | 2.24 | 4.03 | 2.99 | 0.30 | 0.49 |  |
| Average | 3.85 | 5.56 | 4.60 | 4.03 | 2.00 | 2.91 | 0.00 |
| Median | 3.50 | 4.07 | 4.00 | 4.26 | 0.41 |  |  |
| Std Dev | 2.10 | 4.60 | 3.15 | 2.12 | 2.73 | 1.95 | 0.46 |
| Minimum | 0.00 | 0.57 | 0.00 | 0.08 | 0.00 | 0.07 | 0.39 |
| Maximum | 8.00 | 17.37 | 11.00 | 7.70 | 9.00 | 6.46 | 1.00 |
| \% Within Error | $100 \%$ | $100 \%$ | $95 \%$ | $100 \%$ | $85 \%$ | $100 \%$ | $100 \%$ |

Note: 1 individual run of $n \operatorname{sim}=10,000,2$ aggregated runs of $n \operatorname{sim}=200,000$

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

Table 2. MRL summaries for $S 1$

| Expected MRLs | 693 |  | 346 |  | 256 |  | 34 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Simulation | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 |
| 1 | 676 | 675 | 337 | 336 | 248 | 250 | 33 | 34 |
| 2 | 679 | 675 | 337 | 338 | 251 | 251 | 34 | 34 |
| 3 | 683 | 677 | 337 | 338 | 251 | 252 | 34 | 34 |
| 4 | 684 | 681 | 341 | 339 | 251 | 253 | 34 | 34 |
| 5 | 685 | 683 | 341 | 340 | 252 | 254 | 34 | 34 |
| 6 | 685 | 684 | 344 | 341 | 252 | 254 | 35 | 35 |
| 7 | 687 | 687 | 345 | 342 | 253 | 254 | 35 | 35 |
| 8 | 689 | 687 | 345 | 344 | 255 | 254 | 35 | 35 |
| 9 | 690 | 688 | 346 | 344 | 255 | 255 | 35 | 35 |
| 10 | 691 | 690 | 347 | 345 | 256 | 255 | 35 | 35 |
| 11 | 692 | 694 | 347 | 346 | 256 | 256 | 35 | 35 |
| 12 | 694 | 695 | 348 | 347 | 257 | 257 | 35 | 35 |
| 13 | 694 | 696 | 348 | 347 | 259 | 258 | 35 | 35 |
| 14 | 695 | 696 | 349 | 348 | 259 | 258 | 35 | 35 |
| 15 | 697 | 699 | 350 | 348 | 259 | 259 | 35 | 35 |
| 16 | 701 | 703 | 351 | 350 | 260 | 260 | 35 | 35 |
| 17 | 701 | 705 | 352 | 350 | 260 | 260 | 35 | 35 |
| 18 | 702 | 705 | 354 | 352 | 260 | 260 | 35 | 36 |
| 19 | 706 | 712 | 356 | 352 | 262 | 261 | 35 | 36 |
| 20 | 706 | 718 | 358 | 353 | 262 | 262 | 36 | 36 |


| MRL Summary Statistics for 20 Simulation Runs |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Average | 692 | 692 | 347 | 345 | 256 | 256 | 35 | 35 |
| Median | 691 | 692 | 347 | 346 | 256 | 256 | 35 | 35 |
| Std Dev | 8.52 | 12.15 | 6.04 | 5.16 | 4.17 | 3.48 | 0.64 | 0.56 |
| Minimum | 676 | 674.5 | 337 | 336 | 248 | 250 | 33 | 34 |
| Maximum | 706 | 718 | 358 | 353 | 262 | 262 | 36 | 36 |


| ARL Error Summary Statistics for 20 Simulation Runs |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :--- |
| Expected Error1 | 17.33 | 8.65 | 6.40 | 0.85 |  |  |  |  |
| Expected Error2 | 3.87 | 1.93 |  | 1.43 | 0.19 |  |  |  |
| Average | 7.05 | 9.88 | 4.75 | 4.28 | 3.50 | 2.95 | 0.85 | 0.85 |
| Median | 8.00 | 9.75 | 4.50 | 4.00 | 4.00 | 2.50 | 1.00 | 1.00 |
| Std Dev | 4.66 | 6.56 | 3.55 | 2.86 | 2.06 | 1.69 | 0.48 | 0.55 |
| Minimum | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maximum | 17.00 | 25.00 | 12.00 | 10.00 | 8.00 | 6.00 | 2.00 | 2.00 |
| \% Within Error | $100 \%$ | $90 \%$ | $90 \%$ | $95 \%$ | $95 \%$ | $100 \%$ | $95 \%$ | $85 \%$ |

Note: 1 individual run of $n \operatorname{sim}=10,000,2$ aggregated runs of $n \operatorname{sim}=200,000$

## JOHN N. DYER

Table 3. SRL summaries for S1

| Expected SRLs | 1000 |  | 500 |  | 370 |  | 50 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Simulation | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 | Test 1 | Test 2 |
| 1 | 1000 | 984 | 484 | 489 | 361 | 366 | 48 | 48 |
| 2 | 1001 | 984 | 485 | 493 | 363 | 367 | 49 | 49 |
| 3 | 1001 | 985 | 485 | 496 | 364 | 368 | 49 | 49 |
| 4 | 1001 | 985 | 486 | 497 | 365 | 372 | 49 | 49 |
| 5 | 1003 | 988 | 486 | 499 | 366 | 372 | 49 | 49 |
| 6 | 1005 | 989 | 487 | 499 | 367 | 372 | 49 | 49 |
| 7 | 1005 | 992 | 487 | 499 | 368 | 372 | 49 | 49 |
| 8 | 1006 | 992 | 487 | 500 | 368 | 372 | 49 | 49 |
| 9 | 1006 | 995 | 488 | 501 | 369 | 372 | 49 | 49 |
| 10 | 1006 | 1004 | 488 | 503 | 370 | 373 | 50 | 49 |
| 11 | 1007 | 1007 | 489 | 505 | 370 | 375 | 50 | 49 |
| 12 | 1007 | 1008 | 490 | 506 | 371 | 375 | 50 | 49 |
| 13 | 1010 | 1008 | 490 | 506 | 372 | 376 | 50 | 49 |
| 14 | 1010 | 1008 | 491 | 506 | 373 | 376 | 50 | 50 |
| 15 | 1010 | 1009 | 495 | 507 | 374 | 376 | 50 | 50 |
| 16 | 1013 | 1012 | 495 | 509 | 374 | 378 | 50 | 50 |
| 17 | 1013 | 1012 | 495 | 510 | 376 | 378 | 50 | 50 |
| 18 | 1013 | 1012 | 496 | 510 | 376 | 379 | 50 | 50 |
| 19 | 1014 | 1022 | 498 | 511 | 377 | 381 | 50 | 50 |
| 20 | 1014 | 1029 | 498 | 514 | 378 | 389 | 51 | 51 |

SRL Summary Statistics for 20 Simulation Runs

| SRL Summary Statistics for 20 Simulation Runs |  |  |  |  |  |  |  |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Average | 1007 | 1001 | 490 | 503 | 370 | 374 | 50 | 49 |
| Median | 1007 | 1005 | 489 | 504 | 370 | 374 | 50 | 49 |
| Std Dev | 4.57 | 13.48 | 4.47 | 6.46 | 4.88 | 5.18 | 0.66 | 0.70 |
| Minimum | 1000 | 984 | 484 | 489 | 361 | 366 | 48 | 48 |
| Maximum | 1014 | 1029 | 498 | 514 | 378 | 389 | 51 | 51 |


| SRL Error Summary Statistics for $\mathbf{2 0}$ Simulation Runs |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Expected Error1 | 27.73 |  | 13.86 | 10.26 | 0.11 | 0.39 |  |
| Expected Error2 | 6.20 | 0.15 | 5.83 | 4.00 | 5.30 | 0.55 | 0.77 |
| Average | 7.25 | 11.84 | 10.00 | 5.83 | 4.00 | 4.33 | 0.50 |
| Median | 6.50 | 11.49 | 11.50 | 5.84 | 0.58 |  |  |
| Std Dev | 4.56 | 5.82 | 4.44 | 3.80 | 2.61 | 4.15 | 0.59 |
| Minimum | 0.00 | 3.52 | 2.00 | 0.37 | 0.00 | 1.60 | 0.00 |
| Maximum | 14.00 | 29.01 | 16.00 | 13.69 | 9.00 | 18.74 | 2.00 |
| \% Within Error | $100 \%$ | $95 \%$ | $70 \%$ | $95 \%$ | $100 \%$ | $90 \%$ | $95 \%$ |

Note: 1 individual run of $n \operatorname{sim}=10,000,2$ aggregated runs of $n \operatorname{sim}=200,000$

Table 3 displays the same summary data as Table 1 but for SRLs instead. For T1, while the aggregated average SRLs do not all match the expected results, most of individual SRL estimates are relatively close to the expected SRLs. The average errors are significantly larger than those of the ARLs and MRLs. The last row of the table reflects the percent of 20 individual runs for each SRL that are within the expected error. For T1, while expected SRLs 1000, 370 and 50 have $95 \%$ to $100 \%$ of runs within the expected error, $A R L=500$ has only $70 \%$ of runs within the expected error. The average aggregate errors for all T1 SRLs are larger than expected. Results for T2 are more consistent with what is expected. The estimated SRLs are all very close to expected SRLs, and expected SRL 1,000, 500 and 50 have $95 \%$ of runs within the expected error, while $\operatorname{SRL}=370$ has $90 \%$ of runs within expected error.

As previously mentioned, for any given ARL/MRL/SRL estimate and simulation size, any difference between expected errors versus observed errors are due to the choice of RNG, or perhaps implementation and/or numerical precision. While most results are consistent with expected results, and the estimates are relatively adequate, it appears that Excel's implementation of the two RNGs is adequate, and in some cases is superior to the VBA implementation of the Mersenne Twister algorithm and Box-Muller transform methods. Additionally, the T1 design runs about 8-times faster than the T2 design, running on a PC configured with an AMD Phenom II 945 Processor ( 3.00 GHz ), 8 GB of Ram, using Windows 7 (64-bit) and Excel 2010 (32-bit).

## Design Modification, Validation, and Evaluation for the EWMA Control Chart

Assuming the D1 design proposed in this paper has validated the simulation results for the $z \sim \mathrm{~N}(0,1)$ process, we can now modify the program to transform the series of $z \sim \mathrm{~N}(0,1)$ values to a series of EWMA control chart statistics. It is known that the CLs for the EWMA are much narrower than those of the Individuals controlchart. Control-limit equations in the literature relate that the EWMA with parameter $\lambda=0.25$ and desired IC ARL $\approx 500$ has estimated CLs $= \pm 1.134$. The corresponding Shewhart IC ARL $=370$ with $\mathrm{CL}= \pm 3.00$.

The D1 program designed is then modified to evaluate a set of 7 different CL values ranging from $\pm 1.14271$ to $\pm 1.1240$ for which to compare our EWMA statistics. These CLs correspond to Shewhart $z \sim \mathrm{~N}(0,1)$ ARLs from 400 to 340 in steps of 10 , centered on $\mathrm{ARL}=370$. Hence we will modify the MCS validation design to provide estimated ARLs, MRLs, and SRLs for the complete set of EWMA

## JOHN N. DYER

CLs, for both the IC and OC cases. This modified program will thus allow one to validate and then estimate IC and OC RL properties of the EWMA over a wide array of parameter values and any choice of CLs.

The simulation is then designed to accommodate the following three studies.

- $\quad$ Study 1: One individual validation run (Num Runs $=1$ ) of $m=7,000$ and $n \operatorname{sim}=200,000, \lambda=1.00$, for the IC $z \sim \mathrm{~N}(0,1)$ process. The simulation requires no burn-in period. The EWMA with $\lambda=1.00$ corresponds to a Shewhart control-chart, hence the EWMA CLs for $\lambda=1.00$ are shown in Table 4 (2nd row) correspond to desired Shewhart IC ARLs between 400 and 340 . This simulation is to validate the modified design when $\lambda=1.00$. Under this process we expect the error in estimation of Shewhart desired ARL $=370$ to be $E \approx 1$ with $95 \%$ degree of confidence.
- $\quad$ Study 2: One individual validation run (Num Runs $=1$ ) of $m=7,000$ and $n \operatorname{sim}=200,000, \lambda=0.25$, for the IC $z \sim \mathrm{~N}(0,1)$ process. To maintain consistency with methods and results in literature, there is no burn-in period, nor are time-varying control-limits used at start-up. The modified EWMA CLs with $\lambda=0.25$ shown in Table 4 correspond to desired EWMA IC ARLs between about 536 and 462. This simulation is to validate the modified design with expected results when $\lambda=0.25$ with no mean-shift ( $\sigma=0.00$ ).
- $\quad$ Study 3: One individual estimation run (Num Runs $=1$ ) of $m=1,000$ and $n \operatorname{sim}=200,000, \lambda=0.25$, for the $\mathrm{OC} z \sim \mathrm{~N}(\mu$-Shift, 1$)$ process based on CLs in study 2. The simulation uses a burn-in period of 50 . This simulation is to estimate RL properties and generate summary estimates (ARL, MRL, SRL) and compare with expected OC results when $\lambda=0.25$ and mean-shifts of $\sigma=0.50, \sigma=1.00$, and $\sigma=1.50$.

For study 1, the observed ARLs, MRLs and SRLs are consistent with what is expected. While many of the observed ARL and MRL estimates equal the expected results, the maximum error for those that do not equal expected results is never more than $E=1$. While many of the observed SRLs are also equal to expected results, the maximum error never exceeds $E=2$. Recall in the limited validation study in a subsequent section that the SRL is most often marginally inflated/deflated from expected results. Since the simulation result is consistent

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

with what is known, the modified design is validated and expected to be highly accurate.

For study 2, focus on the observed EWMA ARL corresponding to Shewhart IC ARL $=370$. Much of the comparative literature on the EWMA provide results for $\lambda=0.25$ and $\mathrm{ARL}=370$, relating that the expected $\mathrm{EWMA} \operatorname{IC} \mathrm{ARL}=500$.

Many studies based on integral equation, MCMC and MCS studies estimate the actual IC ARL between 501 and 503. This simulation result reveals estimated $A R L=501$, which is largely consistent with the previous finding. Since the simulation result is consistent with the existing literature, the modified design is again validated and expected to be highly accurate.

Table 4. EWMA validation and estimation summaries

| Shewhart IC ARLs | $\mathbf{4 0 0}$ | $\mathbf{3 9 0}$ | $\mathbf{3 8 0}$ | $\mathbf{3 7 0}$ | $\mathbf{3 6 0}$ | $\mathbf{3 5 0}$ | $\mathbf{3 4 0}$ |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| EWMA CLs $\boldsymbol{\lambda}=\mathbf{1 . 0 0}$ | $\mathbf{3 . 0 2 3 3}$ | $\mathbf{3 . 0 1 5 7}$ | $\mathbf{3 . 0 0 7 8}$ | $\mathbf{2 . 9 9 9 7}$ | $\mathbf{2 . 9 9 1 3}$ | $\mathbf{2 . 9 8 2 7}$ | $\mathbf{2 . 9 7 3 8}$ |
| Expected IC ARL | 400 | 390 | 380 | 370 | 360 | 350 | 340 |
| Observed IC ARL | 399 | 389 | 380 | 371 | 360 | 350 | 341 |
| Expected IC MRL | 277 | 270 | 263 | 256 | 249 | 242 | 235 |
| Observed IC MRL | 278 | 270 | 263 | 257 | 250 | 243 | 237 |
| Expected IC SRL | 400 | 390 | 380 | 370 | 360 | 350 | 340 |
| Observed IC SRL | 398 | 389 | 380 | 370 | 359 | 349 | 340 |
|  |  |  |  |  |  |  |  |
| EWMA CLs $\boldsymbol{\lambda}=\mathbf{0 . 2 5}$ | $\mathbf{1 . 1 4 2 7}$ | $\mathbf{1 . 1 3 9 8}$ | $\mathbf{1 . 1 3 6 8}$ | $\mathbf{1 . 1 3 3 8}$ | $\mathbf{1 . 1 3 0 6}$ | $\mathbf{1 . 1 2 7 4}$ | $\mathbf{1 . 1 2 4}$ |


|  | $\boldsymbol{y y y y y y y y}$ | $\boldsymbol{\mu}$-Shift $=\mathbf{0 . 0 0}$ |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Observed IC ARL | 536 | 524 | 512 | 501 | 488 | 475 | 462 |
| Observed IC MRL | 372 | 364 | 355 | 347 | 338 | 330 | 320 |
| Observed IC SRL | 533 | 523 | 510 | 503 | 489 | 475 | 463 |


| $\boldsymbol{\mu}$-Shift $=\mathbf{0 . 5 0}$ |  |  |  |  |  |  |  |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| Observed OC ARL | 50 | 49 | 48 | 48 | 47 | 46 | 46 |
| Observed OC MRL | 36 | 35 | 35 | 34 | 34 | 34 | 33 |
| Observed OC SRL | 46 | 45 | 44 | 44 | 43 | 43 | 42 |


| $\boldsymbol{\mu}$-Shift $=\mathbf{1 . 0 0}$ |  |  |  |  |  |  |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
| Observed OC ARL | 11 | 11 | 11 | 11 | 11 | 11 |  |
| Observed OC MRL | 9 | 9 | 9 | 9 | 9 | 9 |  |
| Observed OC SRL | 8 | 8 | 8 | 8 | 7 | 7 |  |


| $\boldsymbol{\mu}$-Shift $=\mathbf{1 . 5 0}$ |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| Observed OC ARL | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Observed OC MRL | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Observed OC SRL | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

JOHN N. DYER

For study 3, again focus on EWMA IC ARL $=370$ and OC ARLs related to the three specified mean-shifts. The findings are consistent with existing literature regarding OC ARLs (Lucas \& Saccucci, 1990). Although not a part of this study, it is interested to note the same OC ARLs for mean-shifts $\sigma=1.00$, and $\sigma=1.50$, suggesting that across the specified range of IC ARLs one can expect the same OC ARLs over the specified range of EWMA CLs, hence one would benefit most by choosing the wider CLs to increase the IC ARL. Since the simulation result is consistent with the existing literature, the modified design is again validated and expected to be highly accurate, hence one could feel confident using the design over a wider range of EWMA studies.

## Conclusion

Two MCS validation design schemes related to control-charting simulation studies were proposed. The basic design was modified to evaluate the EWMA control-chart. Three EWMA MCS studies were conducted and evaluated, resulting in summaries consistent with existing literature, hence validating the adequacy of the MCS design schemes. Although the MCS design is specific to control-chart evaluation, the basic design and related issues extend to simulation studies in other fields. It is suggested that researchers and practitioners using any MCS design should state results relative to the issues discussed in this paper, including justification of RNGs, simulation size, expected error, burn-in period, and design validation, among others.

## References

Abdel-Aziz, A. S., Abdel Dayem, M., \& Darwis, G. (2008). Network intrusion detection system applying multivariate control charts. Paper presented at INFOS2008, Cairo, Egypt. Retrieved from http://infos2008.fci.cu.edu.eg/infos/NET_02_P011-018.pdf

Abramowitz, M., \& Stegun, I. A. (1972). Handbook of mathematical functions. New York: Dover.

Acklam, P. J. (2014). An algorithm for computing the inverse normal cumulative distribution function. Unpublished manuscript. Retrieved from http://home.online.no/~pjacklam/notes/invnorm/

Ahrens, J. H., \& Dieter, U. (1988). Efficient table-free sampling methods for the exponential, Cauchy, and normal distributions. Communications of the ACM, 31(11), 1330-1337. doi: 10.1145/50087.50094

Annen, K. (2013). WEB_REG_MT19937 [computer software]. Retrieved from http://www.web-reg.de

Benneyan, J., Lloyd, R., \& Plsek, P. (2003). Statistical process control as a tool for research and healthcare improvement. Quality and Safety in Health Care, 12(6), 458-464. doi: 10.1136/qhc.12.6.458

Box, G. E., \& Muller, M. E. (1958). A note on the generation of random normal deviates. The Annals of Mathematical Statistics, 29(2), 610-611. doi: 10.1214/aoms/1177706645

Chen, G. (1997). The mean and standard deviation of the run length distribution of $\bar{X}$ charts when control limits are estimated. Statistica Sinica, 7(3), 789-798. Retrieved from http://www3.stat.sinica.edu.tw/statistica/j7n3/j7n314/j7n314.htm

Daudin, J. J. (1992). Double sampling charts. Journal of Quality Technology, 24(2), 78-87.

Dyer, J. N., Adams, B. M., \& Conerly, M. D. (2003). The reverse moving average control chart and comparisons of forecast based monitoring schemes. Journal of Quality Technology, 35(2), 135-155.

Dyer, J. N., \& Barrett, J. D. (2000). A simple mathematical description of the impact of forecast recovery for AR(2) processes. Palmetto Review, 3, 14-22.

Dyer, J. N., Conerly, M. D., \& Adams, B. M. (2003). A simulation study of the impact of forecast recovery for control charts applied to various ARMA processes. Journal of Modern Applied Statistical Methods, 1(2), 343-353. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol1/iss2/43/

Dyer, J. N., Conerly, M. D., Adams, B. M., \& Barrett J. D. (2002). Multivariate forecast based control charting schemes. Journal of Business, Industry, and Economics, 1(2), 145-156.

Fu, M. C., \& Hu, J. (1999). Efficient design and sensitivity analysis of control charts using Monte Carlo simulation. Management Science, 45(3), 395413. doi: $10.1287 / \mathrm{mnsc} .45 .3 .395$

Gan, F. F. (1993). An optimal design of EWMA control charts based on median run length. Journal of Statistical Computation and Simulation, 45(3-4), 169-184. doi: 10.1080/00949659308811479

Golosnoy, V., \& Schmid, W. (2007). EWMA control charts for monitoring optimal portfolio weights. Sequential Analysis: Design Methods and Applications, 26(2), 195-224. doi: 10.1080/07474940701247099

He, D., Grigoryan, A., \& Sigh, M. (2002). Design of double- and triplesampling X-bar control charts using genetic algorithms. International Journal of Production Research, 40(6), 1387-1404. doi: 10.1080/00207540110118415

Hunter, J. S. (1986). The exponentially weighted moving average. Journal of Quality Technology, 18(4), 203-210.

Irianto, D., \& Shinozaki, N. (1998). An optimal double sampling $\overline{\mathrm{X}}$ control chart. International Journal of Industrial Engineering: Theory, Applications and Practice, 5(3), 226-234.

Jaing, W., Au, T., \& Tsui, K.-L. (2007). A statistical process control approach to business activity monitoring. IIE Transactions, 39(3), 235-249. doi: 10.1080/07408170600743912

Kalgonda, A., Koshti, V., \& Ashokan, K. (2011). Exponentially weighted moving average control chart. Asian Journal of Management Research, 2(1), 253263.

Kinderman, A. J., \& Ramage, J. G. (1976). Computer generation of normal random variables. Journal of the American Statistician, 71(356), 893-898. doi: 10.1080/01621459.1976.10480965

Knusel, L. (2002). On the reliability of Microsoft Excel XP for statistical purposes. Computational Statistics \& Data Analysis, 39(1), 109-110. doi: 10.1016/S0167-9473(02)00035-X

Korn, R., Korn, E., \& Kroisandt, G. (2010). Monte Carlo methods and models in finance and insurance. Boca Raton, FL: CRC Press/Taylor \& Francis.

Lin, W. S., \& Adams, B. N. (1996). Combined control charts for forecastbased monitoring schemes. Journal of Quality Technology, 28(3), 289-302.

Lucas, J. M., \& Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32(1), 112. doi: 10.1080/00401706.1990.10484583

Marsaglia, G., \& Bray, T. A. (1964). A convenient methods for generating normal variables. SIAM Review, 6(3), 260-264. doi: 10.1137/1006063

Marsaglia, G., \& Tsang, W. W. (2000). The ziggurat method for generating random variables. Journal of Statistical Software, 5(8), 1-7. doi:
10.18637/jss.v005.i08

Marsaglia, G., Zaman, A., \& Marsaglia, J. C. (1994). Rapid evaluation of the inverse of the normal distribution function. Statistics \& Probability Letters, 19(4), 259-266. doi: 10.1016/0167-7152(94)90174-0

Matsumoto, M., \& Nishimura, T. (1998). Mersenne twister: A 623dimensionally equidistributed uniform pseudo-random number generator. $A C M$ Transactions on Modeling and Computer Simulation, 8(1), 3-30. doi: 10.1145/272991.272995

McCullough, B. D., \& Heiser, D. A. (2008). On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis, 52(10), 4570-4578. doi: 10.1016/j.csda.2008.03.004

Microsoft. (2011). Description of the NORMSDIST function in Excel. Retrieved from http://support.microsoft.com/kb/827369

Montgomery, D. C. (1996), Introduction to statistical quality control (3rd ed.). New York: Wiley.

Nelson, L. S. (1984). Column: Technical aids: The Shewhart control charttests for special cases. Journal of Quality Technology, 16(4), 238-239.

NIST/SEMATECH. (2012). NIST/SEMATECH e-handbook of statistical methods. Retrieved from http://www.itl.nist.gov/div898/handbook/

Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1/2), 100-115. doi: 10.2307/2333009

Roberts, C., \& Casella, G. (1999). Monte Carlo statistical methods. New York: Springer.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. Technometrics, l(3), 239-250. doi: 10.1080/00401706.1959.10489860

Ryan, T. P. (2000). Statistical methods for quality improvement (2nd ed.). New York: Wiley.

Salleh, S. (2013, June 27). Monte Carlo simulation tips and tricks [Web log post]. Retrieved from http://decision-analytics-blog.lumina.com/blog/monte-carlo-simulation-tips-and-tricks

Severin, T., \& Schmid, W. (1998). Statistical process control and its application in finance. In G. Bol, G. Nakhaeizadeh, \& K.-H. Vollmer (Eds.), Risk measurement, econometrics and neural networks (pp. 83-104). New York: Physica-Verlag. doi: 10.1007/978-3-642-58272-1_7

Shewhart, W. A. (1980). Economic control of quality of manufactured product. Wilwaukee, WI: American Society for Quality. (Original work published 1931)

Shewhart, W. A. (1986). Statistical method from the viewpoint of quality control. Mineola, NY: Dover. (Original work published 1939)

## JOHN N. DYER

Srinivasan, A. (2011). Application of information technology and statistical process control in pharmaceutical quality assurance \& compliance (Master's thesis). Retrieved from DSpace@MIT. (http://hdl.handle.net/1721.1/66047)

Stigler, S. (1973). Studies in the history of probability and statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency. Biometrika, 60(3), 439-445. doi: 10.1093/biomet/60.3.439

Teoh, W. L., \& Khoo, M. B. (2012). Optimal design of the double sampling $\bar{X}$ chart based on median run length. International Journal of Chemical Engineering and Applications, 3(5), 303-306. doi: 10.7763/IJCEA.2012.V3.205

Torng, C., \& Lee, P. (2009, March). A modified statistical design model of double sampling $\bar{X}$ control chart. In S. I. Ao, O. Castillo, C. Douglas, D. D. Feng, \& J.-A. Lee (Eds.), Proceedings of the International Multiconference of Engineers and Computer Scientists 2009 (pp. 1761-1764). Hong Kong: International Association of Engineers. Retrieved from http://iaeng.org/publication/IMECS2009/IMECS2009_pp1761-1764.pdf

Western Electric Company. (1956). Statistical quality control handbook. Indianapolis: Western Electric Co.

Wheeler, D. J., \& Chambers, D. S. (1992). Understanding statistical process control (2nd ed.). Knoxville, TN: SPC Press.

Wichmann, B. A., \& Hill, I. D. (1982). Algorithm AS 183: An efficient and portable pseudo-random number generator. Applied Statistics, 31(2), 188-190. doi: 10.2307/2347988

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

## Appendix A: Design 1 VBA Code

```
Sub Sim()
' 1. Declare Variables
Range("A1"). Select: Dim numRuns As Long, LastRow As Long, LastCol As Long, RLStartRow As Long, RLEndRow As Long
Dim CLStartRow As Long, CLStartCol As Long, CLEndCol As Long
Dim M As Long, nsim As Long, Mu As Double, Sigma As Double, MuShift As Long, NumCL As Long, CL As Double
Dim p1 As Double, p2 As Double, expARL As Double, expMRL As Double, expSRL As Double
Dim Z As Long, a As Long, b As Long, c As Long, d As Long, e As Long, f As Long
Dim zOut() As Variant, CLArray() As Double, RLOut() As Double
Dim wsOut As Range, calcOut As Range, blankOut As Range
Dim wf As WorksheetFunction: Set wf = Application.WorksheetFunction
Dim Path As String: Path = "User Sets Path Here to Save Results": Dim xlsmExt As String: xlsmExt = ".xlsm"
```

```
' 2. Set Number of Simulation Runs
numRuns = Range("B2").Value
'Number
of Simulation Runs (each saved separately to path above)
```

' 3. Start Simulation Runs
For Z = 1 To numRuns: Range("B13") = Z 'Display Run
' 4. Delete Previous Time and Blank Count Outputs Range("B10:B12").Select: Selection.ClearContents 'Delete Previous StartTime, End-Time, \& Run-Time Range("D11").Select: Selection.ClearContents 'Delete Previous Blank Count

[^33]JOHN N. DYER

```
    6. Set Start Time ------------------------------------------------------
StartTime = "=Now()": Range("B10") = StartTime: Range("B10") =
Range("B10") `
```

' 7. Set Variables Typed into Worksheet Inputs ---------------------------
M = Range("B3"). Value
nsim = Range("B4").Value
Mu = Range("B5").Value
Sigma = Range("B6").Value
'IC Std Dev of Normal Distribution
MuShift = Mu + Range("B7").Value * Sigma 'OC Mean of Shifted x~N(Mu+Shift,
1)
' 8. Set Range of Control Limits For ARLs Starting in Range("F3") ------
CLStartRow = 2: CLStartCol $=6$ 'Starting Row/Column of Control Limits
(F3)
NumCL = wf.Count(Range(Cells(1, CLStartCol), Cells(1, LastCol))) 'Count
Number of Control Limits
CLEndCol = CLStartCol + NumCL - 1 'Ending Column of Control Limits
RLStartRow = 10
RLEndRow = RLStartRow + nsim - $1 \quad$ 'Ending Row of Run-Length Output
Set wsOut = Range(Cells(RLStartRow, CLStartCol), Cells(RLEndRow,
CLEndCol)) 'Set Range of RL Output in Worksheet
' 9. CaLculate Control Limits and OC ARLs ----------------------------------ReDim CLArray(1 To 2, 1 To NumCL) 'Re-dimension Control-Limit Array For a = CLStartCol To CLEndCol

CLL = wf.Norm_Inv((1 / Cells(1, a)) / 2, Mu, Sigma) 'CaLculate Lower Control Limit Value CLU = wf.Norm_Inv(1 - (1 / Cells(1, a)) / 2, Mu, Sigma) 'Calculate Upper Control Limit Value

Cells(2, a) = CLU: Cells(3, a) = CLL 'Copy CLs into Worksheet CLArray(1, a - 5) = CLU: CLArray(2, a - 5) = CLL 'Copy CLs into CL Array
'Calculate p1 and p2 for Expected ARL Calculation p1 = 1 - wf.Norm_Dist(Cells(2, a), MuShift, Sigma, True) 'P(z>Upper CL)

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

```
    p2 = wf.Norm_Dist(Cells(3, a), MuShift , Sigma, True)
    'P(z<Lower CL)
'Calculate Expected ARLs (will be the same for the IC Process)
    expARL = (1 / (p1 + p2))
    Cells(4, a) = Round(expARL, 2) 'Copy Expected ARLs into
    Worksheet
'Calculate Expected MRLs (will be the same for the IC Process)
    expMRL = (wf.Ln(0.5)) / (wf.Ln(1 - (1 / Cells(4, a)))
    Cells(5, a) = Round(expMRL, 2) 'Copy Expected MRLs into
    Worksheet
'Calculate Expected SRLs (will be the same for the IC Process)
    expSRL = ((1 - (1 / Cells(4, a))) ^ (1 / 2)) / (1 / Cells(4,
a))
    Cells(6, a) = Round(expSRL, 2) 'Copy Expected SRLs into
    Worksheet
Next a
' 10. Start Simulations -----------------------------------------------------
ReDim RLOut(1 To nsim, 1 To NumCL) 'Re-dimension Run-Length Array For \(\mathbf{b}=1\) To nsim: Application.ScreenUpdating = False 'For Each Simulation
' 11. Fill zOut Array with \(z \sim N(M u, S i g m a) ~ R a n d o m ~ N u m b e r s ~ o f ~ S e r i e s-L e n g t h ~\) M using NormInv Function ---------------------------------------------------ReDim zOut(1 To M, 1 To 2) 'Re-dimension Array of Random Numbers For \(\mathbf{c}=1\) To M 'For Each Random Number to Be Generated in the Array \(p=\) Rnd 'Use Rnd Function to generate value of \(p, 0<p<1\)
```

If $p<=0$ Then
$p=$ Rnd 'Generate new value of $p$ End If
zOut(c, 1) = wf.NormInv(p, MuShift, Sigma) 'Use NormInv Fn to fill Array with Random Value
zOut (c, 2) = c 'Record Location in Array for each Random Value Next c
' 12. Compare each z with Control Limits and Record Run-Lengths in RunLength Array
For $\mathbf{d}=1$ To NumCL
For $\mathbf{e}=1$ To $M$
'For Each CL in Control-Limit Array 'For each Random z-Value in Array zOut

## JOHN N. DYER

```
        If zOut(e, 1) > CLArray(1, d) Or zOut(e, 1) < CLArray(2, d)
        Then 'If z exceeds CLs Then
        RLOut(b, d) = zOut(e, 2) 'Record Run-Length Location
        in Run-Length Array
    Exit For 'Exit and Move to Next CL
    Else: End If
    Next e 'Else Move to Next z
Next d
'-------------------------------------------------------------------------------
```

Next b
13. Copy Run-Length Array into Worksheet
wsOut.Value = RLOut: Application.ScreenUpdating = True 'Copy Run-Length
Array into Worksheet Range
' 14. Calculate Estimated ARLs, MRLs and SRLs
For $\mathbf{f}=$ CLStartCol To CLEndCol: Set calcOut = Range (Cells(RLStartRow, f),
Cells(RLEndRow, f))
Cells(7, f) = wf.Average(calcOut) 'CaLculate ARL
Cells(8, f) = wf.Median(calcOut) 'CaLculate MRL
Cells(9, f) = wf.StDev(calcOut) 'Calculate SRL
Next f
' 15. Count Blank Run-lengths ------------------------------------------------
Set blankOut = Range(Cells(RLStartRow, CLStartCol), Cells(RLEndRow,
CLStartCol))
Range("D11") = wf.CountBlank(blankOut) 'Count Simulations with Blank Run-
Lengths

EndTime = "=Now()": Range("B11") = StartTime: Range("B11") = Range("B11")
Range("B12") = Range("B11") - Range("B10")
17. Save Workbook and Do Next Run Z

Range("A1").Select: ActiveWorkbook.SaveAs Path \& "-M=" \& M \& "-Z=" \& Z \& xlsmExt 'Save Workbook
Next Z 'Next Run
End Sub

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

## Appendix B: Design 2 VBA Code

Sub Sim()
' 1. Declare Variables ----------------------------------------------------Dim LastRow As Long, LastCol As Long, RLStartRow As Integer, RLCountRow As Integer, MaxRow As Long

Dim CLStartCol As Integer, CLEndCol As Integer, CLStartRow As Integer, numRuns As Long, M As Long Dim Mu As Double, Sigma As Double, MuShift As Double, NumCL As Integer, CL As Double, p1 As Double, p2 As Double
Dim exparl As Double, expMrL As Double, expSRL As Double, z As Long, a As Long, c As Long, d As Long, e As Long Dim f As Long, g As Long, h As Long, i As Long, RL2 As Long, RL1 As Long, RLDiff As Long, ZOut() As Variant
Dim RLArray() As Variant, CLArray() As Variant, ROut As Long, MaxCount As Long, MaxRL() As Long, RLCount As Long
Dim calcOut As Range, wf As WorksheetFunction: Set wf = Application.WorksheetFunction
Dim Path As String: Path = "User Sets Path to Save Results": Dim xlsmExt As String: xlsmExt = ".xlsm"
' 2. Set Number of Simulation Runs
numRuns = Range("B2").Value 'Number of Simulation Runs (each saved separately to path above)
' 3. Start Simulation Runs ------------------------------------------------For Z = 1 To numRuns: Range("B13") = Z 'Display Run
' 4. Delete Previous Time and Blank Count Outputs Range("B9:B11").Select: Selection.ClearContents 'Delete Previous StartTime, End-Time, \& Run-Time
' 5. DeLete Previous Estimates and Run-Length Outputs ----------------
LastRow $=$ ActiveSheet.UsedRange.Rows.Count: 'Count Last Row of Worksheet
LastCol = ActiveSheet.UsedRange.Columns.Count 'Count Last Column of

| Worksheet |
| :--- |
| Range(Cells(2, Cells(LastRow, $\quad$ 6), $\quad$ LastCol)).Select: |
| Selection.ClearContents 'DeLete previous Estimates and RL outputs |

## JOHN N. DYER


'7. Set Variables Typed into Worksheet Inputs $\qquad$

M = Range("B3").Value
Mu = Range("B4"). Value
Sigma = Range("B5").Value MuShift = Mu + Range("B6"). Value * Sigma 'OC Mean of Shifted x~N(Mu+Shift, 1)
' 8. Set Range of Control Limits For ARLs Starting in Range("F3") -----CLStartRow $=2:$ CLStartCol $=6$ 'Starting Row/Column of Control Limits (F3)
NumCL = wf.Count(Range(Cells(1, CLStartCol), Cells(1, LastCol))) 'Count Number of Control Limits
CLEndCol = CLStartCol + NumCL - 1 'Ending Column of Control Limits
RLCountRow=10
RLStartRow = 12
MaxRow=1048576
'Series Length
'IC Mean of Normal Distribution 'IC Std Dev of Normal Distribution 'Row of Run-Length Counts
'Starting Row of Run-Length Output
'Last Row in Excel Worksheet
 ReDim CLArray(1 To 2, 1 To NumCL) 'Re-dimension Control-Limit Array For $\mathbf{a}=$ CLStartCol To CLEndCol

```
CLL = wf.Norm_Inv((1 / Cells(1, a)) / 2, Mu, Sigma) 'Calculate
Lower Control Limit Value
CLU = wf.Norm_Inv(1 - (1 / Cells(1, a)) / 2, Mu, Sigma) 'Calculate
Upper Control Limit Value
Cells(2, a) = CLU: Cells(3, a) = CLL 'Copy CLs into Worksheet
CLArray(1, a - 5) = CLU: CLArray(2, a - 5) = CLL 'Copy CLs
into CL Array
'Calculate p1 and p2 for Expected ARL Calculation
p1 = 1 - wf.Norm_Dist(Cells(2, a), MuShift, Sigma, True)
'P(z>Upper CL)
p2 = wf.Norm_Dist(Cells(3, a), MuShift , Sigma, True)
'P(z<Lower CL)
'Calculate Expected ARLs (will be the same for the IC Process)
expARL = (1 / (p1 + p2))
```


## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

```
    Cells(4, a) = Round(expARL, 2) 'Copy Expected ARLs into
    Worksheet
'CaLculate Expected MRLs (will be the same for the IC Process)
    expMRL = (wf.Ln(0.5)) / (wf.Ln(1 - (1 / Cells(4, a)))
    Cells(5, a) = Round(expMRL, 2) 'Copy Expected MRLs into
    Worksheet
'Calculate Expected SRLs (will be the same for the IC Process)
    expSRL = ((1 - (1 / Cells(4, a))) ^ (1 / 2)) / (1 / Cells(4,
a))
    Cells(6, a) = Round(expSRL, 2) 'Copy Expected SRLs into
    Worksheet
```

Next a
10. Start Simulations - Fill zout Array with z~N(Mu,Sigma) Random Numbers of Series-Length M using NormInv Function-----------------------ReDim ZOut(1 To M, 1 To 2): Application.ScreenUpdating = False 'Redimension Run-Length Array

For c = 1 To M 'For Each Random Number to Be Generated in the Array $p=$ Rnd 'Use Rnd Function to generate value of $p, 0<p<1$ If p <= 0 Then
'If $p<\theta$ Then $p$ Rnd 'Generate new value of $p$ End If
zOut(c, 1) = wf.NormInv(p, MuShift, Sigma) 'Use NormInv Fn to fill Array with Random Value
zOut (c, 2) = c 'Record Location in Array for each Random Value Next c
' 11. Compare each z with Control Limits and Record Run-Lengths in RunLength Array -------------------------------------------------------------For $\mathbf{d}=1$ To NumCL

Rout = 11
'For Each CL in Control-Limit Array

For $e=1$ To $M$ 'Set Row Output to Row 11

If ROut = MaxRow Then Length
Exit For 'Exit For Loop If Row Out Exceeds Excel's Max Row Length ElseIf zOut(e, 1) > CLArray(1, d) Or zOut(e, 1) < CLArray(2, d) Then 'If $z$ exceeds CLs Then

Rout = Rout + $1 \quad$ 'Increment Row Output by 1

## JOHN N. DYER

```
    Cells(ROut, d).Value = e 'Record Run-Length Location
        End If
    Next e 'Else Move to Next z-value
Next d
    'Move to Next CL
```

' 12. Calculate the Count of Each Run Length for Each CL \& Store in Row

For $\mathbf{f}=$ CLStartCol To CLEndCol 'For Each CL
RLCount = wf.CountA(Range(Cells(RLStartRow, f), Cells(MaxRow, f)))
'Count Number of Non-Zero Run-Lengths
Cells(RLCountRow, f) = RLCount 'Copy Count into Worksheet Cells
Next f
'Move to Next CL
' 13. Iteratively Subtract Subsequent $R L$ from Previous RL to Calculate
Indexed Run-Lengths ---------------------------------------------------------
LastRow = ActiveSheet.UsedRange.Rows.Count 'Count Last Row of Worksheet
MaxCount = wf.Max(Range(Cells(RLCountRow, CLStartCol), Cells(RLCountRow,
CLEndCol)))
ReDim RLArray(1 To MaxCount, 1 To NumCL) 'Re-dimension Run-Length Array
For $\mathbf{g}=$ CLStartCol To CLEndCol 'For Each CL
RLCount $=$ Cells(RLCountRow, g) 'Set Count from Worksheet Cells
For $\mathbf{h}=1$ To RLCount 'For Each Row/Column Cell
RL2 $=$ Cells $(\mathrm{h}+$ RLCountRow $+1, \mathrm{~g})$.Value 'Set RL2 = to
Subsequent Run-Length Value
RL1 $=$ Cells(h + RLCountRow, g).Value 'Set RL1 = to Previous
Run0-Length Value
RLDiff = RL2 - RL1 'CaLculate Difference RL2-RL1
RLArray(h, g - (CLStartCol - 1)) = RLDiff 'Copy RLDiff into
RLArray
Next h 'Move to Next Row/Column Cell
Next g
'Move to Next CL
' 14. Copy Run-Length Array into Worksheet $\qquad$ Range(Cells(RLStartRow, CLStartCol), Cells(MaxCount + RLCountRow + 1, CLEndCol)) = RLArray
' 15. Find \& Replace 0 Run-Lengths with Null String ---------------------LastRow = ActiveSheet.UsedRange.Rows.Count 'Count Last Row of Worksheet

## MCS DESIGN AND NORMAL-BASED CONTROL CHART PROPERTIES

```
Range(Cells(RLCountRow + 1, CLStartCol), Cells(LastRow, LastCol)).Select
'Find & Replace 0 Run-Lengths
    Selection.Replace What:="0", Replacement:="", LookAt:=xlWhole,
    SearchOrder:=xlByColumns, MatchCase:=True _
    SearchFormat:=False, ReplaceFormat:=False: Range("A1").Select
    16. Calculate Estimated ARLs, MRLs and SRLs ----------------------------
For i= CLStartCol To CLEndCol
RLCount = Cells(RLCountRow, i).Value 'Record RL Count for Each CL
Set calcOut = Range(Cells(RLStartRow, i), Cells(RLCount + RLCountRow + 1,
i)) 'Set Range of Run-Length Output
    Cells(7, i) = wf.Average(calcOut) 'Calculate ARL
    Cells(8, i) = wf.Median(calcOut) 'Calculate MRL
    Cells(9, i) = wf.StDev(calcOut) 'Calculate SRL
Next i
' 17. Set End Time and Calculate Run Time --------------------------------
EndTime = "=Now()": Range("B10") = StartTime: Range("B10") = Range("B10")
Range("B11") = Range("B10") - Range("B9")
' 18. Save Workbook and Do Next Run Z -------------------------------------
Range("A1").Select
ActiveWorkbook.SaveAs Path & "-" & M & "-" & Z & xlsmExt 'Save Workbook
Next Z
                                    'Next Run
End Sub
```


# Latent Variable Model for Weight Gain Prevention Data with Informative Intermittent Missingness 

## Li Qin

Yale University
New Haven, CT

Lisa Weissfeld<br>University of Pittsburgh<br>Pittsburgh, PA

Michele Levine<br>University of Pittsburgh<br>Pittsburgh, PA

Marsha Marcus<br>University of Pittsburgh<br>Pittsburgh, PA

Feng Dai<br>Yale University<br>New Haven, CT

Missing data is a common problem in longitudinal studies because of the characteristics of repeated measurements. Herein is proposed a latent variable model for nonignorable intermittent missing data in which the latent variables are used as random effects in modeling and link longitudinal responses and missingness process. In this methodology, the latent variables are assumed to be normally distributed with zero-mean, and the values of variance-covariance are calculated through maximum likelihood estimations. Parameter estimates and standard errors of the proposed method are compared with the mixed model and the complete-case analysis in the simulations and the application to the weight gain prevention among women (WGPW) data set. In the simulation results with respect to bias, mean squared error, and coverage of confidence interval, the proposed model performs better than the other two methods in different scenarios. Relatively, the proposed latent variable model and the mixed model do a better job for between-subject effects compared to within-subject effects. The converse is true for the complete case analysis. The simulation results also provide support for application of this proposed latent variable model to the WGPW data set.

Keywords: Latent variable, longitudinal study, non-ignorable missing data, weight gain prevention

## Introduction

Missing data is a common issue encountered in the analysis of longitudinal data. In the behavioral intervention setting, missed visits and/or losing to follow up can be extremely problematic. In this area, missed visits are assumed to be a result of

[^34]
## LATENT VARIABLE MODEL IN OBESITY DATA

failure of the intervention, sustained lack of interest in the study, or decreased desire to change the behavior (Qin et al., 2009). For weight loss studies, these are common issues that must be dealt with at the data analysis phase. For example, Levine et al. (2007) conducted a weight gain prevention study among women (WGPW) aged 25 to 45 years old. Participants were assessed for BMI (Body Mass Index) at baseline, year one, two and three. However, the outcomes at follow-ups for some women were missing. Because the missing data might be related to their unobserved BMIs, they were considered as nonignorable, informative, or missing not at random (MNAR) (Rubin, 1976).

To account for informative missingness, a number of model-based approaches were proposed to jointly model the longitudinal outcome and the missingness mechanism. The methodology adopted here is motivated by latent pattern mixture models (Lin, McCulloch, \& Rosenheck, 2004) and latent dropout class models (Roy, 2003). In latent pattern mixture models, the mixture patterns are formed from latent classes that link the longitudinal responses and the missingness process. A non-iterative approach has been proposed, to assess the assumption of the conditional independence between the longitudinal outcomes and the missingness process given the latent classes (Lin et al., 2004). Roy (2003) noted the idea of pattern-mixture models (e.g., Little, 1993) is not appropriate in many circumstances, because there are many reasons for missingness and subjects with the same missingness pattern may not share a common distribution. Roy (2003) assumed the existence of a small number of dropout classes behind the observed dropout times. But for Roy (2003)'s method, it is difficult to decide the number of latent classes ahead of the analysis. It also leads to misclassification because it is difficult to divide subjects into classes due to the variety of reasons for missingness. Some subjects may not belong to any latent classes. So it is reasonable and straightforward to propose a latent variable model in which the latent variable is unobserved and continuous.

The WGPW study data (Levine et al., 2007) provides motivation to adopt the latent pattern mixture model methodology. In this trial, interventions were compared with a control group in preventing weight gain among normal or overweight women. 190 women were randomized to clinic-based group intervention and information-only control condition. For women randomized to the interventions, treatment was provided over a two-year period, with a followup at year three. All women participated in yearly assessment. The primary outcome of interest was body mass index (BMI) calculated from weight assessed yearly and height at baseline. Overall, $81 \%, 76 \%$ and $36 \%$ completed a weight assessment at year one, two and three, respectively. The reasons for this

## QIN ET AL.

incompleteness may be related to their unobserved outcomes. To avoid biased estimations, possible dependence of missingness status on unobserved responses has to be considered.

A latent variable model is proposed for informative intermittent missingness, developed from Henderson, Diggle, and Dobson's (2000) joint modeling of longitudinal measurements and event time data. In the proposed model, longitudinal process and missing data process are linked through a latent bivariate Gaussian process $\boldsymbol{W}(t)=\left\{W_{1}(t), W_{2}(t)\right\}$. An assumption of this latent variable model is that the longitudinal measurements and missing data process are conditional independent given $\boldsymbol{W}(t)$. This assumption simplifies likelihood function. It also increases the strength of the relationship between the missing data process and underlying true outcome process determined by the correlation between $W_{1}(t)$ and $W_{2}(t)$.

The proposed latent variable model and the parameter estimation is described in next section. A simulation study is carried out in the following section, to compare the performance of the latent variable model with mixed model and complete-case analysis. The proposed model is then applied to the WGPW data (Levine et al., 2007) and compared with the mixed model and complete-case analysis, and the assessment of fit of the model is treated. A discussion is provided in the last section.

## Model specification and estimation

Assume the proposed latent variable model is present for the full data. Denoting a normally distributed continuous response variable measured on the $i^{\text {th }}$ subject at the $j^{\text {th }}$ occasion as $Y_{i j}(i=1, \ldots, N ; j=1, \ldots, K)$, the $K$ intended responses are collected into a vector $\boldsymbol{Y}_{i}=\left(Y_{i 1}, \ldots, Y_{i K}\right)$ if there is no missing data.

For various reasons, not all subjects have all $K$ measurements. Here the baseline measure $Y_{i 1}$ is assumed to be observed for every individual. When missingness process occurs as a result of dropout, the response $Y_{i j}$ for subject $i$ is only observed at time points $j=1, \ldots, k_{i}$; where $k_{i} \leq K$. But if the data are subject to intermittent missingness, before time point $k_{i}$, there may be additional missing measurements. A missingness indicator, $R_{i j}$, is used for each of the $K$ measurements, with 1 if $Y_{i j}$ is missing and 0 if $Y_{i j}$ is observed.

In the following, random-effect models are briefly described for the separate analysis of longitudinal data and missingness procedure, and the joint model via a latent zero-mean bivariate Gaussian process.

## LATENT VARIABLE MODEL IN OBESITY DATA

## Longitudinal Responses

The sequence of longitudinal measurements $Y_{i 1}, Y_{i 2}, \ldots, Y_{i K}$ for the $i^{\text {th }}$ subject at times $t_{i 1}, t_{i 2}, \ldots, t_{i K}$ is modeled as

$$
Y_{i j}=\beta^{T} x_{i j}+W_{1 i}\left(t_{i j}\right)+\varepsilon_{i j},
$$

where $\beta^{T} x_{i j}=\mu_{i j}$ is the mean response in which the vector $\beta$ and $x_{i j}$ represent possibly time-varying explanatory variables and their corresponding regression coefficients, respectively; $W_{1 i}\left(t_{i j}\right)$ incorporates subject-specific random effects; and $\varepsilon_{i j} \sim N\left(0, \sigma_{\varepsilon}^{2}\right)$ is a sequence of mutually independent measurement errors corresponding to $Y_{i j}$. The $W_{1 i}\left(t_{i j}\right)$ can be viewed as the actual individual variability of outcome trajectories after they have been adjusted for the overall mean trajectory and other fixed effects.

## Missing Data Procedures

Here $R_{i j}=1$ is defined as $Y_{i j}$ being missing, and $R_{i j}=0$ as $Y_{i j}$ being observed. Letting $\varphi_{i j}$ denote the probability of $R_{i j}=1$, the logistic model for $\varphi_{i j}$ is specified as

$$
\log \frac{i j}{1{ }_{i j}}=\alpha^{T} z_{i j}+W_{2 i}\left(t_{i j}\right) .
$$

where $\alpha$ is a vector of log odds ratios corresponding to $z_{i j} ; z_{i j}$ is a vector of covariates specific to the missingness process for subject $i$; and $W_{2 i}\left(t_{i j}\right)$ represents random effect.

## Latent Variable Model

The dependence between the missingness process and longitudinal responses is characterized by sharing a common random effect vector for the $i^{\text {th }}$ subject, say $\left(W_{1 i}, W_{2 i}\right)^{T}$, which is independent across different subjects. Thus, the stochastic dependence between $W_{1 i}$ and $W_{2 i}$ is critical. It is referred as latent association. Before specifying $\left(W_{1 i}, W_{2 i}\right)^{T}$, the pair of latent variables $\left(U_{1 i}, U_{2 i}\right)^{T}$ are defined with a mean-zero bivariate Gaussian distribution $N(0, \Sigma)$ (Henderson et al., 2000). The $\left(W_{1 i}, W_{2 i}\right)^{T}$ are then modeled as

$$
W_{1 i}(s)=U_{1 i}+U_{2 i} s
$$

## QIN ET AL.

$$
W_{2 i}(t)=\lambda_{1} U_{1 i}+\lambda_{2} U_{2 i} t
$$

Both $W_{1 i}$ and $W_{2 i}$ are represented as random intercept and slope terms; $s$ and $t$ are possibly time-varying explanatory variables; $\lambda_{1}$ and $\lambda_{2}$ are the parameters measuring the association between $W_{1 i}$ and $W_{2 i}$, that is, the association between longitudinal and missing data processes induced through the intercept, slope and current $W_{1}$ value. The derivatives of $W_{2 i}$ are as follows:

$$
\begin{aligned}
W_{2 i}(t) & =\lambda_{1} U_{1 i}+\lambda_{2} U_{2 i} t \\
& =\gamma_{1} U_{1 i}+\gamma_{2} U_{2 i} t+\gamma_{3}\left(U_{1 i}+U_{2 i} t\right) \\
& =\gamma_{1} U_{1 i}+\gamma_{2} U_{2 i} t+\gamma_{3} W_{1 i}(t),
\end{aligned}
$$

where $\lambda_{1}=\gamma_{1}+\gamma_{3}$ and $\lambda_{2}=\gamma_{2}+\gamma_{3}$.
In this way, the traditional Laird-Ware random effects models are combined with a proportionality assumption $W_{2 i}(t) \propto W_{1 i}(t)$. A simple case of this assumption is $W_{2 i}(t)=W_{1 i}(t)$, in which $\gamma_{1}=\gamma_{2}=0$ and $\gamma_{3}=1$. The proportionality assumption allows us to consider more complicated situations in which the association between longitudinal and missing data processes is described in terms of the intercept and slope. In other words, the impact of underlying random effect structure differences between the longitudinal and missing data processes can be assessed. The fixed effects in sub-models mentioned earlier in this section, $x_{i j}$ and $z_{i j}$, may or may not correspond to the same covariates. Actually, the dependence between $Y_{i j}$ and $\varphi_{i j}$ may arise in two ways: through the common fixed effects or through stochastic dependence between $W_{1 i}$ and $W_{2 i}$. Even if $W_{1 i}$ and $W_{2 i}$ are independent, the longitudinal and missing data processes still could be associated through the common fixed effects.

## Estimation

Let $\boldsymbol{y}_{i}, \boldsymbol{y}_{i}{ }^{c}$ and $\boldsymbol{y}_{i}{ }^{m}$ denote the vector of observed, complete and missing longitudinal responses for the $i^{\text {th }}$ subject. Let $\boldsymbol{\psi}^{\mathrm{T}}=\left(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}\right)$ represent the set of parameters of interests; the observed log-likelihood for the joint model is

$$
\begin{aligned}
\log L(; y,, W \mid x, z) & =\sum_{i=1}^{N} \int\left[\log L\left(; y_{i}^{c} \mid x_{i},{ }_{i}, W_{i}\right)+\log L\left(;{ }_{i} \mid z_{i}, W_{i}\right)+\log L\left(; W_{i}\right)\right] d y_{i}^{m} \\
& =\sum_{i=1}^{N}\left[\log L\left(; y_{i} \mid x_{i}, W_{i}\right)+\log L\left(;{ }_{i} \mid z_{i}, W_{i}\right)+\log L\left(; W_{i}\right)\right]
\end{aligned}
$$

where

## LATENT VARIABLE MODEL IN OBESITY DATA

$$
\begin{aligned}
& \log L\left(; y_{i} \mid x_{i}, W_{i}\right)=\left(\frac{1}{2}\right)\left[\begin{array}{rllll}
\log (2 & ) & +\log \left(\begin{array}{cc}
{ }^{2}+ & \\
& 11
\end{array}\right) \\
& +\left(\begin{array}{cc}
{ }^{2}+ & \\
& \\
& 11
\end{array}\right)^{1}\left(\begin{array}{lllll}
y_{i} & { }^{T} x_{i} & W_{1 i}
\end{array}\right)^{T}\left(\begin{array}{llll}
y_{i} & { }^{T} x_{i} & W_{1 i}
\end{array}\right)
\end{array}\right], \\
& \log L\left(;_{i} \mid z_{i}, W_{i}\right)=\sum_{j=1}^{K} R_{i j} \log { }_{i j}=\sum_{j=1}^{K} R_{i j}\left({ }^{T} z_{i j}+W_{2 i}\right) \quad \log \left\{\sum_{j=1}^{K} \exp \left({ }^{T} z_{i j}+W_{2 i}\right)\right\} \text {, } \\
& \log L\left(; W_{i}\right)=\left(\frac{1}{2}\right)\left[\log (2)^{2}+\log | |+\left(\begin{array}{lll}
W_{i} & { }_{i}
\end{array}\right)^{T} \quad{ }^{1}\left(\begin{array}{ll}
W_{i} & i
\end{array}\right)\right] \text {. } \\
& { }_{i}=\binom{U_{1 i}+U_{2 i} s}{{ }_{1} U_{1 i}+{ }_{2} U_{2 i} t}
\end{aligned}
$$

is the mean vector for $W_{i}$. Here let

$$
\log L\left(; y_{i} \mid x_{i},{ }_{i}, W_{i}\right)=\log L\left(; y_{i} \mid x_{i}, W_{i}\right)
$$

that is, given the latent variables $W_{i}$, the outcome $\boldsymbol{Y}_{i}$ is independent of the missingness $\varphi_{i}$. This is an important assumption which reduces the mathematical complexity for estimation. Because $\varphi_{i}$ affects $y_{i}$ through $W_{i}$, the missingness is not ignored in the maximum likelihood inference.

The maximum likelihood estimation of the joint model is obtained by the quasi-Newton method, in which the latent variables are estimated by empirical Bayes and standard errors are estimated using the delta method. Because the likelihood equations for the $L\left(\alpha ; \varphi_{i} \mid z_{i}, W_{i}\right)$ are non-linear (from logistic regression) and do not have closed form maximizers, which may lead to some maximization algorithms having difficulty converging, a modified quasi-Newton algorithm is used for maximizing the likelihood. For example, the current estimate of $\psi$ is updated by

$$
{ }^{(k+1)}=(k) \quad(k)\left\{\frac{\partial^{2} l()}{\partial \partial^{T}}\right\}^{1} \frac{\partial l()}{\partial}
$$

where $l(\boldsymbol{\psi})=\log L(\boldsymbol{\psi} ; y, \varphi, W \mid x, z)$, and $a(k)$ is a small constant with values between 0 and 1 . Generally, $a(k)$ starts from very small (e.g., 0.01 ) toward 1 as $k$ increases. The above algorithm may be repeated for different starting values of $\psi$

## QIN ET AL.

to make sure that it will converge to a global maximum. Here, the starting values are chosen from the estimates of complete-case analysis.

## Sensitivity Analysis

The proposed method assumes that the distribution of the longitudinal responses (both observed and missing) does not depend on the missingness procedure after conditioning to latent zero-mean bivariate Gaussian process. This conditional independence assumption is strong, and neither it nor the missing not at random assumption can be tested just using the observed data. The sensitivity analyses will be considered for these assumptions by comparing the new model with commonly used mixed model and complete-case analysis in the simulation and data analysis sections. Results by the proposed method will be reported with different latent processes $W_{1}(s)$ and $W_{2}(t)$. Akaike's information criterion (AIC) (Akaike, 1981) and the Bayesian information criteria (BIC) (Schwartz, 1978) will be used to assess model fit. It must be kept in mind that the unobserved outcomes cannot be checked in any sensitivity analyses.

## Simulation study

A small simulation study was carried out to compare the performance of the latent variable model with mixed model under MAR assumption and complete-case analysis that discards subjects with missing observations. The data sets were generated by considering two aspects: the complete data structure with outcomes and observable independent variables; and the missingness structure.
Complete data is generated with $N=200$ subjects with $J=4$ time points. It is assumed that there are 2 treatment groups with an equal number of subjects in each group. The following specifications for the longitudinal component are assumed: intercept $=-0.5$; treatment $(T x)=1.0$; time 2 vs. time $1\left(T_{2}-T_{1}\right)=0.5$; time 3 vs. time $1\left(T_{3}-T_{1}\right)=1.0$; time 4 vs. time $1\left(T_{4}-T_{1}\right)=1.5$. Consequently the mean of the dependent variable $Y_{i j}$ can be written as:

$$
E\left(Y_{i j}\right)=\beta_{0}+\beta_{1} T x+\beta_{2}\left(T_{2}-T_{1}\right)+\beta_{3}\left(T_{3}-T_{1}\right)+\beta_{4}\left(T_{4}-T_{1}\right)
$$

where $\beta_{0}=-0.5, \beta_{1}=1.0, \beta_{2}=0.5, \beta_{3}=1.0$, and $\beta_{4}=1.5$ as defined above. $T x$ is the variable for treatment groups with values of 0 or $1 ;\left(T_{2}-T_{1}\right)=1$ if $Y_{i j}$ is observed at time point 2, 0 otherwise; $\left(T_{3}-T_{1}\right)$ and $\left(T_{4}-T_{1}\right)$ are defined similarly with a value of 1 if $Y_{i j}$ is observed at time point 3 or 4 and a value of 0 otherwise.

## LATENT VARIABLE MODEL IN OBESITY DATA

The error term of outcomes $\boldsymbol{Y}_{i}$ follows a compound symmetry structure, with variance 1 and covariance 0.5 .

For missingness component, the assumption of missing not at random (MNAR) will be followed directly: that is, the missingness depends on the unobserved variables. Here let missingness procedure follow a logistic regression with an intercept and current unobserved response as the only covariate. Specifications are assumed as: intercept $\left(\alpha_{0}\right)=-3.0$ and log odds ratio for the current unobserved response $\left(\alpha_{1}\right)=1.5,1.0$ or 0.5 . That is:

$$
\log \frac{i j}{1{ }_{i j}}={ }_{0}+{ }_{1} y_{i j} .
$$

The summary measures for a parameter estimate include: a) mean bias: the mean difference of a sample estimate from the true parameter average over iterations of a simulation run; b) mean squared error: the mean of the squared deviation of a sample estimate from the true parameter averaged over iterations of a simulation run; and c) the coverage of nominal $95 \%$ confidence intervals, obtained by computing the percentage of iterations for which the corresponding nominal $95 \%$ confidence interval included the true parameter (Ten Have, Kunselman, Pulkstenis, \& Landis, 1998). Data are generated 1000 times under each scenario for the proposed model (latent variable model, LVM), a mixed model (MM) for all available data, and a mixed model that discards the missed observations, that is, a complete-case analysis (CC).

The simulation results are presented in Table 1. When missingness strongly depends on the unobserved outcomes ( $\alpha_{1}=1.5$ ), the time effects ( $T_{2}-T_{1}, T_{3}-T_{1}$, and $T_{4}-T_{1}$ ) are underestimated (negative bias) and coverage of $95 \%$ confidence interval is poor under the mixed model. For complete-case analysis, the betweensubject effect (Intercept and Tx) estimates and confidence interval coverage do not exhibit good properties, though the mixed model displays just the opposite, that is, it is accurate in the between-subject effect estimates but not in the withinsubject effect (time effect) estimates. For the proposed method, both within- and between-subject inference are accurate even under the strong dependence on the unobserved outcomes except for the effect of ( $T_{4}-T_{1}$ ), which is due to the small number of observations at $T_{4}$.

## QIN ET AL.

Table 1. Simulation results: mean bias and mean squared error (MSE) for the three models (latent variable model (LVM), mixed model (MM) and complete case analysis (CC)).

| Statistic | Variable | $\alpha_{1}=1.5$ |  |  | $\alpha_{1}=1.0$ |  |  | $\alpha_{1}=0.5$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | LVM | MM | CC | LVM | MM | CC | LVM | MM | CC |
| \% Bias | Intercept | -1.46 | -3.42 | -28.88 | -1.41 | -1.30 | -17.59 | -1.49 | -1.66 | -6.73 |
|  | TX | -11.19 | -15.23 | -33.13 | -5.50 | -8.09 | -17.18 | -1.12 | -1.94 | -4.73 |
|  | $T_{2}-T_{1}$ | -3.72 | -4.48 | -0.92 | -0.91 | -1.43 | -0.56 | 0.03 | 0.12 | -0.24 |
|  | $T_{3}-T_{1}$ | -8.48 | -11.49 | -5.88 | -3.74 | -4.84 | -3.70 | -0.15 | -1.06 | -1.44 |
|  | $T_{4}-T_{1}$ | -14.78 | -18.73 | -10.14 | -5.02 | -6.27 | -3.50 | 0.67 | 0.60 | 0.37 |
| \% Mean Squared Error | Intercept | 0.70 | 0.87 | 9.77 | 0.63 | 0.65 | 3.90 | 0.82 | 0.78 | 1.32 |
|  | TX | 2.40 | 3.39 | 13.42 | 1.22 | 1.71 | 4.55 | 1.20 | 1.15 | 1.77 |
|  | $T_{2}-T_{1}$ | 0.75 | 0.84 | 0.98 | 0.68 | 0.66 | 0.63 | 0.50 | 0.46 | 0.46 |
|  | $T_{3}-T_{1}$ | 1.12 | 1.82 | 1.41 | 0.63 | 0.77 | 0.68 | 0.50 | 0.52 | 0.44 |
|  | $T_{4}-T_{1}$ | 2.81 | 4.08 | 2.01 | 0.73 | 0.88 | 0.82 | 0.48 | 0.47 | 0.50 |
| Coverage of $95 \%$ CI | Intercept | 0.94 | 0.92 | 0.21 | 0.97 | 0.97 | 0.63 | 0.95 | 0.95 | 0.92 |
|  | TX | 0.81 | 0.73 | 0.42 | 0.93 | 0.90 | 0.78 | 0.96 | 0.95 | 0.95 |
|  | $T_{2}-T_{1}$ | 0.90 | 0.87 | 0.95 | 0.92 | 0.92 | 0.99 | 0.93 | 0.97 | 0.97 |
|  | $T_{3}-T_{1}$ | 0.86 | 0.73 | 0.91 | 0.93 | 0.89 | 0.97 | 0.96 | 0.96 | 0.99 |
|  | $T_{4}-T_{1}$ | 0.60 | 0.37 | 0.88 | 0.95 | 0.94 | 0.97 | 0.97 | 0.99 | 0.97 |

## Application to WGPW data

## Data description and model specifications

The proposed latent variable model is applied to an actual data set to illustrate its features and explore issues involved with its implementation. The sensitivity of inference to the model assumption and constraints in model formulation are also considered.

To illustrate the method, a subset of data from a study involving weight gain prevention in women (WGPW) is used. This trial was conducted in the Department of Psychiatry at the University of Pittsburgh Medical Center (Levine et al., 2007), and involved 25- to 45-year-old women at risk for weight gain and future obesity. The primary aim of the trial was to compare the relative efficacy of three approaches to weight gain prevention: a clinic-based group intervention, a mailed, correspondence intervention and an information-only control group. The measurements were taken at baseline, year 1, year 2 and year 3.

For the analysis, 190 women with complete baseline data are focused on and randomized into the clinic-based group and the control group. Women randomized to the clinic-based intervention group were required to attend 15

## LATENT VARIABLE MODEL IN OBESITY DATA

group meetings over a 24 -month period. These sessions were held biweekly for the first 2 months and bimonthly for the next 22 months. Biweekly sessions focused on self-monitoring of energy intake and expenditure, and behavioral strategies for making modest changes in dietary intake and activity level. During the 11 bimonthly clinic-based meetings, participants received lessons on cognitive change strategies, stimulus control techniques, problem solving, goal setting, stress and time management, and relapse prevention. Women belonging to the control group received booklets containing information about the benefits of weight maintenance, low-fat eating, and regular physical activity.

About $70 \%$ of the women did not complete their scheduled assessments (Table 2). It was suspected that this was in part due to reasons related to their weight outcomes. Among women randomized to the intervention group in which treatment was provided over a 2 -year period, $20 \%$ missed the weight assessments at year $1 ; 27 \%$ at year 2; and $63 \%$ at year 3 of the follow-up. For subjects in the control group, $19 \%, 22 \%$ and $66 \%$ missed the weight assessments at year 1,2 and 3. The plot in Figure 1 indicates that at year 2, which is the end of the treatment, the intervention group exhibits a lower BMI than the control group. However the plot of Figure 2 indicates that at year 2, the probability of missingness in the intervention group is a little higher than that of the control group. If only the observed data are used, the conclusion that the intervention group has a smaller BMI at the end of the treatment (year 2) may be reached. But if the missing data mechanism is considered, what will the data tell us?

Table 2. Distribution of the missingness patterns for WPGW data.

| Pattern | Baseline | Year 1 | Year 2 | Year 3 | Frequency (\%) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $\bullet$ | $\bullet$ | $\bullet$ | $\bullet$ | $56(29.5)$ |
| 2 | $\bullet$ | $\bullet$ | $\bullet$ | $\times$ | $77(40.5)$ |
| 3 | $\bullet$ | $\bullet$ | $\times$ | $\bullet$ | $06(03.2)$ |
| 4 | $\bullet$ | $\times$ | $\bullet$ | $\bullet$ | $01(00.5)$ |
| 5 | $\bullet$ | $\bullet$ | $\times$ | $\times$ | $14(07.4)$ |
| 6 | $\bullet$ | $\times$ | $\bullet$ | $\times$ | $10(05.3)$ |
| 7 | $\bullet$ | $\times$ | $\times$ | $\bullet$ | $05(02.6)$ |
| 8 | $\bullet$ | $\times$ | $\times$ | $\times$ | $21(11.1)$ |

Note: $\cdot$ : observed; $\times$ : missingness

## QIN ET AL.



Figure 1. Observed BMI mean (SE) across years for each treatment group


Figure 2. Probability of missingness across years for each treatment group

## LATENT VARIABLE MODEL IN OBESITY DATA

Let $Y_{i j}$ denote the BMI measurement on the $i^{\text {th }}$ patient at the $j^{\text {th }}$ year in the trial, $j=0,1,2$ and 3 . Six explanatory variables are included as main effects in the analysis: treatment $(T x$, intervention $=1$ and control $=0$ ), years in the trial (year), patient age when enrolled (age), dietary restraint (S3FS1, range from 0-21), disinhibition (S3FS2, range from 0-16), and perceived hunger (S3FS3, range from $0-14)$. Among them, dietary restraint, disinhibition and perceived hunger belong to Stunkard Three-Factor Eating Questionnaire, and they are included in the model as time-variant predictors, as is year. The linear random effects model for BMI is specified as

$$
\begin{aligned}
Y_{i j}=\beta_{0}+\beta_{1} \text { year }_{j}+ & \beta_{2} \text { year }_{j} \\
& \times T x_{i}+\beta_{3} \text { age }_{i}+\beta_{4} S 3 F S 1_{i j}+\beta_{5} S 3 F S 2_{i j}+\beta_{6} S 3 F S 3_{i j}+W_{1 i}\left(\text { year }_{j}\right),
\end{aligned}
$$

where $W_{1 i}\left(\right.$ year $\left._{j}\right)$ is the random effect.
Similarly the missingness procedure is modelled with the logistic regression with random effect, $W_{2 i}\left(\right.$ year $\left._{j}\right)$. Let $\varphi_{i j}=\operatorname{Pr}\left(Y_{i j}\right.$ is missing $)$,

$$
\log \frac{\varphi_{i j}}{1-\varphi_{i j}}=\alpha_{0}+\alpha_{1} T x_{i}+W_{2 i}\left(\text { year }_{j}\right) .
$$

To choose the exact forms of $W_{1 i}$ and $W_{2 i}$, Akaike's information criterion (AIC) (Akaike, 1981) and the Bayesian information criterion (BIC) (Schwartz, 1978) are used. The results are given in Table 3: because Model VII emerges with the smallest values of AIC and BIC, it is selected over the others, and also demonstrates the full complexity of $\left(W_{1 i}, W_{2 i}\right)^{T}$ given under the Latent Variable Model, earlier. In Model VII, $W_{1 i}\left(\right.$ year $\left._{j}\right)=U_{1 i}+U_{2 i}$ year $_{j}$. So $W_{1 i}\left(\right.$ year $\left._{j}\right)$ includes random effects for intercept and slope over time, where $U_{i}=\left(U_{1 i}, U_{2 i}\right)^{T^{i i d}} \stackrel{\sim}{\sim} N_{2}(0, \Sigma)$ and variance-covariance structure $\Sigma=\left(\begin{array}{cc}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22}\end{array}\right)$.
This structure of random effects allow that each subject has her own baseline BMI value and time trend of BMIs over years in the trial. And the random effects in the models of missingness procedure are chosen as $W_{2 i}\left(\right.$ year $\left._{j}\right)=r_{1} U_{1 i}+r_{2} U_{2 i} \boldsymbol{y e a r}_{j}+r_{3}\left(U_{1 i}+U_{2 i} \boldsymbol{y} \boldsymbol{e a r} \boldsymbol{r}_{j}\right)$, where $U_{1 i}$ and $U_{2 i}$ are defined as before. In the following application results and interpretations, inferences will be based on these chosen random effect structures.

## QIN ET AL.

Table 3. Descriptive of model fit for different random effect structures for WGPW data.

| Model | $W_{1 i}$ | $W_{2 i}$ | $-2 \log$ <br> likelihood | AIC | BIC |
| :---: | :---: | :---: | :---: | :---: | :---: |
| I | 0 | 0 | 2904.4 | 2936.4 | 3018.0 |
| II | $U_{1 i}$ | 0 | 2656.6 | 2657.6 | 2709.6 |
| III | $U_{1 i}$ | $\gamma_{1} U_{1 i}$ | 2625.3 | 2657.3 | 2709.3 |
| IV | $U_{1 i}+U_{2 i}{ }^{\text {y }}$ ear ${ }_{j}$ | 0 | 2595.9 | 2629.7 | 2679.8 |
| V | $U_{1 i}+U_{2 i}$ year $_{j}$ | $\gamma_{1} U_{1 i}$ | 2595.7 | 2627.7 | 2679.6 |
| VI | $U_{1 i}+U_{2 i}$ year $^{1}$ | $\gamma_{1} U_{1 i}+Y_{2} U_{2 i}$ | 2614.6 | 2656.6 | 2698.6 |
| VII | $U_{1 i}+U_{2 i}$ year $_{j}$ | $\gamma_{1} U_{1,}+\gamma_{2} U_{2 i}+\gamma_{3} W_{1 i}$ | 2534.7 | 2566.7 | 2618.6 |

## Model interpretation

Table 4 details the model estimates of treatment, time, age, dietary restraint, disinhibition and perceived hunger effects on the BMIs. In Table 5, the estimates in the missingness component of the joint model are compared to the analogous estimates from a random effects model, which ignores the BMI outcome, to address the effects of treatment on the missingness status. In both tables, the estimates for variance-covariance structure $\Sigma$ under models for longitudinal responses and missing data procedure, separately and jointly, are discussed.

As shown in Table 4, the mixed model, under the assumption of missing at random, and the proposed joint model yield similar inference for significant effect of year, whereas the complete case analysis under the assumption of missing completely at random does not show any significant time effect. In the proposed model, age effect intends to be significant ( $p$ value $=0.074$ ), although in the other two models, there is no such intention. Under all three models, dietary restraint and disinhibition show strong effects ( $p$ values < .0001). In Table 5, the association parameter in the proposed method, $\gamma_{3}$, is negative and significantly different from zero. It provides a strong evidence of association between the two sub-models of the proposed method, and indicates that the slope of observed BMI values is negatively associated with the missingness status, because of $\lambda_{2}=\gamma_{2}+\gamma_{3}<0$ with $\gamma_{2}=6.779$ and $\gamma_{3}=-26.94$ (Table 5). This may result from patients with larger BMI values having lower probabilities of dropping out, leaving their relatively larger BMI values in the trial.

## Comparisons with simulation results

The relationship between the proposed method and the mixed model in the application to the WGPW data is now checked, and compared with the patterns observed in the simulations. Table 4 reveals that the proposed method and the

## LATENT VARIABLE MODEL IN OBESITY DATA

mixed model yield similar between-subject effect estimates (age, dietary restraint, disinhibition and perceived hunger), but are different in the within-subject inference (year, and year $\times$ treatment). As in the simulation results, the mixed model gives accurate inference in between-subject effect estimates but not in within-subject effect estimates. This congruence in the between-subject effect estimates, and difference in the within-subject effect estimates, provides evidence that the proposed method is a good choice for the WGPW data.

Table 4. Parameter estimates, estimated standard errors and $p$-values for modeling the outcomes, BMI.

| Variable | CC analysis |  |  | Mixed Model |  |  | Latent Variable model |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Estimate | SE | $p$-value | Estimate | SE | $p$-value | Estimate | SE | $p$-value |
| Intercept | 24.3000 | 2.1350 | <0.0001 | 22.8200 | 1.1410 | <0.0001 | 22.8400 | 1.1590 | <0.0001 |
| Year | 0.0770 | 0.1400 | 0.5850 | 0.2030 | 0.0920 | 0.0290 | 0.1520 | 0.0750 | 0.0440 |
| Year $\times$ Treatment | -0.0240 | 0.1920 | 0.9030 | -0.1630 | 0.1280 | 0.2020 | -0.1240 | 0.1030 | 0.2300 |
| Age | 0.0370 | 0.0580 | 0.5250 | 0.0470 | 0.0300 | 0.1180 | 0.0550 | 0.0300 | 0.0740 |
| Dietary Restraint | -0.2090 | 0.0370 | <0.0001 | -0.1200 | 0.0220 | <0.0001 | -0.1320 | 0.0240 | <0.0001 |
| Disinhibition | 0.1650 | 0.0520 | 0.0030 | 0.1800 | 0.0320 | <0.0001 | 0.1780 | 0.0330 | <0.0001 |
| Perceived Hunger | -0.0270 | 0.0450 | 0.5460 | -0.0180 | 0.0300 | 0.5610 | -0.0280 | 0.0320 | 0.3940 |
| $\Sigma_{11}$ | 1.9480 | 0.2300 | <0.0001 | 2.0180 | 0.1300 | <0.0001 | 2.0490 | 0.1250 | <0.0001 |
| $\Sigma_{12}$ | -0.0090 | 0.1120 | 0.9370 | -0.0230 | 0.0750 | 0.7620 | 0.0040 | 0.0090 | 0.6720 |
| $\Sigma_{22}$ | 0.4940 | 0.0940 | <0.0001 | 0.5280 | 0.0690 | <0.0001 | 0.0620 | 0.0410 | 0.1300 |
| $\sigma_{\varepsilon}{ }^{2}$ | 0.9400 | 0.0720 | <0.0001 | 0.8620 | 0.0500 | <0.0001 | 1.0680 | 0.0460 | <0.0001 |

Table 5. Parameter estimates, estimated standard errors and $p$-values for modeling the missingness status, $R$.

|  | Separate Analysis |  |  |  | Latent Variable Model |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Variable | Estimate | SE | $\boldsymbol{p}$-value |  | Estimate | SE | $\boldsymbol{p}$-value |
| Intercept | -1.0620 | 0.1350 | $<0.0001$ |  | -2.5910 | 0.3580 | $<0.0001$ |
| Treatment | 0.0360 | 0.1800 | 0.8410 |  | 0.0700 | 0.3320 | 0.8320 |
| $V_{1}$ | NA | NA | NA |  | 27.0600 | 17.9600 | 0.2400 |
| $Y_{2}$ | NA | NA | NA |  | 6.7790 | 5.7500 | 0.1360 |
| $Y_{3}$ | NA | NA | NA |  | -26.9400 | 17.9700 | $<0.0001$ |

## Conclusion

A latent variable model was proposed to fit longitudinal data with informative intermittent missingness. The main idea is to jointly model the longitudinal process and missing data process via a latent zero-mean bivariate Gaussian

## QIN ET AL.

process on $\left(W_{1}(t), W_{2}(t)\right)^{T}$, with correlation between $W_{1}(t)$ and $W_{2}(t)$, inducing stochastic dependence between the longitudinal and missing data processes. An advantage of this method, compared with other existing methods for informative missing data problems, is its easy implementation. The models in this method can be easily fit after providing the likelihood functions. Thus it avoids the complexity of EM algorithm programming, facilitating use of this proposed method in practice. The specifications and selections of $W_{1}(t)$ and $W_{2}(t)$ can be implemented via AIC and BIC, and the method enables direct comparisons of different specifications.

In the proposed method, the latent variables are also used to induce conditional independence between the responses (both observed and missing) and missingness status, so that the standard likelihood techniques can be used to derive the estimates. This is a strong assumption and it cannot be tested with the available data. For this type of assumption, a sensitivity analysis is the way to investigate the model fit and departure of the assumption. Such an analysis has been attempted by comparing the proposed method with other alternative models in the true data and in simulations.

The proposed method is developed from the joint model proposed by Henderson et al. (2000) for longitudinal and survival processes. In the future, this method should be considered for extension into other applications, through different link functions (e.g. binary or ordinal data) or random effect structures other than zero-mean bivariate Gaussian distribution.

## References

Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics, 16(1), 3-14. doi: 10.1016/0304-4076(81)90071-3

Henderson, R., Diggle, P. J. \& Dobson, A. (2000). Joint modeling of longitudinal measurements and event time Data. Biostatistic,s 1(4), 465-480. doi: 10.1093/biostatistics/1.4.465

Levine, M. D., Klem, M. L., Kalarchian, M. A., Wing, R. R., Weissfeld, L., Qin. L. \& Marcus. M. D. (2007). Weight gain prevention among women. Obesity, 15(5), 1267-1277. doi: 10.1038/oby.2007.148

Lin, H., McCulloch, C. E. \& Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics, 60(2), 295-305. doi: 10.1111/j.0006-341X.2004.00173.x

## LATENT VARIABLE MODEL IN OBESITY DATA

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association, 88(421), 125-134. doi: 10.1080/01621459.1993.10594302

Qin, L., Weissfeld, L. A., Shen, C. \& Levine, M. D. (2009). A two-latentclass model for smoking cessation data with informative dropouts.
Communications in Statistics - Theory and Methods, 38(15), 2604-19. doi: 10.1080/03610920802585849

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics, 59(4), 829-836. doi: 10.1111/j.0006-341X.2003.00097.x

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581590. doi: 10.1093/biomet/63.3.581

Schwartz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6(2), 461-464. doi: 10.1214\%2Faos\%2F1176344136

Ten Have, T. R., Kunselman, A. R., Pulkstenis, E. P. \& Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. Biometrics, 54(1), 367-383. doi: 10.2307/2534023

# E-Bayesian Estimation of the Parameter of the Logarithmic Series Distribution 

Parviz Nasiri<br>University of Payame Noor<br>Tehran, Iran

Hassan Esfandyarifar<br>University of Payame Noor<br>Tehran, Iran

E-Bayesian estimation is introduced to estimate the parameter of logarithmic series distribution. In addition, E-Bayesian, Bayesian and maximum likelihood estimation with through applying mean squared error.

Keywords: Logarithmic series distribution, E-Bayesian estimation, Bayesian estimation, maximum likelihood estimation, means squared error

## Introduction

The logarithmic series distribution (LSD) is obtained by expanding the logarithmic function $-\log (1-\theta)$ as a power series in. Alternatively, it can also be derived as a limiting case of zero-truncated negative binomial distribution as $k$ decreases to zero. In either case, the logarithmic series distribution is a very useful distribution on the positive integers (Nasiri, 2011). Estimation is an important topic in statistical inference. Bayesian approach is an important approach in the estimation of parameter. A suitable prior distribution plays an effective role in reducing error in the estimation. Therefore, the more the prior information is obtained, the more it affects the posterior.

Lindley and Smith (1972) argued hierarchical prior. E-Bayesian estimation is another method introduced by Han and Ding (2004). Han (2005) applied EBayesian estimation for forecast of security investment. He also (2006, 2007) presented hierarchical Bayesian estimation for computing as well as E-Bayesian estimation for transition probability. In this study, maximum likelihood, Bayesian, and E-Bayesian estimations of the parameter of logarithmic series distribution are discussed in detail. This paper considers the maximum likelihood estimation of $\theta$,

## ESTIMATION PARAMETER OF LSD

the Bayesian estimation of $\theta$, and the E-Bayesian estimation of $\theta$; by use of a simulation, all estimations will be compared by MSE.

## Maximum Likelihood Estimation of $\boldsymbol{\theta}$

Let $\mathrm{f}(x)$ be the density of the logarithmic series distribution given by

$$
\begin{equation*}
\mathrm{f}(x)=\frac{-1}{\log (1-\theta)} \cdot \frac{\theta^{x}}{x}, \quad x=1,2,3, \ldots, \quad 0<\theta<1 \tag{1}
\end{equation*}
$$

The maximum likelihood estimation of $\theta$ in the above distribution is derived by i.i.d observations $x_{1}, x_{2}, \ldots x_{n}$. Hence, the likelihood function is given by

$$
\begin{equation*}
1(\theta)=\left(\frac{1}{-\log (1-\theta)}\right)^{n} \cdot \frac{\theta^{n \bar{x}}}{\prod_{i=1}^{n} x_{i}}=(-\log (1-\theta))^{-n} \frac{\theta^{n \bar{x}}}{\prod_{i=1}^{n} x_{i}} \tag{2}
\end{equation*}
$$

Similarly, the logarithm of the likelihood function is given by

$$
\begin{equation*}
\log 1(\theta)=n \bar{x} \log \theta-n \log (-\log (1-\theta))-\sum_{i=1}^{n} \log x_{i} \tag{3}
\end{equation*}
$$

There are two ways to estimate $\theta$. The first is to apply the "optimum" command in R software, and the second is to take the first order derivative of $\log \mathrm{l}(\theta)$ over $\theta$ and set it equal to zero, as in the following:

$$
\begin{aligned}
& \frac{d \log 1(\theta)}{d \theta}=0 \\
& \frac{\bar{x}}{\theta}+\frac{1}{(1-\theta)(\log (1-\theta))}=0 \\
& (1-\theta)^{1-\theta}=\mathrm{e}^{\frac{-\theta}{\bar{x}}}, \quad 0<\theta<1
\end{aligned}
$$

Assume $\mathrm{g}(\theta)=(1-\theta)^{1-\theta}-\mathrm{e}^{\frac{-\theta}{\bar{x}}}$. This equation can be solved via the NewtonRaphson method:

## NASIRI \& ESFANDYARIFAR

$$
\begin{equation*}
\theta_{n}=\theta_{n-1}-\frac{\mathrm{g}\left(\theta_{n-1}\right)}{\dot{\mathrm{g}}\left(\theta_{n-1}\right)} \tag{4}
\end{equation*}
$$

where

$$
\dot{\mathrm{g}}\left(\theta_{n-1}\right)=(1-\theta)^{1-\theta}[-\log (1-\theta)-1]+\frac{\mathrm{e}^{\frac{-\theta}{\bar{x}}}}{\bar{x}},
$$

and Fixed-Point method as

$$
\begin{equation*}
\theta_{n}=\mathrm{h}\left(\theta_{n-1}\right) \tag{5}
\end{equation*}
$$

such that

$$
\mathrm{h}(\theta)=\bar{x}(\theta-1) \log (1-\theta) .
$$

Equations (4) and (5) were solved using the MATLAB software. The "optimum" command was used in R software. There is additional discussion regarding the MLE logarithmic series in Bohning (1983).

## Bayesian Estimation of $\boldsymbol{\theta}$

Let $\pi(\theta)$ be prior density of $\theta$ that has beta distribution:

$$
\begin{gather*}
\pi(\theta)=\frac{1}{\mathrm{~B}(a, b)} \theta^{a-1}(1-\theta)^{b-1}  \tag{6}\\
\mathrm{~B}(a, b)=\frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}, \quad \Gamma(a)=\int_{0}^{\infty} t^{a-1} \mathrm{e}^{-t} d t
\end{gather*}
$$

By using i.i.d. observations $x_{1}, x_{2}, \ldots x_{n}$, the posterior distribution of $\theta$ was calculated as in the following:

$$
\begin{equation*}
\pi(\theta \mid \underline{x})=\frac{1(\theta) \pi(\theta)}{\int_{0}^{1} 1(\theta) \pi(\theta) d \theta}=\frac{(-\log (1-\theta))^{-n} n^{n \pi x+a-1}(1-\theta)^{b-1}}{\int_{0}^{1}(-\log (1-\theta))^{-n} \theta^{n \bar{x}+a-1}(1-\theta)^{b-1} d \theta} \tag{7}
\end{equation*}
$$

## ESTIMATION PARAMETER OF LSD

where $1(\theta)$ is the likelihood function introduced in (2).
Note that $(-\log (1-\theta))^{-n}$ can be expanded as $\theta^{-n} \sum_{m=0}^{\infty} \rho_{m}(-n) \theta^{m}$, where $\rho_{0}(-n)=1, \rho_{m}(-n)=n \psi_{m-1}(m-n-1)$ for $m \geq 1$, and the coefficients $\psi_{m}($.$) , are$ Sterling polynomials given by Castellares and Lemonte (2014).

Consider

$$
(-\log (1-\theta))^{-n} \theta^{n \bar{x}+a-1}(1-\theta)^{b-1}=\sum_{m=0}^{\infty} \rho_{m}(-n) \theta^{m+n(\bar{x}-1)+a-1}(1-\theta)^{b-1}
$$

and suppose

$$
\begin{aligned}
\mathrm{A}(a) & =\sum_{m=0}^{\infty} \rho_{m}(-n) \int_{0}^{1} \theta^{m+n(\bar{x}-1)+a-1}(1-\theta)^{b-1} d \theta \\
& =\sum_{m=0}^{\infty} \rho_{m}(-n) \mathrm{B}(m+n(\bar{x}-1)+a, b)
\end{aligned}
$$

Then

$$
\begin{equation*}
\pi(\theta \mid \underline{x})=\frac{\sum_{m=0}^{\infty} \rho_{m}(-n) \theta^{m+n(\underline{x}-1)+a-1}(1-\theta)^{b-1}}{\mathrm{~A}(a)} \tag{8}
\end{equation*}
$$

The Bayesian estimation of $\theta$ under loss function $\mathrm{l}(\theta, d)=(d-\theta)^{2}$ is $\mathrm{E}(\theta \mid \underline{x})$,

$$
\begin{equation*}
\mathrm{E}(\theta \mid \underline{x})=\frac{\mathrm{A}(a+1)}{\mathrm{A}(a)} \tag{9}
\end{equation*}
$$

$\mathrm{E}(\theta \mid \underline{x})$ is computed by numerical methods using R software.

## E-Bayesian Estimation

Let the prior distribution of $\theta$ be given as:

$$
\begin{equation*}
\pi(\theta \mid a, b)=\frac{1}{\mathrm{~B}(a, b)} \theta^{a-1}(1-\theta)^{b-1}, \quad 0<\theta<1 \tag{10}
\end{equation*}
$$

## NASIRI \& ESFANDYARIFAR

where $a$ and $b$ are super parameters. According to Han (1997) $a$ and $b$ should be selected to guarantee $\pi(\theta \mid a, b)$ is a decreasing function of $\theta$. Therefore, we applied one order derivative of $\pi(\theta \mid a, b)$ over $\theta$ to obtain

$$
\frac{d \pi(\theta \mid a, b)}{d \theta}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \Gamma(\beta)} \theta^{a-2}(1-\theta)^{b-2}[(a-1)(1-\theta)-(b-1) \theta]
$$

Because $a>0, b>0$, and $0<\theta<1$, then $0<a \leq 1, b>1$ result in $\frac{d \pi(\theta \mid a, b)}{d \theta}<0$. Thus, $\pi(\theta \mid a, b)$ is a decreasing function of $\theta$ given $0<a \leq 1, b\rangle 1$.

As $b$ grows larger, the tail of the beta density function grows thinner. However, as far as the robustness of Bayesian estimation is concerned (Berger, 1985), the thinner-tailed prior distribution often leads to the worse robustness of the Bayesian estimate. Accordingly, $b$ should not be too big; it is better to be selected below the given upper bound $c(c>1)$ (see Han \& Ding, 2004). All in all, the super parameters $a$ and $b$ were selected to be in the ranges $0<a \leq 1$ and $1<b \leq c$.

Let $a=1$ and $b$ have density function given by the following:

$$
\begin{equation*}
\mathrm{g}(b)=\frac{1}{c-1}, \quad 1<b \leq c \tag{11}
\end{equation*}
$$

Hence, the prior distribution is given by

$$
\begin{gather*}
\pi(\theta \mid a=1, b)=\pi(\theta \mid b)=b(1-\theta)^{b-1}, \quad 0<\theta<1 \\
\pi(\theta, b)=\pi(\theta \mid b) \mathrm{g}(b)=\frac{b}{c-1}(1-\theta)^{b-1}, \quad 0<\theta<1, \quad 1<b \leq c \tag{12}
\end{gather*}
$$

If the prior distribution is named $\pi_{E}(\theta)$, it is calculated as

$$
\begin{equation*}
\pi_{E}(\theta)=\frac{1}{c-1} \int_{1}^{c} b(1-\theta)^{b-1} d b \tag{13}
\end{equation*}
$$

then $\pi(\theta \mid \underline{x})=\frac{1(\theta) \pi_{E}(\theta)}{\int_{0}^{1} 1(\theta) \pi_{E}(\theta) d \theta}$.
Consider

## ESTIMATION PARAMETER OF LSD

$$
\sum_{m=0}^{\infty} \rho_{m}(-n) \theta^{m+n(\bar{x}-1)+a-1} \int_{1}^{c} b(1-\theta)^{b-1} d b
$$

and suppose $\mathrm{B}(a)=\sum_{m=0}^{\infty}\left(\rho_{m}(-n) \int_{1}^{c} \mathrm{~B}(m+n(\bar{x}-1)+a, b) d b\right)$. Then

$$
\begin{equation*}
\pi_{E}(\theta \mid \underline{x})=\frac{\sum_{m=0}^{\infty} \rho_{m}(-n) \theta^{m+n(\bar{x}-1)+a-1} \int_{1}^{c} b(1-\theta)^{b-1} d b}{\mathrm{~B}(a)} \tag{14}
\end{equation*}
$$

is the posterior distribution of $\theta$ and, under loss function $1(\theta, d)=(d-\theta)^{2}$, the Expected Bayesian (E-Bayesian) estimation is given as

$$
\begin{equation*}
\mathrm{E}(\theta \mid \underline{x})=\frac{\mathrm{B}(a+1)}{\mathrm{B}(a)} \tag{15}
\end{equation*}
$$

$\mathrm{E}(\theta \mid \underline{x})$ is computed by numerical methods using R software.

## Simulation

The simulation logarithmic series distribution is applied and the MSE among these three estimations are compared. The sample sizes chosen are $n=10$ (10)50, 100 from the logarithmic series distribution and then the above sampling is repeated 1000 times. In all the tables below, $a=1, b=1$.

Table 1. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.1, \theta=0.2$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\boldsymbol{\theta}}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\boldsymbol{\theta}}$ | MSE |
| 10 | 0.1754406 | 0.02280529 | 0.268888 | 0.015603929 | 0.2670986 | 0.015603929 |
| 20 | 0.1848736 | 0.01274284 | 0.235789 | 0.009822050 | 0.2349044 | 0.009710344 |
| 30 | 0.1888380 | 0.00885097 | 0.223687 | 0.007274929 | 0.2231012 | 0.007219458 |
| 40 | 0.1915576 | 0.00665076 | 0.217965 | 0.005719935 | 0.2175264 | 0.005686960 |
| 50 | 0.1951806 | 0.00536783 | 0.216286 | 0.004821819 | 0.2159331 | 0.004798442 |
| 100 | 0.1968546 | 0.00251246 | 0.207595 | 0.002366069 | 0.2074195 | 0.002360269 |

## NASIRI \& ESFANDYARIFAR

Table 2. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.5, \theta=0.2$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.1754406 | 0.02280529 | 0.268888 | 0.015603929 | 0.2600561 | 0.013984288 |
| 20 | 0.1848736 | 0.01274284 | 0.235789 | 0.009822050 | 0.2313332 | 0.009281856 |
| 30 | 0.1888380 | 0.00885097 | 0.223687 | 0.007274929 | 0.2207135 | 0.007003590 |
| 40 | 0.1915576 | 0.00665076 | 0.217965 | 0.005719935 | 0.2157289 | 0.005557730 |
| 50 | 0.1951806 | 0.00536783 | 0.216286 | 0.004821819 | 0.2144824 | 0.004706188 |
| 100 | 0.1968546 | 0.00251246 | 0.207595 | 0.002366069 | 0.2066942 | 0.002337212 |

Table 3. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=2, \theta=0.2$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| n | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.1754406 | 0.02280529 | 0.268888 | 0.015603929 | 0.2517434 | 0.012635242 |
| 20 | 0.1848736 | 0.01274284 | 0.235789 | 0.009822050 | 0.2269419 | 0.008802810 |
| 30 | 0.1888380 | 0.00885097 | 0.223687 | 0.007274929 | 0.2177276 | 0.006756307 |
| 40 | 0.1915576 | 0.00665076 | 0.217965 | 0.005719935 | 0.2134603 | 0.005407923 |
| 50 | 0.1951806 | 0.00536783 | 0.216286 | 0.004821819 | 0.2126412 | 0.004597898 |
| 100 | 0.1968546 | 0.00251246 | 0.207595 | 0.002366069 | 0.2057618 | 0.002309824 |

Table 4. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=3, \theta=0.2$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\boldsymbol{\theta}}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\boldsymbol{\theta}}$ | MSE |
| 10 | 0.1754406 | 0.02280529 | 0.268888 | 0.015603929 | 0.2370938 | 0.010660721 |
| 20 | 0.1848736 | 0.01274284 | 0.235789 | 0.009822050 | 0.2187640 | 0.008047442 |
| 30 | 0.1888380 | 0.00885097 | 0.223687 | 0.007274929 | 0.2120299 | 0.006352388 |
| 40 | 0.1915576 | 0.00665076 | 0.217965 | 0.005719935 | 0.2090714 | 0.005159055 |
| 50 | 0.1951806 | 0.00536783 | 0.216286 | 0.004821819 | 0.2090494 | 0.004414182 |
| 100 | 0.1968546 | 0.00251246 | 0.207595 | 0.002366069 | 0.2039068 | 0.002262655 |

Table 5. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=5, \theta=0.2$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| n | $\hat{\boldsymbol{\theta}}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.1754406 | 0.02280529 | 0.268888 | 0.015603929 | 0.2145302 | 0.008617966 |
| 20 | 0.1848736 | 0.01274284 | 0.235789 | 0.009822050 | 0.2051204 | 0.007182372 |
| 30 | 0.1888380 | 0.00885097 | 0.223687 | 0.007274929 | 0.2021512 | 0.005863618 |
| 40 | 0.1915576 | 0.00665076 | 0.217965 | 0.005719935 | 0.2012877 | 0.004851086 |
| 50 | 0.1951806 | 0.00536783 | 0.216286 | 0.004821819 | 0.2025892 | 0.004176082 |
| 100 | 0.1968546 | 0.00251246 | 0.207595 | 0.002366069 | 0.2004542 | 0.002201122 |

## ESTIMATION PARAMETER OF LSD



Figure 1. MSE of MLE, Bayesian, and E-Bayesian estimation for $\theta=0.2$

According to Tables $1-5$ and Figure 1 below, if $\theta$ is close to zero, then the EBayesian estimator will be better than the others. Furthermore, the E-Bayesian estimator for big $c$ is better than for that of small $c$.

Table 6. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.1, \theta=0.5$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.4493031 | 0.036068680 | 0.4727567 | 0.020631200 | 0.4703712 | 0.020689950 |
| 20 | 0.4774472 | 0.015412300 | 0.4856296 | 0.011751570 | 0.4842737 | 0.011769950 |
| 30 | 0.4822146 | 0.011401902 | 0.4872720 | 0.009491975 | 0.4863317 | 0.009503848 |
| 40 | 0.4862952 | 0.007998191 | 0.4897325 | 0.006975903 | 0.4890090 | 0.006984277 |
| 50 | 0.4899989 | 0.006229734 | 0.4925262 | 0.005593061 | 0.4919384 | 0.005597817 |
| 100 | 0.4917995 | 0.003252503 | 0.4929799 | 0.003078263 | 0.4926775 | 0.003081360 |

Table 7. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.5, \theta=0.5$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.4493031 | 0.036068680 | 0.4727567 | 0.020631200 | 0.4611687 | 0.021062690 |
| 20 | 0.4774472 | 0.015412300 | 0.4856296 | 0.011751570 | 0.4789428 | 0.011893800 |
| 30 | 0.4822146 | 0.011401902 | 0.4872720 | 0.009491975 | 0.4826069 | 0.009577576 |
| 40 | 0.4862952 | 0.007998191 | 0.4897325 | 0.006975903 | 0.4861318 | 0.007033491 |
| 50 | 0.4899989 | 0.006229734 | 0.4925262 | 0.005593061 | 0.4895955 | 0.005627388 |
| 100 | 0.4917995 | 0.003252503 | 0.4929799 | 0.003078263 | 0.4914670 | 0.003097037 |

## NASIRI \& ESFANDYARIFAR

Table 8. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=2, \theta=0.5$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.4493031 | 0.036068680 | 0.4727567 | 0.020631200 | 0.4506494 | 0.021772010 |
| 20 | 0.4774472 | 0.015412300 | 0.4856296 | 0.011751570 | 0.4726720 | 0.012145060 |
| 30 | 0.4822146 | 0.011401902 | 0.4872720 | 0.009491975 | 0.4781747 | 0.009721149 |
| 40 | 0.4862952 | 0.007998191 | 0.4897325 | 0.006975903 | 0.4826866 | 0.007126098 |
| 50 | 0.4899989 | 0.006229734 | 0.4925262 | 0.005593061 | 0.4867802 | 0.005685544 |
| 100 | 0.4917995 | 0.003252503 | 0.4929799 | 0.003078263 | 0.4900001 | 0.003122489 |

Table 9. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=3, \theta=0.5$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n$ | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.4493031 | 0.036068680 | 0.4727567 | 0.020631200 | 0.4329851 | 0.023664110 |
| 20 | 0.4774472 | 0.015412300 | 0.4856296 | 0.011751570 | 0.4617896 | 0.012865210 |
| 30 | 0.4822146 | 0.011401902 | 0.4872720 | 0.009491975 | 0.4703766 | 0.010129876 |
| 40 | 0.4862952 | 0.007998191 | 0.4897325 | 0.006975903 | 0.4765777 | 0.007386216 |
| 50 | 0.4899989 | 0.006229734 | 0.4925262 | 0.005593061 | 0.4817677 | 0.005854256 |
| 100 | 0.4917995 | 0.003252503 | 0.4929799 | 0.003078263 | 0.4873595 | 0.003187286 |

Table 10. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=5, \theta=0.5$

| n | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.4493031 | 0.036068680 | 0.4727567 | 0.020631200 | 0.4081765 | 0.027960910 |
| 20 | 0.4774472 | 0.015412300 | 0.4856296 | 0.011751570 | 0.4460818 | 0.014636860 |
| 30 | 0.4822146 | 0.011401902 | 0.4872720 | 0.009491975 | 0.4590000 | 0.011149126 |
| 40 | 0.4862952 | 0.007998191 | 0.4897325 | 0.006975903 | 0.4676159 | 0.008033505 |
| 50 | 0.4899989 | 0.006229734 | 0.4925262 | 0.005593061 | 0.4744028 | 0.006284713 |
| 100 | 0.4917995 | 0.003252503 | 0.4929799 | 0.003078263 | 0.4834468 | 0.003338421 |

## ESTIMATION PARAMETER OF LSD



Figure 2. MSE of MLE, Bayesian, and E-Bayesian estimation for $\theta=0.5$

According to Tables 6-10 and Figure 2 above, if $\theta$ is equal to 0.5 , then the Bayes estimator will be better than the others.

Table 11. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.1, \theta=0.8$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| n | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.7565883 | 0.016160921 | 0.7342053 | 0.016276950 | 0.7320425 | 0.016636190 |
| 20 | 0.7771566 | 0.006733548 | 0.7635732 | 0.007190623 | 0.7623963 | 0.007302307 |
| 30 | 0.7853821 | 0.004234158 | 0.7757980 | 0.004486893 | 0.7749946 | 0.004538776 |
| 40 | 0.7908123 | 0.003149407 | 0.7834313 | 0.003279836 | 0.7828241 | 0.003307885 |
| 50 | 0.7891018 | 0.002431975 | 0.7831037 | 0.002570948 | 0.7826102 | 0.002592876 |
| 100 | 0.7962759 | 0.000991319 | 0.7931716 | 0.001020993 | 0.7929747 | 0.001032422 |

Table 12. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=1.5, \theta=0.8$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.7565883 | 0.016160920 | 0.7342053 | 0.016276950 | 0.7240587 | 0.018094940 |
| 20 | 0.7771566 | 0.006733548 | 0.7635732 | 0.007190623 | 0.7580292 | 0.007758596 |
| 30 | 0.7853821 | 0.004234158 | 0.7757980 | 0.004486893 | 0.7720108 | 0.004751512 |
| 40 | 0.7908123 | 0.003149407 | 0.7834313 | 0.003279836 | 0.7805691 | 0.003423867 |
| 50 | 0.7891018 | 0.002431975 | 0.7831037 | 0.002570948 | 0.7807726 | 0.002682152 |
| 100 | 0.7962759 | 0.000991319 | 0.7931716 | 0.001020993 | 0.7919732 | 0.001041411 |

## NASIRI \& ESFANDYARIFAR

Table 13. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=2, \theta=0.8$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.7565883 | 0.016160920 | 0.7342053 | 0.016276950 | 0.7156517 | 0.019879590 |
| 20 | 0.7771566 | 0.006733548 | 0.7635732 | 0.007190623 | 0.7534337 | 0.008318997 |
| 30 | 0.7853821 | 0.004234158 | 0.7757980 | 0.004486893 | 0.7688793 | 0.005013464 |
| 40 | 0.7908123 | 0.003149407 | 0.7834313 | 0.003279836 | 0.7782094 | 0.003568182 |
| 50 | 0.7891018 | 0.002431975 | 0.7831037 | 0.002570948 | 0.7788405 | 0.002790142 |
| 100 | 0.7962759 | 0.000991319 | 0.7931716 | 0.001020993 | 0.7910271 | 0.001068188 |

Table 14. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=3, \theta=0.8$

|  | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| n | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.7565883 | 0.016160920 | 0.7342053 | 0.016276950 | 0.7034408 | 0.023040870 |
| 20 | 0.7771566 | 0.006733548 | 0.7635732 | 0.007190623 | 0.7469579 | 0.009294010 |
| 30 | 0.7853821 | 0.004234158 | 0.7757980 | 0.004486893 | 0.7645488 | 0.005464724 |
| 40 | 0.7908123 | 0.003149407 | 0.7834313 | 0.003279836 | 0.7749910 | 0.003818176 |
| 50 | 0.7891018 | 0.002431975 | 0.7831037 | 0.002570948 | 0.7762116 | 0.002971987 |
| 100 | 0.7962759 | 0.000991319 | 0.7931716 | 0.001020993 | 0.7897346 | 0.001107678 |

Table 15. MSE of MLE, Bayesian, and E-Bayesian estimation for $c=5, \theta=0.8$

| $n$ | MLE |  | BAYES |  | E-BAYES |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE | $\hat{\theta}$ | MSE |
| 10 | 0.7565883 | 0.016160921 | 0.7342053 | 0.016276950 | 0.6907121 | 0.027424820 |
| 20 | 0.7771566 | 0.006733548 | 0.7635732 | 0.007190623 | 0.7410871 | 0.010504326 |
| 30 | 0.7853821 | 0.004234158 | 0.7757980 | 0.004486893 | 0.7609354 | 0.005989118 |
| 40 | 0.7908123 | 0.003149407 | 0.7834313 | 0.003279836 | 0.7724500 | 0.004101652 |
| 50 | 0.7891018 | 0.002431975 | 0.7831037 | 0.002570948 | 0.7741743 | 0.003164136 |
| 100 | 0.7962759 | 0.000991319 | 0.7931716 | 0.001020993 | 0.7888172 | 0.001145513 |

## ESTIMATION PARAMETER OF LSD



Figure 3. MSE of MLE, Bayesian, and E-Bayesian estimation for $\theta=0.8$

According to Tables 11-15 and Figure 3 above, if $\theta$ is close to 1 , then the maximum likelihood estimator will be better than the others.

## Conclusion

The comparison among the three estimators revealed that with increasing sample size, all three estimators come together and as a result, the error rate is reduced. However, in the small samples according to the value of $\theta$ is superior to any of the rest, of the figures and tables is shown.

## References

Berger, J. O. (1985). Statistical decision theory and Bayesian analysis (2nd ed.). New York, NY: Springer-Verlag.

Bohning, D. (1983). Maximum likelihood estimation of the logarithmic series distribution. Statistics Hefte, 24(1), 121-140. doi: 10.1007/BF02932495

Castellares, F., \& Lemonte, A. J. (2014). A new generalized Weibull distribution generated by gamma random variables. Journal of the Egyptian Mathematical Society, 23(2), 382-390. doi: 10.1016/j.joems.2014.03.009

## NASIRI \& ESFANDYARIFAR

Han, M. (1997). The structure of hierarchical prior distribution and its application. Operations Research and Management Science, 6(3), 31-40. Available from http://en.cnki.com.cn/Journal_en/A-A002-YCGL-1997-03.htm

Han, M. (2005). Expected Bayesian method for forecast of security investment. Operations Research and Management Science, 14(5), 98-102. Available from http://en.cnki.com.cn/Article_en/CJFDTotalYCGL200505020.htm

Han, M. (2006). E-Bayesian estimation and hierarchical Bayesian estimation of estate probability. Operations Research and Management Science, 15(5), 7074. Available from http://en.cnki.com.cn/Article_en/CJFDTOTALYCGL200605013.htm

Han, M. (2007). E-Bayesian estimation of failure probability and its application. Mathematical and Computer Modeling, 45(9-10), 1272-1279. doi: 10.1016/j.mcm.2006.11.007

Han, M. (2009). E-Bayesian estimation and hierarchical Bayesian estimation of rate. Applied Mathematical Modeling, 33(4), 1915-1922. doi: 10.1016/j.apm.2008.03.019

Han, M., \& Ding, Y. (2004). Synthesized expected Bayesian method of parametric estimate. Journal of Systems Science and Systems Engineering, 13(1), 98-111. doi: 10.1007/s11518-006-0156-0

Lindley, D. V., \& Smith, A. F. M. (1972). Bayesian estimates for the linear model. Journal of the Royal Statistical Society. Series B (Methodological), 34(1), 1-41. Available from http://www.jstor.org/stable/2985048

Nasiri, P. (2011). Estimation parameter of zero truncated mixed Poisson models. International Journal of Mathematical Analysis, 5(10), 465-470.
Retrieved from http://www.m-hikari.com/ijma/ijma-2011/ijma-9-12-
2011/nasiriIJMA9-12-2011.pdf

# Bayesian Analysis of Generalized Exponential Distribution 

Saima Naqash<br>University of Kashmir<br>Jammu and Kashmir, India

S. P. Ahmad<br>University of Kashmir<br>Jammu and Kashmir, India

Aquil Ahmed<br>Aligarh Muslim University<br>Aligarh, UP, Inida

Bayesian estimators of unknown parameters of a two parameter generalized exponential distribution are obtained based on non-informative priors using different loss functions.

Keywords: Generalized exponential distribution, Bayesian estimators, loss function, R-software

## Introduction

One of the simplest and most commonly used distributions (and often erroneously overused due to its simplicity) is the exponential distribution. The two-parameter exponential distribution, which is an extension of the exponential distribution, was first introduced by Gupta and Kundu (1999), and is very popular in analyzing lifetime or survival data. Like Weibull and gamma distributions, the generalized exponential distribution can have an increasing, constant, or decreasing hazard function depending on the shape parameter.

It was observed by Gupta and Kundu (2001) that the generalized exponential (GE) distribution and the gamma distribution have very similar properties in many respects, and in some situations the generalized exponential distribution provides a better fit than Gamma and Weibull distributions in terms of maximum likelihood (ML) or minimum chi-square. Sanku Dey (2010) obtained Bayes estimators of the parameters of GE and its associated risk using different loss functions. Raqab (2002), Raqab and Ahsanullah (2001), Raqab and Madi (2005), Jaheen (2004), Kundu and Gupta (2008) extensively studied this distribution. Singh, Singh, Singh, and Singh (2008) studied the estimation problem of the parameters of this

Saima Naqash is in the Department of Statistics. Email them at:
naqashsaima@gmail.com. S. P. Ahmad is in the Department of Statistics. Email them at: sprvz@yahoo.com. Aquil Ahmed is in the Department of Statistics and O. R. Email them at: aqlstat@yahoo.co.in.

## NAQASH ET AL.

distribution under some symmetric and asymmetric loss functions using Lindley's method.

Let $x_{1}, x_{2}, \ldots, x_{n}$ be independently and identically distributed GE random variables with shape parameter $\alpha$ and scale parameter $\lambda(=1)$. Then the C.D.F. of $x$ will become

$$
\begin{equation*}
\mathrm{F}(x, \alpha)=[1-\exp (-x)]^{\alpha}, \quad x>0, \alpha>0 \tag{1}
\end{equation*}
$$

and the corresponding P.D.F. is

$$
\begin{equation*}
\mathrm{f}(x, \alpha)=\alpha[1-\exp (-x)]^{\alpha-1} \exp (-x), \quad x>0, \alpha>0 \tag{2}
\end{equation*}
$$

For $\alpha=1$, the GE distribution reduces to the one parameter (standard) exponential distribution. The GE distribution is unimodal with mode at $z=\log \alpha$, $\alpha>1$, and its median is $M=-\log \left[1-(0.5)^{\frac{1}{\alpha}}\right]$.

## Maximum Likelihood Estimation

Assume that $X=\left(x_{1}, x_{2}, \ldots, x_{n}\right)$ is a random sample from GE distribution. The likelihood function of $\alpha$ for the given sample observation is:

$$
\begin{align*}
\mathrm{L}(\alpha, x) & =\alpha^{n} \prod_{i=1}^{n}\left(1-\exp \left(-x_{i}\right)\right)^{\alpha-1} \exp \left(-\sum_{i=i}^{n} x_{i}\right)  \tag{3}\\
\Rightarrow \log \mathrm{L}(\alpha, x) & =n \log \alpha+(\alpha-1) \sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)-\sum_{i=1}^{n} x_{i}
\end{align*}
$$

the maximum likelihood estimation (MLE) of $\alpha$ is given by

$$
\begin{equation*}
\hat{\alpha}=-\frac{n}{T} \tag{4}
\end{equation*}
$$

where $T=\sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)^{-1}$.

## Prior and Posterior Distributions

Consider that the parameter $\alpha$ has the non-informative Jeffrey's prior and is given by $\mathrm{g}(\alpha) \propto \sqrt{\operatorname{det}(\mathbf{I}(\alpha))}$, where $\mathbf{I}(\alpha)$ is the Fisher Information Matrix given by

$$
\mathbf{I}(\alpha)=-n \mathrm{E}\left[\frac{\partial^{2}}{\partial \alpha^{2}} \log \mathrm{f}(x, \alpha)\right]=\frac{n}{\alpha^{2}}
$$

and Jeffery's prior distribution becomes

$$
\begin{equation*}
\mathrm{g}(\alpha) \propto \frac{1}{\alpha} \tag{5}
\end{equation*}
$$

The posterior distribution is given by

$$
\begin{aligned}
& \pi(\alpha \mid x) \propto \mathrm{g}(\alpha) \mathrm{L}(\alpha) \\
& \Rightarrow \pi(\alpha \mid x) \propto \frac{1}{\alpha}\left\{\alpha^{n} \prod_{i=1}^{n}\left(1-\exp \left(-x_{i}\right)\right)^{(\alpha-1)} \exp \left(-\sum_{i=1}^{n} x_{i}\right)\right\} \\
& =k \alpha^{n-1} \prod_{i=1}^{n}\left(1-\exp \left(-x_{i}\right)^{\alpha-1} \exp \left(-\sum_{i=1}^{n} x_{i}\right)\right) \\
& =k \alpha^{n-1} \exp \left[-(\alpha-1) \sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)^{-1}\right] \exp \left(-\sum_{i=1}^{n} x_{i}\right) \\
& \Rightarrow \pi(\alpha \mid x)=k \alpha^{n-1} \exp [-(\alpha-1) T] \exp \left(-\sum_{i=1}^{n} x_{i}\right) \\
& \quad=k \alpha^{n-1} \exp (\alpha T) \exp (T) \exp \left(-\sum_{i=1}^{n} x_{i}\right)
\end{aligned}
$$

The constant $k$ is determined such that

$$
\int_{0}^{\infty} \pi(\alpha \mid x) d \alpha=1 \Rightarrow k=\frac{T^{n}}{\Gamma(n) \exp \left(-\sum_{i=1}^{n} x_{i}\right) \exp (T)}
$$

With this value of $k$, the posterior distribution of $\alpha$ becomes

NAQASH ET AL.

$$
\begin{equation*}
\pi(\alpha \mid x)=\frac{T^{n}}{\Gamma(n)} \exp (-\alpha T) \alpha^{n-1} \tag{6}
\end{equation*}
$$

which is a Gamma distribution with parameters $n$ and $T$, where $T=\sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)^{-1}$, i.e., $\alpha \sim \mathrm{G}\left(n, \sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)^{-1}\right)$.

The expected value (mean) and variance of the distribution is given by

$$
\begin{equation*}
\mathrm{E}(\pi(\alpha \mid x))=\frac{n}{T} \tag{7}
\end{equation*}
$$

and

$$
\mathrm{V}(\pi(\alpha \mid x))=\frac{n}{T^{2}}
$$

where $T$ is given as above.

## Bayes Estimator under Jeffery's Prior Using Different Loss Functions

## Squared Error Loss Function (SELF)

Consider the following SELF: $1(\hat{\alpha}, \alpha)=c(\hat{\alpha}-\alpha)^{2}$ and obtain the Risk function as:

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi(\alpha \mid x) d \alpha \\
& =c \frac{T^{n}}{\Gamma(n)} \int_{0}^{\infty}\left(\hat{\alpha}^{2}-2 \hat{\alpha} \alpha+\alpha^{2}\right) \exp (-\alpha T) \alpha^{n-1} d \alpha \\
& =c \hat{\alpha}^{2}-2 \hat{\alpha} c \frac{n}{T}+c \frac{n(n+1)}{T^{2}}
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$ will give the Bayes estimator:

$$
\begin{equation*}
\hat{\alpha}_{B_{i}}=\frac{n}{T} \tag{8}
\end{equation*}
$$

which is the same as the MLE of $\alpha$ given in (4).

## Quadratic Loss Function (QLF)

Consider the following QLF:

$$
1(\hat{\alpha}, \alpha)=\left(\frac{\alpha-\hat{\alpha}}{\alpha}\right)^{2}
$$

and obtain the Risk function as

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi(\alpha \mid x) d \alpha \\
& =\int_{0}^{\infty}\left(\frac{\alpha-\hat{\alpha}}{\alpha}\right)^{2} \frac{T^{n}}{\Gamma(n)} \exp (-\alpha T) \alpha^{n-1} d \alpha \\
& =1-2 \hat{\alpha} \frac{T}{n-1}+\hat{\alpha}^{2} \frac{T^{2}}{(n-1)(n-2)}
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, we get the Bayes estimator of $\alpha$ as:

$$
\begin{equation*}
\hat{\alpha}_{B_{2}}=\frac{n-2}{T} \tag{9}
\end{equation*}
$$

## Al-Bayyati's Loss Function

Al-Bayyati's loss function is of the form $1(\hat{\alpha}, \alpha)=\alpha^{c_{2}}(\hat{\alpha}-\alpha)^{2}, c_{2} \in \mathbb{R}^{+}$. This loss function is used to obtain the estimator of the parameter of GE distribution. The risk function is obtained as:

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi(\alpha \mid x) d \alpha=\int_{0}^{\infty} \alpha^{c_{2}}(\hat{\alpha}-\alpha)^{2} \frac{T^{n}}{\Gamma(n)} \exp (-\alpha T) d \alpha \\
& =\frac{1}{T^{c_{2}} \Gamma(n)}\left[\hat{\alpha}^{2} \Gamma\left(n+c_{2}\right)-2 \hat{\alpha} \frac{\Gamma\left(n+c_{2}+1\right)}{T}+\frac{\Gamma\left(n+c_{2}+2\right)}{T^{2}}\right]
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, we get the Bayes estimator of $\alpha$ as:

$$
\begin{equation*}
\hat{\alpha}_{B_{3}}=\frac{n+c_{2}}{T} \tag{10}
\end{equation*}
$$

## Remark 1:

1. For $c_{2}=-2$ in (10), we get $\hat{\alpha}=\frac{n-2}{T}$, which gives the Bayes estimator under QLF using Jeffery's prior.
2. For $c_{2}=0$ in (10), we get $\hat{\alpha}=\frac{n}{T}$, which gives the Bayes estimator under SELF using Jeffery's prior.

## Precautionary Loss Function (PLF)

Consider the following PLF:

$$
1(\hat{\alpha}, \alpha)=\frac{(\alpha-\hat{\alpha})^{2}}{\alpha}
$$

and obtain the Risk function as:

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi(\alpha \mid x) d \alpha \\
& =\int_{0}^{\infty} \frac{(\hat{\alpha}-\alpha)^{2}}{\hat{\alpha}} \frac{T^{n}}{\Gamma(n)} \exp (-\alpha T) \alpha^{n-1} d \alpha \\
& =\hat{\alpha}-2 \frac{n}{T}+\frac{(n+1) n}{\hat{\alpha} T^{2}}
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, the Bayes estimator of $\alpha$ is

$$
\begin{equation*}
\hat{\alpha}_{B_{4}}=\frac{\sqrt{n(n+1)}}{T} \tag{11}
\end{equation*}
$$

## New Extension of Jeffery's Prior Information

The new extension of Jeffrey's prior information is given by:

$$
\begin{gather*}
\mathrm{g}(\alpha) \propto[\mathbf{I}(\alpha)]^{c_{1}}, c_{1} \in \mathbb{R}^{+} \\
\Rightarrow \mathrm{g}(\alpha) \propto\left[\frac{n}{\alpha^{2}}\right]^{c_{1}}=\frac{n^{c_{1}}}{\alpha^{2 c_{1}}} \\
\Rightarrow \mathrm{~g}(\alpha) \propto \frac{1}{\alpha^{2 c_{1}}} \tag{12}
\end{gather*}
$$

The posterior distribution is obtained in a similar way as in the case of Jeffrey's prior information and is given by

$$
\begin{aligned}
& \int_{0}^{\infty} \pi_{1}(\alpha \mid x) d \alpha=1 \\
& \Rightarrow k=\frac{T^{n-2 c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right) \exp \left(-\sum_{i=1}^{n} x_{i}\right) \exp (T)}
\end{aligned}
$$

Hence the posterior distribution of $\alpha$ becomes

$$
\begin{equation*}
\pi_{1}(\alpha \mid x)=\frac{T^{n-c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right)} \alpha^{n-2 c_{1}} \exp (-\alpha T) \tag{13}
\end{equation*}
$$

which is the Gamma distribution with parameters $\left(n-2 c_{1}+1\right)$ and $T$, i.e.

NAQASH ET AL.

$$
\alpha \sim \mathrm{G}\left(n-2 c_{1}+1, \sum_{i=1}^{n} \log \left(1-\exp \left(-x_{i}\right)\right)^{-1}\right)
$$

The expected value (mean) and variance of the distribution is given by

$$
\begin{equation*}
\mathrm{E}\left(\pi_{1}(\alpha \mid x)\right)=\frac{n-2 c_{1}+1}{T} \tag{14}
\end{equation*}
$$

and

$$
\mathrm{V}\left(\pi_{1}(\alpha \mid x)\right)=\frac{n-2 c_{1}+1}{T^{2}}
$$

## Remark 2:

1. For $c_{1}=1 / 2$ in (14), the posterior distribution under the extension of Jeffery's prior reduces to the posterior distribution under the Jeffery's prior.
2. For $c_{1}=3 / 2$ in (14), the posterior distribution under the extension of Jeffery's prior reduces to the posterior distribution under the Hartigan's prior.

## Bayes Estimation under the Extension of Jeffery's Prior using Different Loss Functions

## Squared Error Loss Function

The risk function under SELF is obtained as

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi_{1}(\alpha \mid x) d \alpha \\
& =c \frac{T^{n-c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right)} \int_{0}^{\infty}\left(\hat{\alpha}^{2}-2 \hat{\alpha} \alpha+\alpha^{2}\right) \exp (-\alpha T) \alpha^{n-2 c_{1}} d \alpha \\
& =c \hat{\alpha}^{2}-2 \hat{\alpha} c \frac{n-2 c_{1}+1}{T}+c \frac{\left(n-2 c_{1}+2\right)\left(n-2 c_{1}+1\right)}{T^{2}}
\end{aligned}
$$

## ESTIMATION OF GENERALIZED EXPONENTIAL DISTRIBUTION

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, the Bayes estimator of $\alpha$ is

$$
\begin{equation*}
\hat{\alpha}_{B_{5}}=\frac{n-2 c_{1}+1}{T} \tag{15}
\end{equation*}
$$

Remark 3: For $c_{1}=1 / 2$ in (15), $\hat{\alpha}=n / T$, which gives the Jeffery's estimator under SELF.

## Quadratic Loss Function

Using the QLF,

$$
1(\hat{\alpha}, \alpha)=\left(\frac{\alpha-\hat{\alpha}}{\alpha}\right)^{2}
$$

The risk function under QLF is obtained as

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi_{1}(\alpha \mid x) d \alpha \\
& =\frac{T^{n-2 c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right)} \int_{0}^{\infty}\left(\hat{\alpha}^{2}-2 \hat{\alpha} \alpha+\alpha^{2}\right) \exp (-\alpha T) \alpha^{n-2 c_{1}-2} d \alpha \\
& =1-2 \hat{\alpha} \frac{T}{n-2 c_{1}}+\hat{\alpha}^{2} \frac{T^{2}}{\left(n-2 c_{1}\right)\left(n-2 c_{1}-1\right)}
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, the Bayes estimator of $\alpha$ is

$$
\begin{equation*}
\hat{\alpha}_{B_{6}}=\frac{n-2 c_{1}-1}{T} \tag{16}
\end{equation*}
$$

Remark 4: For $c_{1}=1 / 2$ in (16), $\hat{\alpha}=\frac{n-2}{T}$, which gives the Bayes estimator under QLF using Jeffery's prior.

## Al-Bayyati's Loss Function

Al-Bayyati's loss function is of the form $1(\hat{\alpha}, \alpha)=\alpha^{c_{2}}(\hat{\alpha}-\alpha)^{2}, c_{2} \in \mathbb{R}^{+}$. The risk function is given by

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha)= & \int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi_{1}(\alpha \mid x) d \alpha \\
= & \frac{T^{n-2 c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right)} \int_{0}^{\infty}(\hat{\alpha}-\alpha)^{2} \exp (-\alpha T) \alpha^{n+c_{2}-2 c_{1}} d \alpha \\
= & \frac{1}{T^{c_{2}} \Gamma\left(n-2 c_{1}+1\right)}\left[\hat{\alpha}^{2} \Gamma\left(n+c_{2}-2 c_{1}+1\right)-2 \hat{\alpha} \frac{\Gamma\left(n+c_{2}-2 c_{1}+1\right)}{T}\right. \\
& \left.+\frac{\Gamma\left(n+c_{2}-2 c_{1}+3\right)}{T^{2}}\right]
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, the Bayes estimator of $\alpha$ is

$$
\begin{equation*}
\hat{\alpha}_{B_{7}}=\frac{n+c_{2}-2 c_{1}+1}{T} \tag{17}
\end{equation*}
$$

## Remark 5:

1. For $c_{1}=1 / 2$ and $c_{2}=0$ in (17), $\hat{\alpha}=n / T$, which gives the Bayes' estimator under SELF using Jeffery's prior.
2. For $c_{1}=1 / 2$ and $c_{2}=-2$ in (17), $\hat{\alpha}=\frac{n-2}{T}$, which gives the Bayes' estimator under QLF using Jeffery's prior.

## Precautionary Loss Function

Using the PLF

$$
1(\hat{\alpha}, \alpha)=\frac{(\alpha-\hat{\alpha})^{2}}{\alpha}
$$

obtain the Risk function under PLF as

$$
\begin{aligned}
\mathrm{R}(\hat{\alpha}, \alpha) & =\int_{0}^{\infty} 1(\hat{\alpha}, \alpha) \pi_{1}(\alpha \mid x) d \alpha \\
& =\int_{0}^{\infty} \frac{(\hat{\alpha}-\alpha)^{2}}{\hat{\alpha}} \frac{T^{n-2 c_{1}+1}}{\Gamma\left(n-2 c_{1}+1\right)} \exp (-\alpha T) \alpha^{n-2 c_{1}} d \alpha \\
& =\hat{\alpha}-2 \frac{\left(n-2 c_{1}+1\right)}{T}+\frac{\left(n-2 c_{1}+2\right)\left(n-2 c_{1}+1\right)}{\hat{\alpha} T^{2}}
\end{aligned}
$$

Solving the equation $\frac{\partial}{\partial \hat{\alpha}} \mathrm{R}(\hat{\alpha}, \alpha)=0$, the Bayes' estimator of $\alpha$ is

$$
\begin{equation*}
\hat{\alpha}_{B_{8}}=\frac{\sqrt{\left(n-2 c_{1}+2\right)\left(n-2 c_{1}+1\right)}}{T} \tag{18}
\end{equation*}
$$

Remark 6: For $c_{1}=1 / 2$ in (18), $\hat{\alpha}=\frac{\sqrt{n(n+1)}}{T}$, which gives the Bayes' estimator under PLF using Jeffery's prior.

## Simulation Study of Generalized Exponential Distribution

In the simulation study, sample sizes were chosen at $n=25,50$, and 100 to represent small, medium, and large data sets. The scale parameter is estimated for Generalized Exponential distribution with Maximum Likelihood and Bayesian using Jeffrey's \& extension of Jeffrey's prior methods. For the scale parameter, $\alpha=0.5,1.0$, and 1.5. The values of Jeffrey's extension are chosen as $c_{1}=1.0,1.5$, and 2. The value for the loss parameter $c_{2}= \pm 1.0$ and $\pm 2.0$. This was iterated 5000 times and the scale parameter for each method was calculated. A simulation study was conducted in R-software to examine and compare the performance of the estimates for different sample sizes with different values for Jeffrey's prior and the extension of Jeffrey's prior under different loss functions. The results are presented in tables for different selections of the parameters and c extension of Jeffrey's prior.

In Table 2, Bayes' estimation with Al-Bayyati's Loss function under Jeffrey's prior provides the smallest values in most cases especially when loss parameter $c_{2}$ is $\pm 2.0$. Similarly, in Table 4, Bayes' estimation with Al-Bayyati's Loss function under extension of Jeffrey's prior provides the smallest values in most cases,

## NAQASH ET AL.

especially when loss parameter $c_{2}$ is $\pm 2.0$ whether the extension of Jeffrey's prior is $0.5,1.0$, or 1.5 . Moreover, when the sample size increases from 25 to 100 , the Mean Squared Error decreases quite significantly.

Table 1. Posterior mean for $\hat{\alpha}$ under Jeffery's prior

| $n$ | $\lambda$ | $\alpha$ | $\alpha_{\text {ML }}$ | $\alpha_{\text {SL }}$ | $\alpha_{\text {QL }}$ | $\alpha_{\text {PL }}$ | $\alpha_{\text {AL }}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  | $c_{2}=1.0$ | $c_{2}=-1.0$ | $c_{2}=2.0$ | $C_{2}=-2.0$ |
| 25 | 1.0 | 0.5 | 0.3815 | 0.3815 | 0.3510 | 0.3890 | 0.3967 | 0.3662 | 0.4120 | 0.3510 |
|  |  | 1.0 | 0.9899 | 0.9899 | 0.9107 | 1.0095 | 1.0295 | 0.9503 | 1.0691 | 0.9107 |
|  |  | 1.5 | 2.3398 | 2.3398 | 2.1526 | 2.3861 | 2.4334 | 2.2462 | 2.5270 | 2.1526 |
| 50 | 1.0 | 0.5 | 0.4521 | 0.4521 | 0.4340 | 0.4566 | 0.4512 | 0.4431 | 0.4702 | 0.4340 |
|  |  | 1.0 | 0.8861 | 0.8861 | 0.8507 | 0.8950 | 0.9038 | 0.8684 | 0.9215 | 0.8507 |
|  |  | 1.5 | 1.5164 | 1.5164 | 1.4557 | 1.5315 | 1.5467 | 1.4861 | 1.5770 | 1.4557 |
| 100 | 1.0 | 0.5 | 0.4845 | 0.4845 | 0.4748 | 0.4869 | 0.4893 | 0.4797 | 0.4941 | 0.4748 |
|  |  | 1.0 | 0.8911 | 0.8911 | 0.8733 | 0.8956 | 0.9000 | 0.8822 | 0.9089 | 0.8733 |
|  |  | 1.5 | 1.4541 | 1.4545 | 1.4250 | 1.4613 | 1.4686 | 1.4395 | 1.4831 | 1.4250 |

Note: ML=Maximum Likelihood, SL=Squared Error Loss Function, QL=Quadratic Loss Function, PL=Precautionary Loss Function, AL=Al-Bayyati's Loss Function

Table 2. Mean squared error for $\hat{\alpha}$ under Jeffery's prior

| $n$ | $\lambda$ | $\alpha$ | $\chi_{\text {ML }}$ | $\alpha_{\text {SL }}$ | $\alpha_{\text {QL }}$ | $\alpha_{\text {PL }}$ | $\alpha_{\text {AL }}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  | $c_{2}=1.0$ | $c_{2}=-1.0$ | $C_{2}=2.0$ | $C_{2}=-2.0$ |
| 25 | 1.0 | 0.5 | 0.0258 | 0.0258 | 0.2722 | 0.0246 | 0.0235 | 0.0288 | 0.0215 | 0.0322 |
|  |  | 1.0 | 0.0473 | 0.0473 | 1.0080 | 0.0492 | 0.0519 | 0.0460 | 0.0600 | 0.0479 |
|  |  | 1.5 | 0.8114 | 0.8114 | 2.6759 | 0.8956 | 0.9860 | 0.6546 | 1.1785 | 0.5157 |
| 50 | 1.0 | 0.5 | 0.0077 | 0.0077 | 0.2544 | 0.0074 | 0.0080 | 0.0084 | 0.0068 | 0.0094 |
|  |  | 1.0 | 0.0347 | 0.0347 | 1.0223 | 0.0331 | 0.0319 | 0.0381 | 0.0297 | 0.0423 |
|  |  | 1.5 | 0.0491 | 0.0491 | 2.2520 | 0.0508 | 0.0530 | 0.0471 | 0.0582 | 0.0470 |
| 100 | 1.0 | 0.5 | 0.0028 | 0.0028 | 0.2506 | 0.0028 | 0.0028 | 0.0030 | 0.0027 | 0.0031 |
|  |  | 1.0 | 0.0223 | 0.0223 | 1.0161 | 0.0214 | 0.0206 | 0.0241 | 0.0191 | 0.0261 |
|  |  | 1.5 | 0.0255 | 0.0255 | 2.2556 | 0.0252 | 0.0249 | 0.0267 | 0.0247 | 0.0281 |

Note: ML=Maximum Likelihood, SL=Squared Error Loss Function, QL=Quadratic Loss Function, PL=Precautionary Loss Function, AL=Al-Bayyati's Loss Function

## ESTIMATION OF GENERALIZED EXPONENTIAL DISTRIBUTION

Table 3. Posterior mean for $\hat{\alpha}$ under extension of Jeffery's prior

| $n$ | $\lambda$ | $\alpha$ | $\alpha_{\text {ML }}$ | $\chi_{\text {SL }}$ | $\alpha_{\text {QL }}$ | $\alpha_{\text {PL }}$ | $\alpha_{\text {AL }}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  | $C_{2}=1.0$ | $c_{2}=-1.0$ | $\mathrm{C}_{2}=2.0$ | $c_{2}=-2.0$ |
| 25 | 1.0 | 0.5 | 0.3815 | 0.3662 | 0.3357 | 0.3738 | 0.3815 | 0.3510 | 0.3967 | 0.3357 |
|  |  | 1.0 | 0.9899 | 0.9107 | 0.8315 | 0.9303 | 0.9503 | 0.8711 | 0.9899 | 0.8315 |
|  |  | 1.5 | 2.3398 | 2.0590 | 1.8718 | 2.1053 | 2.1526 | 1.9654 | 2.2462 | 1.8718 |
| 50 | 1.0 | 0.5 | 0.4521 | 0.4431 | 0.4250 | 0.4476 | 0.4521 | 0.4340 | 0.4612 | 0.4250 |
|  |  | 1.0 | 0.8861 | 0.8107 | 0.8152 | 0.8595 | 0.8684 | 0.8330 | 0.8861 | 0.8152 |
|  |  | 1.5 | 1.5164 | 1.4254 | 1.3648 | 1.4557 | 1.4405 | 1.3951 | 1.4861 | 1.3648 |
| 100 | 1.0 | 0.5 | 0.4845 | 0.4797 | 0.4700 | 0.4821 | 0.4845 | 0.4748 | 0.4893 | 0.4700 |
|  |  | 1.0 | 0.8911 | 0.8733 | 0.8555 | 0.8777 | 0.8822 | 0.8644 | 0.8911 | 0.8555 |
|  |  | 1.5 | 1.4541 | 1.4104 | 1.3814 | 1.4177 | 1.4250 | 1.3959 | 1.4395 | 1.3814 |

Note: ML=Maximum Likelihood, SL=Squared Error Loss Function, QL=Quadratic Loss Function, PL=Precautionary Loss Function, AL=Al-Bayyati's Loss Function

Table 4. Mean squared error for $\hat{\alpha}$ under extension of Jeffery's prior

| $n$ | $\lambda$ | $\alpha$ | $\alpha_{\text {ML }}$ | $\chi_{\text {SL }}$ | $\alpha_{\text {QL }}$ | $\alpha_{\text {PL }}$ | $\alpha_{\text {AL }}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |  | $c_{2}=1.0$ | $C_{2}=-1.0$ | $C_{2}=2.0$ | $c_{2}=-2.0$ |
| 25 | 1.0 | 0.5 | 0.0258 | 0.0288 | 0.0361 | 0.0272 | 0.0258 | 0.0322 | 0.0235 | 0.0361 |
|  |  | 1.0 | 0.0473 | 0.0479 | 0.0617 | 0.0466 | 0.0460 | 0.0531 | 0.0473 | 0.0617 |
|  |  | 1.5 | 0.8114 | 0.3947 | 0.2062 | 0.4253 | 0.5157 | 0.2915 | 0.6546 | 0.2061 |
| 50 | 1.0 | 0.5 | 0.0077 | 0.0552 | 0.0104 | 0.0080 | 0.0077 | 0.0094 | 0.0071 | 0.0104 |
|  |  | 1.0 | 0.0347 | 0.0558 | 0.0526 | 0.0401 | 0.0381 | 0.0471 | 0.0347 | 0.0526 |
|  |  | 1.5 | 0.0491 | 0.0487 | 0.0578 | 0.0460 | 0.0485 | 0.0523 | 0.0471 | 0.0578 |
| 100 | 1.0 | 0.5 | 0.0028 | 0.0032 | 0.0033 | 0.0029 | 0.0028 | 0.0031 | 0.0028 | 0.0033 |
|  |  | 1.0 | 0.0223 | 0.0261 | 0.0305 | 0.0251 | 0.0241 | 0.0282 | 0.0223 | 0.0305 |
|  |  | 1.5 | 0.0255 | 0.0300 | 0.0352 | 0.0291 | 0.0281 | 0.0324 | 0.0267 | 0.0352 |

Note: ML=Maximum Likelihood, SL=Squared Error Loss Function, QL=Quadratic Loss Function, PL=Precautionary Loss Function, AL=Al-Bayyati's Loss Function

## Conclusion

The Bayes' estimator of the parameter of the Generalized Exponential distribution was studied under Jeffrey's prior and the extended Jeffrey's prior assuming different loss functions. The extended Jeffrey's prior gives the opportunity of covering wide spectrum of priors to get Bayes' estimates of the parameter -

## NAQASH ET AL.

particular cases of which are Jeffrey's prior and Hartigan's prior. We have also addressed the problem of Bayesian estimation for the Generalized Exponential distribution, under symmetric loss functions and that of Maximum Likelihood Estimation. In most cases, the Bayesian Estimator under Al-Bayyati's Loss function has the smallest Mean Squared Error values for both prior's i.e, Jeffrey's and an extension of Jeffrey's prior information. Moreover, when the sample size increases from 25 to 100, the MSE decreases quite significantly.

## References

Dey, S. (2010). Bayesian estimation of the shape parameter of the generalised exponential distribution under different loss functions. Pakistan Journal of Statistics and Operation Research, 6(2), 163-174. doi: 10.18187/pjsor.v6i2.147

Gupta, R. D., \& Kundu, D. (1999). Theory \& methods: Generalized exponential distribution. Australia and New Zealand Journal of Statistics, 41(2), 173-188. doi: 10.1111/1467-842X.00072

Gupta, R. D., \& Kundu, D. (2001). Generalized exponential family: An alternative to gamma and Weibull distributions. Biometrical Journal, 43(1), 117130. doi: 10.1002/1521-4036(200102)43:1<117::AID-BIMJ117>3.0.CO;2-R

Jaheen, Z. F. (2004). Empirical Bayes inference for generalized exponential distribution based on records. Communications in Statistics - Theory and Methods, 33(8), 1851-1861. doi: 10.1081/STA-120037445

Kundu, D., \& Gupta, R. D. (2008). Generalized exponential distribution: Bayesian estimations. Computational Statistics \& Data Analysis, 52(4), 18731883. doi: 10.1016/j.csda.2007.06.004

Raqab, M. Z. (2002). Inference for generalized exponential distributions based on record statistics. Journal of Statistical Planning and Inference, 104(2), 339-350. doi: 10.1016/S0378-3758(01)00246-4

Raqab, M. Z., \& Ahsanullah, M. (2001). Estimation of the location and scale parameters of generalized exponential distribution based on order statistics. Journal of Statistical Computation and Simulation, 69(2), 109-124. doi: 10.1080/00949650108812085

Raqab, M. Z., \& Madi, M. T. (2005). Bayesian inference for the generalized exponential distribution. Journal of Statistical Computation and Simulation, 75(10), 841-852. doi: 10.1080/00949650412331299166

# ESTIMATION OF GENERALIZED EXPONENTIAL DISTRIBUTION 

Singh, R., Singh, S. K., Singh, U., \& Singh, G. P. (2008). Bayes estimator of generalized-exponential parameters under LINEX loss function using Lindley's approximation. Data Science Journal, 7, 65-75. doi: 10.2481/dsj.7.65

# Hierarchical Bayes Estimation of Reliability Indexes of Cold Standby Series System under General Progressive Type II Censoring Scheme 

D. R. Barot<br>Ahmedabad University<br>Ahmedabad, India

M. N. Patel<br>Gujarat University<br>Ahmedabad, India

In this paper, hierarchical Bayes approach is presented for estimation and prediction of reliability indexes and remaining lifetimes of a cold standby series system under general progressive Type II censoring scheme. A simulation study has been carried out for comparison purpose. The study will help reliability engineers in various industrial series system setups.

Keywords: Cold standby series system, general progressive Type II censoring, hierarchical Bayes estimation, Monte Carlo simulation

## Introduction

A cold standby series system is widely applied to achieve high reliability in various engineering systems used in space exploration and satellite, textile manufacturing and carbon recovery systems. In such a series system, some units are placed in working mode while the rest in cold standby mode. When any unit in the working mode fails, it is replaced by any of the standby units in negligible time to survive the engineering system. The standby system becomes invalid when all standby units are used up, and one of the working units becomes unusable.

Mei, Liao, and Sun (1992) discussed the point estimation of reliability indexes by assuming that the life units in the series system have identical exponential distribution, and the failure rate is a known constant. Under the assumption that the failure rate is a random variable, Su and Gu (2003) derived the Bayes estimates while Bai, Yu , and Hu (1998) derived the multiple Bayes estimates of reliability indexes for the series system. Pham and Turkkan (1994) studied the reliability of

[^35]
## BAROT \& PATEL

the series system with Beta distribution component live. Willits (1997) studied reliability estimation of the series system using small binomial samples. Xu, Kang, and Shi (2002) discussed Bayesian and multiple Bayesian analysis of reliability performances for the series system. Barot and Patel (2014) derived the exact confidence limits of the reliability indexes for a cold standby series system under general progressive Type II censoring scheme using an empirical Bayesian approach.

In a life testing experiment, a censoring scheme that can balance between total times spent, number of units used and efficiency of statistical inference based on the results of an experiment is desirable. For this reason a more general censoring scheme called, general progressive Type II censoring scheme, has received a significant importance in the last few decades. This censoring scheme is extremely useful in both industrial life testing and clinical settings. The numerous articles dealing with inference procedures under this censoring scheme have been found in the journals (e.g., Balakrishnan \& Sandhu, 1996; Fernández, 2004; Kim \& Han, 2009; Barot \& Patel, 2014).

In Bayes approach, the posterior distribution of the parameters of interest given the data is obtained by assuming that the model hyper-parameter is known and then inferences are considered based on this distribution. However, when the information regarding the model hyper-parameter is unknown, empirical Bayes or hierarchical Bayes approaches are used to handle the super parameter structure for the estimation and prediction. In the empirical Bayes approach, the posterior distribution of the parameter of interest given the data is first obtained, assuming that the model hyper-parameters are known. The hyper-parameter is estimated from the marginal distribution of the data, and inferences are then based on the estimated posterior distribution.

However, in the case of non-availability of empirical data, estimates of parameters can be obtained through only an expert consulting. In such situations, hierarchical Bayes approach is more preferable than empirical Bayes approach. In hierarchical Bayes approach, a prior distribution of the hyper-parameter is specified according to expert's opinions, and then the posterior distribution of the parameter of interest is obtained. A parameter of interest is then estimated by its posterior mean and its precision is measured by its posterior variance. The hierarchical Bayes approach is straightforward and clear-cut, but computationally intensive, often involving high dimensional integration. It looks promising, but caution should be exercised in applying this approach. It has been described and applied extensively for various statistical inferences in literature (e.g., Han, 1998; Lehmann \& Casella,

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

1998; Papadopoulos, Tiwari, \& Zalkikar, 1996; Younes, Delampady, MacGibbon, \& Cherkaoui, 2007).

Statistical prediction was the most prevalent form of statistical inference, which is very important in a variety of disciplines such as medicine, engineering, and business. Various authors have studied the prediction problems in reliability and life testing problems (e.g., Dunsmore, 1974; Chhikara \& Guttman, 1982; Ali Mousa, 2001; Ali Mousa \& Jaheen, 2002).

Most of the research on a cold standby series system has focused on the usual Bayes approach. The objective of the present paper is to investigate estimation and prediction of reliability indexes and remaining lifetimes of the series system using a hierarchical Bayes approach under general progressive Type II censoring scheme.

## Bayes Estimation of Reliability Indexes

In reliability and life testing studies, an exponential distribution is one of the most widely used lifetime models, and inference based on this distribution can be used quite effectively. A number of lifetime data have been analyzed, and it was observed that in most of the cases an exponential distribution provides a good fit. This distribution has been used to describe the life span of many items such as electronic tubes, light bulbs and mechanical components.

Suppose that a cold standby series system has $(k+n-1)$ identical units comprising a series of $k$ working units $U_{1}, U_{2}, \ldots, U_{k}$ being in an operational state and $(n-1)$ standby units $S_{1}, S_{2}, \ldots, S_{(n-1)}$ connected in a series. When any unit of the series of $k$ working units fails, any unit of $(n-1)$ standby units replaces it immediately through an alternation switch in negligible time, so that the series system stays operational. Figure 1 shows a functional diagram of the series system. Barot and Patel (2014) have considered such a series system and placed it on a life testing experiment under general progressive Type II censoring scheme by assuming that every unit has the failure rate $k \lambda$ with the probability density and cumulative distribution functions, respectively, as

$$
\begin{equation*}
\mathrm{f}(x \mid \lambda)=\lambda k \mathrm{e}^{-\lambda k x}, \quad \lambda, x, k>0 \tag{1}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{F}(x \mid \lambda)=1-\mathrm{e}^{-\lambda k x} \tag{2}
\end{equation*}
$$

## BAROT \& PATEL



Figure 1. Cold standby series system with $(k+n-1)$ identical units

According to Cao and Cheng (1986), the reliability $\mathrm{R}(t)$ and average life MTTF of the series system are strictly monotonic decreasing functions with respect to and can be given, respectively, by

$$
\begin{equation*}
\mathrm{R}(t)=\sum_{i=0}^{n-1} \frac{\mathrm{e}^{-\lambda k t}(\lambda k t)^{i}}{i!}, \quad M T T F=\frac{n}{\lambda k} \tag{3}
\end{equation*}
$$

Under the general progressive Type II scheme, the lifetimes of the first $s$ units, i.e., $x_{(1)}, x_{(2)}, \ldots, x_{(s)}$ are not observed, and then the lifetimes until the $m^{\text {th }}$ failure, i.e., $x_{(s+1)}, x_{(s+2), \ldots,} x_{(m)}$ are completely observed. At the time of every $i^{\text {th }}$ failure, $r_{i}$ units are randomly removed from the remaining $(n-s-1)$ standby units $(i=s+1$, $s+2, \ldots, m-1$ ). Instead of continuing the test until the entire standby units are used up, the test is terminated at the time of the $m^{\text {th }}$ failure ( $m<n$ ), and all the remaining $r_{m}$ standby units are removed from the test, where $r_{m}$ is given by

$$
r_{m}=n-m-1-\sum_{i=s+1}^{m-1} r_{i}
$$

Following Barot and Patel (2014), the likelihood function based on the general progressive Type II sample $\mathbf{x}=\left(x_{(s+1)}, x_{(s+2)}, \ldots, x_{(m)}\right)$ can be written as

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

$$
\begin{equation*}
\mathrm{L}(\mathbf{x} \mid \lambda)=A \lambda^{m-s} k^{m-s}\left(1-\mathrm{e}^{-\lambda k k_{(s+1)}}\right)^{s} \mathrm{e}^{-\lambda k w} \tag{4}
\end{equation*}
$$

where

$$
A=\binom{n-1}{s}(n-s-2) \prod_{j=s+2}^{m}\left(n-j-1-\sum_{i=s+1}^{j-1} r_{i}\right), \quad w=\sum_{i=s+1}^{m}\left(1+r_{i}\right) x_{(i)}
$$

The concern in Bayesian estimation is an appropriate choice of a prior distribution for a parameter to consider subjective information from experienced experts. An exponential distribution is one of most prominent random probability distributions, and its good mathematical properties facilitate insight and computational reduction. In reliability analysis and life testing, it is preferred over many other distributions due to its richness, computational ease, better fit to the failure data, analytical tractability, and easy interpretability. To ease the computational burden and get computable closed form expression for the posterior distribution, it is assumed that the unknown failure rate $\lambda$ is the realization of a random variable and follows an exponential prior with the probability density function

$$
\begin{equation*}
\pi(\lambda \mid \beta)=\beta \mathrm{e}^{-\beta \lambda}, \quad \lambda>0 \tag{5}
\end{equation*}
$$

The likelihood function (4) and prior distribution (5) can be easily combined to form a posterior distribution that represents total knowledge about the parameter $\lambda$ after the data have been observed. It is

$$
\begin{equation*}
\pi(\lambda \mid \mathbf{x})=\frac{\lambda^{m-s}\left[1-\mathrm{e}^{-\lambda k x_{(s+1)}}\right]^{2} \mathrm{e}^{-\lambda((k w+\beta)}}{\Gamma(m-s+1) D_{s}} \tag{6}
\end{equation*}
$$

where

$$
D_{s}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left[k w_{j}+\beta\right]^{m-s+1}} ; \quad w_{j}=w+x_{(s+1)} j
$$

## BAROT \& PATEL

In Bayesian analysis, a loss function must be specified in order to obtain Bayes estimates. The loss function is a non-negative function of the distance between the estimate and the true value. When decisions become gradually more damaging for large errors, the use of squared error loss function, $L(\hat{\lambda}, \lambda)=(\hat{\lambda}-\lambda)^{2}$, is more appropriate because of its analytical tractability. The Bayes estimate of parameter $\lambda$, reliability $\mathrm{R}(t)$ and $M T T F$ can be obtained under the squared error loss function, respectively, as

$$
\begin{gather*}
\hat{\lambda}_{B}=\frac{(m-s+1) D_{S 1}}{D_{S}}  \tag{7}\\
\hat{\mathrm{R}}(t)_{B}=\frac{1}{\Gamma(m-s+1) D_{S}} \sum_{i=0}^{n-1} \frac{(k t)^{i} \Gamma(m-2+i+1)}{i!} D_{S 2(i)}  \tag{8}\\
M T T F_{B}=\frac{n D_{S 3}}{k(m-s) D_{S}} \tag{9}
\end{gather*}
$$

where

$$
\begin{gathered}
D_{S 1}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left[k w_{j}+\beta\right]^{m-s+2}} ; \quad D_{S 2(i)}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left(k t+k w_{j}+\beta\right)^{m-s+i+1}} ; \\
D_{S 3}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left(k w_{j}+\beta\right)^{m-s}}
\end{gathered}
$$

## Hierarchical Bayesian Analysis

The idea in a Bayesian model is that when you look at a likelihood function and decide right priors for parameters. Instead, it may be more appropriate to use priors depending on other parameters those are not mentioned in a likelihood function. These parameters themselves will require priors and can depend on new ones. This can continue in a hierarchical framework until there are no more parameters to incorporate in the model. In this section, hierarchical Bayes estimates of reliability indexes of the series system are constructed.

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

Due to the complicity of practical problems and uncertainty about the true level of an expert, it is quite difficult to give the exact estimate of a super parameter $\beta$. However, the value of $\beta$ can be obtained in an approximate interval denoted by $(a, b)$ through an expert consulting. As there is no other information on the parameter $\beta$, it is assumed that it has uniform distribution on $(a, b)$ with the probability density function

$$
\mathrm{U}(\beta \mid a, b)=\frac{1}{b-a}
$$

given in Xu et al. (2002). In order to obtain the posterior density of $\beta$ given $\mathbf{x}$,

$$
\begin{aligned}
\int_{a}^{b} \mathrm{U}(\beta \mid a, b) \mathrm{m}(\mathbf{x} \mid \beta) d \beta & =\frac{\Gamma(m-s+1)}{b-a} A k^{m-s} \sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \int_{a}^{b} \frac{\beta}{\left(k w_{j}+\beta\right)^{m-s+1}} d \beta \\
& =\frac{A k^{m-s} \Gamma(m-s-1)}{a-b} D_{S 4}
\end{aligned}
$$

where

$$
D_{S 4}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j}\left[\left(k w_{j}+b\right)^{-m+s}\left(k w_{j}+b m-b s\right)-\left(k w_{j}+a\right)^{-m+s}\left(k w_{j}+a m-a s\right)\right]
$$

From Bayes theorem, the posterior density of $\beta$ given $\mathbf{x}$ can be obtained as

$$
\mathrm{h}(\beta \mid \mathbf{x})=\frac{\mathrm{U}(\beta \mid a, b) \mathrm{m}(\mathbf{x} \mid \beta)}{\int_{a}^{b} \mathrm{U}(\beta \mid a, b) \mathrm{m}(\mathbf{x} \mid \beta) d \beta}=\frac{-\beta(m-s)(m-s-1) D_{S}}{D_{S 4}}
$$

Under the squared error loss function, the Bayes estimate of $\beta$ can be given by

$$
\begin{equation*}
\hat{\beta}_{B}=\mathrm{E}^{h}(\beta \mid \mathbf{x})=\int_{a}^{b} \beta \mathrm{~h}(\beta \mid \mathbf{x}) d \beta=\frac{D_{S 5}-D_{S 6}}{(m-s-2) D_{S 4}} \tag{10}
\end{equation*}
$$

where

## BAROT \& PATEL

$$
\begin{aligned}
& D_{s 5}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j}\left\{\frac{2 k^{2} w_{j}^{2}+2 b k(m-s) w_{j}+(m-s)(m-s-1) b^{2}}{\left(k w_{j}+b\right)^{m-s}}\right\} \\
& D_{S 6}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j}\left\{\frac{2 k^{2} w_{j}^{2}+2 a k(m-s) w_{j}+(m-s)(m-s-1) a^{2}}{\left(k w_{j}+a\right)^{m-s}}\right\}
\end{aligned}
$$

Using (10) in (7), (8), and (9), the hierarchical Bayes estimates of $\lambda, \mathrm{R}(t)$ and MTTF under the squared error loss function can be obtained as follows:

$$
\begin{gather*}
\hat{\lambda}_{H B}=\frac{(m-s+1) D_{S 7}}{D_{S^{\prime}}}  \tag{11}\\
\hat{\mathrm{R}}(t)_{H B}=\frac{1}{\Gamma(m-s+1) D_{S^{\prime}}} \sum_{i=0}^{n-1} \frac{(k t)^{i} \Gamma(m-s+i+1)}{i!} D_{S 8(i)}  \tag{12}\\
M T T F_{H B}=\frac{n D_{S 9}}{k(m-s) D_{S^{\prime}}} \tag{13}
\end{gather*}
$$

where

$$
\begin{aligned}
& D_{S^{\prime}}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left[k w_{j}+\hat{\beta}_{B}\right]^{m-s+1}} ; \quad D_{S 7}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left[k w_{j}+\hat{\beta}_{B}\right]^{m-s+2}} ; \\
& D_{S 8(i)}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left(k t+k w_{j}+\hat{\beta}_{B}\right)^{m-s+i+1}} ; \quad D_{S 9}=\sum_{j=0}^{s}\binom{s}{j}(-1)^{j} \frac{1}{\left[k w_{j}+\hat{\beta}_{B}\right]^{m-s}} ;
\end{aligned}
$$

In Bayesian inference, the $100(1-\alpha) \%$ highest probability density (HPD) interval of the parameter of interest is the shortest interval in parameter space that contains $100(1-\alpha) \%$ of the probable values of the parameter. It is one of the most useful tools to measure posterior uncertainty that includes more probable values and excludes the least probable values of the parameter. Since the posterior

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

distribution (6) is unimodal, the $100(1-\alpha) \%$ Bayes HPD-interval $\left(p_{1}, p_{2}\right)$ for $\lambda$ must simultaneously satisfy the equations

$$
\begin{gather*}
\pi^{*}\left(p_{1} \mid \mathbf{x}\right)=\pi^{*}\left(p_{2} \mid \mathbf{x}\right)  \tag{14}\\
\int_{p_{1}}^{p_{2}} \pi^{*}(\lambda \mid \mathbf{x}) d \lambda=1-\alpha \tag{15}
\end{gather*}
$$

After tedious algebra, the equations (14) and (15) can be written in the form

$$
\begin{equation*}
\left(\frac{p_{1}}{p_{2}}\right)^{m-s} \mathrm{e}^{\left(p_{2}-p_{1}\right)(k w+\beta)}=\left(\frac{1-\mathrm{e}^{-p_{2} k x_{(s+1)}}}{1-\mathrm{e}^{-p_{1} k x_{(s+1)}}}\right) \tag{16}
\end{equation*}
$$

and

$$
\begin{equation*}
\frac{1}{D_{S}} \sum_{j=0}^{s}\binom{s}{j}(-1)^{j} c_{j}^{-m+s-1}\left[\mathrm{e}^{-c_{j} p_{1}} \sum_{j_{1}=0}^{m-s} \frac{\left(c_{j} p_{1}\right)^{j_{1}}}{j_{1}!}-\mathrm{e}^{-c_{j} p_{2}} \sum_{j_{1}=0}^{m-s} \frac{\left(c_{j} p_{2}\right)^{j_{1}}}{j_{1}!}\right]=1-\alpha \tag{17}
\end{equation*}
$$

where $c_{j}=k w_{j}+\beta$.
The $100(1-\alpha) \%$ Bayes HPD-intervals of $\mathrm{R}(t)$ and MTTF can be obtained from (8) and (9). When the super parameter $\beta$ is unknown, the $100(1-\alpha) \%$ hierarchical Bayes HPD-intervals of reliability indexes can be obtained by using the estimate $\hat{\beta}_{B}$ for $\beta$.

## Prediction of Remaining Lifetimes Truncated at $\boldsymbol{X}_{(m)}$

The prediction of remaining lifetimes, based on a current available sample, known as an informative sample, is an important feature in Bayesian analysis. Howlader (1985) presented HPD-prediction intervals for the $z^{\text {th }}$ order statistic of a future sample. Fernández (2000) considered the problem of predicting an independent future sample from the Rayleigh distribution under doubly Type II censoring scheme. Raqab and Madi (2002) considered an estimation of the predictive distribution of the total time on a test up to certain failures in a future sample, as well as that of the remaining testing time until all the units in the original sample have failed.

## BAROT \& PATEL

Let

$$
x_{(l)}, \quad m+1 \leq l \leq n_{1}, n_{1}=n-\sum_{i=s+1}^{m-1} r_{i}
$$

denote the lifetime of the $l^{\text {th }}$ unit to fail. The conditional probability density function of $y=x_{(l)}-x_{(m)}$ from the probability density function truncated at $x_{(m)}$ is given by

$$
\mathrm{f}_{1}(y \mid \lambda)=\frac{[\mathrm{F}(y \mid \lambda)]^{l-m-1}[1-\mathrm{F}(y \mid \lambda)]^{n_{1}-l} \mathrm{f}(y \mid \lambda)}{\mathrm{B}\left(l-m, n_{1}-l+1\right)}, \quad y \geq 0
$$

From (1) and (2), the function $\mathrm{f}_{1}=(y \mid \lambda)$ can be obtained as

$$
\mathrm{f}_{1}(y \mid \lambda)=\frac{\lambda k \mathrm{e}^{-\lambda k y\left(n_{1}-l+1\right)}\left(1-\mathrm{e}^{-\lambda k y}\right)^{l-m-1}}{\mathrm{~B}\left(l-m, n_{1}-l+1\right)}
$$

Based on the general progressive Type II censored sample $\mathbf{x}$, the conditional joint probability density function of $y$ and $\lambda$ can be written as

$$
\begin{aligned}
\mathrm{g}_{1}(y, \lambda \mid \mathbf{x}) & =\mathrm{f}_{1}(y \mid \lambda) \pi^{*}(\lambda \mid \mathbf{x}) \\
& =\frac{k \lambda^{m-s+1} \mathrm{e}^{-\lambda\left[k y\left(n_{1}-l+1\right)+k w+\beta\right]}\left(1-\mathrm{e}^{-\lambda k y}\right)^{l-m+1}\left(1-\mathrm{e}^{-\lambda k x_{(s+1)}}\right)^{s}}{\Gamma(m-s+1) \mathrm{B}\left(l-m, n_{1}-l+1\right) D_{S}}
\end{aligned}
$$

The Bayes predictive density function of $y$ can be obtained as

$$
\begin{align*}
\mathrm{p}(y \mid \mathbf{x}) & =\int_{0}^{\infty} \mathrm{g}_{1}(y, \lambda \mid \mathbf{x}) d \lambda \\
& =k_{1} \Gamma(m-s+2) \sum_{p_{1}=0}^{l-m-1} \sum_{q_{1}=0}^{s} \frac{\binom{l-m-1}{p_{1}}\binom{s}{q_{1}}(-1)^{p_{1}+q_{1}}}{w_{p_{1} q_{1}}^{m-s+2}} \tag{18}
\end{align*}
$$

where

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

$$
\begin{aligned}
k_{1} & =\frac{k}{\Gamma(m-s+1) \mathrm{B}\left(l-m, n_{1}-l+1\right) D_{S}} \\
w_{p_{1} q_{1}} & =k\left(n_{1}-l+1+p_{1}\right) y+k w-k x_{(s+1)} q_{1}+\beta
\end{aligned}
$$

Under the squared error loss function, the Bayes predictive estimate of $y$ can be obtained as

$$
y^{*}=\frac{\sum_{p_{1}=0}^{l-m-1} \sum_{q_{1}=0}^{s}\binom{l-m-1}{p_{1}}\binom{s}{q_{1}}(-1)^{p_{1}+q_{1}} \frac{\left(k w+k x_{(s+1)} q_{1}+\beta\right)^{-m+s}}{\left(n_{1}-l+1+p_{1}\right)^{2}}}{k(m-s) \mathrm{B}\left(l-m, n_{1}-l+1\right) D_{S}}
$$

Thus, the Bayes predictive estimate of $x_{(l)}$ can be given by

$$
\begin{equation*}
x_{(l)}^{*}=x_{(m)}+y^{*} \tag{19}
\end{equation*}
$$

Moreover, the $100(1-\alpha) \%$ Bayes HPD-prediction interval of $y^{*}$ is given by $\left(h_{1}, h_{2}\right)$, where $h_{1}$ and $h_{2}$ are solutions of the equations

$$
\begin{equation*}
\mathrm{p}\left(h_{1} \mid \mathbf{x}\right)=\mathrm{p}\left(h_{2} \mid \mathbf{x}\right) \tag{20}
\end{equation*}
$$

and

$$
\begin{equation*}
\mathrm{p}\left(h_{1}<y<h_{2}\right)=1-\alpha \tag{21}
\end{equation*}
$$

Using (19) in (20) and (21), after tedious algebra, we have

$$
\begin{equation*}
\sum_{p_{1}=0}^{l-m-1} \sum_{q_{1}=0}^{s} \frac{\binom{l-m-1}{p_{1}}\binom{s}{q_{1}}(-1)^{p_{1}+q_{1}}}{\delta_{p_{1} q_{1}}^{m-s+2}}=\sum_{p_{1}=0}^{l-m-1} \sum_{q_{1}=0}^{s} \frac{\binom{l-m-1}{p_{1}}\binom{s}{q_{1}}(-1)^{p_{1}+q_{1}}}{\gamma_{p_{1} 1_{1}}^{m-s+2}} \tag{22}
\end{equation*}
$$

and

## BAROT \& PATEL

$$
\begin{equation*}
\frac{-\sum_{p_{1}=0}^{l-m-1} \sum_{q_{1}=0}^{s}\binom{l-m-1}{p_{1}}\binom{s}{q_{1}}(-1)^{p_{1}+q_{1}} \frac{\left(\gamma_{p_{1} q_{1}}^{-m+s-1}-\delta_{p_{1} q_{1}}^{-m+s-1}\right)}{n_{1}-l+p_{1}+1}}{\mathrm{~B}\left(l-m, n_{1}-l+1\right) D_{S}}=1-\alpha \tag{23}
\end{equation*}
$$

where

$$
\begin{aligned}
& \delta_{p_{1} q_{1}}=k\left(n_{1}-l+p_{1}+1\right) h_{1}+k w+k x_{(s+1)} q_{1}+\beta \\
& \gamma_{p_{1} q_{1}}=k\left(n_{1}-l+p_{1}+1\right) h_{2}+k w+k x_{(s+1)} q_{1}+\beta
\end{aligned}
$$

Hence, the $100(1-\alpha) \%$ Bayes HPD-prediction interval for $x_{(l)}$ is

$$
\begin{equation*}
\left(x_{(m)}+h_{1}, x_{(m)}+h_{2}\right) \tag{24}
\end{equation*}
$$

When the super parameter $\beta$ is unknown, the hierarchical Bayes predictive estimates $x_{(l)}$ and the corresponding $100(1-\alpha) \%$ hierarchical HPD-prediction interval of can be obtained by using the estimate $\hat{\beta}_{B}$ for $\beta$ in (19) and (24).

## Simulation Study

An extensive Monte Carlo simulation study was carried out to illustrate and compare the performance of hierarchical Bayes estimates of reliability indexes of the system with series of $k$ units in working mode and $(n-1)$ units in cold standby mode. The performance is evaluated based on estimated risks and biases for different combinations of sample size ( $n$ ), effective sample size ( $m-s$ ), and general progressive Type II censoring scheme $\mathbf{r}=\left(r_{s+1}, r_{s+2}, \ldots, r_{m}\right)$. The different censoring schemes applied in the simulation study are summarized in Table 1.

For given values $a=0, b=1$ and 100,00,000 generated uniform numbers, two values of $\beta$, one is the true value $\beta_{T}=0.5002$ and another is the expert value $\beta_{E}=0.4999$ were obtained by the Monte Carlo means. The corresponding $\lambda=2.0008$ is brought from the prior (5) and the expert value $\beta_{E}$. Using the generated value of $\lambda$, we have generated a general progressive Type II censored sample $\mathbf{x}=\left(x_{(s+1)}, x_{(s+2), \ldots,} x_{(m)}\right)$ with the censoring scheme $\mathbf{r}$ from the exponential distribution according to the algorithm presented in Balakrishnan and Sandhu (1996) that involves the following steps:

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

1. Generate $V_{m}$ from the Beta distribution with parameters $(n-s)$ and $(s+1)$
2. Independently generate $Z_{s+i}$ from $\mathrm{U}(0,1)$ and then set $V_{s+i}=Z_{s+i}^{\frac{1}{a_{s+i}}}$, $a_{s+i}=i+\sum_{j=m-i+1}^{m} r_{j}$ for $\mathrm{i}=1,2, \ldots,(m-s-1)$
3. Set $U_{s+1}=1-V_{m}$ and $U_{s+i}=1-\left(V_{m-i+1} V_{m-i+2} \ldots V_{m}\right)$, $i=2,3, \ldots,(m-s)$
4. For the generated value of $\lambda$ and given $k, x_{(s+i)}=-\frac{1}{\lambda k} \ln \left(1-U_{s+i}\right)$, $i=1,2, \ldots,(m-s)$ is the required general progressive Type II censored sample of size $(m-s)$ from the exponential distribution

The Bayes estimates, hierarchical Bayes estimates, and the corresponding estimated risks were computed by averaging over 100,000 simulations, and are reported, respectively, in Tables 2-6. From the simulation results, the following points can be drawn:

1) For the fixed sample size $n$ and initial $s$ unobserved failures, as the predetermined number of failures $m$ increases, the estimated risks of estimates of reliability indexes decrease, that is, the performance becomes better in terms of the estimated risks. (Refer to Tables 2-4)
2) For the fixed effective sample size $(m-s)$, the estimated risks of estimates of failure rate $\lambda$ and reliability $\mathrm{R}(t)$ decrease while that of MTTF increase with the increasing sample size $n$. (Refer to Tables 24)
3) For the fixed sample size $n$ and predetermined number of failures $m$, the estimated risks of estimates of failure rate $\lambda$ and reliability $\mathrm{R}(t)$ increase while that of MTTF decrease with the increasing number of initial $s$ unobserved failures. (Refer to Tables 2-4)
4) For the fixed sample size $n$ and effective sample size $(m-s)$, the estimated risks of the estimates of MTTF decrease while that of reliability $\mathrm{R}(t)$ decrease for small sample size and increase for moderate and large sample sizes with increasing number of working units $k$. (Refer to Table 6)
5) It is noted that an increase in $k$ does not have any dampening effect on the estimated risk of failure rate $\lambda$. (Refer to Table 6)

## BAROT \& PATEL

6) The estimated risks of the Bayes estimates of reliability indexes are smaller than the corresponding hierarchical Bayes estimates for all the considered cases. This indicates that Bayes estimates outperform the hierarchical Bayes estimates. (Refer to Tables 2-4)
7) For the fixed effective sample size $(m-s)$, as the sample size $n$ increases, the Bayes and hierarchical Bayes estimates of failure rate $\lambda$ decrease while reliability $\mathrm{R}(t)$ and $M T T F$ increase, i.e., the series system survives for a long period. (Refer to Tables 2-4)
8) For the fixed sample size $n$ and effective sample size $(m-s)$, as the number of working units $k$ increases, the Bayes and hierarchical Bayes estimates of reliability $\mathrm{R}(t)$ and $M T T F$ decrease, i.e., the series system fails frequently. (Refer to Table 5)

Table 1. Progressive Type II censoring schemes (CS) applied to the simulation study

| $\boldsymbol{n}$ | $\boldsymbol{m}$ | $\boldsymbol{s}$ | CS $\mathbf{N o}$ | $\boldsymbol{r}=\left(\boldsymbol{r}_{\boldsymbol{s}+1}, \boldsymbol{r}_{\boldsymbol{s}+2}, \ldots, \boldsymbol{r}_{\boldsymbol{m}}\right)$ |
| ---: | ---: | ---: | ---: | ---: |
| 20 | 8 | 3 | $[1]$ | $(1,0,4,1,6)$ |
|  |  |  | $[2]$ | $(0,0,0,0,12)$ |
|  |  | $[3]$ | $(12,0,0,0,0)$ |  |
|  |  |  |  |  |
|  |  | 4 | $[4]$ | $(2,0,4,6)$ |
|  |  |  | $[5]$ | $(0,0,0,12)$ |
|  |  |  | $[6]$ | $(12,0,0,0)$ |

$\left.\begin{array}{cccr}10 & 3 & {[7]} & (2,0,3,0,1,2,2) \\ & & {[8]} & (0,0,0,0,0,0,10) \\ & & & {[9]}\end{array}\right)$

| $n$ | m | $s$ | CS No. | $\mathbf{r}=\left(\boldsymbol{r}_{s+1}, \boldsymbol{r}_{\text {s+2 }}, \ldots, \boldsymbol{r}_{m}\right)$ |
| :---: | :---: | :---: | :---: | :---: |
| 50 | 10 | 3 | [19] | (6, 8, 10, 4, 3, 7, 2) |
|  |  |  | [20] | (0, 0, 0, 0, 0, 0, 40) |
|  |  |  | [21] | (40, 0, 0, 0, 0, 0, 0) |
|  |  | 4 | [22] | $(6,8,10,4,5,7)$ |
|  |  |  | [23] | (0, 0, 0, 0, 0, 40) |
|  |  |  | [24] | (40, 0, 0, 0, 0, 0) |
| 100 | 8 | 3 | [25] | (16, 12, 20, 14, 30) |
|  |  |  | [26] | (0, 0, 0, 0, 92) |
|  |  |  | [27] | (92, 0, 0, 0, 0) |
|  |  | 4 | [28] | (28, 25,17, 22) |
|  |  |  | [29] | (0, 0, 0, 92) |
|  |  |  | [30] | (92, 0, 0, 0) |
|  | 10 | 3 | [31] | ( $6,13,15,14,8,12,22)$ |
|  |  |  | [32] | (0, 0, 0, 0, 0, 0, 90) |
|  |  |  | [33] | $(90,0,0,0,0,0,0)$ |
|  |  | 4 | [34] | (16, 18, 15, 14, 15, 12) |
|  |  |  | [35] | (0, 0, 0, 0, 0, 90) |
|  |  |  | [36] | 90, 0, 0, 0, 0, 0 |

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

Table 2. Estimates of failure rate of $\lambda$ and their estimated risks

| CS | $\hat{\beta}$ | $\hat{\lambda}_{B}$ | $\hat{\lambda}_{H B}$ | $\operatorname{ER}\left(\hat{\lambda}_{B}\right)$ | $\operatorname{ER}\left(\hat{\lambda}_{\text {HB }}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.5257 | 2.2951 | 2.3031 | 0.6755 | 0.7867 |
| 2 | 0.5273 | 2.2706 | 2.2772 | 0.6508 | 0.7554 |
| 3 | 0.5078 | 2.5724 | 2.6029 | 1.0682 | 1.3812 |
| 4 | 0.5229 | 2.3396 | 2.3504 | 0.7288 | 0.8545 |
| 5 | 0.5250 | 2.3066 | 2.3154 | 0.6917 | 0.8070 |
| 6 | 0.5005 | 2.6863 | 2.7278 | 1.2986 | 1.6107 |
| 7 | 0.5234 | 2.3322 | 2.3365 | 0.6178 | 0.6879 |
| 8 | 0.5280 | 2.2616 | 2.2629 | 0.5479 | 0.6060 |
| 9 | 0.5098 | 2.5440 | 2.5598 | 0.8976 | 1.0207 |
| 10 | 0.5192 | 2.3984 | 2.4059 | 0.6934 | 0.7772 |
| 11 | 0.5246 | 2.3146 | 2.3181 | 0.5986 | 0.6655 |
| 12 | 0.4985 | 2.7215 | 2.7505 | 1.2238 | 1.4206 |
| 13 | 0.5279 | 2.2612 | 2.2676 | 0.6451 | 0.7537 |
| 14 | 0.5300 | 2.2301 | 2.2348 | 0.6168 | 0.7176 |
| 15 | 0.5078 | 2.5717 | 2.6025 | 1.0667 | 1.2980 |
| 16 | 0.5269 | 2.2779 | 2.2850 | 0.6593 | 0.7669 |
| 17 | 0.5292 | 2.2422 | 2.2473 | 0.6250 | 0.7233 |
| 18 | 0.5006 | 2.6853 | 2.7265 | 1.2945 | 1.6056 |
| 19 | 0.5269 | 2.2790 | 2.2808 | 0.5607 | 0.6197 |
| 20 | 0.5317 | 2.2033 | 2.2042 | 0.4952 | 0.5432 |
| 21 | 0.5098 | 2.5438 | 2.5593 | 0.8921 | 1.0120 |
| 22 | 0.5265 | 2.2845 | 2.2865 | 0.5639 | 0.6223 |
| 23 | 0.5307 | 2.2195 | 2.2199 | 0.5060 | 0.5548 |
| 24 | 0.4985 | 2.7209 | 2.7495 | 1.2191 | 1.4100 |
| 25 | 0.5299 | 2.2310 | 2.2358 | 0.6177 | 0.7183 |
| 26 | 0.5305 | 2.2219 | 2.2262 | 0.6038 | 0.6983 |
| 27 | 0.5076 | 2.5706 | 2.6017 | 1.0657 | 1.2978 |
| 28 | 0.5286 | 2.2511 | 2.2564 | 0.6259 | 0.7249 |
| 29 | 0.5301 | 2.2282 | 2.2323 | 0.6054 | 0.6989 |
| 30 | 0.5004 | 2.6824 | 2.7254 | 1.2902 | 1.5969 |
| 31 | 0.5317 | 2.2031 | 2.2040 | 0.4944 | 0.5427 |
| 32 | 0.5326 | 2.1878 | 2.1883 | 0.4845 | 0.5312 |
| 33 | 0.5098 | 2.5424 | 2.5588 | 0.8907 | 1.0105 |
| 34 | 0.5296 | 2.2362 | 2.2365 | 0.5266 | 0.5806 |
| 35 | 0.5321 | 2.1971 | 2.1970 | 0.4951 | 0.5437 |
| 36 | 0.4985 | 2.7203 | 2.7403 | 1.2161 | 1.4044 |

## BAROT \& PATEL

Table 3. Estimates of reliability $\mathrm{R}(t)$ and their estimated risks

| CS | $\mathbf{R}(t)$ | $\hat{\mathbf{R}}(\boldsymbol{t})_{B}$ | $\hat{\mathbf{R}}(\boldsymbol{t})_{\text {нв }}$ | $\operatorname{ER}\left(\hat{\mathbf{R}}(\boldsymbol{t})_{B}\right)$ | $\operatorname{ER}\left(\hat{\mathbf{R}}(\boldsymbol{t})_{\text {Нв }}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.4695 | 0.4554 | 0.4600 | 0.0598 | 0.0623 |
| 2 |  | 0.4638 | 0.4686 | 0.0601 | 0.0617 |
| 3 |  | 0.3692 | 0.3717 | 0.0646 | 0.0661 |
| 4 |  | 0.4410 | 0.4453 | 0.0600 | 0.0625 |
| 5 |  | 0.4520 | 0.4564 | 0.0603 | 0.0626 |
| 6 |  | 0.3410 | 0.3428 | 0.0789 | 0.0815 |
| 7 |  | 0.4249 | 0.4289 | 0.0586 | 0.0604 |
| 8 |  | 0.4505 | 0.4549 | 0.0583 | 0.0601 |
| 9 |  | 0.3552 | 0.3578 | 0.0642 | 0.0659 |
| 10 |  | 0.4019 | 0.4055 | 0.0596 | 0.0613 |
| 11 |  | 0.4311 | 0.4352 | 0.0584 | 0.0602 |
| 12 |  | 0.3060 | 0.3075 | 0.0734 | 0.0752 |
| 13 | 0.9999 | 0.9643 | 0.9676 | 0.0069 | 0.0092 |
| 14 |  | 0.9665 | 0.9696 | 0.0063 | 0.0085 |
| 15 |  | 0.9361 | 0.9429 | 0.0144 | 0.0195 |
| 16 |  | 0.9631 | 0.9666 | 0.0071 | 0.0095 |
| 17 |  | 0.9657 | 0.9689 | 0.0065 | 0.0087 |
| 18 |  | 0.9219 | 0.9307 | 0.0190 | 0.0259 |
| 19 |  | 0.9722 | 0.9745 | 0.0048 | 0.0063 |
| 20 |  | 0.9768 | 0.9786 | 0.0039 | 0.0050 |
| 21 |  | 0.9513 | 0.9558 | 0.0101 | 0.0131 |
| 22 |  | 0.9719 | 0.9742 | 0.0049 | 0.0064 |
| 23 |  | 0.9759 | 0.9778 | 0.0040 | 0.0052 |
| 24 |  | 0.9318 | 0.9387 | 0.0159 | 0.0209 |
| 25 | 1.0000 | 0.9993 | 0.9997 | $1.136 \times 10^{-5}$ | $1.105 \times 10^{-5}$ |
| 26 |  | 0.9993 | 0.9997 | $1.079 \times 10^{-5}$ | $1.172 \times 10^{-5}$ |
| 27 |  | 0.9981 | 0.9991 | $6.165 \times 10^{-5}$ | $3.532 \times 10^{-5}$ |
| 28 |  | 0.9993 | 0.9996 | $1.934 \times 10^{-5}$ | $1.141 \times 10^{-5}$ |
| 29 |  | 0.9993 | 0.9997 | $1.798 \times 10^{-5}$ | $1.262 \times 10^{-5}$ |
| 30 |  | 0.9973 | 0.9988 | $8.788 \times 10^{-5}$ | $5.091 \times 10^{-5}$ |
| 31 |  | 0.9997 | 0.9999 | $6.051 \times 10^{-5}$ | $3.037 \times 10^{-5}$ |
| 32 |  | 0.9997 | 0.9999 | $5.034 \times 10^{-5}$ | $2.081 \times 10^{-5}$ |
| 33 |  | 0.9993 | 0.9996 | $2.221 \times 10^{-5}$ | $1.098 \times 10^{-5}$ |
| 34 |  | 0.9997 | 0.9998 | $6.072 \times 10^{-5}$ | $3.081 \times 10^{-5}$ |
| 35 |  | 0.9997 | 0.9998 | $5.092 \times 10^{-5}$ | $2.594 \times 10^{-5}$ |
| 36 |  | 0.9986 | 0.9993 | $4.714 \times 10^{-5}$ | $2.338 \times 10^{-5}$ |

HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

Table 4. Estimates of MTTF and their estimated risks

| CS | MTTF | MTTF $_{\text {B }}$ | MTTF $_{\text {HB }}$ | ER( MTTF $_{\text {B }}$ ) | $\operatorname{ER}\left(\right.$ MTTF $\left._{\text {нв }}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1.9991 | 2.1663 | 2.1792 | 0.4857 | 0.5213 |
| 2 |  | 2.1901 | 2.2039 | 0.5051 | 0.5417 |
| 3 |  | 1.9329 | 1.9369 | 0.3798 | 0.4136 |
| 4 |  | 2.1270 | 2.1386 | 0.4620 | 0.4965 |
| 5 |  | 2.1576 | 2.1702 | 0.4834 | 0.5190 |
| 6 |  | 1.8599 | 1.8602 | 0.3788 | 0.4116 |
| 7 |  | 2.0519 | 2.0613 | 0.3475 | 0.3687 |
| 8 |  | 2.1164 | 2.1276 | 0.3809 | 0.4039 |
| 9 |  | 1.8815 | 1.8855 | 0.3062 | 0.3255 |
| 10 |  | 1.9954 | 2.0032 | 0.3271 | 0.3473 |
| 11 |  | 2.0677 | 2.0776 | 0.3552 | 0.3769 |
| 12 |  | 1.7626 | 1.7630 | 0.3027 | 0.3239 |
| 13 | 4.9979 | 5.5000 | 5.5351 | 3.2263 | 3.4580 |
| 14 |  | 5.5787 | 5.6162 | 3.4049 | 3.6458 |
| 15 |  | 4.8297 | 4.8395 | 2.3392 | 2.5313 |
| 16 |  | 5.4589 | 5.4926 | 3.1398 | 3.3672 |
| 17 |  | 5.5472 | 5.5838 | 3.3299 | 3.5669 |
| 18 |  | 4.6484 | 4.6492 | 2.3318 | 2.5252 |
| 19 |  | 5.2490 | 5.2760 | 2.3295 | 2.4697 |
| 20 |  | 5.4308 | 5.4626 | 2.6177 | 2.7724 |
| 21 |  | 4.7025 | 4.7124 | 1.9184 | 2.2388 |
| 22 |  | 5.2357 | 5.2623 | 2.3140 | 2.4531 |
| 23 |  | 5.3902 | 5.4209 | 2.5499 | 2.7009 |
| 24 |  | 4.4066 | 4.4072 | 1.8230 | 2.1551 |
| 25 | 9.9958 | 11.1494 | 11.2244 | 13.5400 | 14.4996 |
| 26 |  | 11.1964 | 11.2729 | 13.7659 | 14.7371 |
| 27 |  | 9.6457 | 9.6649 | 9.2994 | 10.1653 |
| 28 |  | 11.0364 | 11.1081 | 12.9450 | 13.8712 |
| 29 |  | 11.1528 | 11.2282 | 13.4762 | 14.4297 |
| 30 |  | 9.2739 | 9.2750 | 9.2443 | 10.1482 |
| 31 |  | 10.8628 | 10.9265 | 10.5038 | 11.1230 |
| 32 |  | 10.9300 | 10.9954 | 10.7570 | 11.3888 |
| 33 |  | 9.4016 | 9.4214 | 7.6860 | 8.1665 |
| 34 |  | 10.7068 | 10.7662 | 10.0085 | 10.6043 |
| 35 |  | 10.8999 | 10.9644 | 10.6881 | 11.3177 |
| 36 |  | 8.8144 | 8.8156 | 7.0966 | 7.6261 |

## BAROT \& PATEL

Table 5. The effect of $k$ on the estimates of reliability indexes

| CS | $k$ | $\hat{\lambda}_{B}$ | $\hat{\lambda}_{\text {HB }}$ | $\hat{\mathrm{R}}(\boldsymbol{t})_{B}$ | $\hat{\mathbf{R}}(\boldsymbol{t})_{\text {нв }}$ | MTTF $_{\text {B }}$ | MTTF $_{\text {HB }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 4 | 2.2951 | 2.3031 | 0.6261 | 0.6282 | 2.7078 | 2.7240 |
|  | 8 |  |  | 0.1543 | 0.1587 | 1.3539 | 1.3620 |
|  | 12 |  |  | 0.0371 | 0.0388 | 0.9026 | 0.9080 |
| 4 | 4 | 2.3396 | 2.3504 | 0.6118 | 0.6136 | 2.6588 | 2.6732 |
|  | 8 |  |  | 0.1464 | 0.1506 | 1.3294 | 1.3366 |
|  | 12 |  |  | 0.0346 | 0.0362 | 0.8863 | 0.8911 |
| 7 | 4 | 2.3322 | 2.3365 | 0.6095 | 0.6115 | 2.5648 | 2.3365 |
|  | 8 |  |  | 0.1210 | 0.1244 | 1.2824 | 1.2883 |
|  | 12 |  |  | 0.0225 | 0.0236 | 0.8549 | 0.8589 |
| 10 | 4 | 2.3984 | 2.4059 | 0.5866 | 0.5881 | 2.4943 | 2.5040 |
|  | 8 |  |  | 0.1099 | 0.1129 | 1.2471 | 1.2520 |
|  | 12 |  |  | 0.0198 | 0.0206 | 0.8314 | 0.8346 |
| 13 | 4 | 2.2612 | 2.2675 | 0.9896 | 0.9873 | 6.8750 | 6.9188 |
|  | 8 |  |  | 0.8067 | 0.8056 | 3.4375 | 3.4594 |
|  | 12 |  |  | 0.5095 | 0.5144 | 2.2917 | 2.3063 |
| 16 | 4 | 2.2779 | 2.2850 | 0.9893 | 0.9869 | 6.8236 | 6.8658 |
|  | 8 |  |  | 0.8022 | 0.8010 | 3.4118 | 3.4329 |
|  | 12 |  |  | 0.5033 | 0.5080 | 2.2745 | 2.2886 |
| 19 | 4 | 2.2790 | 2.2808 | 0.9930 | 0.9916 | 6.5613 | 6.5950 |
|  | 8 |  |  | 0.8123 | 0.8115 | 3.2806 | 3.2975 |
|  | 12 |  |  | 0.4877 | 0.4923 | 2.1871 | 2.1983 |
| 22 | 4 | 2.2845 | 2.2865 | 0.9929 | 0.9915 | 6.5446 | 6.5779 |
|  | 8 |  |  | 0.8108 | 0.8100 | 3.2723 | 3.2889 |
|  | 12 |  |  | 0.4853 | 0.4898 | 2.1815 | 2.1926 |
| 25 | 4 | 2.2310 | 2.2358 | 0.9999 | 0.9998 | 13.9367 | 14.0305 |
|  | 8 |  |  | 0.9919 | 0.9897 | 6.9684 | 7.0152 |
|  | 12 |  |  | 0.9382 | 0.9348 | 4.6456 | 4.6768 |
| 28 | 4 | 2.2511 | 2.2564 | 0.9999 | 0.9998 | 13.7955 | 13.8852 |
|  | 8 |  |  | 0.9917 | 0.9896 | 6.8977 | 6.9426 |
|  | 12 |  |  | 0.9362 | 0.9327 | 4.5985 | 4.6284 |
| 31 | 4 | 2.2031 | 2.2020 | 0.9999 | 0.9999 | 13.5785 | 13.6581 |
|  | 8 |  |  | 0.9955 | 0.9945 | 6.7892 | 6.8290 |
|  | 12 |  |  | 0.9527 | 0.9505 | 4.5261 | 4.5527 |
| 34 | 4 | 2.2362 | 2.2365 | 0.9999 | 0.9999 | 13.3835 | 13.4578 |
|  | 8 |  |  | 0.9949 | 0.9937 | 6.6917 | 6.7289 |
|  | 12 |  |  | 0.9489 | 0.9465 | 4.4611 | 4.4859 |

Table 6. The effect of $k$ on the estimated risks of estimates of reliability indexes

| CS | $k$ | $\operatorname{ER}\left(\hat{\lambda}_{B}\right)$ | $\operatorname{ER}\left(\hat{\lambda}_{\text {HB }}\right)$ | $\operatorname{ER}\left(\hat{\mathbf{R}}\left(t_{B}\right)\right.$ | $\operatorname{ER}\left(\hat{\mathbf{R}}\left(\boldsymbol{t}_{\text {HB }}\right)\right.$ | $\operatorname{ER}\left(\right.$ MTTF $\left._{\text {B }}\right)$ | $\operatorname{ER}\left(\right.$ MTTF $\left._{\text {HB }}\right)$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 4 | 0.6755 | 0.7867 | 0.0927 | 0.0950 | 0.7590 | 0.8146 |
|  | 8 |  |  | 0.0428 | 0.0454 | 0.1897 | 0.2036 |
|  | 12 |  |  | 0.0043 | 0.0047 | 0.0843 | 0.0905 |
| 4 | 4 | 0.7288 | 0.8545 | 0.0993 | 0.1018 | 0.7218 | 0.7758 |
|  | 8 |  |  | 0.0394 | 0.0418 | 0.1804 | 0.1939 |
|  | 12 |  |  | 0.0039 | 0.0042 | 0.0802 | 0.0862 |
| 7 | 4 | 0.6178 | 0.6879 | 0.0985 | 0.1002 | 0.5429 | 0.5761 |
|  | 8 |  |  | 0.0284 | 0.0300 | 0.1357 | 0.1440 |
|  | 12 |  |  | 0.0019 | 0.0021 | 0.0603 | 0.0640 |
| 10 | 4 | 0.6934 | 0.7772 | 0.1092 | 0.1112 | 0.5111 | 0.5427 |
|  | 8 |  |  | 0.0242 | 0.0256 | 0.1278 | 0.1357 |
|  | 12 |  |  | 0.0015 | 0.0017 | 0.0568 | 0.0603 |
| 13 | 4 | 0.6451 | 0.7537 | 0.0015 | 0.0026 | 5.0411 | 5.4031 |
|  | 8 |  |  | 0.0790 | 0.0829 | 1.2603 | 1.3508 |
|  | 12 |  |  | 0.0871 | 0.0891 | 0.5601 | 0.6003 |
| 16 |  | 0.6593 | 0.7669 | 0.0015 | 0.0027 | 4.9060 | 5.2612 |
|  | 8 |  |  | 0.0815 | 0.0856 | 1.2265 | 1.3153 |
|  | 12 |  |  | 0.0882 | 0.0902 | 0.5451 | 0.5846 |
| 19 |  | 0.5607 | 0.6197 | 0.0009 | 0.0014 | 3.6399 | 3.8589 |
|  | 8 |  |  | 0.0752 | 0.0782 | 0.9100 | 0.9647 |
|  | 12 |  |  | 0.0804 | 0.0818 | 0.4044 | 0.4288 |
| 22 | , | 0.5640 | 0.6223 | 0.0009 | 0.0014 | 3.6156 | 3.8330 |
|  | 8 |  |  | 0.0759 | 0.0790 | 0.9039 | 0.9582 |
|  | 12 |  |  | 0.0808 | 0.0821 | 0.4017 | 0.4259 |
| 25 |  | 0.6177 | 0.7183 | $2.071 \times 10^{-6}$ | $3.023 \times 10^{-5}$ | 21.1563 | 22.6557 |
|  | 8 |  |  | 0.0012 | 0.0023 | 5.2891 | 5.6639 |
|  | 12 |  |  | 0.0176 | 0.0210 | 2.3507 | 2.5173 |
| 28 |  | 0.6259 | 0.7249 | $1.565 \times 10^{-6}$ | $2.568 \times 10^{-5}$ | 20.2266 | 21.6737 |
|  | 8 |  |  | 0.0012 | 0.0022 | 5.0566 | 5.4184 |
|  | 12 |  |  | 0.0180 | 0.0214 | 2.2474 | 2.4082 |
| 31 |  | 0.4944 | 0.5427 | $4.014 \times 10^{-7}$ | $5.444 \times 10^{-6}$ | 16.4122 | 17.3797 |
|  | 8 |  |  | 0.0006 | 0.0010 | 4.1030 | 4.3449 |
|  | 12 |  |  | 0.0121 | 0.0141 | 1.8236 | 1.9311 |
| 34 | 4 | 0.5266 | 0.5806 | $3.553 \times 10^{-7}$ | $4.667 \times 10^{-6}$ | 15.6383 | 16.5692 |
|  | 8 |  |  | 0.0007 | 0.0011 | 3.9096 | 4.1423 |
|  | 12 |  |  | 0.0136 | 0.0159 | 1.7376 | 1.8410 |

## Numerical Examples

Two numerical examples are presented to illustrate how the data support the developed model and how to employ the proposed method for estimation of reliability indexes of the series system. Examples 1 and 2 consider the artificial

## BAROT \& PATEL

general progressive Type II censored samples generated from the real data set provided by Nelson (1982) and the computer simulation, respectively.

## Example 1. Real Life Data

As a numerical illustration, a system comprising a series of 2 working units and 18 cold standby units was considered. This series system is equivalent to a cold standby series system of 19 identical and independent units. The lifetimes of such 19 units were observed until failure during the life test experiment in which specimens of a type of electrical insulating fluid were subject to a constant voltage stress ( $34 \mathrm{KV} /$ minutes). The 19 failure times were obtained as follows:

| 0.19 | 0.78 | 0.96 | 1.31 | 2.78 | 3.16 | 4.15 | 4.67 | 4.85 | 6.50 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 7.35 | 8.01 | 8.27 | 12.06 | 31.75 | 32.52 | 33.94 | 36.71 | 72.89 |  |

Asgharzadeh and Valiollahi (2009) checked the validity of an exponential model with mean $=14.2857$ and indicated that the exponential model is adequate for this data set. To generate an artificial general progressive Type II censored sample from the given real data set, it is assumed that the lifetimes of the first two failures are lost without observation, and then lifetimes were observed until the eighth failure. At each failure from $3^{\text {rd }}$ failure to $8^{\text {th }}$ failure, units were randomly withdrawn according to the general progressive Type II censoring scheme $\mathbf{r}=\left(r_{3}, r_{4}, \ldots, r_{8}\right)$ $=(2,0,1,2,1,5)$. The life test was terminated at the eighth failure, and the vector of observed lifetimes was found to be $\mathbf{x}=\left(x_{(3)}, x_{(4)}, \ldots, x_{(8)}\right)$ $=(0.96,1.31,2.78,4.85,6.50,8.01)$.

Table 7. Estimates of reliability indexes and their $(1-\alpha) \%$ HPD-intervals for Example 1

|  | Parameter | Estimate | 95\% HPD-interval | 99\% HPD-interval |
| ---: | ---: | ---: | ---: | ---: |
| Bayes Estimation |  | 0.0519 | $(0.0209,0.0866)$ | $(0.0107,0.1239)$ |
|  | $\mathrm{R}(\mathrm{t})$ | 0.9415 | $(0.6257,0.9999)$ | $(0.0992,0.9999)$ |
|  | MTTF | 205.5657 | $(109.7102,454.5454)$ | $(76.6798,887.8505)$ |
| Hierarchical |  |  |  |  |
| Bayes Estimation | $\mathrm{R}(\mathrm{t})$ | 0.0519 | $(0.0203,0.0880)$ | $(0.0120,0.1172)$ |
|  | MTTF | 205.7603 | $(0.5988,0.9999)$ | $(0.1526,0.9999)$ |
|  |  |  |  | $(81.0286,791.6666)$ |

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

Table 8. Predictive estimates of the remaining lifetimes and their $(1-\alpha) \%$ HPDprediction intervals for Example 1

|  | $\boldsymbol{I}$ | $\boldsymbol{X}_{(\boldsymbol{l}}$ | 95\% HPD-interval | 99\% HPD-interval |
| ---: | ---: | ---: | ---: | ---: |
| Bayes Estimation | 9 | 10.1738 | $(8.0262,15.3444)$ | $(8.0206,22.2503)$ |
|  | 10 | 12.8786 | $(8.0218,20.7168)$ | $(8.0202,27.9367)$ |
|  | 11 | 16.4851 | $(8.0418,28.2399)$ | $(8.0251,38.7317)$ |
|  | 12 | 21.8947 | $(8.1446,39.6961)$ | $(8.0996,55.4946)$ |
|  | 13 | 32.7139 | $(8.0212,46.9190)$ | $(8.0201,61.6929)$ |
| Hierarchical Bayes | 9 | 10.1758 | $(8.0470,16.1734)$ | $(8.0244,24.3241)$ |
| Estimation | 10 | 12.8833 | $(8.0267,20.7303)$ | $(8.0377,27.9939)$ |
|  | 11 | 16.4931 | $(8.0355,28.2591)$ | $(8.0298,38.7608)$ |
|  | 12 | 21.9078 | $(8.0907,39.7260)$ | $(8.0971,55.5395)$ |
|  | 13 | 32.7373 | $(8.0261,46.9559)$ | $(8.0241,61.7342)$ |

The Bayes and hierarchical Bayes estimates of failure rate $\lambda$, reliability $\mathrm{R}(t)$, and MTTF and the corresponding HPD-intervals at $t=100$ have been computed, and are reported in Table 7. The $95 \%$ and $99 \%$ Bayes and hierarchical Bayes predictive estimates and the corresponding HPD-prediction intervals for the each of the remaining $l$ lifetimes $(9 \leq l \leq)$ have also been computed, and are reported in Table 8.

## Example 2. Simulated Data

As a numerical illustration, a system initiated with the series of 5 working units being in an operational state is placed on a life test along with the other 19 standby units connected in a series. This series system is equivalent to a cold standby series system of 20 identical and independent units. Under a general progressive Type II censoring scheme, the lifetimes of the first two failures are not observed and then the lifetimes are completely observed until the eighth failure. Using the algorithm presented in the previous section, the general progressive Type II censored sample $\mathbf{x}=(0.01250,0.01531,0.02063,0.02679,0.03062,0.05251)$ has been generated with the censoring scheme $\mathbf{r}=(1,0,2,1,2,6)$. For this sample, Bayes and hierarchical Bayes estimates of failure rate $\lambda$, reliability $\mathrm{R}(t)$, and $M T T F$, and the corresponding HPD intervals at $t=2$, have been computed and are reported in Table 9. Moreover, the $95 \%$ and $99 \%$ Bayes and hierarchical Bayes predictive estimates and the corresponding HPD-prediction intervals for each of the remaining $l$ lifetimes ( $9 \leq l \leq$ ) have also been computed, and are reported in Table 10.

## BAROT \& PATEL

Table 9. Estimates of reliability indexes and their $(1-\alpha) \%$ HPD-intervals for Example 2

|  | Parameter | Estimate | 95\% HPD-interval | 99\% HPD-interval |
| ---: | ---: | ---: | ---: | ---: |
| Bayes Estimation | $\lambda$ | 2.4747 | $(0.9950,4.1210)$ | $(0.7440,4.8827)$ |
|  | $\mathrm{R}(t)$ | 0.3201 | $\left(9.053 \times 10^{-5}, 0.9967\right)$ | $\left(9.9 \times 10^{-7}, 0.9999\right)$ |
|  | $M T T F$ | 1.8184 | $(0.9706,4.0201)$ | $(0.8192,5.3763)$ |
| Hierarchical Bayes |  |  |  |  |
| Estimation | $\mathrm{R}(t)$ | 2.4650 | $(0.9910,4.1050)$ | $(0.74800,4.8390)$ |
|  | MTTF | 1.8256 | $\left(9.895 \times 10^{-5}, 0.9968\right)$ | $\left(1.31 \times 10^{-6}, 0.9999\right)$ |
|  |  |  | $(0.9744,4.0363)$ | $(0.82661,5.3476)$ |

Table 10. Predictive estimates of the remaining lifetimes and their $(1-\alpha) \%$ HPDprediction intervals for Example 2

|  | $\boldsymbol{I}$ | $\boldsymbol{x}_{(i)}$ | 95\% HPD-interval | 99\% HPD-interval |
| ---: | ---: | ---: | ---: | ---: |
| Bayes Estimation | 9 | 0.0676 | $(0.0527,0.1079)$ | $(0.0526,0.1661)$ |
|  | 10 | 0.0858 | $(0.0526,0.1394)$ | $(0.0526,0.1888)$ |
|  | 11 | 0.1086 | $(0.0530,0.1859)$ | $(0.0526,0.2546)$ |
|  | 12 | 0.1389 | $(0.0526,0.2477)$ | $(0.0526,0.3431)$ |
|  | 13 | 0.1843 | $(0.0562,0.3429)$ | $(0.0526,0.4821)$ |
| Hierarchical Bayes | 9 | 0.0677 |  |  |
| Estimation | 10 | 0.0860 | $(0.0529,0.1201)$ | $(0.0526,0.1662)$ |
|  | 11 | 0.1088 | $(0.0527,0.1398)$ | $(0.0526,0.1893)$ |
|  | 12 | 0.1392 | $(0.0526,0.1864)$ | $(0.0526,0.2554)$ |
|  | 13 | 0.1848 | $(0.0526,0.3440)$ | $(0.0526,0.3442)$ |
|  |  |  | $(0.0526,0.4838)$ |  |

From the results presented in Tables 7-10, it is observed that the hierarchical Bayes estimates and predictors are very close to the Bayes estimates and predictors for both the considered real and simulated data. Furthermore, the Bayes and hierarchical Bayes predictive estimates and the length of the HPD-prediction interval increases as $l$ increases. This implies that the prediction is less precise as a large $l$ is considered.

## Conclusion

This purpose of this study was to study hierarchical Bayes estimation and prediction of reliability indexes and remaining lifetimes of a cold standby series system consisting a series of $k$ working units and $(n-1)$ cold standby units under general progressive Type II censoring scheme. The Bayes and hierarchical Bayes estimates as well as an HPD interval for reliability indexes of the series system are derived. In addition, we have derived the Bayes and hierarchical Bayes predictive estimates,

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

and HPD-prediction interval for the remaining lifetimes based on an informative sample. We have presented two numerical examples to illustrate the proposed estimation and prediction methods. The Monte Carlo simulation study is carried out to examine and compare the performance of the Bayes and hierarchical Bayes estimates. The simulation results indicated Bayes estimation should be preferred over the hierarchical Bayes estimation for estimation of reliability indexes of the series system. Furthermore, the number of components in the working condition should be less and the number of components in the cold standby mode should be large to run the series system for a long period.

## References

Ali Mousa, M. A. M. (2001). Inference and prediction for Pareto progressively censored data. Journal of Statistical Computation and Simulation, 71(2), 163-181. doi: 10.1080/00949650108812140

Ali Mousa, M. A. M., \& Jaheen, Z. F. (2002). Bayesian prediction for progressively censored data from the Burr model. Statistical Papers, 43, 587-593. doi: 10.1007/s00362-002-0126-7

Asgharzadeh, A., \& Valiollahi, R. (2009). Inference for the proportional hazards family under progressive Type-II censoring. Journal of the Iranian Statistical Society, 8(1-2), 35-53. Retrieved from http://jirss.irstat.ir/article-1-75en.pdf

Bai, C. G., Yu, M. H., \& Hu, S. X. (1998). Multiple Bayes estimates of the reliability performance for cold standby system. Journal of Zhejiang University, 1998(3), 325-328. Available from http://en.cnki.com.cn/Article_en/CJFDTotalZDZC803.011.htm

Balakrishnan, N., \& Sandhu, R. A. (1996). Best linear unbiased and maximum likelihood estimation for exponential distributions under general progressive Type II censored samples. Sankhyā: The Indian Journal of Statistics, Series B, 58(1), 1-9. Available from http://www.jstor.org/stable/25052923

Barot, D. R., \& Patel M. N. (2014). Bayesian estimation of reliability indexes for cold standby system under general progressive Type II censored data. International Journal Quality \& Reliability Management, 31(3), 311-343. doi: 10.1108/IJQRM-08-2012-0114

Cao, J., \& Cheng, K. (1986). Ke kao xing shu xue yin lun [The theory of reliability mathematics]. Beijing: Science Press.

## BAROT \& PATEL

Chhikara, R. S., \& Guttman, I. (1982). Prediction limits for the inverse Gaussian distribution. Technometrics, 24(4), 319-324. doi: 10.1080/00401706.1982.10487794

Dunsmore, I. R. (1974). The Bayesian predictive distribution in life testing models. Technometrics, 16(3), 455-460. doi: 10.1080/00401706.1974.10489216

Fernández, A. J. (2000). Bayesian inference from type II doubly censored Rayleigh data. Statistics \& Probability Letters, 48(4), 393-399. doi: 10.1016/S0167-7152(00)00021-3

Fernández, A. J. (2004). On estimating exponential parameters with general Type II progressive censoring. Journal of Statistical Planning and Inference, 121(1), 135-147. doi: 10.1016/S0378-3758(03)00083-1

Han, M. (1998). Wu shi xiao shu ju de duo ceng Bayes ke kao xing fen xi [The hierarchical Bayesian reliability analysis of zero-failure data]. Mathematica Applicata, 11(2), 131-134.

Howlader, H. A. (1985). HPD prediction intervals for Rayleigh distribution. IEEE Transactions on Reliability, R-34(2), 121-123. doi: 10.1109/TR.1985.5221968

Kim, C., \& Han, K. (2009). Estimation of the scale parameter of the Rayleigh distribution under general progressive censoring. Journal of the Korean Statistical Society, 38(3), 239-246. doi: 10.1016/j.jkss.2008.10.005

Lehmann, E. L., \& Casella, G. (1998). Theory of point estimation. New York: Springer.

Mei, Q., Liao, J., \& Sun, H. (1992). Kěkào xìng gōngchéng de gāi xìtǒng de jīchǔ [The foundation of reliability engineering for the system]. Beijing: Science Press.

Nelson, W. (1982). Applied life data analysis. New York: Wiley.
Papadopoulos, A. S., Tiwari, R. C., \& Zalkikar, J. N. (1996). Hierarchical Bayes estimation for the exponential-multinomial model in reliability and competing risks. IEEE Transactions on Reliability, 45(3), 477-484. doi: 10.1109/24.537019

Pham, T. G., \& Turkkan, N. (1994). Reliability of a standby system with beta-distributed component lives. IEEE Transactions on Reliability, 43(1), 71-75. doi: 10.1109/24.285114

Raqab, M. Z., \& Madi, M. T. (2002). Bayesian prediction of the total time on test using doubly censored Rayleigh data. Journal of Statistical Computation and Simulation, 72(10), 781-789. doi: 10.1080/00949650214670

## HIERARCHICAL BAYES ESTIMATION OF RELIABILITY INDEXES

$\mathrm{Su}, \mathrm{Y} ., \& \mathrm{Gu}, \mathrm{G} .(2003)$. Bayes estimation for reliability indexes of cold standby system. Journal of North China Electric Power University, 30(2), 96-99. Available from http://en.cnki.com.cn/Article_en/CJFDTOTALHBDL200302022.htm

Willits, C. J. (1997). Series-system reliability-estimation using very small binomial samples. IEEE Transactions on Reliability, 46(2), 296-301. doi: 10.1109/24.589960

Xu, Y., Kang, H., \& Shi, Y. (2002). Bayesian and multiple Bayesian analysis of the reliability performances for series system with cold standby units. Chinese quarterly Journal of Mathematics, 17(2), 26-30. Retrieved from http://166.111.121.20:9080/mathjournal/SXJK200202/sxjk200202005.caj.pdf

Younes, H., Delampady, M., MacGibbon, B., \& Cherkaoui, O. (2007). A hierarchical Bayesian approach to the estimation of monotone hazard rates in the random right censorship model. Journal of Statistical Research, 41(2), 35-62. Retrieved from
http://www.isrt.ac.bd/sites/default/files/jsrissues/v41n2/v41n2p35.pdf

# Bayesian Analysis of Discrete Skewed Laplace Distribution 

A. Hossianzadeh<br>Islamic Azad University, Marvdasht Branch<br>Marvdasht, Iran

K. Zare<br>Islamic Azad University, Marvdasht Branch<br>Marvdasht, Iran

The discrete skewed Laplace distribution is a flexible distribution with integer domain and simple closed form that can be applied to model count data. Parameters are estimated under empirical Bayes (EB) analysis and comparison are made between the Bayesian parameter estimation and classical parameter estimation, i.e. the maximum likelihood (ML) approach. The results show that the Bayesian parameter estimations are preferable.

Keywords: Empirical Bayes, discrete skewed Laplace distribution, Bayesian parameter

## Introduction

Skewed distributions are a non-normality of interest (Azzalini, 1985). For example, discrete skewed distributions could be used to model count data. One way to produce discrete skewed distribution is based on the survival function to corresponded continuous case (Roy \& Dasgupta, 2001; Roy, 2003). The discretization for continuous distributions based on the positive real numbers produce discreet distributions on the positive integer numbers, such as the discrete Gamma, Weibull, and negative binomial distributions in Chakraborty and Chakravarty (2012). Roy (2004) presented the discrete Rayleigh distribution and Krishna and Sing (2009) investigated discrete Burr and Pareto. One of the flexible discrete skewed distributions that is defined by Barbiero (2014) on the whole integer numbers is the discrete skewed Laplace distribution. A main advantage for this distribution is the closed and simple forms of its probability function, distribution function, mathematical expectation, and variance. The purpose of this study is to present empirical Bayesian analysis for the parameter estimation.

[^36]
## HOSSIANZADEH \& ZARE

## Discrete Skewed Laplace Distribution

Yu and Zhang (2005) and Kozubowski and Nadarajah (2010) defined different forms for the discrete skewed Laplace distribution. Here, a simple closed form based of the difference between survival functions is used as a way to create a discrete distribution based on the continuous one (Barbiero, 2014).

So, let the continuous skewed Laplace distribution be as follows:

$$
\mathrm{f}(x ; p, q)=\frac{-\log (p) \log (q)}{\log (p)} \begin{cases}p^{x} & x \geq 0 \\ q^{-x} & x<0\end{cases}
$$

such that $0<p, q<1$ are unknown parameters. The survival function for this distribution is defined as

$$
\mathrm{S}(x ; p, q)=\frac{1}{\log (p q)} \begin{cases}\frac{1}{\log (p)} q^{-x} & x=\ldots,-2,-1 \\ \frac{1}{\log (q)} p^{x} & x=0,1,2, \ldots\end{cases}
$$

Now, using the instruction rule to construct a discrete distribution based on the differences between survival functions of the continuous one, i.e.

$$
\varnothing(x)= \begin{cases}\mathrm{S}(x)-\mathrm{S}(x+1) & x \in \mathbb{Z} \\ 0 & \text { o.w }\end{cases}
$$

we have

$$
\begin{aligned}
\varnothing(x ; p, q) & =\frac{1}{\log (p q)} \begin{cases}\frac{1}{\log (p)}\left[q^{-(x+1)}-q^{-x}\right] \\
\frac{1}{\log (q)}\left[p^{x}-p^{x+1}\right]\end{cases} \\
& =\frac{1}{\log (p q)} \begin{cases}\log (p)\left[q^{-(x+1)}(1-q)\right] & x=\ldots,-2,-1 \\
\log (q)\left[p^{x}(1-p)\right] & x=0,1,2, \ldots\end{cases}
\end{aligned}
$$

## DISCRETE SKEWED LAPLACE DISTRIBUTION

This is the simple closed form for the discrete skewed Laplace distribution with $0<p, q<1$ that known as

$$
\mathbf{X}_{d} \sim \operatorname{ADSLaplace}(p, q)
$$

Now, using the iid sample $\mathbf{X}=\left(X_{1}, \ldots, X_{n}\right)$, the maximum likelihood (ML) function is defined as

$$
\begin{aligned}
1(p, q ; \mathbf{X})= & \log \prod_{i=1}^{n} \varnothing\left(p, q ; X_{i}\right) \\
= & s^{-} \log \left(\frac{\log (p)}{\log (p q)}\right)+s^{+} \log \left(\frac{\log (q)}{\log (p q)}\right)-\left(s^{-}+\sum_{X_{i}<0} X_{i}\right) \log (q) \\
& \quad+s^{-} \log (1-q)+\sum_{X_{i}<0} X_{i} \log (p)+s^{+} \log (1-p)
\end{aligned}
$$

such that $s^{-}$and $s^{+}$are defined as number of the negative and positive samples, respectively, i.e. as

$$
s^{-}=\sum_{i=1}^{n} 1_{x_{i}<0}, \quad s^{+}=\sum_{i=1}^{n} 1_{x_{i} \geq 0}
$$

Now, using first order derivative of the likelihood function, the ML estimation for the desired parameters are the solutions to the following equations:

$$
\begin{aligned}
\frac{\partial 1(p, q ; \mathbf{x})}{\partial p}= & s^{-} \frac{\log (p q)}{\log (p)} \frac{1}{p} \frac{1}{(\log (p q))^{2}}-s^{+} \frac{\log (p q)}{\log (q)} \frac{1}{p} \frac{\log q}{(\log (p q))^{2}} \\
& \quad+\sum_{x_{i} \geq 0} \frac{x_{i}}{p}-\frac{s^{+}}{1-p} \\
= & \frac{s^{-}}{p} \frac{\log (q)}{\log (p)} \frac{1}{\log (p q)}-\frac{s^{+}}{p} \frac{1}{\log (p q)}+\sum_{x_{i} \geq 0} \frac{x_{i}}{p}-\frac{s^{+}}{1-p} \\
= & \frac{s^{-} \log q+s^{-} \log p-s^{-} \log p-s^{+} \log p}{p \log p \log (p q)}+\sum_{x_{i} \geq 0} \frac{x_{i}}{p}-\frac{s^{+}}{1-p}
\end{aligned}
$$

$$
=\frac{s^{-}}{p \log (p)}-\frac{n}{p \log (p q)}+\sum_{x_{i} \geq 0} \frac{x_{i}}{p}-\frac{s^{+}}{1-p}
$$

So we have

$$
\frac{\partial \mathrm{l}(p, q ; \mathbf{x})}{\partial p}=s^{-} \frac{1}{p \log (p)}-n \frac{1}{p \log (p q)}+\sum_{x_{i} \geq 0} \frac{x_{i}}{p}-\frac{s^{+}}{1-p}
$$

In a similar way,

$$
\frac{\partial 1(p, q ; \mathbf{x})}{\partial q}=s^{+} \frac{1}{q \log (q)}-n \frac{1}{q \log (p q)}-\sum_{x_{i}<0} \frac{x_{i}}{q}-\frac{s^{-}}{q(1-q)}
$$

So the solutions of these equations lead to the $\left(\hat{p}_{M L}, \hat{q}_{M L}\right)$ such that we did not have closed form and should solve analytically.

## Bayesian Analysis

Let $\boldsymbol{\theta}=(p, q)$ be the parameters of the discrete skewed Laplace distribution with the prior distribution $\pi(\boldsymbol{\theta})=\pi(p) \pi(q \mid p)$. Note that we assume $p$ and $q$ are independent, so $\pi(\boldsymbol{\theta})=\pi(p) \pi(q)$. Also, the prior distribution for $p$ and $q$ are the noninformative prior $\mathrm{U}(0,1)$, the uniform distribution. If $\mathrm{f}(\mathbf{x} \mid \boldsymbol{\theta})$ is the desired distribution, then the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{x}$ is as follows:

$$
\pi(\boldsymbol{\theta} \mid \mathbf{x})=\frac{\mathrm{f}(\mathbf{x} \mid \boldsymbol{\theta})}{\int_{\theta} \mathrm{f}(\mathbf{x} \mid \boldsymbol{\theta}) d \boldsymbol{\theta}}
$$

Note that, under the square integrable loss function, the Bayes estimator for $\boldsymbol{\theta}=(p, q)$ is as follows:

$$
\boldsymbol{\theta}^{\mathrm{B}}=\frac{\int_{\boldsymbol{\theta}} \boldsymbol{\theta} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) d \boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) d \boldsymbol{\theta}}
$$

That leads to

## DISCRETE SKEWED LAPLACE DISTRIBUTION

$$
\boldsymbol{\theta}^{\mathrm{B}}=\frac{\mathrm{E}\left[\boldsymbol{\theta} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta})\right]}{\mathrm{E}\left[\prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta})\right]}
$$

Now, if $\theta_{1}, \theta_{2}, \ldots, \theta_{m}$ are $m$ iid samples from the prior distribution $\pi(\boldsymbol{\theta})$, we have

$$
\frac{1}{m} \sum_{j=1}^{m} \theta_{j} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \theta_{j}\right) \xrightarrow{\text { a.s. }} \mathrm{E}\left[\boldsymbol{\theta} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right)\right], \quad m \rightarrow \infty
$$

and

$$
\frac{1}{m} \sum_{j=1}^{m} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \theta_{j}\right) \xrightarrow{\text { a.s. }} \mathrm{E}\left[\prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right)\right], \quad m \rightarrow \infty
$$

So the empirical Bayes (EB) estimator $\boldsymbol{\theta}^{\mathrm{EB}}$ is as follows:

$$
\boldsymbol{\theta}^{\mathrm{EB}}=\frac{\frac{1}{m} \sum_{j=1}^{m} \theta_{j} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right)}{\frac{1}{m} \sum_{j=1}^{m} \prod_{i=1}^{n} \mathrm{f}\left(x_{i} \mid \boldsymbol{\theta}\right)}
$$

and finally the Bayes estimators $\hat{p}_{\mathrm{EB}}$ and $\hat{q}_{\mathrm{EB}}$ can be easily found.

## Simulation Study

Now, to validate our estimation method presented in this paper, we simulate 1000 samples for different combinations of the parameters $(p, q)$ and compare the ML estimator with the EB method. Note that this can be easily achieved through the R package DiscreteLaplace (Barbiero, 2014; Barbiero \& Inchingolo, 2016). As Table 1 shows, the differences between the considered values for $(p, q)$ and their EB estimators are less than that of the estimators provided by the ML method.

Table 1. Simulation study for empirical Bayes and maximum likelihood methods

| $(\boldsymbol{p}, \boldsymbol{q})$ | $\hat{\boldsymbol{p}}_{\text {ML }}$ | $\hat{\boldsymbol{q}}_{\text {ML }}$ | $\hat{\boldsymbol{p}}_{\text {EB }}$ | $\hat{\boldsymbol{q}}_{\text {EB }}$ |
| ---: | ---: | ---: | ---: | ---: |
| $(0.25,0.25)$ | 0.211 | 0.229 | 0.231 | 0.238 |
| $(0.25,0.50)$ | 0.221 | 0.472 | 0.231 | 0.495 |
| $(0.50,0.25)$ | 0.482 | 0.232 | 0.491 | 0.239 |
| $(0.75,0.75)$ | 0.724 | 0.722 | 0.731 | 0.739 |

## Numerical Example

Kappenman (1975) and later Barbiero (2014) considered the following data:
$1.96,1.96,3.60,3.80,4.79,5.66,5.76,5.78,6.27,6.30,6.76,7.65$, $7.84,7.99,8.51,9.18,10.13,10.24,10.25,10.43,11.45,11.48,11.75$, $11.81,12.34,12.78,13.06,13.29,13.98,14.18,14.40,16.22,17.06$
which are assumed to represent a random sample of size $n=33$ from a symmetrical Laplace distribution with a location parameter. Before employing these data for our purposes, we transform them by subtracting their median, 10.13, and then take its integer part. We expect that these final values can be modeled through our proposed discrete distribution. We then apply our estimation methods discussed above and compare these estimators to those which the maximum likelihood method provides: $\hat{p}_{\mathrm{ML}}=0.7165$ and $\hat{q}_{\mathrm{ML}}=0.7657$, while $\hat{p}_{\mathrm{EB}}=0.7005$ and $\hat{q}_{\mathrm{EB}}=0.7511$. The Bayesian information criteria (BIC) for the ML estimation are -211.75, while this criteria for empirical Bayes method is -236.54 ; this shows that the EB estimators are more efficient than classical estimators.

## Conclusion

The presented paper investigates Bayesian analysis for the discrete skewed Laplace distribution and compares it to the classical estimation method, the maximum likelihood estimator. The BIC criteria show that the empirical Bayes estimators are preferable.

## DISCRETE SKEWED LAPLACE DISTRIBUTION

## References

Azzalini, B. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12(2), 171-178. Available from http://www.jstor.org/stable/4615982

Barbiero, A. (2014). An alternative discrete skew Laplace distribution. Statistical Methodology, 16, 47-67. doi: 10.1016/j.stamet.2013.07.002

Barbiero, A., \& Inchingolo, R. (2016). DiscreteLaplace: Discrete Laplace distributions [computer software]. Retrieved from http://CRAN.Rproject.org/package=DiscreteLaplace

Chakraborty, C., \& Chakravarty, D. (2012). Discrete gamma distributions: Properties and parameter estimations. Communications in Statistics - Theory and Methods, 41(18), 3301-3324. doi: 10.1080/03610926.2011.563014

Kappenman, R. F. (1975). Conditional confidence intervals for double exponential distribution parameters. Technometrics, 17(2), 233-235. doi: 10.1080/00401706.1975.10489307

Kozubowski, T. J., \& Nadarajah, S. (2010). Multitude of Laplace distributions. Statistical Papers, 51, 127-148. doi: 10.1007/s00362-008-0127-2

Krishna, H., \& Singh, P. (2009). Discrete Burr and discrete Pareto distributions. Statistical Methodology, 6(2), 177-188. doi: 10.1016/j.stamet.2008.07.001

Roy, D. (2003). The discrete normal distribution. Communications in Statistics - Theory and Methods, 32(10), 1871-1883. doi: 10.1081/STA120023256

Roy, D. (2004). Discrete Rayleigh distribution. IEEE Transactions on Reliability, 53(2), 145-150. doi: 10.1109/TR.2004.829161

Roy, D., \& Dasgupta, T. (2001). A discretizing approach for evaluating reliability of complex systems under stress-strength model. IEEE Transactions on Reliability, 50(2), 145-150. doi: 10.1109/24.963121

Yu, K., \& Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics - Theory and Methods, 34(9-10), 1867-1879. doi: 10.1080/03610920500199018

# Designing of Bayesian Skip Lot Sampling Plan under Destructive Testing 

K. K. Suresh<br>Bharathiar University<br>Tamil Nadu, India

S. Umamaheswari<br>Bharathiar University<br>Tamil Nadu, India

Skip-lot sampling plan serves as a cost-effective technique to manage the cost of performing frequent product inspections. As a powerful tool within a real-time quality management system, the ability to collect data which an optimize skip-lot sampling parameters affords manufacturers the luxury of lowering inspection expenses in various manufacturing units. The good quality of product can be produced in continuous improvement of production process in excellent quality history for suppliers. The procedures and necessary tables are provided for finding the respective plans for which sum of producer and consumer risks are minimized with acceptable and limiting quality levels which accounts for the prior distribution of process state for each lot and revenue received appreciably which reduces destructive testing.

Keywords: Bayesian sampling plan, gamma-Poisson distribution, producer's quality level, consumer's quality level, weighted risk

## Introduction

Quality has been an internal part of all products and services. It has become one of the most important consumer decision factors in the selection among competing product and services. The modern quality control methods are developed to growing awareness of needs and demands of the consumer. The method of quality control is mainly refers to a spectrum of managerial methods for attempting to maintain the quality of products at a desired level.

Acceptance sampling is a statistical procedure for accepting or rejecting a batch of merchandise or documents involves determining the maximum of defects discovered in a sample before the entire batch is rejected. The sampling procedure is defined on the inspection and classification of sample of units selected at random from a larger batch or lot and the ultimate decision about disposition of

Dr. K. K. Suresh is Professor and Head of the Department of Statistics. Email him at: sureshkk1@rediffmail.com. S. Umamaheswari is a Research Scholar in the Department of Statistics. Email her at: umasoundhar@ gmail.com.

## SURESH \& UMAMAHESWARI

the lot is made. Acceptance sampling is the specific plan that states the sample size or sizes to be used and associated with acceptance and rejection criteria. This method has rapidly gained wide application in industry, particularly in the following stages of manufacturing: incoming materials inspection, on line production control and finished product quality auditing.

Acceptance sampling is concerned with the risks of decision making. In industry it is used to take decision on lots, whether accept or reject a lot of some product or to accept or reject process. The rejection of a lot means return the lot to supplier or its submission to 100 percent inspection. The risks are classified as two namely, producer's risk and consumer's risk. The producer's risk implies that a good lot may be rejected by a sampling plan and the consumer's risk implies that a bad lot may be accepted by a sampling approach. Sampling plans are usually designed to control one or both of these risks.

The theory of acceptance sampling offers various inspection procedures, termed as sampling plans, which are categorized under four types, namely, (i) lot-by-lot sampling by attributes, in which each unit in a sample is inspected on a go-on-go basis for one or more characteristics, (ii) lot-by-lot sampling by variables, in which each unit in a sample is measured for single characteristics, (iii) sampling plans for continuous production by method of attributes and (iv) special purpose plans. Lot-by-lot sampling by attributes, in particular, comprises plans such as single sampling plans, double sampling plans, multiple sampling plans and sequential sampling plans.

A sampling plan is usually specified by one or more parameters such as sample size ( $n$ ) and acceptance number ( $c$ ) and associates with itself an important measure of performance, called operating characteristic function. The determination of the parameters of a sampling plan is prescribed the conditions on its operating characteristic curve providing protection to the producer and consumer is called designing of the sampling plan.

Acceptance sampling by attributes each item tested is classified as conforming or non-conforming. A sample is taken and it contains too many nonconforming items, then the batch is rejected, otherwise it is accepted. For this method to be effective, batches containing some non-conforming items must be acceptable. If the only acceptable percentage of non-conforming items is zero, this can only be achieved by examine every item and removing the item which are non-conforming. This is known as $100 \%$ inspection.

Effective acceptance sampling involves effective selection and the application of specific rules for lot inspection. The acceptance sampling plan applied on a lot-by-lot basis becomes an element in the overall approach to

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

maximize quality at minimum cost. Since different sampling plans may be statistically valid at different times during the life of a process, therefore all sampling plans should be periodically reviewed.

Many quality characteristics of a product can be measured by their performance measures. In such situations each product can be inspected and classified as either satisfying or non-satisfying a given set of specifications. Thus the products can be classified as defectives or non-defectives otherwise good or bad which based on inspections. Such quality characteristics are called attributes. This kind of inspection procedure is known as inspection by attributes and the respective plan is called as attribute sampling plan. In attribute sampling plan, decision is taken by comparing the number of defectives found on inspection with a stated acceptance number.

## Bayesian Acceptance Sampling Plan

The Bayesian approach provides a formal mechanism for taking sample of preferences for striking an economical balance between the cost of sampling and the expectation of loss due to accepting an insufficiently reliable product or rejecting a sufficiently reliable one. The assumption underlying the theory of acceptance sampling is that the production process from which lots are formed is stable and the lot quality defined in terms of fraction nonconforming is a fixed constant. The sampling inspection procedures defined under such assumptions are considered as conventional sampling plans.

However, in practice, the production processes are not always stable and the lots coming from such processes will have quality variations which may occur due to random fluctuations. The quality variations in the lots are separated into two types, viz., within-lot (sampling) variation and between-lot (sampling and process) variation. If these two sources of variation are equal and implying more process variation, the dispersion of process about the process average is zero, and each lot can be considered as a random sample drawn from a process with a constant probability of producing a non-conforming unit. This is the premise behind conventional acceptance sampling. In frequently, between-lot variation is greater than within-lot variation, which indicating that process variation exists and the probability of obtaining non-conforming unit varies continually. In such situations, the decisions on the submitted lots should be made with the consideration of between-lot variations and the lot quality will be treated as a random variable rather than a fixed quantity.

## SURESH \& UMAMAHESWARI

Further, Bayesian acceptance sampling considers both sources of variation. Thus the distinction between conventional and Bayesian approach is associated with the variations in lot quality and it can be studied by an appropriate prior distribution based on process history or knowledge in the selection of distribution to describe the random fluctuations involved in sampling plans. Sampling plans which use prior distribution for the lot quality together with the sampling distribution of sample information for making decisions such as accepting or rejecting the submitted lots are termed as Bayesian acceptance sampling plans and which are treated as alternative to conventional sampling plans. [See, Calvin (1990)].

The procedures for Bayesian plans which are introduce an economic considerations and prior results into the sampling equation. These procedures are suited to the sampling lots from process or assembly operations that contain assignable causes. These causes may be unknown and awaiting for isolation, known but unremovable due to state-of-the limitations, or known but uneconomical to remove. Conventional acceptance sampling assumes these assignable causes have been eliminated. Thus, the distinction between conventional and Bayesian approach is associated with the utilization of prior process history or knowledge in the selection of a distribution to describe the random fluctuations involved in acceptance sampling (Calvin, 1990).

Wetherill and Chiu (1975) noted the economic schemes based on Bayesian theory is more precise and scientific, leaving much less to judgement than those based on classical theory. The objective of the paper is to develop a Bayesian acceptance sampling plan with fixed acceptance numbers, when the number of defects in a unit can be described by a Poisson distribution with parameter $\lambda$ and the prior distribution of $\lambda$ takes the form of a gamma or non-informative function.

The gamma distribution was selected for utilization as prior knowledge because of two inherent characteristics: which are (i) The Poisson natural conjugate prior and (ii) It possesses sufficient productivity in distribution form, varying its parameters, which allows a reasonable representation of the specific prior knowledge. The first aspect leads to mathematical compatibility; a convenient attribute which obtained facilitates the computations. The second point implies that the gamma distribution, which provides a variety of distribution forms ranging from the positively skewed exponential distribution to an approximately symmetrical distribution shape.

The non-informative function used in the absence of specific prior knowledge corresponds to Jeffrey's non-informative prior (Box \& Tiao, 1992). The relationship between defectives in sample and defectives in remaining lot for

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

each of prior distributions can be exploring the idea of Bayesian methodology. Further it observed that the use of a binomial prior renders sampling useless and unsuitable. These serve to make the designers and users of Bayesian sampling plans more aware of the consequence associated with selection of particular prior distribution (See Case \& Keats, 1982). Phelps (1982) derived sampling procedure for skip-lot model using Bayesian approach under destructive testing. The model is developed for the purpose of (i) To maximize the expected return per lot produced items with non-conformities. (ii) To determine the inspection duration of preceding or succeeding lots, sample size ( $n$ ), and acceptance number ( $c$ ). The problem is generated with the help of posterior distribution of success state for each lot and it may reduce the constructive of sampling plan.

## Designing of Skip-lot Sampling Plan

The theory behind skip-lot is that continuous lots should be of high quality. In a skip-lot inspection, quality management recruits only inspect a small percentage of very high-conforming lots. Once companies develop a reference plan based on historical data of consumers' risks and producers' risks from inspections proceed to lot-by-lot. However, once a specified number of consecutive high-conforming lots have passed inspection, firms only inspect a fraction of subsequent lots at random. This skip-lot process continues until a lot does not pass, which then reverts to lot-by-lot inspection until products pass the skip-lot threshold again. The continuous inspection procedures which are optimum for a specified income function and a production model which can be only in of two states, which are states of repair, and known transition probabilities. The Markov process, generated by the model and class of decision procedures, approaches a limiting distribution.

Dodge (1955) presented an extension of continuous sampling plans for individual units to a skipping lot sampling plan that are applicable to bulk materials or products produced in successive lots or batches and designates the inspecting plan. The skipping inspection has specific rules based on the record of lot acceptance and rejections, for switching back and forth between normal inspection and skipping inspection.

Perry $(1970,1973)$ was concerned with the development and evaluation of a system of lot inspection sampling plans in which the provision are made for inspecting only some fraction of the submitted lots when the quality of the submitted product is good as demonstrated by the quality history of the product. A good proportion of the ideas and concepts of the skip-lot sampling plan SkSP-2

## SURESH \& UMAMAHESWARI

has modified from the continuous sampling plans for individual units or items of production. A continuous sampling inspection plan used to inspect a product which consisting of individual units and manufactured by an essentially in continuous process. This plan proposes that when quality is good, only a fraction $f$ of the submitted units need to be inspected [see Dodge (1943)].

Carr (1982) extended the procedures of CSP-M type plan and developed with a system of skip-lot plans designated as CSP-MSkSP. Carr (1982) noted inspection errors can have a severe impact on an attributes single-sampling plan for lot acceptance due to misclassification of units as defectives or nondefectives. However, with estimates of errors, the plan can be adjusted to preserve the desired operating characteristic curve for specified sampling plan. The skip-lot sampling plan have been developed at various situations such as cost, MIL_STD_105D, error of inspection, which are qualified by Schilling (1982), Hsu (1977), Okada (1967), Stephens (1979), Cox (1980), and Carr (1982). Aslam, Balamurali, Jun, and Ahmad (2010) established the designing methodology to determine the parameters for system of skip-lot sampling plan with corresponding to two points on the operating characteristic curve and also to minimize the average sample number with the help of binomial distribution.

The SkSP-2 plan is described as one that uses a given lot inspection plan by method of attributes called 'reference plan' together with the following rules.

| Rule 1. | Start with normal inspection (inspecting every lot) using <br> reference plan |
| :--- | :--- |
| Rule 2. | When $i$ consecutive lots are accepted on normal inspection, <br> switch to skipping inspection and inspect only a fraction $f$ of <br> the lots. |
| Rule 3. | When a lot is rejected on skipping inspection, return to <br> normal inspection |

The positive integer $i$ and sampling fraction $f$ are the parameters of SkSP-2. Here $0<f<1$. When $f=1$ the plan reduces to original reference plan. The probability of acceptance of the plan SkSP-2 is denoted by $P_{a}(f, i)$. Using Markov-chain technique one can find the probability of acceptance of SkSP-2 plan. A Markov process represents the observations of system which satisfying the condition that the probability of physical system will be given a state at time $t_{2}$ may be deduced from knowledge of its state at earlier time $t_{1}$. A Markov chain is a special case of Markov process in which the set of states or state space is discrete. A more complete characterization of the one step transitions of a Markov chain

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

with their corresponding probabilities provided in a matrix form is called the transition matrix (see Parzen, 1964).

The technique of Markov chains to evaluate the sampling plans, a trial corresponds to the drawing of sample from a lot which is under consideration. The results and outcomes of these trials, called states of chain, will depend upon the sampling plan. In some instances, the outcomes of these trials are either accepting or rejecting a particular lot while in others, the outcomes are more involved. The sates for the Markov chain of the skip-lot plan of type 2 can be categorized into two main classifications which are (i) normal inspection states (ii) skipping inspection states.
$P_{a}(f, i)$ can be determined by Markov- Chain Technique as follows:
$\mathrm{NR}=$ State where lot is rejected under normal inspection
$\mathrm{Nj}=$ State under normal inspection representing the number of consecutively accepted lots $j$
SA $=$ State where lots accepted during skipping inspection
$\mathrm{SR}=$ State where lots rejected during skipping inspection
SN = State where lot is skipped
$\mathrm{P}=$ Probability of acceptance of a lot according to the reference plan
$\mathrm{Q}=1-\mathrm{P}$

Because the Markov Chain is finite, recurrent and irreducible and periodic the stationary probabilities $\pi_{i}$ for each state can be obtained from the system

$$
\pi_{i}=\sum_{\text {alli }} \pi_{j} P_{j i} \text { For all states } i
$$

$P_{i j}=$ one step transition probability of state from $i$ to state $j$.

$$
\begin{equation*}
\text { and } \sum_{\text {alli }} \pi_{i}=1 \text { thus } P_{a}(f, i)=\frac{f P+(1-f) P^{i}}{f+(1-f) P^{i}} \tag{1}
\end{equation*}
$$

The properties of SkSP-2 are (i) for $f_{2}<f_{1}$, fixed $i$ and given reference plan, which implies that $P_{a}\left(f_{1}, i\right) \leq P_{a}\left(f_{2}, i\right)$, (ii) for integers $i<j$, fixed $f$ given reference plan, which implies that $P_{a}(f, i) \leq P_{a}(f, i)$ and (iii) $P_{a}(f, i) \geq P$ developed by Perry (1973).

## Selection Procedure for SkSP-2 with Repetitive Group Sampling Plan as Reference Plan

When sampling plans are set up for product characteristics that can involve costly and destructive testing by attributes, and the samples are required small acceptance numbers such as $c_{1}=0$ and $c_{2}=1$. The operating characteristics curve with $c_{1}=0$ and $c_{2}=1$ which leads to conflicting interest between the producer and consumer. The plan with acceptance number 0 favors to consumer and 1 favor to producer. Such conflict can be overcome, if the design of plan having both $c_{1}=0$ and $c_{2}=1$. In such situations Repetitive Group Sampling Plan with acceptance numbers 0 and 1 (with rejection number 2 ) can be used.

In the Repetitive Group Sampling Plan (RGSP), derived by Sherman (1965), a sample is drawn and the number of defectives is counted. According to fixed criterion the lot is either accepted or rejected. This is continued until the fixed criterion, the lot is either accepted or rejected or the sample is completely disregarded and one begins with another new sample, which is employed for making decision about an isolated lot of finished items. The RGS plan gives minimum sample size as well as the desired protection. Furthermore, RGS Plans are not nearly as efficient as the sequential sampling plans but they are usually more efficient than single sampling plan.

This plan gives an intermediate value in sample size efficiency between single sampling and sequential sampling plans. The RGS plan is used to improve operating characteristics curve with zero acceptance number. To increase, discriminating power of this curve, one way is to increase the sample size, an alternative way to use the RGS plan for attribute inspection. The RGSP plays a dominant role in industries to achieve high standard of quality as well as satisfaction of consumer. It is known that the sampling inspection has two principal effects namely filtering and incentive effects. The classified solution of sampling plans seems to be reasonable when filter is aim; but it seems unjustified when incentive is the main purpose. The selection of sampling plan with an index which is a simple function of derivative. Suresh (1993) has studied single sampling plan with the producers takes into both filtering and incentive effectives simultaneously.

Calvin (1990) derived the procedures which are suited to the sampling of lots from process or assembly operations, which contain assignable causes. These causes may be unknown and awaiting for isolation or known and irremovable due to the state limitations, or known and it has removed for uneconomical situations. Further considered the Bayesian sampling, in which, primary concern with the

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

process average function of non-conforming is $\bar{P}$, with lot fraction nonconforming is $p$ and its limitations being discussed. Further suggested that the posterior beta distribution for lot fraction non-conforming requires a family stable process with infrequent shifts. Theoretically, any major shifts would require reassessment of the sampling plan that accurate sampling risks were to be maintained. The RGS plan under Bayesian methodology could be developed by past history of the lot quality based on prior distribution of sample information, which is termed as BRGS Plan.

If the number of nonconforming units, $d$, then the sample follows a binomial model under attribute sampling with characteristic function from a finite lot with replacement. This can be used under the sampling an attribute characteristic from a finite lot without replacement for the case of non-conforming units $d$, whenever $n / N \leq 0.10$, which is based on two parameters namely, sample size $n$ and lot size $N$. Under hypergeometric model, the case of non-conforming units $d$, can be determined from a finite lot without replacement.

The Poisson model can be used whenever $n / N \leq 0.10, \mathrm{n}$ is large and $p$ is small such that $n p<5$ under the situations of attribute characteristic from finite lot without replacement. However the case of non-conforming units can be used whenever $n$ is large and $p$ is small such that $n p<5$ under finite lot with replacement. The Poisson model permits operating characteristics function of all attribute sampling plans simply as function of the product np for given acceptance and rejection numbers. The OC function remains some various combinations of $n$ and $p$ provided their product of given acceptance and rejection numbers. To develop compact table for the selection of sampling plans as only one parameter is considered in place of two parameters viz., $n$ and $p$.

However, when the Poisson model is assumed, the sampling plans are constructed by tables are necessarily the plans with risks are greater than the specified limits. The values will be close, but differences occurs in sample size and which meet the specified risks, the results found from tables and start to search for the appropriate plan. The gamma distribution is a natural conjugate prior for the sampling from a Poisson distribution. When the sample items are drawn randomly from a process, the number of defects in the sample is distributed according to Poisson, the gamma distribution is conjugate prior to the average number of defects per items as its parameters, denoted by $\lambda$. The Poisson distribution is defined with reference to the fixed parameter $\lambda$, representing the expected number of defects per unit. When $\lambda$ is assumed to vary at random from lot-to-lot, the gamma distribution is a suitable prior distribution for $\lambda$. According to Hald (1981), the production process produces output in a continuous stream

## SURESH \& UMAMAHESWARI

and observed number of defects in the sample from this process is distributed as Poisson used as an approximation to the binomial distribution for small values of $p$, which denoted as $p<0.10$. The Poisson distribution is an appropriate distribution for the case of
(i) Number of nonconforming items in the samples, when $p<0.10$.
(ii) Number of nonconformities in the sample.

The operating characteristics function for RGS plan by attributes under Poisson distribution is expressed by

$$
\begin{equation*}
P_{a}(p)=P_{a} / P_{a}+P_{r}, \tag{2}
\end{equation*}
$$

where $P_{a}$ and $P_{r}$ are the probability of acceptance and rejection of a lot respectively when the fraction is nonconforming. (i.e.) $P_{a}=P\left(d \leq c_{1} / P\right)$ and $P_{r}=P\left(d>c_{2} / P\right)$. Where $c_{1}$ and $c_{2}$ are the acceptance numbers. According to gamma distribution, the natural conjugate prior for sampling from the Poisson distribution, the function of prior distribution $p$ is denoted by

$$
\begin{equation*}
f(p / a, m)=\frac{e^{-a p} a^{m} p^{m-1}}{\Gamma m}, 0 \leq p<\infty, a, m>0 \tag{3}
\end{equation*}
$$

where $a$ is scale parameter and $m$ is shape parameter. Here $m$ is specified from the prior information about the production process. The posterior distribution for nonconformities is reduced under gamma-Poisson distribution. When the production is unstable, the nonconforming item $(d)$ and average number of defects $p$ are independently distributed. According to Hald (1981), the nonconforming items ( $d$ ) can be developed under the process average $\bar{P}<0.1, \frac{\bar{P}}{m}<0.2$ is given by

$$
\begin{equation*}
P(d ; n \bar{P}, m)=\frac{(m+d-1)!}{d!(m-1)!}\left(\frac{n \bar{P}}{m+n \bar{P}}\right)^{d}\left(\frac{m}{m+n \bar{P}}\right)^{m}, d=0,1,2, \ldots \tag{4}
\end{equation*}
$$

A design is presented for skip-lot sampling inspection plans with conditional repetitive group sampling plan as reference plan, to reduce the sample size and minimize the producer and consumer risks using repetitively selection of group of

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

samples. Further fixing the acceptance numbers $c_{1}=0$ and $c_{2}=1$ as the reference plan is advantage for the situations of costly or destructive testing.

The operating procedures for SkSRGSP are

1. At the outset, select a random sample of size n from each lot and find the number of defectives $d$.
2. If $d=0$, accept the lot

If $d>1$, reject the lot
If $d=1$, repeat the steps 1 and 2 .
3. When $i$ consecutive lots are accepted on normal inspection, switch to skipping inspection.
4. When a lot is rejected on skipping inspection, inspect next $i$ lots are produced.
5. When a lot is rejected while inspecting $i$ lots, switch to normal inspection.
6. When all $i$ lots are accepted, proceed as in step 3.
7. Screen each rejected lot and correct or replace all the nonconforming units.

The purpose of this study is to design a sampling plan, which is useful to save time and cost of the experimenters (producer and consumer). The product to be inspected comprises a series of successive lots produced by an essentially continuing process and the size of the lots is taken to be sufficiently large. Under the normal conditions the lots are expected to be essentially the same quality and the product comes from a source in which the consumer has confidence as good. This goal is achieved if we find a minimum/optimal sample size, $n$, that satisfies either both risks or only the consumer's risk. These procedures are useful to minimize the sample size of required sampling plan and increase the production level at minimum cost.

As the rapid advancement of manufacturing technology, supplier require their products to be high quality with low fraction of defectives often measured in parts per million. Unfortunately, traditional methods in some particular situations fail to find out a minute defect in the product. In order to overcome these problems the Bayesian methodology can be used to develop the sampling plan with minimum cost of inspection.

The attribute sampling plans have been developed for the situations where one of the parameters either the sample size $n$ or the acceptance number $c$ is prefixed. The method for obtaining this plan is to minimize the sum of the

## SURESH \& UMAMAHESWARI

producer's risk and the consumer's risk. In single sampling plan which minimize the sum of weighted risk fixed acceptance numbers developed by Vijayathilakan (1982) for the Poisson model.

## Procedure for Determination of Sample Size

When the sum of risk is minimized, the individual values of producer's and consumer's risk are taken into account and the decision of plan may be advantage. The method is developed to minimize the sum of risks with different weights for the producer's and consumer's risk. The sum is minimized when both risks have equal weights. If the consumer risk has larger weight, then it can be assigned to the consumer's rather than producer's risk. Suppose $w_{1}$ and $w_{2}$ are the weights such that $\left(w_{1}+w_{2}\right)=1$, which implies $\left(w_{1} \alpha+w_{2} \beta\right)$ can be minimized to obtain the necessary plan.

Minimizing $\left(w_{1} \alpha+w_{2} \beta\right)$ which is same as minimizing $\alpha+\left(w_{2} / w_{1}\right) \beta$. ( $w_{2} / w_{1}$ ) can be referred to the index of relative importance to given consumer's risk with the comparison of producer's risk and it will be denoted by $w$. The weights of the plan has two properties which are (i) when $w$ is greater than one, the plan will be more favorable to consumer while compared to equal weights of plan. (ii) When $w$ is less than one, it will be more favorable to producer. The Poisson model can be used to minimize the sum of weighted risks with fixed acceptance numbers it is obtained from

$$
\begin{equation*}
\alpha+w \beta=P_{a}(p)(R)+w P_{a}(p)(A) \tag{5}
\end{equation*}
$$

The expression derived from Poisson model is given by

$$
\begin{equation*}
\left[w \sum_{r=0}^{c} \exp \left(-n \mu_{2}\right)\left(n \mu_{2}\right)^{r} / r!\right]-\left[\sum_{r=0}^{c} \exp \left(-n \mu_{1}\right)\left(n \mu_{1}\right)^{r} / r!\right], \tag{6}
\end{equation*}
$$

which has imposing the conditions as $\left[\mu_{2} / \mu_{1}\right]^{c}<\frac{1}{w} \exp \left[n\left(\mu_{2}-\mu_{1}\right)\right]<\left[\mu_{2}-\mu_{1}\right]^{c+1}$. On the simplification for expression of value $c$ as the integral part is given by

$$
\begin{equation*}
\frac{\operatorname{In}\left(w^{-1}\right)}{\operatorname{In}\left(\mu_{2} / \mu_{1}\right)}+\frac{n\left(\mu_{2}-\mu_{1}\right)}{\operatorname{In}\left(\mu_{2} / \mu_{1}\right)} . \tag{7}
\end{equation*}
$$

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

The equation (7) can be modified in terms of $\mu_{1}, \mu_{2}$, and $n$, which becomes

$$
\begin{equation*}
\exp \left[n\left(\mu_{2}-\mu_{1}\right)\right]\left(\mu_{1} / \mu_{2}\right)^{n}<w<\exp \left[n\left(\mu_{2}-\mu_{1}\right)\right] \tag{8}
\end{equation*}
$$

Using (7) expression, Soundararajan (1981) has tabulated the value of $n$ which minimize the weighted risks for $c=0$ and 1 over different combinations of AQL and LQL. The fixed acceptance numbers is useful in the area of compliance testing where strict adherence to quality is important. Plans with fixed small acceptance numbers will have better control over acceptance of lots with more defectives. For any given sample size, it is known that acceptance numbers of zero and one will reduce the consumer's risk-(i.e.) the chance of accepting the lot with more than LQL percent defective will be reduced. Such plans are necessary while dealing with defence products.

## Numerical Study for Proposed Sampling Plan

1. Given $\mathrm{AQL}=5$ percent and $\mathrm{LQL}=15$ percent, one can find the values of sample size $n$ which minimize the risks for given value of desired distribution. The value of $N=10$ and $w=0.5$.
(i) Substituting $\mu_{1}=0.05, \mu_{2}=0.15$ and $m=0$ in Table 1 , one can find the value of $n$ is 4
(ii) Substituting $\mu_{1}=0.05, \mu_{2}=0.15$ and $m=5$ in Table 2, one can find the value of $n$ is 6 .
2. Given $\mathrm{AQL}=12$ percent and $\mathrm{LQL}=25$ percent, one can find the values of sample size $n$ which minimize the weighted risks for given value of desired distribution. The value of $N=25$ and $w=1$.
(i) Substituting $\mu_{1}=0.12$ and $\mu_{2}=0.25$ and $m=10$ in Table 3, one can find the value of $n$ is 6 .
(ii) Substituting $\mu_{1}=0.12$ and $\mu_{2}=0.25$ and $m=10$ in Table 4, one can find the value of $n$ is 9 from given value of $N=25$, $m=15$ and $w=1.5$.

From above examples, the expression for $n$ may be obtained by using desire distribution, which gives the values of $n$ given by the exact tables and the large number of tables required for various combinations of the lot size $N$ and the acceptance number $c$ may be dispensed with approximating expression can be

## SURESH \& UMAMAHESWARI

used instead. Tables have been prepared for various combinations of AQL and LQL from required sample size $n$ can be found out easily for given acceptance number $c$.

The tables are constructed with the help of $n$ values which minimise the sum of risks for fixed acceptance number $c_{1}=0$ and $c_{2}=1$ based on different combinations of AQL and LQL. The tables constructed as follows:

Table 1, $N=10, m=0, \mathrm{AQL}=3(1) 20$ and $\mathrm{LQL}=8(1) 45$
Table 2, $N=10, m=5, \mathrm{AQL}=5(1) 22$ and $\mathrm{LQL}=10(1) 47$
Table 3, $N=25, m=10$, $\mathrm{AQL}=11(1) 28$ and $\mathrm{LQL}=1(1) 38$
Table $4, N=25, m=15, \mathrm{AQL}=2(1) 19$ and $\mathrm{LQL}=15(1) 52$

## Conclusion

A new procedure of weighted risk techniques adapted on skip-lot sampling plan with repetitive group sampling plan, designed as SkSPRGSP has been proposed in this paper. The interest of performance measure is derived to minimize the sample number along with smaller acceptance number such as $c_{1}=0$ and $c_{2}=1$, which is advantage for small sample situations and also costly or destructive testing. Using Bayesian methodology the proposed plan provides better protection to the consumer and producer than the conventional sampling plans. The proposed approach can be applied to any variants of a skip-lot sampling plan to design a more economical plan. The new approach plays an important role in industries to achieve high standard of quality as well as satisfaction of consumer. Each received lot has been inspecting in a time-consuming endeavor, especially if lots are large size. Raw materials are one example of an ideal explorer for skip-lot techniques. Products with critical parameters may still require a more thorough inspection process, but skip-lot inspection protocols serve as a way to offset the cost of inspecting high-conforming products. Conducting business with a supplier of proven record is another ideal condition for skip-lot strategies.

Table 1. Obtain the sample size $n$ which minimizing ( $\alpha+0.5 \beta$ ), when fixed acceptance number $m=0, N=10$

| Acceptable Quality Levels in Percent Defective ( $\mu_{1}$ ) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|  | 8 | 6 | - | - | - | - | - | - | - | - | - | - |  | - | - | - | - | - | - |
|  | 9 | 7 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 10 | 7 | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 11 | 7 | 4 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 12 | 8 | 5 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 13 | 8 | 5 | 3 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 14 | 8 | 5 | 4 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 15 | 8 | 6 | 4 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 16 | 7 | 6 | 4 | 3 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 17 | 7 | 6 | 4 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| $\stackrel{\text { ® }}{ }$ | 18 | 7 | 6 | 4 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| 3 | 19 | 7 | 6 | 4 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - | - |
| $\stackrel{0}{\square}$ | 20 | 7 | 6 | 5 | 4 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
| O | 21 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - |
| $\stackrel{\omega}{0}$ | 22 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - | - |
|  | 23 | 7 | 5 | 5 | 4 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - | - |
| ¢ | 24 | 7 | 5 | 5 | 4 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - | - |
| - | 25 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - |
| $\stackrel{\square}{\square}$ | 26 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - |
| $\cdots$ | 27 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - |
| $\stackrel{\square 0}{\stackrel{0}{0}}$ | 28 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - | - | - |
| $\underset{\sim}{\Delta}$ | 29 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | 1 | - | - | - | - | - | - | - | - |
| $\geq$ | 30 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | - | - | - | - | - | - | - | - |
| $\overline{\bar{T}}$ | 31 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | - | - | - | - | - | - | - | - |
| $\bigcirc$ | 32 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - | - |
| . | 33 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - | - |
| 言 | 34 | 6 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - | - |
|  | 35 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - |
|  | 36 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - |
|  | 37 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - | - |
|  | 38 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | - | - | - |
|  | 39 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | - | - | - |
|  | 40 | 5 | 4 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | - | - | - |
|  | 41 | 5 | 4 | 4 | 4 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - | - | - |
|  | 42 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | - | - |
|  | 43 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | - | - |
|  | 44 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - | - | \% | - |
|  | 45 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |

*Key: $\alpha$-Producer Risk, $\beta$-Consumer Risk

## SURESH \& UMAMAHESWARI

Table 2. Obtain the sample size $n$ which minimizing ( $\alpha+0.5 \beta$ ), when fixed acceptance number $m=5, N=10$

| Acceptable Quality Levels in Percent Defective ( $\mu_{1}$ ) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
|  | 10 | 7 | 6 | 5 | 3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 11 | 7 | 6 | 5 | 4 | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 12 | 6 | 6 | 5 | 4 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | - | - |
|  | 13 | 6 | 6 | 5 | 4 | 4 | 3 | 1 | - | - | - | - | - | - | - | - | - | - | - |
|  | 14 | 6 | 5 | 5 | 4 | 4 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - | - |
|  | 15 | 6 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 1 | - | - | - | - | - | - | - | - | - |
|  | 16 | 6 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | - |
|  | 17 | 5 | 5 | 5 | 4 | 4 | 3 | 3 | 3 | 2 | 1 | - | - | - | - | - | - | - | - |
|  | 18 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 2 | 2 | 1 | - | - | - | - | - | - | - |
|  | 19 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 1 | - | - | - | - | - | - |
|  | 20 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | - | - | - | - | - |
| $\cdots$ | 21 | 5 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 1 | - | - | - | - |
| $\stackrel{\text { ® }}{ \pm}$ | 22 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - |
| O | 23 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | - | - | - |
| $\stackrel{\circ}{\circ}$ | 24 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | - | - |
| $\stackrel{\square}{\square}$ | 25 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | - |
| $\stackrel{\text { ¢ }}{ }$ | 26 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| $\stackrel{\rightharpoonup}{0}$ | 27 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| $\stackrel{\square}{\square}$ | 28 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| $\cdots$ | 29 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| $\stackrel{\square}{0}$ | 30 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| $\stackrel{1}{\square}$ | 31 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| ㄹ | 32 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| \% | 33 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| $\bigcirc$ | 34 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 은 | 35 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 产 | 36 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| - | 37 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 38 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 39 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 40 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 41 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 42 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 43 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 44 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 45 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 46 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
|  | 47 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |

*Key: $\alpha$-Producer Risk, $\beta$-Consumer Risk

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

Table 3. Obtain the sample size $n$ which minimizing ( $\alpha+1 \beta$ ), when fixed acceptance number $m=10, N=25$

| Acceptable Quality Levels in Percent Defective ( $\mu_{1}$ ) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
|  | 1 | 23 | 21 | 21 | 20 | 18 | 18 | 17 | 16 | 16 | 15 | 14 | 14 | 14 | 13 | 13 | 12 | 12 | 12 |
|  | 2 | 19 | 18 | 17 | 16 | 16 | 15 | 14 | 14 | 13 | 13 | 13 | 12 | 12 | 11 | 11 | 11 | 11 | 10 |
|  | 3 | 17 | 16 | 15 | 14 | 14 | 13 | 13 | 12 | 12 | 12 | 11 | 11 | 11 | 10 | 10 | 10 | 9 | 9 |
|  | 4 | 15 | 14 | 14 | 13 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 8 |
|  | 5 | 14 | 13 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 |
|  | 6 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 7 |
|  | 7 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 |
|  | 8 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 |
|  | 9 | 11 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 6 |
| ก | 10 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 |
| $\bigcirc$ | 11 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 |
| ? | 12 | 9 | 9 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| O | 13 | 9 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 5 |
| $\stackrel{1}{6}$ | 14 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 |
| $\bigcirc$ | 15 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 |
| $\stackrel{\square}{2}$ | 16 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 |
| ¢ | 17 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\frac{\square}{0}$ | 18 | 8 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\simeq$ | 19 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\cdots$ | 20 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 |
| $\stackrel{0}{0}$ | 21 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 |
| ব | 22 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 |
|  | 23 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 |
| 忈 | 24 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 |
| $\frac{\mathfrak{T}}{2}$ | 25 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 |
| $\overrightarrow{0}$ | 26 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
|  | 27 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 言 | 28 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| $\underset{\underline{E}}{\underline{E}}$ | 29 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| - | 30 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
|  | 31 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
|  | 32 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
|  | 33 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 |
|  | 34 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 |
|  | 35 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 |
|  | 36 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 |
|  | 37 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 |
|  | 38 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 |

*Key: $\alpha$-Producer Risk, $\beta$-Consumer Risk

## SURESH \& UMAMAHESWARI

Table 4. Obtain the sample size $n$ which minimizing ( $\alpha+1.5 \beta$ ), when fixed acceptance number $m=15, N=25$

| Acceptable Quality Levels in Percent Defective ( $\mu_{1}$ ) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|  | 15 | 25 | 22 | 19 | 18 | 17 | 16 | 15 | 14 | 14 | 13 | 13 | 13 | 15 | -- | -- | -- | -- | -- |
|  | 16 | 24 | 21 | 19 | 17 | 16 | 15 | 14 | 14 | 13 | 13 | 12 | 12 | 13 | 14 | -- | -- | -- | -- |
|  | 17 | 23 | 20 | 18 | 17 | 15 | 15 | 14 | 13 | 13 | 12 | 12 | 12 | 12 | 12 | 14 | -- | -- | -- |
|  | 18 | 22 | 19 | 17 | 16 | 15 | 14 | 13 | 13 | 12 | 12 | 11 | 11 | 11 | 11 | 11 | 13 | -- | -- |
|  | 19 | 21 | 19 | 17 | 15 | 14 | 14 | 13 | 12 | 12 | 11 | 11 | 11 | 10 | 10 | 10 | 11 | 13 | -- |
|  | 20 | 20 | 18 | 16 | 15 | 14 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 10 | 9 | 10 | 10 | 12 |
|  | 21 | 20 | 17 | 16 | 15 | 14 | 13 | 12 | 12 | 11 | 10 | 10 | 10 | 10 | 10 | 9 | 9 | 9 | 10 |
|  | 22 | 19 | 17 | 15 | 14 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 9 | 9 |
|  | 23 | 19 | 16 | 15 | 14 | 13 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| 3 | 24 | 18 | 16 | 14 | 13 | 12 | 12 | 11 | 11 | 10 | 10 | 9 | 9 | 9 | 9 | 9 | 8 | 8 | 8 |
| 0 | 25 | 18 | 16 | 14 | 13 | 12 | 11 | 11 | 10 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 |
| \% | 26 | 17 | 15 | 14 | 13 | 12 | 11 | 11 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 8 |
| U | 27 | 17 | 15 | 13 | 12 | 12 | 11 | 10 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 7 |
| $\stackrel{\square}{0}$ | 28 | 16 | 14 | 13 | 12 | 11 | 11 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 7 | 7 |
| $\bigcirc$ | 29 | 16 | 14 | 13 | 12 | 11 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 8 | 7 | 7 | 7 |
| $\stackrel{\square}{\square}$ | 30 | 15 | 14 | 12 | 12 | 11 | 10 | 9 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 |
| U | 31 | 15 | 13 | 12 | 11 | 11 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 |
| $\begin{gathered} \frac{1}{0} \\ 0 \end{gathered}$ | 32 | 14 | 13 | 12 | 11 | 10 | 10 | 9 | 9 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 |
| $\pm$ | 33 | 14 | 13 | 12 | 11 | 10 | 10 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 |
| $\cdots$ | 34 | 14 | 13 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 7 | 6 |
| (1) | 35 | 14 | 12 | 11 | 10 | 10 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 |
| d | 36 | 13 | 12 | 11 | 10 | 10 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 |
|  | 37 | 13 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 |
| $\pm$ | 38 | 13 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 |
| $\stackrel{\text { ® }}{\sim}$ | 39 | 13 | 11 | 10 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 |
| $\bigcirc$ | 40 | 13 | 11 | 10 | 10 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 |
| $\underset{\cong}{\text { O)}}$ | 41 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 |
| 言 | 42 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 |
| $. \underline{E}$ | 43 | 12 | 11 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 6 |
|  | 44 | 12 | 10 | 10 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 |
|  | 45 | 12 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 |
|  | 46 | 11 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 |
|  | 47 | 11 | 10 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 |
|  | 48 | 11 | 10 | 9 | 8 | 8 | 8 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 |
|  | 49 | 11 | 10 | 9 | 8 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 |
|  | 50 | 11 | 10 | 9 | 8 | 8 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 |
|  | 51 | 11 | 9 | 9 | 8 | 8 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 |
|  | 52 | 10 | 9 | 9 | 8 | 7 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 |

*Key: $\alpha$-Producer Risk, $\beta$-Consumer Risk

## DESIGNING OF BAYESIAN SKIP LOT SAMPLING PLAN

## Acknowledgments

The authors are thankful to the unknown referee for his comments towards revision of this paper. The first author is thankful to University Grants Commission, New Delhi for providing UGC-BSR-OTG. The second author is thankful to Department of Science and Technology, New Delhi towards providing DST-INSPIRE Fellowship for carrying out the research work.

## References

Aslam, M., Balamurali, S., Jun, C. H., \& Ahmad, M. (2010). Optimal designing of a skip-lot sampling plan by two point method. Pakistan Journal of Statistics, 26(4), 585-592.

Box, G. E. P., \& Tiao, G. C. (1992). Bayesian inference in statistical analysis. New York, NY: Wiley.

Calvin, T. W. (1990). How and when to perform Bayesian acceptance sampling. Milwaukee, WI: American Society for Quality Control.

Carr, W. E. (1982). Sampling plan adjustment for inspection error and skiplot plan. Journal of Quality Technology, 14(3), 105-116.

Case, K. E., \& Keats, J. B. (1982). On the selection of a prior distribution in Bayesian acceptance sampling. Journal of Quality Technology, 14(1), 10-18.

Cox, D.C. (1980). Skip-lot sampling plans. Quality, 21(8), 26-27.
Dodge, H. F. (1943). A sampling inspection plan for continuous production. Annals of Mathematical Statistics, 14(3), 264-279. doi:10.1214/aoms/1177731420

Dodge, H. F. (1955). Skip-lot sampling plan. Industrial Quality Control, 11, 3-5.

Hald, A. (1981). Statistical theory of sampling inspection by attributes. London: Academic Press.

Hsu, J. I. S. (1977). A cost model for skip-lot destructive testing. IEEE Transactions on Reliability, $R$-26(1), 70-72. doi:10.1109/TR.1977.5215081

Okada, H. (1967). Skip-lot sampling inspection plan combined with MIL-STD-105D. Reports of Statistical Application Research, 14(4), 13-20.

Parzen, E. (1964). Stochastic processes. San Francisco, CA: Holden-Day.
Perry, R. L. (1970). A system of skip-lot sampling plan for lot inspection (Unpublished doctoral thesis). Rutgers, The State University of New Jersey, New Brunswick, NJ.

## SURESH \& UMAMAHESWARI

Perry, R. L. (1973). Skip-lot sampling plans. Journal of Quality Technology, 5(3), 123-130.

Phelps, R. I. (1982). Skip-lot destructive sampling with Bayesian inference, IEEE Transactions on Reliability, R-31(2), 191-193. doi:10.1109/TR.1982.5221295

Schilling, E. G. (1982). Acceptance sampling in quality control. Milwaukee, WI: American Society for Quality.

Sherman, R. E. (1965). Design and evaluation of a repetitive group sampling plan. Technometrics, 7(1), 11-21. doi:10.1080/00401706.1965.10490222

Soundararajan, V. (1981). Single sampling attributes plan indexed by AQL and AOQL. Journal of Quality Technology, 13(3), 195-200.

Stephens, K. S. (1979). How to perform continuous sampling. Milwaukee, WI: American Society for Quality Control.

Suresh, K. K. (1993). A study on acceptance sampling using acceptable and limiting quality levels (Unpublished doctoral thesis). Bharathiar University, Tamil Nadu, India.

Vijayathilakan, J. P. (1982). Studies on lot acceptance procedures (Unpublished doctoral thesis). Madras University, Tamil Nadu, India.

Wetherill, G. B., \& Chiu, W. K. (1975). A review of acceptance sampling schemes with emphasis on the economic aspect. International Statistical Review / Revue Internationale de Statistique, 43(2), 191-210. doi:10.2307/1402898

# The $b r^{2}$-weighting Method for Estimating the Effects of Air Pollution on Population Health 

Goran Krstić
Fraser Health Authority
New Westminster, British Columbia

Nikolas S. Krstić
University of British Columbia
Vancouver, British Columbia

Mauricio Zambrano-Bigiarini<br>Universidad de La Frontera<br>Temuco, Chile

Uncertainties, limitations and biases may impede the correct application of concentrationresponse linear functions to estimate the effects of air pollution exposure on population health. The reliability of a prediction depends largely on the strength of the linear correlation between the studied variables. This work proposes the joint use of the coefficient of determination, $r^{2}$, with the regression slope, $b$, as an improved measure of the strength of the linear relation between air pollution and its effects on population health. The proposed $b r^{2}$-weighting method offers more reliable inferences about the potential effects of air pollution on population health, and can be applied universally to other fields of research.

Keywords: Linear regression coefficients, uncertainty analysis, concentrationresponse function, air pollution, population health

## Introduction

Inherent uncertainties associated with the application of relative risks (RR), hazard ratios (HR) and concentration-response (C-R) functions derived from the epidemiological studies on air pollution exposure vs. population mortality/morbidity have been discussed in the published literature (Burnett et al., 2014; Fann, Gilmore, \&Walker, 2013; Fann et al., 2011; Krewski et al., 2009; Environmental Protection Agency, 2006; Post, Watts, Al-Hussainy, \& Neubig, 2005; Lipfert \& Wyzga, 1995). Considering that confounding factors not controlled or accounted for could affect our ability to predict reliably the effects attributed to a variable of interest (e.g., effects of $\mathrm{PM}_{2.5}$ on population health), epidemiological studies often include adjustments for potential impacts from various environmental, behavioral, genetic, and socio-economic health risk factors.

[^37]
## KRSTIĆ ET AL.

The coefficient of correlation $(r)$ has been developed in its current format by Pearson in 1895 (Rodgers \& Nicewander, 1988). The squared value of $r$ is defined as the coefficient of determination $\left(r^{2}\right)$, which provides an estimated proportion of the variation in a dependent/response variable $y$ that could be explained by the variation in an independent/explanatory variable $x$. In linear least squares regression with an estimated intercept term, the $r^{2}$ can be calculated with the following equation:

$$
\begin{equation*}
r^{2}=\left(\frac{\sum_{i=1}^{n}\left(O_{i}-\bar{O}\right)\left(P_{i}-\bar{P}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_{i}-\bar{O}\right)^{2}} \sqrt{\sum_{i=1}^{n}\left(P_{i}-\bar{P}\right)^{2}}}\right)^{2} \tag{1}
\end{equation*}
$$

where $O$ are the observed and P the predicted values (Krause, Boyle, \& Bäse, 2005).

When used for regression between an environmental risk factor vs. population health, the $r^{2}$ provides a statistical estimate of how well the regression line approximates the real observations. The $r^{2}$ provides an estimate of the combined dispersion against the single dispersion of the observed and predicted series, with values in the range 0 to 1 , where $r^{2}=1$ indicates a perfect linear correlation (i.e., the dispersion values of the observation and the prediction are equal) and $r^{2}=0$ indicates absence of a linear correlation between the studied variables. Refer to Rodgers and Nicewander, (1988) for a set of different ways to express $r$ and conversely the $r^{2}$.

The coefficient of determination $\left(r^{2}\right)$ is sensitive to outliers and extreme dataset values, which may lead to a "bias toward the extreme events if correlationbased measures are employed in model evaluation" (Legates \& McCabe, 1999, p. 234). Arnold et al. (2012) indicated the use of $r^{2}$ without the regression coefficients could be associated with an over-estimation bias and that "if $\mathrm{r}^{2}$ is the primary statistical measure, it should always be used with slope and intercept to ensure that means are reasonable (slope $=1$ ) and bias is low" $(p .1495)$.

The study by Pope, Ezzati, and Dockery, (2009) could be used as an example to illustrate the importance of $r^{2}$-value as well as the slope in predicting the effects of $\mathrm{PM}_{2.5}$ on population health. Pope, Ezzati, and Dockery $(2009,2012)$ suggested a reduction in $\mathrm{PM}_{2.5}$ concentration observed over the period 1980s 2000s is responsible for a statistically significant improvement of life expectancy in the metropolitan areas of the United States. However, the observed correlation

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

with/without the influential observations is very weak ( $r^{2} \sim 0.05$ ) and Pope et al. (2012) acknowledged that "given that there are other determinates of life expectancy that may have changed in correlation with changes in air pollution" ( p . 234) their analyses "cannot fully eliminate the potential of some residual confounding" (p.234). This indicates in statistical terms that only approximately $5 \%$ of the variation in a change of life expectancy could be explained by the variation in a change of $\mathrm{PM}_{2.5}$ concentration and that the remaining $95 \%$ could be attributed to a set of selected explanatory variables including income and proxy smoking or other environmental, behavioral, genetic and socio-economic health risk factors not controlled or accounted for in the presented study (e.g., medical practice improvement, public health expenditure change, ambient air temperature).

The focus of the current study is on improving the interpretation of statistical linear regression analyses between air pollution vs. population health. Krause et al., (2005) introduced the application of the regression slope (b) as a weighing factor of the coefficient of determination $\left(r^{2}\right)$ to address potential underor over-estimates of model predictions. The proposed method has been used extensively by other researchers in the field of hydrology (Malagò, Pagliero, Bouraoui, \& Franchini , 2014; Feaster et al., 2014; Arnold et al., 2012; ZambranoBigiarini, 2010; Bellocchi, Rivington, Donatelli, \& Matthews, 2009). However, application of this approach in the field of environmental health has been limited (Krstić, 2012; Young \& Xia, 2013).

## Methodology

In a comparison of different efficiency criteria for hydrological model assessment, Krause et al., (2005) consider that $r^{2}$ alone may be limited in its ability to explain the relationship between the response and the explanatory variables, as it quantifies only the dispersion, where "a model which systematically over- or under-predicts all the time will still result in good $r^{2}$ values close to 1.0 even if all predictions were wrong" (p. 90). Hence, they argue that "for a proper model assessment the gradient b should always be discussed together with $r^{2 "}$ (p. 90), and proposed the following model of a weighted coefficient of determination $\left(w r^{2}\right)$ (Krause et al., 2005):

$$
{ }_{w} r^{2}=\left\{\begin{array}{l}
|b| \cdot r^{2} \text { for } b \leq 1  \tag{2}\\
|b|^{-1} \cdot r^{2} \text { for } b>1
\end{array}\right.
$$

## KRSTIĆ ET AL.

The weighted coefficient of determination $\left(w r^{2}\right)$ quantifies under- or overpredictions from both the $r^{2}$ and the slope or gradient of the regression line (b) for a more comprehensive representation of the variable dynamics and model results. In a recently developed R package ( R Core Team, 2015) on goodness-of-fit functions for comparison of simulated and observed hydrological time series ("hydroGOF"), Zambrano-Bigiarini (2014) indicates "the $\mathrm{br}^{2}$ coefficient allows accounting for the discrepancy in the magnitude of two signals (depicted by ' $b$ ') as well as their dynamics (depicted by $\mathrm{r}^{2}$ )" (p. 6). Hence, the commutative product of $|b|$ and $r^{2}$ presented above in (2) can be considered also from the opposite perspective, where $r^{2}$ is used for weighting the slope/gradient (b) to take into account the strength of the linear correlation between the studied variables.

For example, a weak correlation model (e.g., $r^{2}<0.1$ ) cannot be considered the same as a model with near perfect correlation (i.e., $r^{2}$ value close to 1.0 ), which should be taken into account for the interpretation of linear regression analyses by adjusting the slope/gradient (b) accordingly:

$$
{ }_{w} b=\left\{\begin{array}{l}
r^{2} \cdot|b| \text { for } b \leq 1  \tag{3}\\
r^{2} \cdot|b|^{-1} \text { for } b>1
\end{array}\right.
$$

where $w b$ represents a weighted slope/gradient ( $b$ ) of the regression line. If $r^{2}=1.0$, in a hypothetical situation of a perfect linear correlation, then $w b=|b|$ or $w=|b|^{-1}$ (i.e., $r^{2}$ - neutral).

In case of $|b| \leq 1$, the limit of $r^{2}|b|$ equals 0 if both $|b|$ and $r^{2}$ approach 0 . The same result for the limit of $r^{2}|b|$ is obtained if $|b| \rightarrow 0$ and $r^{2} \rightarrow 1$ as well as if $|b| \rightarrow 1$ and $r^{2} \rightarrow 0:$

$$
\begin{equation*}
\lim _{\left(|b|, r^{2}\right) \rightarrow(0,0)}\left(r^{2} \cdot|b|\right)=\lim _{\left(|b|, r^{2}\right) \rightarrow(0,1)}\left(r^{2} \cdot|b|\right)=\lim _{\left(|b|, r^{2}\right) \rightarrow(1,0)}\left(r^{2} \cdot|b|\right)=0 \tag{4}
\end{equation*}
$$

The limit of $r^{2}|b|$ equals 1 when both $|b|$ and $r^{2}$ approach 1 :

$$
\begin{equation*}
\lim _{\left(|b|, r^{2}\right) \rightarrow(1,1)}\left(r^{2} \cdot|b|\right)=1 \tag{5}
\end{equation*}
$$

In case of $|b|>1$, the limit of $r^{2}|b|^{-1}$ equals 0 if $|b| \rightarrow 1$ and $r^{2} \rightarrow 0$ or if $|b| \rightarrow \infty$ and $r^{2} \rightarrow 0$ or if $|b| \rightarrow \infty$ and $r^{2} \rightarrow 1$ :

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

$$
\begin{equation*}
\lim _{\left(b \mid, r^{2}\right) \rightarrow(1,0)}\left(r^{2} \cdot|b|^{-1}\right)=\lim _{\left(b \mid, r^{2}\right) \rightarrow(\infty, 0)}\left(r^{2} \cdot|b|^{-1}\right)=\lim _{\left(b \mid, r^{2}\right) \rightarrow(\infty, 1)}\left(r^{2} \cdot|b|^{-1}\right)=0 \tag{6}
\end{equation*}
$$

As in the case of $w b=r^{2}|b|$, the limit of $r^{2}|b|^{-1}$ equals 1 when both $|b|$ and $r^{2}$ approach 1:

$$
\begin{equation*}
\lim _{\left.\left(|b| r^{2}\right) \rightarrow(1,1)\right)}\left(r^{2} \cdot|b|^{-1}\right)=1 \tag{7}
\end{equation*}
$$

The results of linear regression analyses models can be used to make predictions about the effects of exposure to environmental/socio-economic factors on population health. A linear dose-response model or a linear concentrationresponse (C-R) function is typically assumed:

$$
\begin{equation*}
y=a+b x \tag{8}
\end{equation*}
$$

where $y$ is the dependent/response variable, $x$ - independent/explanatory variable, $a$ - the $y$-axis intercept, and $b$ - the slope/gradient of the line. However, it needs to be taken into consideration that the reliability of a prediction made with the aforementioned model depends largely on the strength of the linear correlation between the studied variables, where $r^{2}$-values greater than $\sim 0.5$ indicate a strong relationship with high reliability and $r^{2}$-values less than $\sim 0.1$ indicate a weak relationship with low reliability of model predictions. This is where the weighted slope/gradient $(w b)$ can be used for a more robust procedure to assess the potential effects of exposure to environmental and/or socio-economic factors on population health.

Using the methodology for particulate matter risk analysis described by the U.S. Environmental Protection Agency (US EPA), Environmental Protection Authority of Victoria (Australia) developed the equations for dose-response or concentration-response (C-R) functions. The authors estimate health outcome changes and calculate the health-endpoint-specific effect coefficient $(\beta)$ on the basis of available dose-response data (Burgers \& Walsh, 2002).

The C-R functions can be estimated from epidemiological studies using a Poisson regression where the natural base logarithm of a health endpoint or an effect is presented as a linear function of air pollution (e.g., $\mathrm{PM}_{2.5}$ ) concentration (Environmental Protection Agency, 2010a):

## KRSTIĆ ET AL.

$$
\begin{equation*}
y_{1}=y_{0} \cdot e^{\beta x_{1}}, \tag{9}
\end{equation*}
$$

where $y_{1}$ is the incidence rate of a specific health endpoint of interest at the ambient air pollution concentration $\left(x_{1}\right), e$ - the base of natural logarithm (ln or $\log _{e}$ ), $\beta$ - the health effect coefficient of ambient air pollution derived from epidemiological studies, and $y_{o}$ - the baseline incidence rate in hypothetical absence of ambient air pollution, provided that there is no threshold concentration (i.e., level of air pollution below which there is no significant health effect).

The change in the number of cases for a specific health endpoint (e.g., lung cancer incidence or mortality rate) $\Delta y=y_{1}-y_{o}$ or $y_{1}=\Delta y+y_{o}$, corresponding to a given change in ambient air pollution levels relative to the background ( $\Delta x=\mathrm{x}_{1}-\mathrm{X}_{\mathrm{o}}$ or $x_{1}=\Delta x+x_{o}$ ), can be calculated from the C-R function in (9) presented above using the following equation:

$$
\begin{equation*}
\Delta y=y_{o}\left(e^{\beta\left(\Delta x+x_{o}\right)}-1\right), \tag{10}
\end{equation*}
$$

where $\beta$ is the health-endpoint-specific effect coefficient representing an incremental change in the health outcome associated with a unit change in air pollution $(\Delta x)$. In a hypothetical situation where the background air pollution $x_{o}=0,(10)$ can be presented as following:

$$
\begin{equation*}
\Delta y=y_{o}\left(e^{\beta \Delta x}-1\right) \text { or } \Delta y=y_{o}\left(R R_{\Delta x}-1\right) \tag{11}
\end{equation*}
$$

where the term $e^{\beta \Delta x}$ is also known as the relative risk $\left(R R_{\Delta x}\right)$ associated with the change in $\Delta x$. If $e^{\beta \Delta x}=R R_{\Delta x}$ then $\beta \Delta x=\ln \left(R R_{\Delta x}\right)$, and $\beta=\ln \left(R R_{\Delta x}\right) / \Delta x$.

The percentage change in the number of cases of a given health endpoint $\left(z_{p}\right)$, corresponding to a given change in air pollution concentration $(\Delta x)$, can be calculated from (Burgers \& Walsh, 2002):

$$
\begin{equation*}
z_{p}=\frac{\left(y_{1}-y_{o}\right)}{y_{o}} \cdot 100 \text { or } z_{p}=\frac{\Delta y}{y_{o}} \cdot 100 \tag{12}
\end{equation*}
$$

Then, combining and rearranging (11) and (12) provides the equation to calculate $\beta$ for different health endpoints on the basis of available dose-response data from epidemiological studies for a $1 \mu \mathrm{~g} / \mathrm{m}^{3}$ change in air pollution:

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

$$
\begin{equation*}
e^{\beta \Delta x}=1+z_{p} / 100 \text { and } \beta=\frac{\ln \left(1+z_{p} / 100\right)}{\Delta x} \tag{13}
\end{equation*}
$$

Finally, an impact of air pollution on a health endpoint can be calculated from the following equation (Fann et al., 2011):

$$
\begin{equation*}
\Delta y=y_{o}\left(e^{\beta \Delta x}-1\right) \cdot p o p, \tag{14}
\end{equation*}
$$

where pop is population size of a particular group exposed to air pollution.
Case study data used in the current paper are obtained from Vinikoor-Imler, Davis, and Luben (2011), the National Center for Environmental Assessment of the U.S. EPA, who studied an association between air pollution and population health in North Carolina. They reported the following slopes for $\mathrm{PM}_{2.5}$ vs. lung cancer mortality and incidence after adjusting for the neighborhood socioeconomic status and the prevalence of cigarette smoking: $b=0.96$ per $1 \mu \mathrm{~g} / \mathrm{m}^{3}$ $\mathrm{PM}_{2.5}$ for lung cancer mortality ( $95 \% \mathrm{CI}: 0.34,1.59$, $p$-value $<0.01 ; r^{2}=0.18$; y -axis intercept, $a=40.96$ ) and $b=1.35$ per $1 \mu \mathrm{~g} / \mathrm{m}^{3} \mathrm{PM}_{2.5}$ for lung cancer incidence ( $95 \% \mathrm{CI}: 0.36,2.35, p$-value $0.01 ; r^{2}=0.09$; y-axis intercept, $a=44.36$ ).

## Results

## Case Study Worked Example Calculations: Lung Cancer Mortality

Vinikoor-Imler et al., (2011) provided an adjusted slope of 0.96 lung cancer mortality cases per 100,000 population per $1 \mu \mathrm{~g} / \mathrm{m}^{3}$ change in $\mathrm{PM}_{2.5}$ $\left(b=0.96 \cdot 10^{-5}\right)$, a $y$-axis intercept $(a)$ or an estimated baseline lung cancer mortality rate at $x_{o}=0$ of 40.96 cases per 100,000 population $\left(y_{o}=40.96 \cdot 10^{-5}\right)$, and lung cancer mortality rate per 100,000 population associated with an incremental $1 \mu \mathrm{~g} / \mathrm{m}^{3}$ increase in $\mathrm{PM}_{2.5}\left(y_{1}=0.96 \cdot 10^{-5}+40.96 \cdot 10^{-5}=41.92 \cdot 10^{-5}\right)$. Using (12) the value of $z_{p}$ is calculated at $2.344 \%$. Considering that $y_{1}=b x_{1}+a$ and $y_{o}=b x_{o}+a$, the same calculation can be obtained on the basis of the relationship: $y_{1}-y_{o}=\left(b x_{1}+a\right)-\left(b x_{o}+a\right)$ or $\Delta y=b \Delta x$, where if $\Delta x=1 \mu \mathrm{~g} / \mathrm{m}^{3}$ then $\Delta y=b$ (i.e., 0.96 cases per 100.000 population per $1 \mu \mathrm{~g} / \mathrm{m}^{3}$ ):

$$
\begin{equation*}
z_{p}=\frac{b}{y_{o}} \cdot 100 \tag{15}
\end{equation*}
$$

## KRSTIĆ ET AL.

The C-R coefficient $\beta$ can be then calculated using (13):

$$
\beta=\frac{\ln (1+2.344 / 100)}{1\left(\mu g / m^{3}\right)}=0.0232\left(\mu g / m^{3}\right)^{-1} .
$$

On the basis of the analysis presented by Vinikoor-Imler et al., (2011), using (14), it is estimated that incremental $10 \mu \mathrm{~g} / \mathrm{m}^{3}$ increase in $\mathrm{PM}_{2.5}$ concentration could be associated with additional 10.68 cases of lung cancer mortality per 100,000 population (i.e., 34,710 additional cases in $\sim 325$ million U.S. population).

In the following estimate of the coefficient $\beta$, the slope of the regression line (b) is adjusted for the observed strength of the association between $\mathrm{PM}_{2.5}$ exposure and lung cancer mortality ( $r^{2}$ ) using (13) and (15) with (3), where $|b|=0.96 \cdot 10^{-5}$ and $r^{2}=0.18$ for a weighted slope/gradient $w b=1.728 \cdot 10^{-6}$ per $\mu \mathrm{g} / \mathrm{m}^{3}$ and where $\Delta x=1 \mu \mathrm{~g} / \mathrm{m}^{3}$ for $\Delta y={ }_{w} b$ :

$$
\begin{align*}
& z_{w}=\frac{|b| \cdot r^{2}}{y_{o}} \cdot 100 \text { or } z_{w}=\frac{{ }_{w} b}{y_{o}} \cdot 100  \tag{16}\\
& z_{w}=\frac{0.96 \cdot 10^{-5} \cdot 0.18}{40.96 \cdot 10^{-5}} \cdot 100=0.422 \%
\end{align*}
$$

A weighted coefficient $\beta_{w}$ can be then calculated using a weighted percentage increase in the number of cases of a given health endpoint $z_{w}$ in the following equation:

$$
\begin{align*}
& \beta_{w}=\frac{\ln \left(1+z_{w} / 100\right)}{\Delta x} \\
& \beta_{w}=\frac{\ln (1+0.422 / 100)}{1\left(\mu \mathrm{~g} / \mathrm{m}^{3}\right)}=0.0042\left(\mu \mathrm{~g} / \mathrm{m}^{3}\right)^{-1} \tag{17}
\end{align*}
$$

Hence, adjusting for the neighborhood socio-economic status, cigarette smoking, and the $r^{2}$ between $\mathrm{PM}_{2.5}$ concentration and lung cancer mortality yields a weighted C-R coefficient $\beta_{w}$ of 0.0042 per $\mu \mathrm{g} / \mathrm{m}^{3}$. Using (14), it is estimated that an incremental $10 \mu \mathrm{~g} / \mathrm{m}^{3}$ increase in $\mathrm{PM}_{2.5}$ concentration could be associated with additional 1.76 cases of lung cancer mortality per 100,000 population or 5,720 additional cases if applied to $\sim 325$ million U.S. population, which is much lower

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

than the 34,710 additional cases foreseen by using the unadjusted slope coefficient $b$.

## Case Study Worked Example Calculations: Lung Cancer Incidence

Using the approach described above and the data from Vinikoor-Imler et al., (2011), a weighted C-R coefficient $\beta_{w}$ is calculated for the cancer incidence where the slope/gradient $b>1$ (i.e., $\mathrm{b}=1.35$ ), $r^{2}=0.09$ and an estimated baseline lung cancer incidence rate $y_{o}=44.36$ per 100,000 population at $x_{o}=0$. Hence, from (3) a weighted slope/gradient is $w b=r^{2} \cdot|b|^{-1}=0.09 \cdot 0.7407 \cdot 10^{-5}=6.666 \cdot 10^{-7}$ per $\mu \mathrm{g} / \mathrm{m}^{3}$ and $z_{w}$ can be calculated using a modified version of (16) to reflect that $b>1$ :

$$
\begin{align*}
& z_{w}=\frac{r^{2} \cdot|b|^{-1}}{y_{o}} \cdot 100  \tag{18}\\
& z_{w}=\frac{0.09 \cdot 0.7407 \cdot 10^{-5}}{44.36 \cdot 10^{-5}} \cdot 100=0.1503 \%
\end{align*}
$$

A weighted C-R coefficient $\beta_{w}$ is calculated using (17):

$$
\beta_{w}=\frac{\ln (1+0.1503 / 100)}{1\left(\mu g / m^{3}\right)}=0.0015\left(\mu g / m^{3}\right)^{-1}
$$

then, using (14), an incremental $10 \mu \mathrm{~g} / \mathrm{m}^{3}$ increase in $\mathrm{PM}_{2.5}$ concentration could be associated with additional 0.67 cases of lung cancer incidence per 100,000 population or 2,178 additional cases if applied to $\sim 325$ million U.S. population.

## Discussion and Conclusion

Some of the key uncertainties and limitations of currently accepted approach in assessing the effects of air pollution on population health stem from the quality and reliability of epidemiological studies (e.g., study design, exposure assessment, confounding factors, statistical model assumptions, risk characterization, potential errors and biases). The assumptions required for a valid least-squares regression are often not possible to satisfy completely in epidemiological study designs. It should be emphasized that regression coefficient/slope $b$ becomes meaningless and should not be used to make linear inferences/predictions if the $r^{2}$ approaches

## KRSTIĆ ET AL.

0 (e.g., $r^{2}<0.1$ ) even in situations where it may appear to be statistically significant.

It is also important to consider available evidence for a plausible biological mechanism of toxicity and for a slope and shape of the dose-response relationship at low to very low levels of air pollution (Vedal, Brauer, White, \& Petkau, 2003). There is no universal agreement among the researchers for an assumed linear nothreshold effect of air pollution on population health. Specifically regarding PM 2.5 -related mortality the U.S. EPA indicated "a review of the time-series and cohort studies may lead to the conclusion that although a threshold is not apparent at commonly observed concentrations, one may exist at lower levels" (Environmental Protection Agency, 2010b, p. 23). Uncertainties associated with the evidence for and likelihood of causality should be acknowledged. In addition, there is variability in the estimated C-R functions and the magnitude of potential effects of air pollution on population health as reported by different research groups (Environmental Protection Agency, 2010a).

The described methodological approach, first proposed by Krause et al., (2005) in the context of hydrology, was applied by Krstić, (2012) and accepted by Young \& Xia, (2013) from the National Institute of Statistical Sciences (NISS) to adjust the predicted population health effects in the context of ambient air pollution. The analyses presented in the current paper on the basis of epidemiological and environmental data from Vinikoor-Imler et al., (2011) showed that inclusion of the $r^{2}$ in the calculation is expected to yield better estimates of the predicted effects of air pollution on population health, which reflect more accurately the strength of the real linear correlation between the air pollution and the specified population health endpoint.

The proposed $b r^{2}$-weighting method is sensitive to extreme values of both $|b|$ and $r^{2}$ where model prediction reliability increases if $|b|$ and $r^{2}$ approach 1 and decreases if $|b|$ departs from 1 in either direction (i.e., $|b| \rightarrow \infty$ or $|b| \rightarrow 0$ ) and/or if $r^{2}$ departs from 1 and approaches 0 . The method identifies situations of maximum prediction ability as those of $|b| \leq 1$ as well as for $|b|>1$, provided that both $|b|$ and $r^{2}$ are close to 1 . This is in agreement with theoretical/ideal conditions in linear regression where a perfect correlation requires that $r=1,|b|=1$ and $y$-intercept $a=0$ if the relationship between the studied variables is truly linear in nature, resulting in a $45^{\circ}$ angle for the regression line as the best fit of the leastsquares estimator (Nau, 2014; Legendre, 2014).

The least-squares regression coefficient $b$ is considered as an unbiased prediction estimator under the assumptions of a perfect correlation between the studied variables (Legendre, \& Legendre, 1998). The estimated $r^{2}$-values closer to

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

1 allow more direct and reliable application of $b$ in making inferences and predictions. On the other hand, $r^{2}$-values closer to 0 indicate a necessity to adjust the slope $b$ for the observed reduction in model prediction ability. In situations of very low $r^{2}$-values, it becomes increasingly more likely even for the $95 \%$ confidence interval of the slope $b$ not to include the ideal $45^{\circ}$ angle line of the best regression fit (Mesplé et al., 1996; Legendre, 2014).

The presented analyses illustrate the importance of weighting the slope of the regression ( $b$ ) by the coefficient of determination $\left(r^{2}\right)$ to obtain more reliable inferences in projecting potential effects of air pollution on population health. The proposed $b r^{2}$-weighting method could be applied universally in studies of other environmental, behavioral, genetic or socio-economic risk factors for more comprehensive health impact estimates with lower potential bias and better decision-making.

## References

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., ... Jha, M. K. (2012). SWAT: Model use, calibration, and validation. Transactions of the ASABE, 55(4), 1491-1508.
doi: 10.13031/2013.42256
Bellocchi, G., Rivington, M., Donatelli, M., \& Matthews, K. (2009). Validation of biophysical models: Issues and methodologies. A review. Agronomy for Sustainable Development, 30(1), 109-130. doi: 10.1051/agro/2009001.

Burgers, M., \& Walsh, S. (2002). Exposure assessment and risk characterisation for the development of PM2.5 standard. Environment Protection Authority of Victoria. Retrieved from:
http://www.scew.gov.au/system/files/resources/9947318f-af8c-0b24-d928-04e4d3a4b25c/files/aaq-pm25-rpt-exposure-assessment-and-risk-characterisation-final-200209.pdf (accessed in May 2014).

Burnett, R. T., Pope, C. A., III, Ezzati, M., Olives, C., Lim, S. S., Mehta, S, ... Cohen, A. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environmental Health Perspectives, 122(4), 397-403. doi: 10.1289/ehp. 1307049.

Environmental Protection Agency. (2006). Expanded expert judgment assessment of the concentration-response relationship between PM 2.5 exposure and mortality. Research Triangle Park, NC: U.S. Environmental Protection

## KRSTIĆ ET AL.

Agency. Retrieved from:
http://www.epa.gov/ttn/ecas/regdata/Uncertainty/pm_ee_report.pdf (accessed in May 2014).

Environmental Protection Agency. (2010a). Quantitative health risk assessment for particulate matter (EPA Publication No. 452/R-10-005). Research Triangle Park, NC: U.S. Environmental Protection Agency. Retrieved from: http://www.epa.gov/ttn/naaqs/standards/pm/data/PM_RA_FINAL_June_2010.pdf (accessed in May 2014).

Environmental Protection Agency. (2010b). Summary of expert opinions on the existence of a threshold in the concentration-response function for $P M_{2.5-}$ related mortality (Technical Support Document (TSD)). Research Triangle Park, NC: U.S. Environmental Protection Agency. Retrieved from:
http://www.epa.gov/ttn/ecas/regdata/Benefits/thresholdstsd.pdf (accessed in July 2014).

Fann, N., Gilmore, E. A., \& Walker, K. (2013). Characterizing the longterm $P M_{2.5}$ concentration response function: comparing the strengths and weaknesses of research synthesis approaches (Working paper prepared for: Methods for Research Synthesis: A Cross-Disciplinary Workshop). Harvard Center for Risk Analysis. Retrieved from:
http://www.hsph.harvard.edu/hcra/files/2013/09/Fann-Gilmore-Walker-Sept20131.pdf (accessed in May 2014). (I found it at https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1273/2013/09/Fann-Gilmore-Walker-Sept-20131.pdf)

Fann, N., Lamson, A. D., Anenberg, S. C., Wesson, K., Risley, D., \& Hubbell, B. J. (2011). Estimating the national public health burden associated with exposure to ambient PM2.5 and ozone. Risk Analysis, 32(1), 81-95. doi: 10.1111/j.1539-6924.2011.01630.x.

Feaster, T. D., Benedict, S. T., Clark, J. M., Bradley, P. M., \& Conrads, P. A. (2014). Scaling up watershed model parameters-Flow and load simulations of the Edisto River Basin, South Carolina, 2007-09 (Scientific Investigations Report No. 2014-5104). U.S. Geological Survey. doi: 10.3133/sir20145104.

Krause, P., Boyle, D. P., \& Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences, 5, 89-97.

Krewski, D., Jerrett, M., Burnett, R. T., Ma, R., Hughes, E., Shi, Y., ... Thun, M. J. (2009). Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality (Research

## $b r^{2}$ COEFFICIENT FOR ESTIMATING HEALTH EFFECTS

Report 140). Boston, MA: HEI. Retrieved from:
http://pubs.healtheffects.org/getfile.php?u=478 (accessed in May 2014).
Krstić, G. (2012). A reanalysis of fine particulate matter air pollution versus life expectancy in the United States. Journal of the Air \& Waste Management Association, 62(9), 989-991. doi:10.1080/10962247.2012.697445.

Legates, D. R., \& McCabe, G. J., Jr. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1), 233-241.doi: 10.1029/1998WR900018

Legendre, P. \& Legendre, L. (1998). Numerical ecology (2 $2^{\text {nd }}$ ed.). Amsterdam: Elsevier.

Legendre, P. (2014). Model II regression user's guide, R edition. The Comprehensive R Archive Network (CRAN). Retrieved from: http://cran.rproject.org/web/packages/lmodel2/vignettes/mod2user.pdf (accessed in December 2014).

Lipfert, F. W., \& Wyzga, R. E. (1995). Air pollution and mortality: Issues and uncertainties. Journal of the Air \& Waste Management Association, 45(12), 949-966. doi: 10.1080/10473289.1995.10467427

Malagò, A., Pagliero, L., Bouraoui, F., \& Franchini, M. (2014). Comparing calibrated parameter sets of the SWAT model for the Scandinavian and Iberian Peninsulas. Hydrological Sciences Journal, 60(5), 949-967. doi: 10.1080/02626667.2014.978332

Mesplé, F., Troussellier, M., Casellas, C., \& Legendre, P. (1996). Evaluation of simple statistical criteria to qualify a simulation. Ecological Modelling, 88(1/3), 9-18. doi: 10.1016/0304-3800(95)00033-X

Nau, R. F. (2014). Introduction to linear regression analysis. Durham, NC: Fuqua School of Business, Duke University. Retrieved from: http://people.duke.edu/~rnau/regintro.htm (accessed in December 2014).

Pope, C. A., III, Ezzati, M., \& Dockery, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. The New England Journal of Medicine, 360, 376-386. doi: 10.1056/NEJMsa0805646

Pope, C. A., III, Ezzati, M., \& Dockery, D. W. (2012). Validity of observational studies in accountability analyses: The case of air pollution and life expectancy. Air Quality, Atmosphere \& Health, 5(2), 231-235. doi: 10.1007/s11869-010-0130-3

Post, E., Watts, K., Al-Hussainy, E., \& Neubig, E. (2005). Particulate matter health risk assessment for selected urban areas (EPA Publication No. 452/R-05-

## KRSTIĆ ET AL.

007A). Rockville, MD: U.S. Environmental Protection Agency. Retrieved from: http://www.epa.gov/ttn/naaqs/standards/pm/data/PMrisk20051220.pdf (accessed in July 2014).

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org/.

Rodgers, J. L., \& Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1), 59-66.
doi: 10.1080/00031305.1988.10475524
Vedal, S., Brauer, M., White, R., \& Petkau, J. (2003). Air pollution and daily mortality in a city with low levels of pollution. Environmental Health Perspectives, 111(1), 45-51.

Vinikoor-Imler, L. C., Davis, J. A., \& Luben, T. J. (2011). An ecologic analysis of county-level $\mathrm{PM}_{2.5}$ concentrations and lung cancer incidence and mortality. International Journal of Environmental Research and Public Health, 8(6), 1865-1871. doi: 10.3390/ijerph8061865

Young, S. S., \& Xia, J. Q. (2013). Assessing geographic heterogeneity and variable importance in an air pollution data set. Statistical Analysis and Data Mining, 6(4), 375-386. doi: 10.1002/sam. 11202

Zambrano-Bigiarini, M. (2010). On the effects of hydrological uncertainty in assessing the impacts of climate change on water resources (Doctoral thesis). University of Trento, Trento, Italy. Retrieved from: http://eprints-phd.biblio.unitn.it/284/1/MZB-PhD_Thesis-UT-05Ago2010.pdf (accessed in December 2014).

Zambrano-Bigiarini, M. (2014). Package "hydroGOF": Goodness-of-fit functions for comparison of simulated and observed hydrological time series. Comprehensive R Archive Network. Retrieved from: http://cran.rproject.org/web/packages/hydroGOF/hydroGOF.pdf (accessed in December 2014)

## Algorithms and Code

# Monte Carlo Simulations for Structural Equation Modelling (Revolution R) 

Sarah A. Rose<br>Wayne State University<br>Detroit, MI

Barry Markman<br>Wayne State University<br>Detroit, MI

Revolution R code is presented to setup Structural Equation Model (SEM) for a Monte Carlo study. The example is a comparison of different fit indices.

Keywords: Revolution R, structural equation model, SEM, fit indices, RMSEA, SRMR, CFI, chi-squared

Revolution R Lavaan package contains algorithms for performing SEM analytics. It has the ability to perform Monte Carlo simulations. The example considered here is for comparing model fit indices. This algorithm was verified using two wellknown SEM computer software programs by extracting the last repetition of the Monte Carlo simulation in Revolution R and comparing the output results with IBM SPSS Amos Graphics and with Mplus Version 5.1.

The algorithm was developed for a $4 \times 4$ correlation matrix of random values constrained within a specific range. Lines 4 and 5 of the code indicate the minimum and maximum values of the correlation value range, respectively. These values are defined as variable "bl" for the lower limit and "c1" for the upper limit. Line 6, variable "var", specifies the variant range that increases in magnitude based on the location of the correlation value within the correlation matrix. The algorithm defaults the variant range to zero, so all random correlation values for the matrix are within the same range.

The SEM model was designed for four variables: $X_{1}, X_{2}, X_{3}$, and $X_{4}$. The first three are exogenous variables, and $X_{4}$ is the only endogenous variable. The $4 \times 4$ correlation matrix is specified on line 75 . To modify the correlation matrix to

[^38]
## ROSE \& MARKMAN

another matrix size (i.e. $3 \times 3$ matrix, or $5 \times 5$ matrix) the variables on lines 7 through 16 would need to be redefined. Furthermore, the SEM model specification within the Monte Carlo command loop would require appropriate modifications on lines 69 through 80.

The model is specified with no correlation relationships, and no direct paths between variables $X_{1}$ and $X_{2}, X_{1}$ and $X_{3}$, and $X_{2}$ and $X_{3}$. Relationships can be added or modified by deleting or modifying lines 78 through 80 , which are currently designed to force relationship values to zero.

The algorithm was designed to provide six output files for six sample sizes ( $50,100,150,200,300$, and 500 ). To change the sample sizes, modify the variable names and values on lines 18 through 23 . The new variable names will have to be modified accordingly on lines 82 and 140.

The code provided in this article only contains the Monte Carlo simulation for one sample size of 50 . Repetition of the algorithm logic for the other five sample sizes has been deleted to more concisely display the algorithm in this article. The repeated code should be copied and pasted, as appropriate, for compilation of all six output files. To do this, copy lines 68 through 156 for each of the various sample sizes required. The variable "ss50" should be replaced with the appropriate variable names that were specified in lines 19 through 23 (e.g. "ss100", "ss150", "ss200", "ss300", and "ss500"). These variable names should replace "ss50" in the copied version of lines 82,140 , and 156 .

To modify the number of repetitions for the Monte Carlo simulation loop, modify the variable value of "rep" on line 25.

The model fit indices calculated and provided in the output files are the ChiSquared ("baseline.pvalue" as specified by Revolution R), Root Mean Square Error Approximation (RMSEA), Standardized Root Mean Square Residual (SRMR), and Comparative Fit Index (CFI). The output files indicate the sample size, number of repetitions, mean degrees of freedom for all simulations, specified correlation range, and the percentage of times the SEM resulted in model fit index values above or below a critical value.

The Lavaan Package contains additional model fit indices that can be added to the output file. These include Tucker-Lewis Index (TLI), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Normalized Fit Index (NFI), and Nonnormalized Fit Index (NNFI). To add these fit indices, copy lines 98 through 102 and paste them at the end of the repetition loop. Replace references to "srmr" with "tli", "gfi", "agfi", "nfi", or "nnfi", as appropriate. Prepare the output by copying and modifying lines 118 through 120, lines 133 and 134, and lines 148 and 149 appropriately. The new variables created for these model fit indices would

## MONTE CARLO SIMULATIONS FOR SEM (REVOLUTION R)

require specification of variable length. Insert the appropriate length and zero constant values, after lines 39 and 66, by copying the previous lines and modifying the variable names as appropriate.

Figure 1. R code and Lavaan package for conducting SEM fit indices

```
#Load SEM Lavaan Package
library(lavaan)
#Specify the Correlation Matrix Value Range
b1=0.06
c1=0.07
var=0
b2=b1+var
c2=c1+var
b3=b2+var
c3=c2+var
b4=b3+var
c4=c3+var
b5=b4+var
c5=c4+var
b6=b5+var
c6=c5+var
#set sampe size
ss50=50
ss100=100
ss150=150
ss200=200
ss300=300
ss500=500
#set number of repetition
rep=1000
#create arrays of specified length
reja5=numeric(length=rep)
reja1=numeric(length=rep)
reja01=numeric(length=rep)
acceptb=numeric(length=rep)
rejectb=numeric(length=rep)
acceptc=numeric(length=rep)
rejectc=numeric(length=rep)
acceptd=numeric(length=rep)
rejectd=numeric(length=rep)
close9=numeric(length=rep)
close75=numeric(length=rep)
close5=numeric(length=rep)
dof=numeric(length=rep)
#initialize constants to zero
totreja5=NULL
```


## ROSE \& MARKMAN

```
totreja1=NULL
totreja01=NULL
totacceptb=NULL
totrejectb=NULL
totacceptc=NULL
totrejectc=NULL
totacceptd=NULL
totrejectd=NULL
totlessthane9=NULL
ttotlessthane75=NULL
totlessthane5=NULL
totdof=NULL
tle5a=NULL
tle1a=NULL
tle01a=NULL
tleaccb=NULL
tlerejb=NULL
tleaccc=NULL
tlerejc=NULL
tleaccd=NULL
tlerejd=NULL
tlecl9=NULL
tlecl75=NULL
tlecl5=NULL
meandof=NULL
#start loop
for (i in 1:rep) {
    d1=runif(1,b1,c1)
    d2=runif(1,b2,c2)
    d3=runif(1,b3,c3)
    d4=runif(1,b4,c4)
    d5=runif(1,b5,c5)
    d6=runif(1,b6,c6)
    data.cor=lav_matrix_lower2full(c(1,d1,1,d2,d3,1,d4, d5, d6,1))
    rownames(data.cor)=colnames(data.cor)=c("z", "x1", "x2", "x3")
    model='z~x1+x2+x3
    x1~~0*x2
    x1~~0*x3
    x2~~0*x3
    fit=sem(model, sample.cov=data.cor,sample.nobs=ss50, fixed.x=FALSE)
    doff=fitMeasures(fit,"df")
    dof[i]<-doff
    chisqpvalue=fitMeasures(fit,"baseline.pvalue")
# If chisqpvalue is significant at various levels, increment counter
    if (chisqpvalue <= .05) reja5[i] <- 1
    if (chisqpvalue > .05) reja1[i] <- 1
    if (chisqpvalue < .001) reja01[i] <- 1
    rmsealower=fitMeasures(fit,"rmsea.ci.lower")
```

```
# If RMSEA value is less than 0.05
    if (rmsealower <= .05) acceptb[i] <- 1
    if (rmsealower > .05) rejectb[i] <- 1
    rmseaupper=fitMeasures(fit,"rmsea.ci.upper")
# If rmseaupper value is less than 0.1
    if (rmseaupper <= .1) acceptc[i] <- 1
    if (rmseaupper > .1) rejectc[i] <- 1
    srmr=fitMeasures(fit,"srmr")
# If srmr value is less than 0.09
    if (srmr <= .09) acceptd[i] <- 1
    if (srmr > .09) rejectd[i] <- 1
    cfi=fitMeasures(fit,"cfi")
# If cfi value is close to 1.0
    if (cfi < .9) close9[i] <- 1
    if (cfi < .75) close75[i] <- 1
    if (cfi >= .9) close5[i] <- 1
    }
#A sum the number of rejections
totreja5=sum(reja5)
totreja1=sum(reja1)
totreja01=sum(reja01)
#B sum the number of rejections
totacceptb=sum(acceptb)
totrejectb=sum(rejectb)
#C sum the number of rejections
totacceptc=sum(acceptc)
totrejectc=sum(rejectc)
#D sum the number of rejections
totacceptd=sum(acceptd)
totrejectd=sum(rejectd)
#E sum the number of rejections
totlessthane9=sum(close9)
totlessthane75=sum(close75)
totlessthane5=sum(close5)
totdof=sum(dof)
#divide the sum by number of repetitions
tle5a=totreja5/rep
tle1a=totreja1/rep
tleaccb=totacceptb/rep
tlerejb=totrejectb/rep
tleaccc=totacceptc/rep
tlerejc=totrejectc/rep
tleaccd=totacceptd/rep
tlerejd=totrejectd/rep
tlecl9=totlessthane9/rep
tlecl75=totlessthane75/rep
tlecl5=totlessthane5/rep
meandof=totdof/rep
# Summarize results in output file
```


## ROSE \& MARKMAN

```
results <- c("sample size=",ss50,"repetitions=",rep,
    "mean DOF=",meandof, "correlation=",b1,"-",c1,
    "chisq a<=0.05=",tle5a,"-> This means that p-chi-squared was less
    than 0.05 % of the time",
    "chisq a>0.05=",tle1a,"-> This means that p-chi-squared was
    greater than 0.05 % of the time",
    "RMSEA (lower)<=0.05=",tleaccb,"-> This means that RMSEA lower is
    less than 0.05 % of the time",
    "RMSEA (lower)>0.05=",tlerejb,"-> This means that RMSEA lower is
    greater than 0.05 % of the time",
    "RMSEA (upper)<=0.1=",tleaccc,"-> This means that RMSEA upper is
    less than 0.1 % of the time",
    "RMSEA (upper)>0.1=",tlerejc,"-> This means that RMSEA upper is
    greater than 0.1 % of the time",
    "SRMR<=0.09=",tleaccd,"-> This means that SRMR is less than
    0.09 % of the time",
    "SRMR>0.09=",tlerejd,"-> This means that SRMR is greater than
    0.09 % of the time",
    "CFI<0.9=",tlecl9,"-> This means that CFI is less than 0.9 % of
    the time",
    "CFI<0.75=",tlecl75,"-> This means that CFI is less than 0.75 %
    of the time",
    "CFI>=0.9=",tlecl5,"-> This means that CFI is greater than 0.9 %
    of the time")
# Write results to the hard disk.
# To write the results to your computer
# change file path to the folder you created on the C drive.
cat(results,sep="\n",file="c:/Users/Sarah/2cor_results-4variables-
50.txt",append=TRUE)#
#
#
#Repeat Lines 68 through 156 with appropriate modifications of variable
#"ss50" on lines 82 & 140 and output file name on line 156.
```


# An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach 

Wan Muhamad Amir W Ahmad
University of Science, Malaysia
Kelantan, Malaysia
Nor Azlida Aleng
University Mayalasia Terengganu
Kuala Terengganu, Malaysia

Mohamad Arif Awang Nawi<br>University Mayalasia Terengganu<br>Kuala Terengganu, Malaysia<br>Mohamad Shafiq<br>University of Science, Malaysia<br>Kelantan, Malaysia

Research on modeling is becoming popular nowadays, there are several of analyses used in research for modeling and one of them is known as applied multiple linear regressions (MLR). To obtain a bootstrap, robust and fuzzy multiple linear regressions, an experienced researchers should be aware the correct method of statistical analysis in order to get a better improved result. The main idea of bootstrapping is to approximate the entire sampling distribution of some estimator. To achieve this is by resampling from our original sample. In this paper, we emphasized on combining and modeling using bootstrapping, robust and fuzzy regression methodology. An algorithm for combining method is given by SAS language. We also provided some technical example of application of method discussed by using SAS computer software. The visualizing output of the analysis is discussed in detail.

Keywords: Multiple linear regression, robust regression, bootstrap method

## Introduction

Multiple linear regression (MLR) is an extension of simple linear regression. The random error term is added to make the model probabilistic rather than deterministic. The value of the coefficient $\beta_{i}$ determines the contribution of the independent variables $x_{i}$, and $\beta_{0}$ is the $y$-intercept (Ngo \& La Puente, 2012; Amir,

[^39]
## AHMAD ET AL.

Shafiq, Rahim, Liza, \& Aleng, 2016). A fuzzy regression model corresponding to equation (1) can be stated as:

$$
\begin{equation*}
y=A_{0}+A_{1} x_{1}+A_{2} x_{2}+\ldots+A_{k} x_{k} \tag{1}
\end{equation*}
$$

Explanation variables $x_{i}$ 's are assumed to be precise. However, response variable $Y$ is not crisp; it is fuzzy in nature. That means the parameters are also fuzzy in nature. Hence, the objective is to estimate these parameters.

Assume $A_{i}$ 's are assumes symmetric fuzzy numbers which can be presented by interval. For example, $A_{i}$ can be expressed as a fuzzy set given by $A_{i}=\left\langle a_{1 c}, a_{1 w}\right\rangle$ where $a_{i c}$ is center and $a_{i w}$ is radius or has associated vagueness. The fuzzy set reflects the confidence in the regression coefficients around $a_{i c}$ in terms of symmetric triangular memberships function. Application of this method should be given more attention when the underlying phenomenon is fuzzy which means that the response variable is fuzzy. Thus, the relationship is also considered to be fuzzy.
$A_{i}=\left\langle a_{1 c}, a_{1 w}\right\rangle$ can be written as $A_{i}=\left[a_{1 L}, a_{1 R}\right]$ with $a_{1 L}=a_{1 c}-a_{1 w}$ and $a_{1 R}=a_{1 c}-a_{1 w}$ (Kacprzyk \& Fedrizzi, 1992). In fuzzy regression methodology, parameters are estimated by minimizing total vagueness in the model.

$$
\begin{equation*}
y_{j}=A_{0}+A_{1} x_{1 j}+A_{2} x_{2 j}+\ldots+A_{k} x_{k j} \tag{2}
\end{equation*}
$$

Using $A_{i}=\left\langle a_{1 c}, a_{1 w}\right\rangle$ write

$$
\begin{equation*}
y_{j}=<a_{0 c}, a_{0 w}>+<a_{1 c}, a_{1 w}>x_{1 j}+\ldots+<a_{n c}, a_{n w}>x_{n j}=<a_{j c}, a_{j w}> \tag{3}
\end{equation*}
$$

Thus,

$$
\begin{gather*}
y_{j c}=a_{0 c}+a_{1 c} x_{1 j}+\cdots+a_{n c} x_{n j}  \tag{4}\\
y_{j w}=a_{0 w}+a_{1 w}\left|x_{1 j}\right|+\cdots+a_{n w}\left|x_{n j}\right| \tag{5}
\end{gather*}
$$

As $y_{j w}$ represent radius and so cannot be negative, therefore on the righthand side of equation $y_{j w}=a_{0 w}+a_{1 w}\left|x_{1 j}\right|+\ldots+a_{n w}\left|x_{n j}\right|$, absolute values of $x_{i j}$ are taken. Suppose there m data point, each comprising $a(n+1)$ - row vector. Then parameters $A_{i}$ are estimated by minimizing the quantity, which is total vagueness

## ALTERNATIVE MULTIPLE LINEAR MODEL REGRESSION MODELING

of the model-data set combination, subject to the constraint that each data point must fall within estimated value of response variable. This can be visualized as the following linear programming problem.

$$
\text { Minimized } \sum_{j=1}^{m}\left(a_{0 w}+a_{1 w}\left|x_{1 j}\right|+\cdots+a_{n w}\left|x_{n j}\right|\right)
$$

Subject to

$$
\begin{aligned}
& \left\{\left(a_{0 c}+\sum_{i=1}^{n} a_{i c} x_{i j}\right)-\left(a_{0 w}+\sum_{i=1}^{n} a_{i w} x_{i j}\right)\right\} \leq Y_{j} \\
& \left\{\left(a_{0 c}+\sum_{i=1}^{n} a_{i c} x_{i j}\right)+\left(a_{0 w}+\sum_{i=1}^{n} a_{i w} x_{i j}\right)\right\} \geq Y_{j}
\end{aligned}
$$

and $a_{i w} \geq 0$. Simplex procedure is generally employed in order to solve the linear programming problem.

## Calculation for linear Regression using SAS

```
/* First do Multiple Linear regression */
procreg data=temp1;
model y=x1 x2;
run;
```


## Approach the MM-Estimation Procedure for Robust Regression

```
/* Then do robust regression, in this case, MM-estimation */
ods graphics on;
procrobustreg method = MM fwls data=biostatistics plot=fitplot(nolimits)
plots=all;
model y = x1 x2 / diagnostics itprint;
output out=resids out=robout r=residual weight=weight outlier=outlier sr=stdres;
run;
ods graphics off;
```


## AHMAD ET AL.



Figure 1. Flow Chart of Robust, Bootstrap and Fuzzy Regression

## Procedure for Bootstrap with Case Resampling $\boldsymbol{n}=1000$

```
/* And finally, use a bootstrap with case resampling */
ods listing close;
procsurveyselect data=temp1 out=boot1 method=urs
samprate=1 outhits rep=1000;
run;
```


## Procedure for bootstrap into fuzzy regression Model

```
/*Combination of Bootstrap Technique with Fuzzy Regression*/
ods listing close;
procoptmodel;
set j= 1..8;
numberFish{j}, weight{j}, height{j};
read data boot1 into [_n_] Fishweight height;
/*Print Fishweight height*/
printFishweight height;
number n init 8; /*Total of Observations*/
/* Decision Variables bounded or not bounded*/
/*Theses three variables are bounded*/
var aw{1..3}>=0;
/*These three variables are not bounded*/
var ac{1..3};
/* Objective Function*/
min z1= aw[1]*n + sum{i in j} weight[i]*aw[2]+sum{i in j} height[i]*aw[3];
/*Linear Constraints*/
con c{i in 1..n}:
ac[1]+weight[i]*ac[2]+height[i]*ac[3]-aw[1]-weight[i]*aw[2]-height[i]*aw[3] <=
Fish[i];
con c1{i in 1..n}:
ac[1]+ weight[i]*ac[2]+ height[i]*ac[3]+aw[1]+ weight[i]*aw[2]+ height[i]*aw[3]
>= Fish[i];
```

AHMAD ET AL.

```
expand;/* This provides all equations */
solve;
print ac aw;
quit;
ods rtf close;
```


## An Illustration of a Biostatistics Case

## A Case Study of Aquaculture

Table 1. Description of the Variables

| Variables | Code | Description |
| ---: | ---: | ---: |
| Fish | Y | Number of Fish Caught |
| Weight | X1 | Weight in $(\mathrm{g})$ |
| Height | X2 | Height in $(\mathrm{cm})$ |

*(Talib, Jaafar, \& Sirwar, 2007)

## Full Algorithm for Alternative Multiple Linear Regression Modelling

```
Title 'Alternative Linear programming with combining robust and bootstrap';
data Biostatistics;
input Fish weigh height;
datalines;
\begin{tabular}{lll}
97.32 & 110.41 & 103.74
\end{tabular}
\(174.52 \quad 111.08 \quad 104.80\)
\(214.56 \quad 114.98 \quad 105.71\)
\(178.44 \quad 114.16 \quad 105.27\)
199.48 112.99 105.45
189.92 115.20 105.34
170.48 113.24 105.11
207.16 117.19 105.66
;
run;
ods rtf file='result_ex1.rtf' ;
/*The next step is performing the procedure of modeling Linear
```


## ALTERNATIVE MULTIPLE LINEAR MODEL REGRESSION MODELING

```
regression model */
procreg data = biostatistics;
modelFish =weigh height;
run;
/* Then do robust regression, in this case, MM-estimation */
ods graphics on;
procrobustreg method = MM fwls data= biostatistics plot=fitplot(nolimits)
plots=all;
modelFish =weigh height/ diagnostics itprint;
output out=resids out=robout r=residual weight=weight outlier=outlier sr=stdres;
run;
ods graphics off;
/* And finally use a bootstrap with case resampling */
ods listing close;
procsurveyselect data = biostatistics out = boot1 method = urs
samprate =1 outhits rep=1000;
run;
/*Combination of Bootstrap Technique with Fuzzy Regression*/
ods listing close;
procoptmodel;
set j= 1..8;
numberFish{j}, weigh{j}, height{j};
read data boot1 into [_n_] Fish weigh height;
/*Print Fish weight height*/
printFish weigh height;
/*Total of Observations*/
number n init 8;
/*Theses three variables are bounded*/
var aw{1..3}>=0;
/*These three variables are not bounded*/
var ac{1..3};
/* Objective Function*/
```


## AHMAD ET AL.

```
min z1= aw[1]*n + sum{i in j} weigh[i]*aw[2]+sum{i in j} height[i]*aw[3];
/*Linear Constraints*/
con c{i in 1..n}:
ac[1]+ weigh[i]*ac[2]+height[i]*ac[3]-aw[1]-weigh[i]*aw[2]-
height[i]*aw[3] <= Fish[i];
con c1{i in 1..n}:
    ac[1]+ weigh[i]*ac[2]+ height[i]*ac[3]+aw[1]+ weigh[i]*aw[2]+
height[i]*aw[3] >= Fish[i];
expand; /* This provides all equations */
solve;
print ac aw;
quit;
ods rtf close;
```


## Results

A higher R-squared value indicated how well the data fit the model and indicates a better model.

Table 2. Goodness-of-fit

| Goodness-of-Fit |  |
| ---: | ---: |
| Statistic | Value |
| R-Square | 0.8199 |
| AICR | 5.5323 |
| BICR | 9.4456 |
| Deviance | 234.4750 |

Method of Multiple linear regression (MLR), we obtained the result as shown in Table 3

Table 4 shows the results by using bootstrapping method for fuzzy regression with $n=1000$. The aim of bootstrapping procedure is to approximate the entire sampling distribution of some estimator by resampling (simple random sampling with replacement) from the original data (Yaffee, 2002). Table 4 summarizes the findings of the calculated parameter.

## ALTERNATIVE MULTIPLE LINEAR MODEL REGRESSION MODELING

Table 3. Parameter Estimates for Final Weighted Least Squares Fit

| Parameter Estimates for Final Weighted Least Squares Fit |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Parameter | DF | Estimate | Standard Error | $\begin{array}{r} \text { 95\% Cor } \\ \text { Lim } \end{array}$ | idence <br> s | Chi-Square | Pr $>$ ChiSq |
| Intercept | 1 | -6334.91 | 608.3789 | -7527.31 | -5142.51 | 108.43 | <. 0001 |
| x 1 | 1 | -3.0164 | 2.1608 | -7.2516 | 1.2188 | 1.95 | 0.1627 |
| x2 | 1 | 65.2183 | 7.5704 | 50.3807 | 80.0559 | 74.22 | <. 0001 |
| Scale | 0 | 7.1356 |  |  |  |  |  |

## Method of Fuzzy Regression (FR) (OPTMODEL)

Table 4. Value of ac and aw

|  | ac | aw |
| ---: | ---: | ---: |
| 1 | -5764.1545 | 0.000000 |
| 2 | -3.0958 | 0.000000 |
| 3 | 59.8722 | 0.075811 |

While using bootstrap procedure, different output for the ac and aw will be obtained:

$$
\begin{aligned}
& \text { ac1 }=-5764.1545 \\
& \text { ac2 }=-3.0958 \\
& \text { ac3 }=59.8722 \\
& \text { aw1 }=0 \\
& \text { aw2 }=0 \\
& \text { aw3 }=0.075811 .
\end{aligned}
$$

The next step is to compare the performance of multiple linear regression and fuzzy regression.

The Fitted Model for Multiple Linear Regressions

$$
\begin{equation*}
Y=-6334.91-3.0164 \text { weight }+62.21 \text { height } \tag{6}
\end{equation*}
$$

Standard Error (608.3789) (2.1608)

## AHMAD ET AL.

The upper limits of prediction interval are computed by coefficient plus standard error

$$
\begin{aligned}
& Y=(-6334.91+608.3789)+(-3.0164+2.1608) \text { weight }+(65.21+7.5704) \text { height } \\
& Y=(-5726.53)+(-0.86) \text { weight }+(72.78) \text { height }
\end{aligned}
$$

The lower limits of prediction interval are computed by coefficient minus standard error

$$
\begin{aligned}
& Y=(-6334.91-608.3789)+(-3.0164-2.1608) \text { weight }+(65.21-7.5704) \text { height } \\
& Y=(-6943.29)+(-5.1772) \text { weight }+(57.6396) \text { height }
\end{aligned}
$$

## The Fitted Model for Fuzzy bootstrap Regression Is

$$
\begin{equation*}
Y=[-5764.1545,0]+[-3.0958,0] \text { weight }+[59.8722,0.075811] \text { height } \tag{7}
\end{equation*}
$$

The upper limits of prediction interval are computed by coefficient plus standard error

$$
\begin{aligned}
& Y=[-5764.1545+0]+[-3.0958+0] \text { weight }+[59.8722+0.075811] \text { height } \\
& Y=[-5764.15]+[-3.10] \text { weight }+[60.00] \text { height }
\end{aligned}
$$

The lower limits of prediction interval are computed by coefficient minus standard error

$$
\begin{aligned}
& Y=[-5764.1545-0]+[-3.0958-0] \text { weight }+[59.8722-0.075811] \text { height } \\
& Y=[-5764.15]+[-3.10] \text { weight }+[59.80] \text { height }
\end{aligned}
$$

The width of prediction intervals in respect of multiple linear regression model and fuzzy regression model corresponding to each set of observed explanatory variables is computed manually.

## ALTERNATIVE MULTIPLE LINEAR MODEL REGRESSION MODELING

Table 5. Average Width for Former Multiple Linear Regression model and Fuzzy Bootstrap Regression Model

| Multiple Linear Regression model |  |  | Fuzzy Bootstrap Regression Model |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
| Lower limit | Upper limit | Width | Lower limit | Upper limit | Width |
| -1535.37 | 1728.71 | 3264.09 | 97.23 | 117.98 | 20.75 |
| -1477.74 | 1800.92 | 3278.66 | 154.95 | 179.50 | 24.55 |
| -1445.48 | 1868.16 | 3313.64 | 200.87 | 222.01 | 21.14 |
| -1466.60 | 1836.84 | 3303.44 | 177.10 | 1988.15 | 21.05 |
| -1450.17 | 1850.95 | 3301.12 | 191.49 | 212.58 | 21.09 |
| -1467.93 | 1841.04 | 3308.99 | 178.06 | 199.13 | 21.07 |
| -1471.06 | 1825.99 | 3297.05 | 170.38 | 191.41 | 21.02 |
| -1459.81 | 1862.62 | 3322.43 | 191.03 | 212.16 | 21.13 |
|  | Average | 3298.68 |  | Average | 21.48 |

From Table 5, average width for former multiple regression was found to be 3298.68 while using fuzzy regression, the average width is 21.48 this indicate that the superiority of fuzzy regression methodology. From this analysis, the most efficient method to obtained relationship between response and explanatory variable is to apply fuzzy regression method compared to linear regression method.

## Conclusion

It was explained how to combine an algorithm between robust, fuzzy regression and the bootstrap method. A small sample size (8 observations only) was used
(a) to apply a bootstrap method in order to achieve an adequate of sample size.
(b) to compare the efficiency between original method and with the bootstrap method.
(c) to give a better understanding on how the algorithm works

According to biostatistics history, all the independent variables that we used in this case were significant to the number of fish caught. Without using bootstrapping, the result shows that two out of eight were significant. It is surprising that, using bootstrapping method (with $n=1000$ ) the entire significant variable are included in the model as the finding from the biostatistics record. This algorithm provides us with the improved understanding of the modified method and underlying of relative contributions. For further study, it is possible to

## AHMAD ET AL.

approach response surface methodology for every each of significant variables in single algorithm.

## References

Amir, W. M., Shafiq, M., Rahim, H. A., Liza, P., \& Aleng, A. (2016). Algorithm for combining robust and bootstrap in multiple linear model regression (SAS). Journal of Modern Applied Statistical Methods, 15(1), 884-892. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol15/iss1/44

Kacprzyk, J., \& Fedrizzi, M. (1992). Fuzzy regression analysis. Warsaw: Omnitech Press.

Ngo, T. H. D., \& La Puente C. A. (2012). The steps to follow in a multiple regression analysis. In Proceedings of the SAS Global Forum 2012 Conference, Orlando, Florida, April 22-25. Retrieved from http://support.sas.com/resources/papers/proceedings12/333-2012.pdf

Talib, B. A., Jaafar, A. H., \& Siwar, C. (2007). Penangkapan ikan oleh nelayan Negeri Sembilan: Satu kajian empirikal. International Journal of Management Studies, 14(1), 219-230. Retrieved from http://myais.fsktm.um.edu.my/10620/

Yaffee, R. A. (2002). Robust regression analysis: Some popular statistical package options. ITS Statistics, Social Science, and Mapping Group, New York State University.

# An Alternative Algorithm and $\mathbf{R}$ Programming Implementation for Least Absolute Deviation Estimator of the Linear Regression Models 

Suraju Olaniyi Ogundele<br>Federal University of Petroleum Resources<br>Effurun, Nigeria.

J. I. Mbegbu<br>University of Benin<br>Benin-City, Nigeria

C. R. Nwosu<br>Nnamdi Azikiwe University<br>Awka, Nigeria

We propose a least absolute deviation estimation method that produced a least absolute deviation estimator of parameter of the linear regression model. The method is as accurate as existing method.

Keywords: Linear regression model, least absolute deviation (LAD), equation of a line, R statistical programming and algorithm

## Introduction

Regression is a statistical methodology that is use to relate a variable of interest, which is called the dependent variable or response variable, to one or more predictors (independent/regressors) variables. The objective of regression analysis is to build a regression model or prediction equation that helps us to describe, predict and control the dependent variable on the basis of the independent variable(s). When we predict the dependent variable for a particular set of values of the independent variables, we wish to place a bound on the error of prediction. The goal is to build a regression model that produces an error bound that will be small enough to meet our needs.

In the simple linear regression model, the Population Regression Function (PRF) is given by:

$$
\begin{equation*}
y=\beta_{0}+\beta_{1} x+\varepsilon \tag{1}
\end{equation*}
$$

Mr. Ogundele is an Assistant Lecturer in the Department of Mathematics / Computer Science. Email him at: ogundele.olaniyi@fupre.edu.ng. J. I. Mbegbu is a Professor in the Department of Mathematics. C. R. Nwosu is an Associate Professor in the Department of Statistics.

## ALGORITHM AND CODE FOR LAD ESTIMATOR

In this model there is only one factor $x$ to explain $y$. All the other factors that affect $y$ are jointly captured by the error term denoted by $\varepsilon$. We typically refer to y as the endogenous or dependent variable and $x$ as the exogenous or independent variable.

The idea of the regression model is to estimate the population parameters, $\beta_{0}$ and $\beta_{1}$ from a given sample. The Sample Regression Function (SRF) is the sample counterpart of the population regression function (PRF). Since the SRF is obtained for a given sample, a new sample will generate different estimates. The SRF, which is an estimation of the PRF is given by:

$$
\begin{equation*}
\hat{y}_{i}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{i} \tag{2}
\end{equation*}
$$

Equation (2) is used to calculate the fitted value $\left(\hat{y}_{i}\right)$ for $y$ when $\mathrm{x}=x_{i}$. In the SRF $\hat{\beta}_{0}$ and $\hat{\beta}_{1}$ are estimators of the parameters $\beta_{0}$ and $\beta_{1}$. For each $x_{i}$ we have an observed value $\left(y_{i}\right)$ and a fitted value $\left(\hat{y}_{i}\right)$. The difference between $y_{i}$ and $\hat{y}_{i}$ is called the residual $\hat{\varepsilon}_{i}$ given by:

$$
\begin{equation*}
\hat{\varepsilon}_{i}=y_{i}-\hat{y}_{i} \tag{3}
\end{equation*}
$$

The ordinary least squares (OLS) method is the most widely used method of parameter estimation. The OLS criteria is to minimize the sum of squared error of prediction

$$
\begin{equation*}
\hat{\varepsilon}_{i}^{2}=\left(y_{i}-\hat{y}_{i}\right)^{2} \tag{4}
\end{equation*}
$$

OLS regression yields estimates for the parameters that have the desirable property of being minimum variance unbiased estimators (Chatterjee \& Hadi, 2006).

Ordinary least squares estimation places certain restrictive assumptions on the random component in the model, the errors of prediction. OLS estimation assumes, among others, that the errors of prediction are normally distributed, with a common error variance at all levels of $\mathrm{X}\left[\varepsilon \sim \mathrm{N}\left(0, \sigma^{2}\right)\right]$. The normality assumption is frequently untenable in practice. Violation of this assumption is often manifested by the presence of outliers in the observed data (Nevitt \& Tam, 1998). Thus data containing outlying values may reflect non-normal error distributions with heavy tails or normal error distributions containing observations
atypical of the usual normal distribution with larger variance than the assumed $\sigma^{2}$. It is well demonstrated that outliers in the sample data heavily influence estimates using OLS regression, sometimes even in the presence of one outlier (Rousseeuw \& Leroy, 1987).

If the assumption that the uncertainties (i.e., errors) in the data are uncorrelated and normally distributed are valid for the data at hand, then for most quantitative experiments, the method of least squares is the "best" analytical technique for extracting information from a set of data. The method is best in the sense that the parameters determined by the least squares analysis are normally distributed about the true parameters with the least possible standard deviations (Wolberg, 2006).

However, the assumption of the general applicability of the normal law of errors has been under attack from the very beginning of the development of linear regression and, in particular, the least squares analysis hinges critically on the existence of the second moment of the error distribution. Thus, if it must assumed the error distribution follows, for instance, a Cauchy distribution or any longtailed distribution having no finite second moment, then the elegant arguments made in favour of the least squares regression estimators become invalid, and thus, it may become mandatory to look for other criteria to find best estimators for the linear regression model (Giloni \& Padberg, 2002). For situations in which the underlying assumptions of OLS estimation are not tenable, the choice of method for parameter estimation is not clearly defined. Thus, the choice of estimation method under non-ideal conditions has been a long-standing problem for methodological researchers (Nevitt \& Tam, 1998). The history of this problem is lengthy with many alternative estimation methods having been proposed and investigated (Birkes \& Dodge, 1993).

Robust estimation refers to the ability of a procedure to produce highly insensitive estimates to model misspecifications. Hence, robust estimates should be good under wide range of possible data generating distributions. In the regression context, under normality with identically and independently distributed errors, the least squares is the most efficient among the unbiased estimation methods. However, when the normality assumption not feasible, it is frequently possible to find estimation methods that are more efficient than the traditionaI least squares. This occurs when the data generating process has fat tails resulting to several outliers compared to the normal distribution. In these cases the least squares becomes highly unstable and sample dependent because of the quadratic weighting, which makes the procedure very sensitive to outlying observations (Pynnonen \& Salmi, 1994). Examples of this type of robust estimation are Huber

## ALGORITHM AND CODE FOR LAD ESTIMATOR

M-estimation, the method of Least Median of Squares, and the method of Least Absolute Deviations (LAD).

The robust LAD estimator is investigated in the present study and so we provide a brief description of the method. LAD was developed by Roger Joseph Boscovich in 1757, nearly 50 years before OLS estimation (see Birkes \& Dodge, 1993 for a review and historical citations). In contrast to OLS estimation which defines the loss function on the residuals as $\Sigma e_{i}{ }^{2}$, LAD finds the slope and $Y$ intercept that minimize the sum of the absolute values of the residuals, $\Sigma\left|e_{i}\right|$.

Although the concept of LAD is not more diffucult than the concept of the OLS estimation but due to computational difficulties in obtaining LAD estimates and lack of exact sampling theory based on such estimates, the LAD method lay in the background and the LS method became popular (Rao \& Toutenburg, 1999). Since there are no exact formulas for LAD estimates, an algorithm is used to iteratively obtain the estimate of the parameters.

## Methodology

## Propose Method

Using the LAD criterion, the model $\hat{y}_{i}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{i}$ should be constructed by two pairs of data points that yield the minimum sum of absolute deviation, our approach is to investigate the sum of absolute deviation generated by all possible different combinations of data points and then select the two data points that produced the least absolute deviation to find the least absolute deviation estimator.

The two points $\left(x_{i}, y_{i}\right)$ and $\left(x_{j}, y_{j}\right)$ yield the following system of equations:

$$
\begin{align*}
y_{i} & =\hat{\beta}_{0}+\hat{\beta}_{1} x_{i}  \tag{5}\\
y_{j} & =\hat{\beta}_{0}+\hat{\beta}_{1} x_{j}
\end{align*}
$$

The solution of equation (5) yield the value of $\hat{\beta}_{0}$ and $\hat{\beta}_{1}$, for the selected pair of data. Subtituting the value of $\hat{\beta}_{1}$ and $\hat{\beta}_{0}$ obtained from (5) into (1), we have

$$
\begin{equation*}
\hat{y}_{i}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{i} \tag{6}
\end{equation*}
$$

## OGUNDELE ET AL.

(6) is then used to determine the sum of absolute deviation of all other data points from the line joining $\left(x_{i}, y_{i}\right)$ and $\left(x_{j}, y_{j}\right)$ by substituting for $x_{i}(i=1,2, \ldots, n)$ of all data points and calculating

$$
\begin{equation*}
\sum_{i=1}^{n}\left|y_{i}-\hat{y}_{i}\right| . \tag{7}
\end{equation*}
$$

This procedure will be repeated for all other combinations of data points and the two data points that yielded the least absolute deviation determine the least absolute deviation estimator.

The presence of personal computers make it possible to evaluate repeated process using any programming language of choice. The R program is employed for all calculations, and the program yields the least absolute deviation estimate for the data.

## Algorithm for Simple Linear Regression

INPUT: Observations of $x$ and $y$ as vectors X and Y .
OUTPUT: Slope and intercept.
Step 1. $\quad$ Set $i=1$ and $j=2$
Step 2. While $i \leq(n-1) ; j \leq n ; i \neq j$, do steps 3 to 4
Step 3. Select the pairs $\left(x_{i}, y_{i}\right)$ and $\left(x_{j}, y_{j}\right)$ from the data and calculate the values of $\hat{\beta}_{1}$ and $\hat{\beta}_{0}$ from the system of equations given by (5).
Step 4. For $i=1$ to $n$, determine the estimated value of $y(\hat{y})$ by substituting for $\left(x_{i}, y_{i}\right)$ in (6) and calculate

$$
\operatorname{AbsDev}(i, j)=\sum_{i=1}^{n}\left|y_{i}-\hat{y}_{i}\right|
$$

Step 5. Determine the minimum value among all $\operatorname{AbsDev}(i, j)$ and select the two data points that produced the minimum value.
Step 6. Print out the values of $\hat{\beta}_{1}$ and $\hat{\beta}_{0}$ that correspond to the two selected data points.
Step 7. If $i>(n-1)$ and $j>n$, then OUTPUT; stop.

## ALGORITHM AND CODE FOR LAD ESTIMATOR

Step 8. Set $i=i+1 ; j+1$ and go to step 2.
Step 9. OUTPUT "Method failed after $i>(n-1)$ and $j>n$ ".

This algorithm is improved upon for multiple linear regression and can be scaled to accommodate the numbers of independent variables present in the data. We provide the algorithm and the R program for the simplest form of the multiple regression with two independent variables

## Algorithm for Multiple Linear Regression

INPUT: Observations of $x_{1}, x_{2}$ and $y$ as vectors $\mathrm{X} 1, \mathrm{X} 2$ and Y .
OUTPUT: Intercept, first parameter and second parameter.

Step 1. $\quad$ Set $i=1, j=2$ and $k=3$
Step 2. While $i \leq(n-2) ; j \leq(n-1) ; k \leq n ; i \neq j \neq k$, do steps 3 to 4
Step 3. $\quad$ Select the pairs $\left(x_{i}, y_{i}\right),\left(x_{j}, y_{j}\right)$ and $\left(x_{k}, y_{k}\right)$ from the data and calculate the values of $\hat{\beta}_{2}, \hat{\beta}_{1}$ and $\hat{\beta}_{0}$ from the system of equations

$$
\begin{align*}
& y_{i}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{1 i}+\hat{\beta}_{2} x_{2 i} \\
& y_{j}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{1 j}+\hat{\beta}_{2} x_{2 j}  \tag{8}\\
& y_{k}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{1 k}+\hat{\beta}_{2} x_{2 k}
\end{align*}
$$

Step 4. For $i=1$ to $n$, determine the estimated value of $y(\hat{y})$ by substituting for $\left(x_{1 i}, x_{2 i}, y_{i}\right)$ in

$$
\begin{equation*}
\hat{y}_{i}=\hat{\beta}_{0}+\hat{\beta}_{1} x_{1 i}+\hat{\beta}_{2} x_{2 i} \tag{9}
\end{equation*}
$$

and calculate

$$
\operatorname{AbsDev}(i, j, k)=\sum_{i=1}^{n}\left|y_{i}-\hat{y}_{i}\right|
$$

Step 5. Determine the minimum value among all AbsDev $(i, j, k)$ and select the three data points that produced the minimum value.

## OGUNDELE ET AL.

Step 6. Print out the values of $\hat{\beta}_{2}, \hat{\beta}_{1}$ and $\hat{\beta}_{0}$ that correspond to the three selected data points.
Step 7. If $i>(n-2), j>(n-1)$ and $k>n$, then OUTPUT; stop.
Step 8. $\quad$ Set $i=i+1 ; j=j+1 ; k=k+1$ and go to step 2.
Step 9. OUTPUT "Method failed after $i \leq(n-2), j \leq(n-1)$ and $k \leq n "$.

## Results

The R program designed by applying this algorithm is presented in the Appendix. Application of the R program written for the algorithm to the data in Table 1 yielded the same results with the iterative method by Birkes and Dodge (1993). The best two data points are given to be at $\left(x_{5}, y_{5}\right)$ and $\left(x_{14}, y_{14}\right)$. The least absolute deviation estimate of the model parameters are

$$
\begin{aligned}
& \hat{\beta}_{0(L A D)}=46.38444 . \\
& \hat{\beta}_{1(L A D)}=-0.53778
\end{aligned}
$$

Then, the LAD regression line is

$$
\hat{y}_{i}=46.38444-0.53778 x_{i}
$$

Table 1. Birth Rate Data

| Country | Birth Rate $\left(\mathbf{y}_{\mathbf{i}}\right)$ | Urban Percentage ( $\mathbf{x}_{\mathbf{i}}$ ) |
| ---: | ---: | ---: |
| Canada | 16.2 | 55 |
| Costa Rica | 30.5 | 27.3 |
| Cuba | 16.9 | 33.3 |
| Dominican Republic | 33.1 | 11.1 |
| El Salvador | 40.2 | 14.2 |
| Guatemala | 38.4 | 13.9 |
| Haiti | 41.3 | 19 |
| Honduras | 43.9 | 33.1 |
| Jamaica | 28.3 | 43.2 |
| Mexico | 33.9 | 28.5 |
| Nicaragua | 44.2 | 37.7 |
| Panama | 28 | 6.8 |
| Trinidad-Tobago | 24.6 | 56.5 |
| United States | 16 |  |

## ALGORITHM AND CODE FOR LAD ESTIMATOR

The R program written for the multiple regression implementation of the proposed method was applied to find the least absolute deviation estimate of the parameter of subset of the supervisor data (Chatterjee \& Hadi, 2006) which includes $\mathrm{Y}, \mathrm{X}_{1}$ and $\mathrm{X}_{2}$. The data is presented in Table 2. The best three data points are given to be at $\left(x_{8}, y_{8}\right),\left(x_{9}, y_{9}\right)$ and $\left(x_{21}, y_{21}\right)$.

The program gives the least absolute deviation estimate of the parameters as

$$
\begin{aligned}
& \hat{\beta}_{0(L A D)}=28.33487 \\
& \hat{\beta}_{1(L A D)}=0.6835637 \\
& \hat{\beta}_{2(L A D)}=-0.172043
\end{aligned}
$$

Table 2. Subset of Supervisor Data

| Y | X 1 | X 2 |
| ---: | ---: | ---: |
| 43 | 51 | 30 |
| 63 | 64 | 51 |
| 71 | 70 | 68 |
| 61 | 63 | 45 |
| 81 | 78 | 56 |
| 43 | 67 | 49 |
| 58 | 75 | 50 |
| 71 | 82 | 72 |
| 72 | 61 | 45 |
| 67 | 53 | 53 |
| 64 | 60 | 47 |
| 67 | 82 | 57 |
| 69 | 77 | 83 |
| 68 | 90 | 54 |
| 77 | 85 | 50 |
| 81 | 60 | 64 |
| 74 | 70 | 65 |
| 65 | 58 | 46 |
| 65 | 40 | 68 |
| 50 | 61 | 33 |
| 50 | 66 | 52 |
| 64 | 37 | 52 |
| 53 | 54 | 42 |
| 40 | 77 | 42 |
| 63 | 75 | 66 |
| 66 | 57 | 58 |
| 78 | 85 | 44 |
| 48 | 82 | 71 |
| 85 |  | 39 |
| 82 |  |  |

## OGUNDELE ET AL.

## Conclusion

The proposed method produced a least absolute deviation estimate that is the same as the one provided by the iterative method by Birkes and Dodge (1993) and other existing methods

## References

Birkes, D., \& Dodge, Y. (1993). Alternative methods of regression. New York, NY: John Wiley \& Sons Inc.

Chatterjee, S., \& Hadi, A. S. (2006). Regression analysis by example. New York, NY: John Wiley \& Sons Inc.

Giloni A., \& Padberg, M. (2002). Alternative methods of linear regression. Mathematical and Computer Modelling, 35(3/4), 361-374.
doi: 10.1016/S0895-7177(01)00170-4
Nevitt, J., \& Tam, H. P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model. Multiple Linear Regression Viewpoints, 25, 54-69.

Pynnonen, S. \& Salmi, T. (1994). A report on least absolute deviation regression with ordinary linear programming. The Finnish Journal of Business Economics, (1), 33-49.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Australia. Retrieved from http://www.R-project.org/.

Rao, R. C. \& Toutenburg, H. (1999). Linear models: Least squares and alternatives ( $2^{\text {nd }}$ ed.). New York, NY: Springer-Verlag, Inc.

Rousseeuw, P. J. \& Leroy, A. M. (1987). Robust regression and outlier detection. New York, NY: John Wiley \& Sons.

Wolberg, J. (2006). Data analysis using the method of least squares. Berlin Heidelberg: Springer-Verlag.

## ALGORITHM AND CODE FOR LAD ESTIMATOR

## Appendix

```
# Simple Linear Regression
# Least Absolute Deviation Estimator (LAD)
Y<-c(16.2,30.5,16.9,33.1,40.2,38.4,41.3,43.9, 28.3,33.9,44.2, 28, 24.6,16)
X<-c(55,27.3,33.3,37.1,11.5,14.2,13.9,19,33.1,43.2, 28.5,37.7,6.8,56.5)
p<-1
n<-length(Y)
Const<-rep(1,(p+1))
l<-0;AbsError<-0;EstError<-0;VecLAD<-c();Vecj<-c();Veck<-c();Vecx<-c();Vecy<-
c();VecB0<-c();VecB1<-c()
TrialModel1<-function(Xi,Yi,B0,B1){(abs(Yi-(B0+(B1*Xi))))}
LAD1<-function(n,X,Y,B0,B1){
    for (i in 1:n){Xi<-X[i];Yi<-Y[i]
        EstError<-TrialModel1(Xi,Yi,B0,B1)
        AbsError<-AbsError+EstError
    }
    EstLAD<-AbsError
}
Kstart<-2
for (j in 1:(n-1)){xj<-X[j];yj<-Y[j];
    for (k in Kstart:n){xk<-X[k];yk<-Y[k]
        if(k==n){Kstart<-Kstart+1}
        Vecx<-c(xj,xk)
        Vecy<-c(yj,yk)
        A<-cbind(Const,Vecx)
        Det<-round(det(A),4)
        if(Det!=0){
            Beta<-solve(A,Vecy)
            B0<-Beta[1]
            B1<-Beta[2]
            LADEst<-LAD1(n,X,Y,B0,B1)
            l<-l+1
                VecLAD[l]<-c(LADEst)
                    Vecj[l]<-c(j)
                    Veck[l]<-c(k)
                    VecB0[l]<-c(B0)
                    VecB1[l]<-c(B1)
```


## OGUNDELE ET AL.

```
            }
    }
}
LADLoc<-sort.int(VecLAD,index.return=TRUE)
PairLoc<-LADLoc$ix[1]
Pair1<-Vecj[PairLoc]
Pair2<-Veck[PairLoc]
Intercept<-round(VecB0[PairLoc],5)
Slope<-round(VecB1[PairLoc],5)
Label1<-"Best two data points for LAD estimate are";Label2<-"and"
Label3<-"The intercept of the LAD regression is"
Label4<-"The slope of LAD regression is"
Label1;Pair1;Label2;Pair2;Label3;Intercept;Label4;Slope
# Multiple Linear Regression
# Least Absolute Deviation Estimator (LAD)
Y<-
c(43, 63, 71, 61, 81, 43, 58, 71, 72, 67, 64, 67, 69, 68, 77, 81, 74, 65, 65, 50, 50, 64, 53, 40, 63, 66,
78,48,
85, 82)
X1<-
c(51, 64, 70, 63, 78, 55, 67, 75, 82, 61, 53, 60, 62, 83, 77, 90, 85, 60, 70, 58, 40, 61, 66, 37,54,77,
75,57,
85,82)
X2<-
c(30, 51, 68, 45, 56, 49, 42, 50, 72, 45, 53, 47, 57, 83, 54, 50, 64, 65, 46, 68, 33, 52, 52, 42, 42, 66,
58,44,
71,39)
p<-2
n<-length(Y)
Const<-rep(1, (p+1))
l<-0;AbsError<-0; EstError<-0;VecLAD<-c();VecB0<-c();VecB1<-c();VecB2<-
c();Vecx1<-c();Vecx2<-c();Vecy<-c();Vecj<-c()
Veck<-c();Vecv<-c()
TrialModel2<-function(Yi,X1i,X2i,B0,B1,B2){(abs(Yi-(B0+(B1*X1i)+(B2*X2i))))}
LAD2<-function(n, Yi, X1i, X2i, B0, B1, B2){
for (i in 1:n){X1i<-X1[i];X2i<-X2[i];Yi<-Y[i]
```


# ALGORITHM AND CODE FOR LAD ESTIMATOR 

```
    EstError<-TrialModel2(Yi,X1i,X2i,B0,B1,B2)
    AbsError<-AbsError+EstError
    }
    EstLAD<-AbsError
}
Kstart<-1;Vstart<-2;EndCount<-3
for (j in 1:(n-2)){x1j<-X1[j];x2j<-X2[j];yj<-Y[j];
    Kstart<-Kstart+1
    for (k in Kstart:(n-1)){x1k<-X1[k];x2k<-X2[k];yk<-Y[k]
        if(Vstart<n){
            Vstart<-Vstart+1
        }else{
            EndCount<-(EndCount+1)
            Vstart<-EndCount
        }
            for (v in Vstart:n){x1v<-X1[v];x2v<-X2[v];yv<-Y[v]
            Vecx1<-c(x1j,x1k,x1v)
            Vecx2<-c(x2j,x2k,x2v)
            Vecy<-c(yj,yk,yv)
            A<-cbind(Const,Vecx1,Vecx2)
            Det<-round(det(A),4)
            if(Det!=0){
                Beta<-solve(A,Vecy)
                B0<-Beta[1]
                B1<-Beta[2]
                B2<-Beta[3]
                LADEst<-LAD2(n,Yi,X1i,X2i,B0,B1,B2)
                l<-l+1
                VecLAD[l]<-c(LADEst)
                VecB0[l]<-c(B0)
                VecB1[l]<-c(B1)
                VecB2[1]<-c(B2)
                Vecj[l]<-c(j)
                Veck[l]<-c(k)
                Vecv[l]<-c(v)
            }
        }
}
```


## OGUNDELE ET AL.

```
}
LADLoc<-sort.int(VecLAD,index.return=TRUE)
PairLoc<-LADLoc$ix[1]
FirstPoint<-Vecj[PairLoc]
SecondPoint<-Veck[PairLoc]
ThirdPoint<-Vecv[PairLoc]
Constant<-VecB0[PairLoc]
Beta1<-VecB1[PairLoc]
Beta2<-VecB2[PairLoc]
Label1<-"The first LAD data point is";Label2<-"The second data point
is";Label3<-"The third data point is"
Label4<-"The constant of LAD regression model is";Label5<-"The first parameter
is"
Label6<-"The second parameter is"
Label1;FirstPoint;Label2;SecondPoint;Label3;ThirdPoint;Label4;Constant;Label5;Be
ta1;Label6;Beta2
```


## Statistical Software <br> Applications and Review

# Fitting Flexible Parametric Regression Models with GLDreg in R 

Steve Su

Covance
Sydney, Australia

This article outlines the functionality of the GLDreg package in R which fits parametric regression models using generalized lambda distributions via maximum likelihood estimation and L moment matching. The main advantage of GLDreg is the provision of robust regression lines and smooth regression quantiles beyond the capabilities of existing known methods.

Keywords: Regression model, quantile regression, generalized lambda distributions, GLDreg, GLDEX, R

## Package GLDreq in R

The GLDreg package in R is designed to implement the Generalized Lambda Distribution (G $\lambda \mathrm{D}$ ) regression model outlined in Su (2015) with some extensions. Currently, it is possible to fit G $\lambda \mathrm{D}$ regression to data using maximum likelihood estimation (MLE) (Su, 2007a; b) and L moment matching (Asquith, 2007; Karvanen \& Nuutinen, 2008). Users may also chose initial values to start the model building process, or use the default searching algorithm using the ordinary least square regression model (Su, 2015). The GLDreg package also allows user to fit quantile regressions parametrically and non-parametrically by: 1) fixing the intercept, 2) fixing coefficients other than the intercept, and 3) allowing all coefficients to vary.

The GLDreg package requires GLDEX (Su, 2007a; 2010). The GLDEX 2.0.0.1 package has a faster implementation of the $G \lambda D$ fitting algorithm compared to its predecessors. This is because a number of frequently used codes have been written in C. In addition, the GLDEX 2.0.0.1 package has faster maximum likelihood fitting functions fun.RMFMKL.ml.m, fun.RPRS.ml.m for FKML

## STEVE SU

(Freimer, Kollia, Mudholkar, \& Lin, 1988) and RS (Ramberg \& Schmeiser, 1974) G $\lambda$ Ds.

## Background

Traditionally, the Pearson and Johnson systems are considered to be the standard approaches to identifying and fitting different types of statistical distributions to data. However, these systems require different mathematical functions to cover a range of symmetric and asymmetric distributions. Tukey's lambda distribution contains a class of symmetric distributions which can approximate a number of common distributions such as Normal and Cauchy distributions, and provide an indication as to whether heavy tailed distributions are needed. Ramberg and Schmeiser (1974) then generalized Tukey's lambda distribution to include both symmetric and asymmetric distributions, and this became known as the G $\lambda \mathrm{D}$. Although defined as a single mathematical function, $\mathrm{G} \lambda \mathrm{D}$ can cover a broad range of statistical distributions which is much more efficient than the use of Pearson and Johnson systems involving several mathematical functions. Since then, the flexibility of $\mathrm{G} \lambda \mathrm{D}$ has attracted a number of researchers. Today, along with the increased computation power and the introduction of dedicated packages for $G \lambda D$ in $R$, it is now possible to fit $G \lambda D$ to data and extend the use of $G \lambda D$ in many areas of statistical analysis.

## Generalized Lambda Distributions

The RS G $\lambda$ D (Ramberg \& Schmeiser, 1974) is defined by its inverse distribution function:

$$
\begin{equation*}
\mathrm{F}^{-1}(u)=\lambda_{1}+\frac{u^{\lambda_{3}}-(1-u)^{\lambda_{4}}}{\lambda_{2}}, \quad 0 \leq u \leq 1 \tag{1}
\end{equation*}
$$

From (1), $\lambda_{1}, \lambda_{2}, \lambda_{3}, \lambda_{4}$ are respectively the location, inverse scale, and shape parameters of generalized lambda distribution $\operatorname{G} \lambda \mathrm{D}\left(\lambda_{1}, \lambda_{2}, \lambda_{3}, \lambda_{4}\right)$. Note that $\lambda_{3}, \lambda_{4}$ are both shape parameters. Karian, Dudewicz, and McDonald (1996) noted that $G \lambda D$ is defined only if

$$
\frac{\lambda_{2}}{\lambda_{4}(1-u)^{\lambda_{4}-1}+\lambda_{3} u^{\lambda_{3}-1}} \geq 0
$$

## GLDreg IN R

where $0 \leq \mathrm{u} \leq 1$.
Freimer et al. (1988) described another distribution known as FKML/FMKL $G \lambda D$. This distribution is slightly different to RS G $\lambda \mathrm{D}$ with respect to formulation. The FKML/FMKL G $\lambda \mathrm{D}$ is defined as:

$$
\begin{equation*}
\mathrm{F}^{-1}(u)=\lambda_{1} \frac{\frac{u^{\lambda_{3}}-1}{\lambda_{3}}-\frac{(1-u)^{\lambda_{4}}-1}{\lambda_{4}}}{\lambda_{2}}, \quad 0 \leq u \leq 1 \tag{2}
\end{equation*}
$$

Under (2), $\lambda_{1}, \lambda_{2}, \lambda_{3}, \lambda_{4}$ are respectively the location, scale, and shape parameters of the generalized lambda distribution. Again, both $\lambda_{3}, \lambda_{4}$ are shape parameters. Technically, the correct abbreviation for this distribution is FKML distribution, based on the correct ordering of authorship in the original paper. However, the use of the term FMKL distribution had been widespread in the literature and this package, along with GLDEX 2.0.0.1, will allow both "fkml" and "fmkl" specifications in the implementation of FKML distribution.

The fundamental motivation for the development of FKML G $\lambda$ D is that the distribution is defined over all $\lambda_{3}$ and $\lambda_{4}$ (Freimer et al., 1988). The only restriction on FKML G $\lambda \mathrm{D}$ is $\lambda_{2}>0$.

An extensive discussion of the shapes and properties of $\mathrm{G} \lambda \mathrm{Ds}$ can be found in the original papers (Ramberg \& Schmeiser, 1974; Freimer et al., 1988), as well as in subsequent works such as in Su (2015). Note the probability density functions of $\mathrm{G} \lambda \mathrm{Ds}$ is obtained by observing $\mathrm{F}^{-1}(u)=x$ and so

$$
\mathrm{f}(x)=\frac{1}{\frac{d \mathrm{~F}^{-1}(u)}{d u}}
$$

This is calculated using the Newton-Raphson method in GLDEX.

## Sketch of the Regression Fitting Mechanism

The full fitting algorithm of $\mathrm{G} \lambda \mathrm{D}$ regression models is provided in Su (2015). This section sketches the procedure in building these models. Consider the following regression model

$$
\begin{equation*}
Y=X \beta+\varepsilon \tag{3}
\end{equation*}
$$

## STEVE SU

The goal is to find estimated coefficients for $\beta$ under the condition $\varepsilon \sim \mathrm{G} \lambda \mathrm{D}\left(\lambda_{1}, \lambda_{2}, \lambda_{3}, \lambda_{4}\right)$, with the property such that $\mathrm{E}(\varepsilon)=0$. Consequently, the average deviation of actual and estimated values should be zero.

When deriving the probability density function of $\mathrm{G} \lambda \mathrm{Ds}, \lambda_{1}$ will vanish, which means this parameter has no bearing on the overall shape of the distribution, only its location. The consequence is that it is possible to shift the G $\lambda$ Ds to have exactly zero mean by calculating $\lambda_{1}$ after estimating $\lambda_{2}, \lambda_{3}, \lambda_{4}{ }^{1}$

The idea of fitting the $\mathrm{G} \lambda \mathrm{D}$ regression in $\mathrm{Su}(2015)$ is to find a set of $\hat{\beta}$ by modelling the residuals using $G \lambda D$ through maximum likelihood estimation or L moment matching. The zero mean residual line is achieved by only allowing $\lambda_{2}, \lambda_{3}$, $\lambda_{4}$ to vary in the optimization process, and the intercept of the line is adjusted to ensure the residuals add up to zero. Statistical properties of G $\lambda \mathrm{D}$ regression coefficients are obtained by recreating actual values $y_{k}=\hat{y}_{k}+\hat{\varepsilon}_{k}$ for $k=1,2,3, \ldots$, $n$ observations by simulating $\hat{\varepsilon}_{k} \sim G \lambda D\left(\hat{\lambda}_{1}, \hat{\lambda}_{2}, \hat{\lambda}_{3}, \hat{\lambda}_{4}\right)$ and refitting the entire model and repeating the process, say 1000 times. This can all be done using the GLD.lm.full function and the goodness of fit of the model is assessed using the Kolmogorov-Smirnoff test and QQ plots. To run a simple model without any simulations, the function $G L D . l m$ can be used instead.

Once a reference $G \lambda D$ regression line is found, quantile regression can be obtained by 1) fixing the intercept, 2) fixing coefficients other than the intercept, and 3) allowing all coefficients to vary. Case 1 and 3 are designed to fit non parallel lines for heteroskedastic data, and case 2 is primarily designed to fit parallel lines for homoskedastic data.

GLDreg allows non-parametric and parametric fitting of quantile regression lines. The non-parametric approach uses the least squares approach to find a $q$-th quantile $G \lambda D$ line such that the percentage of observations below the line corresponds to the $q$-th quantile. The parametric approach uses the least squares approach to find a $q$-th quantile $\mathrm{G} \lambda \mathrm{D}$ line such that the percentage of observations below the line (under a G $\lambda \mathrm{D}$ fit) corresponds to the $q$-th quantile. In the case of 1 ) and 3), the Nelder-Mead simplex algorithm from optim is used in the optimization process. In the case of 2), the Brent method from optim is used instead. The initial values for non-parametric quantile regression optimization are taken from the sample quantile of simulated regression coefficients obtained during the model building process. For the parametric approach, the initial values are taken from the non-parametric quantile regression. The estimates of both parametric and nonparametric $\mathrm{G} \lambda \mathrm{D}$ quantile regression can be obtained using a single wrapper function (GLD.quantreg).

## GLDreg IN R

## Using the Package

GLDreg can be installed from CRAN ${ }^{2}$ through the R interface. Once installed, the following command will load the package.

```
> Library(GLDreg)
Loading required package: GLDEX
Loading required package: cluster
```

As usual, ?GLDreg will take user to the main help menu. Table 1 shows the main functions available under GLDreg.

Table 1. List of main functions

| Purpose | Function |
| :---: | :---: |
| Fit G $\lambda$ D Regression only | G入D.Im |
| Fit G $\lambda$ D Regression and obtain statistical properties through simulation | GLD.Im.full |
| Fit G ${ }^{\text {d }}$ quantile regression | GLD.quantreg |
| Plot summary graphics of G ${ }^{\text {d }}$ d regression | summaryGraphics.G入D.Im |
| Plot quantile regression lines | fun.plot.q |

Table 2. Fitting RS or FKML GAD regression using MLE or L moment matching under GLD.Im or GLD.Im.full

| Type of GXD | Type of estimation | Param | fun |
| ---: | ---: | ---: | ---: |
| $R S$ | $M L E$ | "rs" | fun.RPRS.ml.m or fun.RPRS.ml |
| $R S$ | $L$ moment matching | "rs" | fun.RPRS.Im |
| FKML | $M L E$ | "fmkl" or "fkml" | fun.RMFMKL.ml.m or fun.RMFMKL.ml |
| FKML | $L$ moment matching | "fmkl" or "fkml" | fun.RMFMKL.Im |

GLDreg currently implements MLE and L moment matching for G $\lambda \mathrm{D}$ regression. The associated param and fun inputs for GLD.lm and GLD.lm.full are given in Table 2. The fun.RPRS.ml.m and fun.RMFMKL.ml.m functions are faster implementations of MLE than the previous fun.RPRS.ml and fun.RMFMKL.ml functions.

## STEVE SU

## Engel Dataset

This example uses the well-known Engel dataset from the quantreg package in R . The following fits a full FKML G $\lambda \mathrm{D}$ regression model using MLE, along with simulations (default is 1000 runs), and then fits a quantile regression for quantiles from 0.1 to 0.9 in 9 equal spacings with fixed intercept and varying slopes.

```
> Library(quantreg)
> data(engel)
> set.seed(1000)
> engel.fit.all<-GLD.LmfulL(foodexp~income, data=engel,
+ param="'fmkl",fun=fun.RMFMKL.mL.m,summary.plot=F)
> result<-GLD.quantreg(seg(0.1,.9,length=9),
+engel.fit.all,intercept="fixed")
```

There are warnings associated with using the Nelder-Mead algorithm for single parameter optimization problems. To check whether optimization has indeed been achieved for quantile regression, the easiest way is to check whether the proportion of the fitted quantile line below the response variable corresponds to the level of quantile specified. The multiplication of 100 is to convert the quantiles into percentages, and the aim is to see how close these are to $10,20,30,40,50,60,70$, 80 and 90.

```
> sapply(1:9,function(i) sum(engel$foodexp-
cbind(1,engel$income)%*%(result[1:2,i])<0)/nrow(engel)*100)
[1] 8.085106 22.127660 30.638298 38.723404 51.489362 58.723404
67.659574 82.978723 90.638298
```

The results are reasonably close to the intended quantiles. Note differences occur because the optimization here is based on matching the quantile of fitted $G \lambda D$ rather than the empirical data, the idea being that if the fitted $G \lambda D$ is close to the actual distribution, the quantile lines obtained here will be more robust to changes in empirical data.

Alternatively, quantile lines could be obtained using the non-parametric approach using the following, and checked whether the desired quantile line has

## GLDreg IN R

been achieved despite warnings from the optim function. Obviously, the match is better, simply because optimization was done based on empirical data.

Once the quantile lines are obtained, it is possible to plot them; in this case, the parametric $G \lambda D$ quantile lines are plotted in Figure 1.

```
> fun.plot.q(x=engel$income, y=engel$foodexp,fit=engel.fit.all[[1]],
```

result)


Figure 1. FKML G $\lambda$ D quantile regression for Engel dataset

It is also possible to speed up GLD.lm.full by reducing the number of simulations to 100 by setting $n=100$, if the primary purpose is to obtain quantile regression rather than looking at the statistical properties of the coefficients. GLDreg also provides a graphic summary of regression coefficients based on object obtained from GLD.lm.full, using the summaryGraphics.gld.lm function, and the graphics are shown in Figure 2 and Figure 3.

## STEVE SU

> summaryGraphics.gld.lm(engel.fit.all)

Shown in Figure 2 is the statistical distribution of coefficients obtained using simulation, and the $95 \%$ interval is obtained directly from simulated results. Shown in Figure 3 is how well FKML G $\lambda \mathrm{D}$ fits the data; in this case the fit is quite good, with Kolmogorov-Smirnoff goodness of fit $p$-value exceeding 5\% and the QQ plot indicating a close fit.

| Parameter | Estimate | Density | Summary 95\% interval: |
| :---: | :---: | :---: | :---: |
| (Intercept) | 32.2 |  | -20.5, 85.1 |
| income | 0.603 |  | 0.55, 0.649 |

Figure 2. FKML G AD quantile regression for Engel dataset coefficient summary plot

## GLDreg IN R



Figure 3. $\operatorname{FKML}$ G $\lambda$ D quantile regression for Engel dataset - QQ plots


Figure 4. Comparison of different regression techniques for Engel dataset

## STEVE SU

Note the FKML G $\lambda$ D regression line is the most robust to outliers compared to linear regression, robust regression, and quantile regression at median. This can be seen using the following code, which is displayed in Figure 4.

```
> Library(quantreg)
> Library(MASS)
> par(mfrow=c(1,1))
> plot(foodexp~income,data=engel,xLab="income",ylab="food expense",
+ main="Belgian Engel Dataset")
> abline(Lm(foodexp~income,data=engel),lty=2,lwd=3)
> abline(rlm(foodexp~income,data=engel),lty=4,lwd=3)
> abline(rq(foodexp~income,data=engel),lty=3, lwd=2)
> abline(engel.fit.all[[1]][[3]][1:2],Lty=1,lwd=3)
> Legend("bottomright",c("Standard Regression","Robust Regression",
+ "Quantile Regression","GLD Regression"),Lty=c(2,4,3,1),Lwd=c(3,3,2,3))
```


## Simulated Motorcycle Accident Dataset

The mcycle dataset is a simulated motorcycle accident dataset from MASS library in R. It is possible to fit splines to this dataset, and a reasonable strategy is to fit time before 15 seconds separately as was done in Su (2015). In the same fashion as in Engel dataset analysis, the full $G \lambda D$ regression is fitted first, followed by evaluation of quantile regression lines, for which parallel lines appear to be quite suitable for this dataset. The only difference is that the modelling is split into two parts. In this case, the first part (before 15 seconds) was fitted using splines with 8 degrees of freedom using RS G $\lambda \mathrm{D}$ via maximum likelihood estimation. The second part (greater or equal to 15 seconds) was fitted using splines with 15 degrees of freedom using FKML G $\lambda$ D via maximum likelihood estimation. The code to produce Figure 5 is given below.

```
> Library(MASS)
> Library(splines)
> cutoff<-15
> mcycle.part1<-mcycle[mcycle$times<cutoff,]
```


## GLDreg IN R

```
> mcycle.part2<-mcycle[mcycle$times>=cutoff,]
> mcycle.p1.fit<-GLD.lm.full(accel~bs(times,df=8),
+ data=mcycle.part1,param="rs",fun=fun.RPRS.ml.m, summary.plot=F,n=100)
> mcycle.p2.fit<-GLD.lm.fulL(accel~bs(times, df=15),
+ data=mcycle.part2,param="fmkL",fun=fun.RMFMKL.mL.m,summary.plot=F,
n=100)
> mcycle.p1.fit.quant<-GLD.quantreg(seq(0.1,.9, Length=9),
+ mcycle.p1.fit,slope="fixed")
> mcycle.p2.fit.quant<-GLD.quantreg(seq(0.1,.9,length=9),
+ mcycle.p2.fit,slope="fixed")
> plot(mcycle,ylim=c(-150,75))
> sapply(1:9,function(i)
+ Lines(mcycle$times,c(
+ cbind(1,bs(mcycle[which(mcycle$times<cutoff),]$times,
+ df=8))%*%mcycle.p1.fit.quant[1:9,i],
+ cbind(1,bs(mcycle[which(mcycle$times>=cutoff),]$times,
+ df=15))%*%mcycle.p2.fit.quant[1:16,i])))
```



Figure 5. Simulated motorcycle accident dataset

## STEVE SU

## Modified Crime Data

A fair exposition of $G \lambda D$ regression needs to include a discussion of its weaknesses. Being a regression model involving numerical computations, the initial values play an important role in deriving the final model results. While all numerical optimization problems suffer from bad choice of initial values and it is a general limitation of the Nelder-Mead optimization method, it is possible to still attain the attractive robust property of $G \lambda D$ regression by simply choosing a better set of initial values.

This example used the crime dataset from UCLA website ${ }^{3}$ and examined the relationship between crime rate and percentage of single family parents. To examine the behavior of regression models under extreme outliers, two data points were altered in this illustration. The coding involved in extracting and altering the dataset is given below.

```
> require(foreign)
> require(MASS)
> require(GLDreg)
> require(quantreg)
> cdata <- read.dta("http://www.ats.ucLa.edu/stat/data/crime.dta")
> mcdata<-subset(cdata,select=c("crime", "single"))
# Altering the data to create extreme outliers
> mcdata[51,1]<-10
>mcdata[25,2]<-22
```

If the modeler begins with RS G $\lambda \mathrm{D}$ using MLE using the default initial values, the resulting $G \lambda D$ regression line is shown in panel $A$ of Figure 6 , which is perhaps the worst model among all others (linear regression, robust regression, and quantile regression). This happens because the default initial values was based on linear regression and a less-than-optimal line was found using the Nelder-Mead simplex algorithm.

```
> \(\operatorname{par}(m f r o w=c(2,2))\)
> plot(crime~single,data=mcdata,ylab="Violent crimes per 100,000 people",
```


## GLDreg IN R

```
+ xlab="Percentage of population that are single parents",
+ main="GLD Regression fails due to
+ improper initial value selection (A)")
> abline(rlm(crime~single,data=mcdata),lty=2,lwd=3)
> abline(Lm(crime~single,data=mcdata),Lty=4,Lwd=3)
> abline(rq(crime~single,data=mcdata),Lty=3, Lwd=2)
> abline(GLD.Lm(crime~single,data=mcdata,
+ param="rs",fun=fun.RPRS.mL.m, diagnostics=FALSE)[[3]][1:2],
+ lty=1,lwd=3)
> Legend("topright", c("Standard Regression", "Robust Regression",
+ "Quantile Regression", "GLD Regression"),
+ Lty=c(2,4,3,1),lwd=c(3,3,2,3),bg="white")
```

It is possible to improve the quality of fit by using different initial values. In the following, initl uses quantile regression coefficients, init2 uses robust regression coefficients, and init 3 uses $G \lambda D$ regression coefficients obtained by removing the outliers at the bottom right corner of the graph.

```
> init1<-rq(crime~single,data=mcdata)$coeff
init2<-rlm(crime~single, data=mcdata)$coeff
init3<-GLD.lm(crime~single,
data=rbind(mcdata[1:24,],mcdata[26:50,]),
param="rs",fun=fun.RPRS.ml.m,diagnostics=FALSE)[[3]][1:2]
```

The modeler can then refit RS G $\lambda \mathrm{D}$ regression model using maximum likelihood estimation by using these initial values using the following code.

```
> plot(crime~single,data=mcdata,
+ ylab="Violent crimes per 100,000 people",
+ xlab="Percentage of population that are single parents",
+ main="GLD Regression with quantile regression coefficients
+ as initial values (B)")
> abline(rlm(crime~single,data=mcdata),lty=2,lwd=3)
> abline(Lm(crime~single,data=mcdata),lty=4, lwd=3)
> abline(rq(crime~single,data=mcdata),Lty=3,lwd=2)
```


## STEVE SU

```
> abline(GLD.lm(crime~single,data=mcdata,
+ param="rs",fun=fun.RPRS.mL.m, diagnostics=FALSE,
+ init=init1)[[3]][1:2],lty=1,lwd=3)
> plot(crime~single,data=mcdata,
+ ylab="Violent crimes per 100,000 people",
+ xlab="Percentage of population that are single parents",
+ main="GLD Regression with robust regression coefficients
+ as initial values (C)")
> abline(rlm(crime~single,data=mcdata),lty=2,lwd=3)
> abline(Lm(crime~single,data=mcdata),Lty=4,lwd=3)
> abline(rq(crime~single,data=mcdata),Lty=3,lwd=2)
> abline(GLD.lm(crime~single,data=mcdata,
+ param="rs",fun=fun.RPRS.ml.m,diagnostics=FALSE,
+ init=init2)[[3]][1:2],lty=1,lwd=3)
plot(crime~single, data=mcdata,
+ ylab="Violent crimes per 100,000 people",
+ xlab="Percentage of population that are single parents",
+ main=" GLD Regression using modified data
fitted by GLD regression as initial values (D)")
> abline(rlm(crime~single,data=mcdata),lty=2,lwd=3)
> abline(Lm(crime~single,data=mcdata),lty=4, Lwd=3)
> abline(rq(crime~single,data=mcdata),Lty=3,lwd=2)
> abline(GLD.lm(crime~single,data=mcdata,
+ param="rs",fun=fun.RPRS.mL.m, diagnostics=FALSE,
+ init=init3)[[3]][1:2],\ty=1,lwd=3)
```

The results are shown in Panel B to D in Figure 6. In these cases, the G $\lambda \mathrm{D}$ regression line is now the most robust among all regression lines. The importance of checking the resulting fit and trying out different sets of initial values for optimization is highlighted in this example.

## GLDreg IN R



Figure 6. Modified crime data set

## Occupation Dataset

The last example involves modelling beyond a single explanatory variable on the Duncan dataset from car in R. In this case, a FKML G $\lambda \mathrm{D}$ regression using L moment estimation was fitted to the dataset with initial values taken from robust linear regression. It is worthwhile to plot the fitted values against actual values, and to see whether the observations correspond to a 45 degree line from the origin. Large deviations from a linear trend would suggest the model does not fit well with respect to the response variable. The coding to carry the above tasks is given below.

The resulting fit is similar to the robust regression result, and QQ plots and high $p$-value (Figure 7) suggest the G $\lambda \mathrm{D}$ regression fit is a good one. Statistical properties obtained (Figure 8) suggest that the most important variable related to income is prestige. Figure 9 shows that, while the fitted values are not extremely accurate with respect to actual values, the general linear trend is still observed, indicating that the linear form of the model is appropriate.

As a final remark, it can be worthwhile to fit quantile regression lines for fitted values against actual values to give a range of likely actual values that could be obtained using the linear model.

[^40]
## STEVE SU

```
> Library(car)
>data(Duncan)
> job.fit.fulL<-GLD.lm.fulL(income~education+prestige,
+ data=Duncan, param="fkmL",fun=fun.RMFMKL.Lm,
+ init=rlm(income~education+prestige,data=Duncan)$coeff,
+ summary.plot=F)
> summaryGraphics.gld.lm(job.fit.full)
    # Plot actual vs fitted observations:
> plot(job.fit.fulL[[1]]$Fitted,
+ job.fit.fulL[[1]]$y,xLab="Fitted",ylab="Actual")
> abline(0,1)
```



Figure 7. Occupation data set modelling - QQ plots

## GLDreg IN R

| Parameter | Estimate | Density | Summary <br> $95 \%$ interval: |
| :---: | :---: | :---: | :---: |
| (Intercept) | 7.69 | 0.109 |  |
| education |  |  | $-3.07,19.8$ |

Figure 8. Occupation data set modelling - summary plots


Figure 9. Occupation data set modelling - actual vs. fitted values

## STEVE SU

## Conclusion

The flexibility of $\mathrm{G} \lambda \mathrm{D}$ regression models was illustrated; rather than confining the regression model to only examining the mean or median as is the case of linear regression or classic quantile regression model, the $G \lambda D$ regression models attempt to fit a line that represent a typical value for the dataset which may or may not correspond to standard measures such as the mean or median. Some extensions were also shown beyond the methodology described in Su (2015), and two different optimization schemes (L moment matching and maximum likelihood estimation) were implemented to increase the versatility of $G \lambda D$ regression in different modelling situations. Potential shortcomings of G $\lambda \mathrm{D}$ regression in initial value selection were illustrated, as well as how different initial values could lead to better regression fits, which is a problem for many model fitting problems involving numerical computations. Since its inception, G $\lambda$ Ds have been used to model a wide range of empirical data, and the flexibility and robustness of $G \lambda D$ regression is particularly attractive either as a check for standard results or as a replacement. It is hoped that the introduction of this work would encourage future researchers to develop new methodological improvements to further enhance the usability of $G \lambda D s$ in practice.

## References

Asquith, W. (2007). L-moments and TL-moments of the generalized lambda distribution. Computational Statistics and Data Analysis, 51(9), 4484-4496. doi: 10.1016/j.csda.2006.07.016

Freimer, M., Kollia, G., Mudholkar, G. S., \& Lin, C. T. (1988). A Study of the generalized Tukey lambda family. Communications in Statistics - Theory and Methods, 17(10), 3547-3567. doi: 10.1080/03610928808829820

Karian, Z., Dudewicz, E., \& McDonald, P. (1996). The extended generalized lambda distribution systems for fitting distributions to data: history, completion of theory, tables, applications, the "final word" on moment fits. Communications in Statistics - Computation and Simulation, 25(3), 611-642. doi: 10.1080/03610919608813333

Karvanen, J., \& Nuutinen, A. (2008). Characterizing the generalized lambda distribution by L-moments. Computational Statistics and Data Analysis, 52(4), 1971-1983. doi: 10.1016/j.csda.2007.06.021

## GLDreg IN R

Ramberg, J., \& Schmeiser, B. (1974). An approximate method for generating asymmetric random variables. Communications of the Association for Computing Machinery, 17(2), 78-82. doi: 10.1145/360827.360840
$\mathrm{Su}, \mathrm{S} .(2007 \mathrm{a})$. Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R. Journal of Statistical Software, 21(9), 1-17. doi: 10.18637/jss.v021.i09
$\mathrm{Su}, \mathrm{S} .(2007 \mathrm{~b})$. Numerical maximum log likelihood estimation for generalized lambda distributions. Computational Statistics and Data Analysis, 51(8), 3983-3998. doi: 10.1016/j.csda.2006.06.008
$\mathrm{Su}, \mathrm{S}$. (2010). Chapter 15: Fitting GLD to data using the GLDEX 1.0.4 in R. In Z. Karian \& E. Dudewicz (Eds.), Handbook of distribution fitting methods with $R$ (585-608). Boca Raton: CRC Press/Taylor \& Francis.
$\mathrm{Su}, \mathrm{S}$. (2015). Flexible parametric quantile regression model. Statistics and Computing, 25(3), 635-650. doi: 10.1007/s11222-014-9457-1

## STEVE SU

## Footnotes

1. The theoretical mean for RS G $\lambda \mathrm{D}$ is

$$
\lambda_{1}+\frac{\frac{1}{\lambda_{3}+1}-\frac{1}{\lambda_{4}+1}}{\lambda_{2}}
$$

and the theoretical mean for FKML G $\lambda \mathrm{D}$ is

$$
\lambda_{1}-\frac{\frac{1}{\lambda_{3}+1}-\frac{1}{\lambda_{4}+4}}{\lambda_{2}}
$$

2. CRAN can be accessed at http://cran.r-project.org/
3. The UCLA crime dataset can be accessed at http://www.ats.ucla.edu/stat/data/crime.dta

## Emerging Scholars

# A Generalization of the Weibull Distribution with Applications 

Maalee Almheidat<br>University of Petra<br>Amman, Jordan

Carl Lee<br>Central Michigan University<br>Mount Pleasant, MI

Felix Famoye<br>Central Michigan University<br>Mount Pleasant, MI


#### Abstract

The Lomax-Weibull distribution, a generalization of the Weibull distribution, is characterized by four parameters that describe the shape and scale properties. The distribution is found to be unimodal or bimodal and it can be skewed to the right or left. Results for the non-central moments, limiting behavior, mean deviations, quantile function, and the mode(s) are obtained. The relationships between the parameters and the mean, variance, skewness, and kurtosis are provided. The method of maximum likelihood is proposed for estimating the distribution parameters. The applicability of this distribution to modeling real life data is illustrated by three examples and the results of comparisons to other distributions in modeling the data are also presented.


Keywords: Estimation, moments, quantile function, Shannon's entropy, TWeibull $\{Y\}$ family

## Introduction

The Weibull distribution is a popular distribution for modeling phenomena with monotonic failure rates (Weibull, 1939; 1951). It is used to model lifetime data. However, it cannot capture the behavior of lifetime data sets that exhibit bathtub or upside-down bathtub (unimodal) failure rate, often encountered in reliability and engineering studies. A number of new distributions were developed as generalizations or modifications of the Weibull distribution. Xie and Lai (1995) introduced the additive Weibull model, which was obtained by adding two Weibull survival functions. Mudholkar and Srivastava (1993) proposed the exponentiated Weibull distribution. Xie, Tang, and Goh (2002) studied the modified Weibull extension. Bebbington, Lai, and Zitikis (2007) proposed a flexible Weibull

[^41]
## ALMHEIDAT ET AL.

distribution and discussed its properties. For a review of some generalized Weibull distributions, one may refer to Lai (2014).

Different methods to generate probability distributions continue to appear. Eugene, Lee, and Famoye (2002) introduced the beta-generated family and some properties of the family were studied by Jones (2004). Many beta-generated distributions were studied (e.g., Eugene et al., 2002; Nadarajah \& Kotz, 2004; Famoye, Lee, \& Eugene, 2004; Famoye, Lee, \& Olumolade, 2005; Nadarajah \& Kotz, 2006; Akinsete, Famoye, \& Lee, 2008; Barreto-Souza, Santos, \& Cordeiro, 2010; Mahmoudi, 2011; Alshawarbeh, Lee, \& Famoye, 2012). For a review of betagenerated distributions and other generalizations, see Lee, Famoye, and Alzaatreh (2013).

Alzaatreh, Lee, and Famoye (2013) extended the idea of beta-generated distributions to using any continuous random variable $T$ with probability density function (PDF) $\mathrm{r}(t)$ as a generator and developed a new class of distributions called the ' $T$ - $X$ family'. Given a random variable $X$ with cumulative distribution function (CDF) $\mathrm{F}(x)$, the CDF of the $T-X$ family of distributions is defined by Alzaatreh, Lee, and Famoye (2013) as

$$
\begin{equation*}
\mathrm{G}(x)=\int_{a}^{\mathrm{W}(\mathrm{~F}(x))} \mathrm{r}(t) d t \tag{1}
\end{equation*}
$$

where $\mathrm{W}(\mathrm{F}(x))$ is a monotonic and absolutely continuous function of the $\operatorname{CDF} \mathrm{F}(x)$. Alzaatreh, Lee, and Famoye (2013) studied in details the case when $\mathrm{W}(\mathrm{F}(x))=-\log (1-\mathrm{F}(x))$. Some members of the family have been investigated, including gamma-Pareto distribution (Alzaatreh, Famoye, \& Lee, 2012), WeibullPareto distribution (Alzaatreh, Famoye, \& Lee, 2013), and gamma-normal distribution (Alzaatreh, Famoye, \& Lee, 2014a).

Aljarrah, Lee, and Famoye (2014) used the quantile function $\mathrm{Q}_{Y}$ of a random variable $Y$ to define the transformation $\mathrm{W}($.$) in the T$ - $X$ family in (1) and called it the $T-R\{Y\}$ family. Following the notation proposed by Alzaatreh, Famoye, and Lee (2014b), the CDF of the $T-R\{Y\}$ family, as defined by Aljarrah et al. (2014), is given by

$$
\begin{equation*}
\mathrm{F}_{X}(x)=\int_{a}^{\mathrm{Q}_{Y}\left(\mathrm{~F}_{R}(x)\right)} \mathrm{f}_{T}(t) d t=\mathrm{F}_{T}\left\{\mathrm{Q}_{Y}\left(\mathrm{~F}_{R}(x)\right)\right\} \tag{2}
\end{equation*}
$$

where $\mathrm{F}_{T}(x), \mathrm{F}_{R}(\mathrm{x})$, and $\mathrm{F}_{Y}(x)$ are, respectively, the CDFs of the random variables $T, R$, and $Y$. The PDF corresponding to (2) is

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$$
\begin{equation*}
\mathrm{f}_{X}(x)=\frac{\mathrm{f}_{R}(x)}{\mathrm{f}_{Y}\left\{\mathrm{Q}_{Y}\left(\mathrm{~F}_{R}(x)\right)\right\}} \mathrm{f}_{T}\left\{\mathrm{Q}_{Y}\left(\mathrm{~F}_{R}(x)\right)\right\} \tag{3}
\end{equation*}
$$

Almheidat, Famoye, and Lee (2015) used the $T-R\{Y\}$ framework to define and study different approaches to the generalization of the Weibull distribution, the $T$-Weibull $\{Y\}$ family. The authors defined the $T$-Weibull $\{Y\}$ family by taking $R$ in (2) to be a Weibull random variable with $\operatorname{CDF~}_{R}(x)=1-\mathrm{e}^{-(x / \lambda)^{k}}$ and using the quantile function of the random variable $Y$, where $Y$ has uniform, exponential, loglogistic, Fréchet, logistic, or extreme value distribution. When $Y$ follows loglogistic distribution with parameters $\theta$ and $\beta$, the CDF and PDF of the $T$-Weibull\{log-logistic ( $T$-Weibull\{LL\}) family are, respectively, given by

$$
\begin{gather*}
\mathrm{F}_{X}(x)=\mathrm{F}_{T}\left\{\theta\left(\frac{\mathrm{~F}_{R}(x)}{1-\mathrm{F}_{R}(x)}\right)^{1 / \beta}\right\}  \tag{4}\\
\mathrm{f}_{X}(x)=\frac{\theta \mathrm{f}_{R}(x)}{\beta \mathrm{F}_{R}^{(\beta-1) / \beta}(x)\left(1-\mathrm{F}_{R}(x)\right)^{(\beta+1) / \beta}} \mathrm{f}_{T}\left\{\theta\left(\frac{\mathrm{~F}_{R}(x)}{1-\mathrm{F}_{R}(x)}\right)^{1 / \beta}\right\} \tag{5}
\end{gather*}
$$

Setting $\beta=1=\theta$ and taking $T$ in (4) to be a Lomax random variable with CDF $\mathrm{F}_{T}(x)=1-(1+(x / \theta))^{-\alpha}$, Almheidat et al. (2015) defined the Lomax-Weibull\{LL\} distribution (LWD) as an example of $T$-Weibull\{LL\} family.

The purpose of this study is to investigate the LWD as a generalization of the Weibull distribution and a member of $T$-Weibull $\{Y\}$ family.

## Definition and Some Properties of the LWD

The CDF of the LWD defined in Almheidat et al. (2015) is given by

$$
\begin{equation*}
\mathrm{F}_{X}(x)=1-\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{-\alpha} \tag{6}
\end{equation*}
$$

and the PDF corresponding to (6) is

## ALMHEIDAT ET AL.

$$
\begin{equation*}
\mathrm{f}_{X}(x)=\frac{k \alpha}{\theta \lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{(x / \lambda)^{k}}\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{-(\alpha+1)}, \quad x>0, \alpha, \theta, k, \lambda>0 \tag{7}
\end{equation*}
$$

Special cases of the LWD are as follows:

- when $\theta=\alpha=1$, the LWD reduces to the Weibull distribution with parameters $k$ and $\lambda$.
- when $\theta=k=1$, the LWD reduces to the exponential distribution with mean $\lambda / \alpha$.
- when $\alpha=1 / 2, \theta=1$, and $k=2$, the LWD reduces to the Rayleigh distribution with parameter $\lambda$.


## Lemma 1: (Transformations)

1. If a random variable $T$ follows a Lomax distribution with parameters $\alpha$ and $\theta$, then the random variable $X=\lambda\{\ln (T+1)\}^{1 / k}$ follows the LWD.
2. If a random variable $T$ follows an exponential distribution with mean $1 / \alpha$, then the random variable $X=\lambda\left\{\ln \left(\theta \mathrm{e}^{T}-\theta+1\right)\right\}^{1 / k}$ follows the LWD.
3. If a random variable $T$ follows a standard uniform distribution, then the random variable $X=\lambda\left\{\ln \left[\theta(1-T)^{-1 / \alpha}-\theta+1\right)\right\}^{1 / k}$ follows the LWD.

Proof: Using the transformation technique, it is easy to show that the random variable $X$ follows the LWD as given in (7).

## Hazard Function

The hazard function associated with the LWD in (7) is

$$
\begin{equation*}
\mathrm{h}_{X}(x)=\frac{\mathrm{f}_{X}(x)}{1-\mathrm{F}_{X}(x)}=\frac{k \alpha}{\theta \lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{(x / \lambda)^{k}}\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{-1} \tag{8}
\end{equation*}
$$

The following Lemma addresses the limiting behaviors of the hazard function in (8).

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

Lemma 2: $\quad$ The limits of the LWD hazard function as $x \rightarrow 0$ and as $x \rightarrow \infty$ are, respectively, given by

$$
\lim _{x \rightarrow 0} \mathrm{~h}_{X}(x)=\left\{\begin{array}{ll}
0, & k>1  \tag{9}\\
\frac{\alpha}{\theta \lambda}, & k=1, \\
\infty, & k<1
\end{array} \quad \lim _{x \rightarrow \infty} \mathrm{~h}_{X}(x)= \begin{cases}\infty, & k>1 \\
\frac{\alpha}{\lambda}, & k=1 \\
0, & k<1\end{cases}\right.
$$

Proof: This result is obtained by taking the limit of the hazard function in (8).
The following theorem is on the limiting behaviors of the PDF in (7).

Theorem 1: $\quad$ The limit of the LWD as $x \rightarrow \infty$ is 0 and the limit as $x \rightarrow 0$ is given by

$$
\lim _{x \rightarrow 0} \mathrm{f}_{X}(x)= \begin{cases}0, & k>1  \tag{10}\\ \frac{\alpha}{\theta \lambda}, & k=1 \\ \infty, & k<1\end{cases}
$$



Figure 1. The PDFs of LWD for various values of $\alpha, \theta, k$, and $\lambda$

## ALMHEIDAT ET AL.

Proof: The $\lim _{x \rightarrow \infty} \mathrm{f}_{X}(x)=0$. If $k \leq 1$, the result follows from Lemma 2 and the fact that $\mathrm{f}_{X}(x)=\mathrm{h} X(x)\left(1-\mathrm{F}_{X}(x)\right)$. If $k>1$, using L'Hôpital's rule, we have

$$
\begin{aligned}
\lim _{x \rightarrow \infty} \mathrm{f}_{X}(x) & =\lim _{x \rightarrow \infty} \frac{k \alpha(x / \lambda)^{k-1} \mathrm{e}^{(x / \lambda)^{k}}}{\theta \lambda\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{\alpha+1}} \\
& =\frac{\alpha}{\lambda(\alpha+1)} \lim _{x \rightarrow \infty}\left\{\frac{k(x / \lambda)^{k-1}}{\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{\alpha}}+\frac{k-1}{\left.(x / \lambda)\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{\alpha}\right]}\right\} \\
& =\frac{\alpha}{\lambda(\alpha+1)} \lim _{x \rightarrow \infty}\left\{\frac{k(x / \lambda)^{k-1}}{\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{\alpha}}+0\right\} \\
& =\frac{\alpha}{\lambda(\alpha+1)} \lim _{x \rightarrow \infty}\left\{\frac{\theta(k-1)}{\left.\alpha(x / \lambda) \mathrm{e}^{(x / \lambda)^{k}}\left[1+\left(\mathrm{e}^{(x / \lambda)^{k}}-1\right) / \theta\right]^{\alpha-1}\right\}=0}\right\}
\end{aligned}
$$

This completes the proof of the limit as $x \rightarrow \infty$. The result in (10) follows directly by taking the limit of the LWD.

In Figures 1 and 2, various graphs of $\mathrm{f}_{X}(x)$ are provided for different values of the parameters. The graphs in Figure 1 indicate that the LWD is unimodal with different shapes such as left-skewed, right-skewed with long right tail, or monotonically decreasing (reversed J- shape). The graphs in Figure 2 show that the LWD can be bimodal with two positive modal points (when $k>1$ ) or one positive mode and the other mode at zero (when $k<1$ ). The parameters $\alpha$ and $k$ are shape parameters which characterize the skewness, kurtosis, and bimodality of the distribution. However, the parameter $\lambda$ is a scale parameter and the parameter $\theta$ is a shape and scale parameter.

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION



Figure 2. The PDFs of LWD for various values of $\alpha, \theta$, and $k$ when $\lambda=1$


Figure 3. Hazard function of LWD for various values of $\alpha, \theta, k$, and $\lambda$

Displayed in Figure 3 are different graphs of the hazard function related to the LWD for various values of $\alpha, \theta, k$ and $\lambda$. When $k=1$, the LWD failure rate is either constant (when $\theta=1$ ) or first increases (when $\theta>1$ ) or decreases (when $\theta<1$ ) and then becomes a constant. When $k<1$, the failure rate of the LWD is either monotonically decreasing or decreasing followed by unimodal (reflected N shape). When $k>1$, the failure rate of the LWD is either increasing or unimodal followed by increasing ( N -shape). These different failure rate shapes provide more flexibility to the LWD over the Weibull distribution, which has only increasing, decreasing, or constant failure rate.

## ALMHEIDAT ET AL.

## Quantile Function

The quantile function is commonly used in general statistics (Steinbrecher \& Shaw, 2008). Many distributions do not have a closed form quantile function. For the LWD, the quantile function has a closed form as given in the following lemma.

Lemma 3: The quantile function of the LWD is given by

$$
\begin{equation*}
\mathrm{Q}_{X}(p)=\lambda\left\{\ln \left[\theta(1-p)^{-1 / \alpha}-\theta+1\right]\right\}^{1 / k}, \quad 0<p<1 \tag{11}
\end{equation*}
$$

Proof: The result follows directly by using part (iii) of Lemma 2 in Almheidat et al. (2015) when the random variable $T$ follows a Lomax distribution.

Using the formula in (11), the quantile function of the LWD is

- $\quad$ an increasing function of $\lambda$ when $\alpha, \theta$, and $k$ are held fixed.
- a decreasing function of $\alpha$ when $\theta, \lambda$, and $k$ are held fixed.
- $\quad$ an increasing function of $\theta$ when $\alpha, k$, and $\lambda$ are held fixed.
- a decreasing (increasing, or constant) function of $k$, if $\theta<B(\theta>B$, or $\theta=B)$, when $\alpha, \theta$, and $\lambda$ are held fixed, where $B=(\mathrm{e}-1) /\left[(1-p)^{-}\right.$ $\left.{ }^{(1 / \alpha)}-1\right]$.

The closed form quantile function in (11) makes simulating the LWD random variates straightforward. If $U$ is a uniform random variate on the unit interval $(0,1)$, then the random variable $X=\mathrm{Q} x(U)$ follows the LWD. Note that the median $(M)$ can be calculated by setting $p=0.5$ in the quantile function in (11). The median of the LWD is given by $M=\mathrm{Q}(0.5)=\lambda\left\{\ln \left[\theta(0.5)^{-1 / \alpha}-\theta+1\right]\right\}^{1 / k}$.

## Mode(s)

From Almheidat et al. (2015), the mode(s) of $T$-Weibull\{LL\} family satisfy the implicit equation

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$$
x= \begin{cases}\lambda\left[\frac { k } { ( k - 1 ) } \left(\frac{-\mathrm{f}_{T}^{\prime}\left(\mathrm{F}_{R}(x) / \overline{\mathrm{F}}_{R}(x)\right)}{\left.\left.\overline{\overline{\mathrm{F}}_{R}(x) \mathrm{f}_{T}\left(\mathrm{~F}_{R}(x) / \overline{\mathrm{F}}_{R}(x)\right)}-1\right)\right]^{-1 / k},} \quad\right.\right. & k \neq 1  \tag{12}\\ \lambda \log \left[\frac{1}{\overline{\mathrm{~F}}_{R}^{2}(x)}\left(\frac{\mathrm{f}_{T}^{\prime}\left(\mathrm{F}_{R}(x) / \overline{\mathrm{F}}_{R}(x)\right)}{\mathrm{f}_{T}\left(\mathrm{~F}_{R}(x) / \overline{\mathrm{F}}_{R}(x)\right)}+2 \overline{\mathrm{~F}}_{R}(x)\right)\right], & k=1\end{cases}
$$

where $\mathrm{F}_{R}(x)$ and $\overline{\mathrm{F}}_{R}(x)$ are, respectively, the CDF and the survival function of the Weibull distribution. When $T$ is a Lomax random variable, (12) can be simplified to

$$
x=\left\{\begin{array}{cc}
\lambda\left[\frac{k\left(\alpha+1-\theta \overline{\mathrm{F}}_{R}(x)-\mathrm{F}_{R}(x)\right)}{(k-1)\left(\theta \overline{\mathrm{F}}_{R}(x)+\mathrm{F}_{R}(x)\right)}\right]^{-1 / k}, & k \neq 1  \tag{13}\\
\lambda \log \left[\frac{-(\alpha+1)+2 \theta \overline{\mathrm{~F}}_{R}(x)+2 \mathrm{~F}_{R}(x)}{\overline{\mathrm{F}}_{R}(x)\left(\theta \overline{\mathrm{F}}_{R}(x)+\mathrm{F}_{R}(x)\right)}\right], & k=1
\end{array}\right.
$$

Thus, the mode(s) of the LWD satisfy (13). Consider the variational behavior with respect to changes in the parameter values. When $k \neq 1$, (13) can be simplified to

$$
\begin{equation*}
x=\lambda\left[\frac{(k-1)\left(1+(\theta-1) \mathrm{e}^{-(x / \lambda)^{k}}\right)}{k\left(\alpha-(\theta-1) \mathrm{e}^{-(x / \lambda)^{k}}\right)}\right]^{1 / k} \tag{14}
\end{equation*}
$$

Rewriting (14),

$$
\begin{equation*}
(x / \lambda)^{k}=\frac{(k-1)\left(1+(\theta-1) \mathrm{e}^{-(x / \lambda)^{k}}\right)}{k\left(\alpha-(\theta-1) \mathrm{e}^{-(x / \lambda)^{k}}\right)} \tag{15}
\end{equation*}
$$

Setting $u=(x / \lambda)^{k}$ in (15),

## ALMHEIDAT ET AL.

$$
\begin{equation*}
u=\frac{(k-1)\left(1+(\theta-1) \mathrm{e}^{-u}\right)}{k\left(\alpha-(\theta-1) \mathrm{e}^{-u}\right)} \tag{16}
\end{equation*}
$$

Both $x$ and $u$ have the same variational behaviors with respect to changes in the parameters $\alpha$ and $\theta$. The first derivatives of $u$ with respect to $\alpha$ and $\theta$ are, respectively, given by

$$
\begin{equation*}
\frac{\partial u}{\partial \alpha}=\frac{-(k-1)\left(1+(\theta-1) \mathrm{e}^{-u}\right)}{k\left(\alpha-(\theta-1) \mathrm{e}^{-u}\right)^{2}}, \quad \frac{\partial u}{\partial \theta}=\frac{(k-1)(\alpha+1) \mathrm{e}^{-u}}{k\left(\alpha-(\theta-1) \mathrm{e}^{-u}\right)^{2}} \tag{17}
\end{equation*}
$$

From (17), the mode is a decreasing function of $\alpha$ when $k>1$ and an increasing function of $\alpha$ when $k<1$. On the other hand, the mode is an increasing function of $\theta$ when $k>1$ and a decreasing function of $\theta$ when $k<1$. When $k=1$, (13) can be simplified as

$$
\begin{equation*}
x=\lambda \log \left[\frac{-\alpha+1+2(\theta-1) \mathrm{e}^{-x / \lambda}}{\mathrm{e}^{-x / \lambda}\left((\theta-1) \mathrm{e}^{-x / \lambda}+1\right)}\right] \tag{18}
\end{equation*}
$$

or, equivalently,

$$
\mathrm{e}^{x / \lambda}=\frac{-\alpha+1+2(\theta-1) \mathrm{e}^{-x / \lambda}}{\mathrm{e}^{-x / \lambda}\left((\theta-1) \mathrm{e}^{-x / \lambda}+1\right)}
$$

On simplifying (18),

$$
\begin{equation*}
x=\lambda \log [(\theta-1) / \alpha] \tag{19}
\end{equation*}
$$

Therefore, when $k=1$, the mode is an increasing function of $\theta$ and a decreasing function of $\alpha$. The mode is an increasing function of the scale parameter $\lambda$. However, it is not easy to determine increasing/decreasing behavior of the mode with respect to changes in parameter $k$.

From Figures 1 and 2, the LWD can be unimodal or bimodal depending on the parameter values. This property gives more flexibility to the LWD over the

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

Weibull distribution, which is only unimodal. The following theorem shows some cases when the LWD is only unimodal.

Theorem 2: $\quad$ The LWD is unimodal whenever (i) $k=1$ or (ii) $k<1$ and $\theta \leq 1$.
i) If $k=1$, then the mode is at the point $x=0$ whenever $\theta-1 \leq \alpha$ and the mode is at the point $x=\lambda \ln [(\theta-1) / \alpha]$ whenever $\theta-1>\alpha$.
ii) If $k<1$ and $\theta \leq 1$, the mode is at the point $x=0$.

Proof: The derivative with respect to $x$ of the PDF in (7) is given by

$$
\begin{equation*}
\mathrm{f}_{X}^{\prime}(x)=\frac{k \alpha}{(\lambda \theta)^{2}}\left(\frac{x}{\lambda}\right)^{k-2} \mathrm{e}^{(x / \lambda)^{k}}\left[1+\frac{\mathrm{e}^{(x / \lambda)^{k}}-1}{\theta}\right]^{-(\alpha+2)} \mathrm{m}(x) \tag{20}
\end{equation*}
$$

where

$$
\begin{equation*}
\mathrm{m}(x)=(k-1)\left[\theta-1+\mathrm{e}^{(x / \lambda)^{k}}\right]+k(x / \lambda)^{k}\left[\theta-1-\alpha \mathrm{e}^{(x / \lambda)^{k}}\right] \tag{21}
\end{equation*}
$$

By using (20) when $k \leq 1$, the critical points of $\mathrm{f}_{x}(x)$ are $x=0$ and $x=x_{0}$ where $\mathrm{m}\left(x_{0}\right)=0$. Hence, if there is a mode of the LWD, then it will be either at $x=0$ or at $x=x_{0}$ where $\mathrm{m}\left(x_{0}\right)=0$. Note that the signal of $\mathrm{f}_{x}^{\prime}(x)$ is the same as that of $\mathrm{m}(x)$.

If $k=1$, then $\mathrm{m}(x)=(\theta-1)-\alpha \mathrm{e}^{(x / \lambda)}$. Equating $\mathrm{m}(x)$ to zero and solving for $x$ we get $x=\lambda \log [(\theta-1) / \alpha]$, the same result we obtained in (19). If $\theta-1>\alpha$, then the modal point is at $x=\lambda \log [(\theta-1) / \alpha]$, otherwise the mode is at $x=0$. If $k<1$, it is easy to see that $\mathrm{m}(x)<0$ whenever $\theta \leq 1$, therefore $\mathrm{f}_{X}^{\prime}(x)<0$, so $\mathrm{f}_{x}(x)$ is strictly decreasing. From Theorem $1, \lim _{x \rightarrow 0} \mathrm{f}_{X}(x)=\infty$ and $\lim _{x \rightarrow \infty} \mathrm{f}_{X}(x)=0$. Thus $\mathrm{f}_{X}(x)$ has a unique mode at $x=0$.

Graphical displays of the LWD for many combinations of the parameters when $k<1$ and $\theta>1$, and when $k>1$ indicate that the LWD is unimodal or bimodal depending on the parameter values. However, no analytical method has been used to show when the distribution is unimodal or bimodal.

Numerical methods are applied to study the regions of unimodality and bimodality. To study the modes of the LWD, the number of turning points of $\mathrm{f}_{X}(x)$

## ALMHEIDAT ET AL.

in (7) is examined, which is equivalent to examining the sign of $\mathrm{f}_{X}^{\prime}(x)$. This is equivalent to studying the sign of the equation $\mathrm{m}(x)$ in (21).

Consider the situation when $k<1$. Select a fixed value of $k<1(k=0.5,0.7$, 0.9 ) and allow the values of $\alpha$ and $\theta$ to change from 0.001 to 15 at an increment of 0.001 and the values of $x$ to change from $10^{-6}$ to 30 at an increment of 0.001 .

A matrix $\mathbf{M}_{\mathbf{1}}$ is constructed with two entries $\{0,2\}$ which indicates the number of turning points of $\mathrm{f}_{X}(x)$. For each combination of $\alpha$ and $\theta$, if the sign of $\mathrm{m}(x)$ is negative for all values of $x$ between $10^{-6}$ and 30 , then it is indicated by 0 in the matrix $\mathbf{M}_{1}$. If the sign of $\mathrm{m}(x)$ starts as being negative, turns positive, then turns negative, it is indicated by 2 in the matrix $\mathbf{M}_{\mathbf{1}}$. This leads to the following two regions: In the first region (the values corresponding to 0 in the matrix $\mathbf{M}_{1}$ ), $\mathrm{f}_{X}(x)$ contains no turning points. This region indicates that the distribution has only one mode, which is at zero (reversed J-shape). In the second region (corresponding to 2 in the matrix $\left.\mathbf{M}_{1}\right), \mathrm{f}_{X}(x)$ contains two turning points. This region indicates that the distribution has two modes (one of them at zero). By using the boundary between the two regions, we draw a regression line which is a linear function relating $\alpha$ to $\theta$ for each value of $k$ in the set $\{0.5,0.7,0.9\}$. The regression lines all have $R^{2}=100 \%$.

Shown in Figure 4 is the region when LWD is unimodal or bimodal for different values of $k$ and three PDFs for the bimodal case when $k$ is $0.5,0.7$, and 0.9. Values of $k<1, k=0.1$ to 0.9 are also considered at an increment of 0.1 , and the relationship between $\alpha$ and $\theta$ on the boundary points of the bimodality region remains linear.

For the case $k>1$, a matrix $\mathbf{M}_{2}$ is constructed with entries $\{1,3\}$. If the sign of $\mathrm{m}(x)$ starts as being positive then turns negative for $x$ values between $10^{-6}$ and 30 , then it is indicated by 1 in the matrix $\mathbf{M}_{2}$. If the sign of $\mathrm{m}(x)$ starts as being positive, turns negative, then turns positive again and finally becomes negative, it is indicated by 3 in the matrix $\mathbf{M}_{2}$.

This leads to the following regions: In the first region (where the values in the matrix $\mathbf{M}_{\mathbf{2}}$ are 1), $\mathrm{f}_{X}(x)$ contains one turning point. This region indicates that the distribution has only one positive mode. In the second region (where the value in the matrix $\mathbf{M}_{\mathbf{2}}$ are 3 ), $\mathrm{f}_{X}(x)$ contains three turning points. This region indicates that the distribution has two positive modes. By using the boundary between the two regions, we draw two regression lines which are non-linear functions relating $\alpha$ to $\theta$ for each value of $k$ in the set $\{2,4,6\}$. Each regression line has $\mathrm{R}^{2}=100 \%$.

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION



Figure 4. Regions of modality of LWD when $\lambda=1$ and $k=0.5$ (a); $k=0.7$ (b); $k=0.9$ (c); Some PDFs of LWD when $\lambda=1$ and $k=\{0.5,0.7,0.9\}$ (d)

Shown in Figure 5 are the regions when LWD is unimodal or bimodal and three PDFs for the bimodal case when $k$ is 2, 4 and 6 . Note that, from Figures 4 and 5, the bimodal region increases as $k$ increases when $k<1$ and the bimodal region decreases as $k$ increases when $k>1$. Notice when $k$ is large ( $k>20$ ), the region of bimodality does not change with respect to changes in the value of parameter $k$.

## ALMHEIDAT ET AL.



Figure 5. Regions of modality of LWD when $\lambda=1$ and $k=2$ (a); $k=4$ (b); $k=6$ (c); Some PDFs of LWD when $\lambda=1$ and $k=\{2,4,6\}$ (d)

## Moments, Mean Deviations, and Shannon's Entropy

## Moments

The $n^{\text {th }}$ non-central moment $\mathrm{E}\left(X^{n}\right)$ of the LWD can be computed by using an infinite sum as shown in the following theorem:

Theorem 3: The $n^{\text {th }}$ non-central moment of the LWD is given by the expression

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$$
\begin{align*}
& \mathrm{E}\left(X^{n}\right)=\lambda^{n} \alpha \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{\frac{1}{\theta^{i+1}}\left[\sum_{j_{1}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}} w_{i, j_{1}}\right]\right. \\
&\left.+(-1)^{i} \theta^{i+\alpha}\left[\sum_{j_{2}=0}^{\infty} \frac{(\alpha+i+1)_{j_{2}} \Gamma_{i, j_{2}}}{j_{2}!\left(\alpha+i+j_{2}\right)^{n / k}}\right]\right\} \tag{22}
\end{align*}
$$

where $\Gamma_{i, j_{2}}=\Gamma\left(n / k+1,\left(\alpha+i+j_{2}\right) \log (\theta+1)\right),(a)_{r}=a(a+1) \ldots(a+r-1)$ is the ascending factorial, $\Gamma(a, x)$ is the incomplete gamma function given in Abramowitz and Stegun (1972) by

$$
\Gamma(a, x)=\int_{x}^{\infty} t^{a-1} \mathrm{e}^{-t} d t
$$

and

$$
w_{i, j_{1}}=\sum_{m=0}^{\infty} \frac{\left(j_{1}+1\right)^{m}}{m!\left(\frac{n}{k}+m+1\right)}(\log (\theta+1))^{(n / k)+m+1}
$$

Proof: By definition,

$$
\begin{align*}
\mathrm{E}\left(X^{n}\right) & =\int_{0}^{\infty} x^{n} \mathrm{f}_{X}(x) d x \\
& =\left(\frac{\lambda^{n-1} \alpha k}{\theta}\right) \int_{0}^{\infty}(x / \lambda)^{n+k-1} \mathrm{e}^{(x / \lambda)^{k}}\left[1+\frac{\mathrm{e}^{(x / \lambda)^{k}}-1}{\theta}\right]^{-(\alpha+1)} d x \tag{23}
\end{align*}
$$

Using the substitution $u=(x / \lambda)^{k}$, the integral in (23) can be simplified as

## ALMHEIDAT ET AL.

$$
\begin{align*}
\mathrm{E}\left(X^{n}\right)= & \frac{\lambda^{n} \alpha}{\theta} \int_{0}^{\infty} u^{n / k} \mathrm{e}^{u}\left[1+\frac{e^{u}-1}{\theta}\right]^{-(\alpha+1)} d u \\
= & \frac{\lambda^{n} \alpha}{\theta}\left\{\int_{0}^{\log (\theta+1)} u^{n / k} \mathrm{e}^{u}\left[1+\frac{e^{u}-1}{\theta}\right]^{-(\alpha+1)} d u\right.  \tag{24}\\
& \left.\quad+\theta^{\alpha+1} \int_{\log (\theta+1)}^{\infty} u^{n / k} \mathrm{e}^{u}\left(\mathrm{e}^{u}-1\right)^{-(\alpha+1)}\left[1+\frac{\theta}{e^{u}-1}\right]^{-(\alpha+1)} d u\right\} \\
= & \frac{\lambda^{n} \alpha}{\theta}\left\{I_{1}+I_{2}\right\}
\end{align*}
$$

Using the generalized binomial expansion

$$
\left[1+\frac{e^{u}-1}{\theta}\right]^{-(\alpha+1)}=\sum_{i=0}^{\infty} \frac{(-1)^{i}(\alpha+1)_{i}}{i!}\left(\frac{\mathrm{e}^{u}-1}{\theta}\right)^{i}
$$

the integral $I_{1}$ in (24) reduces to

$$
\begin{equation*}
I_{1}=\sum_{i=0}^{\infty} \frac{(-1)^{i}(\alpha+1)_{i}}{i!}\left\{\frac{1}{\theta^{i}} \int_{0}^{\log (\theta+1)} u^{n / k} \mathrm{e}^{u}\left(\mathrm{e}^{u}-1\right)^{i} d u\right\} \tag{25}
\end{equation*}
$$

where $(\alpha+1)_{i}$ is the ascending factorial. Using the binomial expansion

$$
\left(e^{u}-1\right)^{i}=\sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{i-j_{1}} \mathrm{e}^{j_{1} u}
$$

equation (25) can be simplified as

$$
\begin{equation*}
I_{1}=\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{\left(\frac{1}{\theta^{i}}\right) \sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}} \int_{0}^{\log (\theta+1)} u^{n / k} \mathrm{e}^{\left(j_{1}+1\right) u} d u\right\} \tag{26}
\end{equation*}
$$

On using the series representation for the exponential function

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$$
\mathrm{e}^{\left(j_{i}+1\right) u}=\sum_{m=0}^{\infty} \frac{\left(\left(j_{1}+1\right) u\right)^{m}}{m!}
$$

equation (26) becomes

$$
\begin{align*}
I_{1} & =\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{\left(\frac{1}{\theta^{i}}\right)\left[\sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}}\left(\sum_{m=0}^{\infty} \frac{\left(j_{1}+1\right)^{m}}{m!} \int_{0}^{\log (\theta+1)} u^{\frac{n}{k}+m} d u\right)\right]\right\} \\
& =\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{\left(\frac{1}{\theta^{i}}\right)\left[\sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}} \sum_{m=0}^{\infty} \frac{\left(j_{1}+1\right)^{m}}{m!\left(\frac{n}{k}+m+1\right)^{2}} \log (\theta+1)^{\frac{n}{k}+m+1}\right]\right\}  \tag{27}\\
& =\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{\left(\frac{1}{\theta^{i}}\right)\left[\sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}} w_{i, j_{1}}\right]\right\}
\end{align*}
$$

where $w_{i, j_{1}}$ is as defined after equation (22) in Theorem 3.
By using the generalized binomial expansion

$$
\left[1+\frac{\theta}{\mathrm{e}^{u}-1}\right]^{-(\alpha+1)}=\sum_{i=0}^{\infty} \frac{(-1)^{i}(\alpha+1)_{i}}{i!}\left(\frac{\mathrm{e}^{u}-1}{\theta}\right)^{-i}
$$

the integral $I_{2}$ in (24) reduces to

$$
\begin{equation*}
I_{2}=\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{(-1)^{i} \theta^{(i+\alpha+1)} \int_{\log (\theta+1)}^{\infty} u^{n / k} \mathrm{e}^{-(\alpha+i) u}\left(1-\mathrm{e}^{u}\right)^{-(\alpha+i+1)} d u\right\} \tag{28}
\end{equation*}
$$

Using the generalized binomial expansion

$$
\left(1-e^{-u}\right)^{-(\alpha+i+1)}=\sum_{j_{2}=0}^{\infty} \frac{(\alpha+i+1)_{j_{2}}}{j_{2}!} \mathrm{e}^{-j_{2} u}
$$

equation (28) reduces to

## ALMHEIDAT ET AL.

$$
\begin{align*}
I_{2} & =\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{(-1)^{i} \theta^{(i+\alpha+1)}\left[\sum_{j_{2}=0}^{\infty} \frac{(\alpha+i+1)_{j_{2}}}{j_{2}!} \int_{\log (\theta+1)}^{\infty} u^{n / k} \mathrm{e}^{-\left(\alpha+i+j_{2}\right) u} d u\right]\right\} \\
& =\sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{(-1)^{i} \theta^{(i+\alpha+1)}\left[\sum_{j_{2}=0}^{\infty}\left(\frac{(\alpha+i+1)_{j_{2}} \Gamma_{i, j_{2}}}{j_{2}!\left(\alpha+i+j_{2}\right)^{n / k}}\right)\right]\right\} \tag{29}
\end{align*}
$$

where $\Gamma_{i, j_{2}}$ is as defined after equation (22) in Theorem 3. Substituting $I_{1}$ given by (27) and $I_{2}$ given by (29) into (24) completes the proof of the result in (22).

Table 1. Mean and variance of LWD for some values of $\alpha, \theta$, and $k$

|  |  | $k=0.5$ |  | $k=1.0$ |  | $k=7.0$ |  | $k=10.0$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\theta$ | $\alpha$ | Mean | Var | Mean | Var | Mean | Var | Mean | Var |
| 0.5 | 0.5 | 2.9207 | 59.5720 | 0.7854 | 0.8435 | 0.4894 | 0.0087 | 0.4906 | 0.0044 |
|  | 0.7 | 1.3477 | 13.7330 | 0.5290 | 0.3941 | 0.4628 | 0.0077 | 0.4718 | 0.0040 |
|  | 1.0 | 0.5822 | 2.7659 | 0.3466 | 0.1710 | 0.4364 | 0.0066 | 0.4529 | 0.0036 |
|  | 5.0 | 0.0132 | 0.0013 | 0.0549 | 0.0036 | 0.3392 | 0.0035 | 0.3798 | 0.0022 |
|  | 7.0 | 0.0063 | 0.0003 | 0.0382 | 0.0017 | 0.3226 | 0.0031 | 0.3667 | 0.0020 |
| 1.0 | 0.5 | 4.0000 | 80.0000 | 1.0000 | 1.0000 | 0.5164 | 0.0075 | 0.5098 | 0.0037 |
|  | 0.7 | 2.0408 | 20.8240 | 0.7143 | 0.5102 | 0.4922 | 0.0068 | 0.4929 | 0.0035 |
|  | 1.0 | 1.0000 | 5.0000 | 0.5000 | 0.2500 | 0.4677 | 0.0062 | 0.4757 | 0.0032 |
|  | 5.0 | 0.0400 | 0.0080 | 0.1000 | 0.0100 | 0.3716 | 0.0039 | 0.4050 | 0.0023 |
|  | 7.0 | 0.0204 | 0.0021 | 0.0714 | 0.0051 | 0.3542 | 0.0035 | 0.3916 | 0.0022 |
| 5.0 | 0.5 | 7.8417 | 149.0700 | 1.6140 | 1.3158 | 0.5704 | 0.0048 | 0.5474 | 0.0023 |
|  | 0.7 | 4.8301 | 49.3640 | 1.2789 | 0.7795 | 0.5518 | 0.0046 | 0.5348 | 0.0022 |
|  | 1.0 | 2.9624 | 16.5910 | 1.0059 | 0.4694 | 0.5327 | 0.0044 | 0.5218 | 0.0022 |
|  | 5.0 | 0.3447 | 0.2425 | 0.3302 | 0.0633 | 0.4493 | 0.0042 | 0.4628 | 0.0023 |
|  | 7.0 | 0.2123 | 0.1009 | 0.2555 | 0.0409 | 0.4319 | 0.0041 | 0.4501 | 0.0023 |
| 7.0 | 0.5 | 8.9207 | 167.7400 | 1.7588 | 1.3671 | 0.5800 | 0.0043 | 0.5539 | 0.0020 |
|  | 0.7 | 5.6703 | 57.9240 | 1.4168 | 0.8279 | 0.5624 | 0.0041 | 0.5421 | 0.0020 |
|  | 1.0 | 3.6049 | 20.5970 | 1.1351 | 0.5140 | 0.5445 | 0.0040 | 0.5299 | 0.0020 |
|  | 5.0 | 0.4999 | 0.4265 | 0.4070 | 0.0844 | 0.4649 | 0.0041 | 0.4742 | 0.0022 |
|  | 7.0 | 0.3196 | 0.1925 | 0.3206 | 0.0570 | 0.4480 | 0.0041 | 0.4619 | 0.0023 |

Given in Table 1 are the mean and the variance of LWD for various combinations of $\alpha, \theta$, and $k$ when $\lambda=0.5$. Many parameter combinations were used but, to save space, only a few of them are reported in Table 1 . For fixed $\theta$ and $k$, the mean is a decreasing function of $\alpha$. The mean is an increasing function of $\theta$ when

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$\alpha$ and $k$ are fixed. For fixed $\alpha$ and $\theta$, the mean decreases first and then increases as $k$ increases. However, there is no clear pattern for the variance with respect to changes in the parameter values.

The skewness ( Sk ) and kurtosis $(\mathrm{Ku})$ of LWD are given in Table 2 for some values of $\alpha, \theta$, and $k$. For fixed $\alpha$ and $\theta$ the skewness of LWD decreases as $k$ increases. For fixed values of $\alpha$ and $k$, the skewness of LWD decreases as $\theta$ increases. Note that when $\theta=1$, at which the LWD reduces to the Weibull distribution with shape parameter $k$ and scale parameter $\lambda \alpha^{-1 / k}$, the skewness and the kurtosis do not depend on $\alpha$. However, there is no clear pattern for the kurtosis with respect to changes in the parameter values.

Table 2. Skewness and kurtosis of LWD for some values of $\alpha, \theta$, and $k$

| $\theta$ |  | $k=0.5$ |  | $k=1.0$ |  | $k=7.0$ |  | $k=10.0$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\alpha$ | Sk | Ku | Sk | Ku | Sk | Ku | Sk | Ku |
| 0.5 | 0.5 | 7.6717 | 116.0500 | 2.3412 | 10.9980 | -0.0566 | 2.6407 | -0.2648 | 2.8196 |
|  | 0.7 | 8.1298 | 130.0700 | 2.4656 | 11.9600 | -0.0263 | 2.7001 | -0.2391 | 2.8675 |
|  | 1.0 | 8.7698 | 152.0000 | 2.5946 | 13.1370 | -0.0146 | 2.7708 | -0.2321 | 2.9339 |
|  | 5.0 | 10.1060 | 233.5900 | 2.4961 | 13.3190 | -0.1419 | 2.8924 | -0.3596 | 3.1300 |
|  | 7.0 | 9.4602 | 206.0800 | 2.3840 | 12.3250 | -0.1696 | 2.8902 | -0.3853 | 3.1450 |
| 1.0 | 0.5 | 6.6188 | 87.7200 | 2.0000 | 9.0000 | -0.2541 | 2.8803 | -0.4632 | 3.1872 |
|  | 0.7 | 6.6188 | 87.7200 | 2.0000 | 9.0000 | -0.2541 | 2.8803 | -0.4632 | 3.1872 |
|  | 1.0 | 6.6188 | 87.7200 | 2.0000 | 9.0000 | -0.2541 | 2.8803 | -0.4632 | 3.1872 |
|  | 5.0 | 6.6188 | 87.7200 | 2.0000 | 9.0000 | -0.2541 | 2.8803 | -0.4632 | 3.1872 |
|  | 7.0 | 6.6188 | 87.7200 | 2.0000 | 9.0000 | -0.2541 | 2.8803 | -0.4632 | 3.1872 |
| 5.0 | 0.5 | 4.8668 | 49.7400 | 1.4377 | 6.5163 | -0.6390 | 3.9785 | -0.8628 | 4.6405 |
|  | 0.7 | 4.3383 | 40.2360 | 1.2518 | 5.6651 | -0.7372 | 4.0188 | -0.9569 | 4.7131 |
|  | 1.0 | 3.7677 | 30.7260 | 1.0804 | 4.8592 | -0.7984 | 3.9727 | -1.0117 | 4.6712 |
|  | 5.0 | 2.9128 | 16.2010 | 1.0097 | 3.9304 | -0.6615 | 3.2898 | -0.8568 | 3.8290 |
|  | 7.0 | 3.0655 | 17.4180 | 1.0952 | 4.1609 | -0.6009 | 3.1715 | -0.7960 | 3.6733 |
| 7.0 | 0.5 | 4.5966 | 44.8970 | 1.3539 | 6.2417 | -0.6997 | 4.2784 | -0.9275 | 5.0251 |
|  | 0.7 | 4.0149 | 35.1060 | 1.1402 | 5.3288 | -0.8240 | 4.3688 | -1.0481 | 5.1612 |
|  | 1.0 | 3.4034 | 25.7610 | 0.9440 | 4.4858 | -0.9047 | 4.3469 | -1.1222 | 5.1480 |
|  | 5.0 | 2.5039 | 12.4240 | 0.8373 | 3.4387 | -0.7689 | 3.5053 | -0.9640 | 4.1107 |
|  | 7.0 | 2.6408 | 13.3050 | 0.9238 | 3.6191 | -0.6991 | 3.3404 | -0.8931 | 3.8997 |

## ALMHEIDAT ET AL.



Figure 6. Galton's skewness and Moors' kurtosis for LWD when $\alpha=0.5$

A measure of skewness and kurtosis, based on the quantile function, is obtained by using Galton's skewness (Galton, 1883) and Moors' kurtosis (Moors, 1988). By using the quantile function defined in (11), Galton's skewness and Moors' kurtosis for LWD, respectively, are given by

$$
\begin{equation*}
S=\frac{\mathrm{Q}\left(\frac{6}{8}\right)-2 \mathrm{Q}\left(\frac{4}{8}\right)+\mathrm{Q}\left(\frac{2}{8}\right)}{\mathrm{Q}\left(\frac{6}{8}\right)-\mathrm{Q}\left(\frac{2}{8}\right)}, \quad K=\frac{\mathrm{Q}\left(\frac{7}{8}\right)-\mathrm{Q}\left(\frac{5}{8}\right)+\mathrm{Q}\left(\frac{3}{8}\right)-\mathrm{Q}\left(\frac{1}{8}\right)}{\mathrm{Q}\left(\frac{6}{8}\right)-\mathrm{Q}\left(\frac{2}{8}\right)} \tag{30}
\end{equation*}
$$

Presented in Figure 6 are three dimensional graphs of Galton's skewness and Moors' kurtosis for the same parameter values as in Table 2. To save space, these values are not reported but are compared with the values in Table 2. The results show similar patterns to those in Table 2.

## Mean Deviations

Let $X$ be a random variable with mean $\mu$ and median $M$. The mean deviation from the mean is defined as

$$
\begin{align*}
D_{\mu} & =\mathrm{E}(|X-\mu|)=\int_{-\infty}^{\infty}|x-\mu| \mathrm{f}_{X}(x) d x \\
& =\int_{-\infty}^{\mu}(\mu-x) \mathrm{f}_{X}(x) d x+\int_{\mu}^{\infty}(x-\mu) \mathrm{f}_{X}(x) d x  \tag{31}\\
& =2 \mu \mathrm{~F}_{X}(\mu)-2 \int_{-\infty}^{\mu} x \mathrm{f}_{X}(x) d x
\end{align*}
$$

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

where $\mathrm{F}_{X}(\mu)=\int_{-\infty}^{\mu} \mathrm{f}_{X}(x) d x$ can be calculated using (6). Similarly, the mean deviation from the median can be defined as

$$
\begin{align*}
D_{M} & =\mathrm{E}(|X-M|)=\int_{-\infty}^{\infty}|x-M| \mathrm{f}_{X}(x) d x \\
& =\int_{-\infty}^{M}(M-x) \mathrm{f}_{X}(x) d x+\int_{M}^{\infty}(x-M) \mathrm{f}_{X}(x) d x  \tag{32}\\
& =\mu-2 \int_{-\infty}^{M} x \mathrm{f}_{X}(x) d x
\end{align*}
$$

The integrals $\int_{0}^{\mu} x \mathrm{f}_{X}(x) d x$ and $\int_{0}^{M} x \mathrm{f}_{X}(x) d x$ from (31) and (32), respectively, can be obtained as follows: Let $u=(x / \lambda)^{k}$. Then

$$
\begin{equation*}
\int_{0}^{\mu} x \mathrm{f}_{X}(x) d x=\frac{\alpha \lambda}{\theta} \int_{0}^{(\mu / \lambda)^{k}} u^{1 / k} \mathrm{e}^{u}\left[1+\frac{\mathrm{e}^{u}-1}{\theta}\right]^{-(\alpha+1)} d u \tag{33}
\end{equation*}
$$

If $\frac{\mathrm{e}^{(\mu / \lambda)^{k}}-1}{\theta} \leq 1$, using a similar approach as in Theorem 3, (33) reduces to

$$
\int_{0}^{\mu} x \mathrm{f}_{X}(x) d x
$$

$$
\left.=\alpha \lambda \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!\theta^{i+1}}\left\{\begin{array}{r}
\left(\frac{1}{\theta^{i+1}} \sum_{j_{i}=0}^{i}\left[\binom{i}{j_{1}}(-1)^{j_{1}} \sum_{m=0}^{\infty}\left(\frac{\left(j_{1}+1\right)^{m}}{m!}\right)\right\}\right.  \tag{34}\\
\left.\left.\quad \times \int_{0}^{(\mu / \lambda)^{k}} u^{(1 / k)+m} d u\right)\right]
\end{array}\right)\right\}
$$

$$
=\alpha \lambda \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!\theta^{i+1}}\left\{\sum_{j_{i}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}} \sum_{m=0}^{\infty}\left[\frac{\left(j_{1}+1\right)^{m}}{m!\left(\frac{1}{k}+m+1\right)}\left(\frac{\mu}{\lambda}\right)^{k+m k+1}\right]\right\}
$$

If $\frac{\mathrm{e}^{(\mu / \lambda)^{k}}-1}{\theta}>1$, then

## ALMHEIDAT ET AL.

$$
\begin{align*}
\int_{0}^{\mu} x \mathrm{f}_{X}(x) d x= & \frac{\alpha \lambda}{\theta} \int_{0}^{(\mu / \lambda)^{k}} u^{1 / k} \mathrm{e}^{u}\left[1+\frac{\mathrm{e}^{u}-1}{\theta}\right]^{-(\alpha+1)} d u \\
= & \frac{\alpha \lambda}{\theta} \int_{0}^{\log (\theta+1)} u^{1 / k} \mathrm{e}^{u}\left[1+\frac{\mathrm{e}^{u}-1}{\theta}\right]^{-(\alpha+1)} d u  \tag{35}\\
& \quad+\frac{\alpha \lambda}{\theta} \int_{\log (\theta+1)}^{(\mu /)^{k}} u^{1 / k} \mathrm{e}^{u}\left[1+\frac{\mathrm{e}^{u}-1}{\theta}\right]^{-(\alpha+1)} d u \\
= & I_{1}^{*}+I_{2}^{*}
\end{align*}
$$

Again, using the approach in Theorem 3, the integrals $I_{1}^{*}$ and $I_{2}^{*}$ can be simplified as

$$
\begin{aligned}
I_{1}^{*} & =\alpha \lambda \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!\theta^{i+1}}\left\{\sum_{j_{1}=0}^{i}\binom{i}{j_{1}}(-1)^{j_{1}}\left[\sum_{m=0}^{\infty} \frac{\left(j_{1}+1\right)^{m}}{m!\left(\frac{1}{k}+m+1\right)^{2}}\left[\log (\theta+1)^{\frac{1}{k}+m+1}\right]\right\}\right. \\
I_{2}^{*} & =\alpha \lambda \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{(-1)^{i} \theta^{(i+\alpha)}\left[\sum_{j_{2}=0}^{\infty} \frac{(\alpha+i+1)_{j_{2}}}{j_{2}!} \int_{\log (\theta+1)}^{(\mu / \lambda)^{k}} u^{1 / k} \mathrm{e}^{-\left(\alpha+i+j_{2}\right) u} d u\right]\right\} \\
& =\alpha \lambda \sum_{i=0}^{\infty} \frac{(\alpha+1)_{i}}{i!}\left\{(-1)^{i} \theta^{(i+\alpha)}\left[\sum_{j_{2}=0}^{\infty} \frac{(\alpha+i+1)_{j_{2}}}{j_{2}!\left(\alpha+i+j_{2}\right)} \Gamma_{i, j_{2}}^{*}\right]\right\}
\end{aligned}
$$

where

$$
\Gamma_{i, j_{2}}^{*}=\Gamma\left(\frac{1}{k}+1,\left(\alpha+i+j_{2}\right) \log (\theta+1)\right)-\Gamma\left(\frac{1}{k}+1,\left(\alpha+i+j_{2}\right)\left(\frac{\mu}{k}\right)^{k}\right)
$$

The integral $\int_{0}^{M} x \mathrm{f}_{X}(x) d x$ can be obtained in a similar fashion.

## Shannon's Entropy

The entropy of a random variable $X$ is a measure of variation of uncertainty. Shannon (1948) defined the entropy of a random variable $X$ with PDF $g(x)$ to be

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$\eta_{X}=\mathrm{E}[-\ln \mathrm{g}(X)]$. Entropy has various applications in many fields including science, engineering, and economics. Using Theorem 2 in Almheidat et al. (2015), the Shannon's entropy of LWD is given by

$$
\begin{equation*}
\eta_{X}=\log \left(\frac{\lambda}{k}\right)-\frac{k-1}{k} \mathrm{E}\{\log [\log (T+1)]\}-\left(\frac{1}{\lambda^{k}}\right) \mu_{k}^{\prime}+\eta_{T} \tag{36}
\end{equation*}
$$

where $\mu_{k}^{\prime}$ is the $k^{\text {th }}$ non-central moment of the LWD and $\eta_{T}=\ln (\theta / \alpha)+(1 / \alpha)+1$ is the Shannon's entropy of the Lomax random variable. Thus, from (36), the Shannon's entropy of LWD can be simplified as

$$
\begin{equation*}
\eta_{X}=\ln \left(\frac{\theta \lambda}{\alpha k}\right)+\frac{1}{\alpha}+1-\left(\frac{1}{\lambda^{k}}\right) \mu_{k}^{\prime}-\frac{k-1}{k} \mathrm{I}(\alpha, \theta) \tag{37}
\end{equation*}
$$

where

$$
\mathrm{I}(\alpha, \theta)=\frac{\alpha}{\theta} \int_{0}^{\infty} \ln [\ln (1+t)]\left(1+\frac{t}{\theta}\right)^{-(\alpha+1)} d t
$$

## Parameter Estimation

Let $X_{1}, X_{2}, \ldots, X_{n}$ be a random sample from LWD with parameters $\alpha, \theta, k$, and $\lambda$. The log-likelihood function $\ell=\ell(\alpha, \theta, k, \lambda)$ for the PDF in (7) is given by

$$
\begin{align*}
& \ell=n(\log k+\log \alpha-\log \theta-\log \lambda) \\
&+\sum_{j=1}^{n}\left\{(k-1) \log \left(\frac{x_{j}}{\lambda}\right)+\left(\frac{x_{j}}{\lambda}\right)^{k}-(\alpha+1) \log \left[1+\frac{e^{\left(x_{j} / \lambda\right)^{k}}-1}{\theta}\right]\right\} \tag{38}
\end{align*}
$$

On taking the first partial derivatives of the log-likelihood function in (38) with respect to the parameters $\alpha, \theta, k$, and $\lambda$,

$$
\begin{equation*}
\frac{\partial \ell}{\partial \alpha}=\frac{n}{\alpha}-\sum_{j=1}^{n} \log \left[1+\frac{\mathrm{e}^{\left(e_{j} / \lambda\right)^{k}}-1}{\theta}\right] \tag{39}
\end{equation*}
$$

## ALMHEIDAT ET AL.

$$
\begin{gather*}
\frac{\partial \ell}{\partial \theta}=\frac{-n}{\theta}+\frac{(\alpha+1)}{\theta} \sum_{j=1}^{n}\left[1+\theta\left(\mathrm{e}^{\left(x_{j} / \lambda\right)^{k}}-1\right)^{-1}\right]^{-1}  \tag{40}\\
\frac{\partial \ell}{\partial k}=\frac{n}{k}+\sum_{j=1}^{n}\left\{1+\left[1-(\alpha+1) \mathrm{e}^{\left(x_{j} / \lambda\right)^{k}}\left(\theta+\mathrm{e}^{\left(x_{j} / \lambda\right)^{k}}-1\right)^{-1}\right]\left(\frac{x_{j}}{\lambda}\right)^{k}\right\} \log \left(\frac{x_{j}}{\lambda}\right)  \tag{41}\\
\frac{\partial \ell}{\partial \lambda}=\frac{-n k}{\lambda}+\frac{k}{\lambda} \sum_{j=1}^{n}\left\{(\alpha+1) \mathrm{e}^{\left(x_{j} / \lambda\right)}\left(\theta+\mathrm{e}^{\left(x_{j} / \lambda\right)^{k}}-1\right)^{-1}-1\right\}\left(\frac{x_{j}}{\lambda}\right)^{k} \tag{42}
\end{gather*}
$$

By setting (39) to (42) equal to zero and solving them simultaneously, obtain $\hat{\alpha}, \hat{\theta}, \hat{k}$, and $\hat{\lambda}$, the maximum likelihood estimates (MLEs) for the parameters $\alpha, \theta$, $k$, and $\lambda$, are respectively obtained. The computations are done using the NLMIXED procedure in SAS. In this procedure the initial estimates of $\alpha, \theta, k$, and $\lambda$ can be obtained as follows: First, assume that the sample data $\left(x_{1}, x_{2}, \ldots, x_{n}\right)$ is from a Weibull distribution. The parameter estimates given in Johnson, Kotz, and Balakrishnan (1994, pp. 635-643) are used for $k$ and $\lambda$ as the initial estimates, which are

$$
k_{0}=\frac{\pi}{s_{w} \sqrt{6}}, \quad \lambda_{0}=\exp \left(\bar{w}+\frac{\gamma}{k_{0}}\right)
$$

where $w_{i}=\log \left(x_{i}\right), \bar{w}$ and $s_{w}$ are respectively the mean and the standard deviation of $w$ random sample, and $\gamma=-\Gamma(1) \approx 0.57722$ is the Euler's constant. By using Lemma 1 , the sample data ( $x_{1}, x_{2}, \ldots, x_{n}$ ) can be transformed to a data set from Lomax distribution by using

$$
y_{i}=\exp \left(\frac{x}{\lambda_{0}}\right)^{k_{0}}-1
$$

The initial estimates for $\alpha$ and $\theta$ are the moment estimates of $\alpha$ and $\theta$ from the Lomax distribution and they are given by

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

$$
\alpha_{0}=\frac{2 v_{y}}{v_{y}-\bar{y}^{2}}, \quad \theta_{0}=\bar{y}\left(\alpha_{0}-1\right)
$$

where $\bar{y}$ and $v_{y}$ are, respectively, the mean and the variance of $\left(y_{1}, y_{2}, \ldots, y_{n}\right)$.

## Applications

Three applications of the LWD using real life data sets are considered. Each of the three data sets exhibits right skewed, left skewed, or bimodal distribution shape. In these applications, the maximum likelihood estimates of the parameters of the fitted distributions are obtained. The LWD is compared with other distributions based on the maximized log-likelihood, the Kolmogorov-Smirnov (K-S) test along with the corresponding $p$-value, and Akaike Information Criterion (AIC). In addition, the histogram of the data and the PDFs of the fitted models are presented for graphical illustration of the goodness of fit.

## Wheaton River Data

The data set in Table 3, from Choulakian and Stephens (2001), is the exceedances of flood peaks (in $m^{3} / s$ ) of Wheaton River, Yukon Territory, Canada. The data consists of 72 exceedances for the years 1958-1984, rounded to one decimal place. It is a right-skewed data (skewness $=1.5$ and kurtosis $=3.19$ ) with a long right tail.

The data set was analyzed using several distributions. Akinsete et al. (2008) used this data set as an application of beta-Pareto distribution (BPD). Alshawarbeh et al. (2012) fitted the data set to beta-Cauchy distribution (BCD). It was also used by Al-Aqtash, Famoye, and Lee (2014) to illustrate the flexibility of GumbelWeibull distribution (GWD) to fit different data sets. We fit the LWD to the data set. The MLEs and the goodness of fit statistics are presented in Table 4. The results for BPD, BCD and GWD are taken from Al-Aqtash et al. (2014).

The goodness of fit statistics indicate that the BCD, GWD, and LWD provide good fit based on the $p$-value of K-S statistic. But the LWD seems to provide the best fit among these distributions in Table 4, since it has the smallest AIC and K-S statistics and the largest log-likelihood value. The LWD seems to be very competitive to other distributions in fitting the data. This suggests that LWD fits highly right-skewed data with a long tail very well. Figure 7 contains the histogram of the data with the fitted distribution and supports the results in Table 4.

## ALMHEIDAT ET AL.

Table 3. Exceedances of the Wheaton River data

| 1.7 | 2.2 | 14.4 | 1.1 | 0.4 | 20.6 | 5.3 | 0.7 | 1.9 | 13.0 | 12.0 | 9.3 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1.4 | 18.7 | 8.5 | 25.5 | 11.6 | 14.1 | 22.1 | 1.1 | 2.5 | 14.4 | 1.7 | 37.6 |
| 0.6 | 2.2 | 39.0 | 0.3 | 15.0 | 11.0 | 7.3 | 22.9 | 1.7 | 0.1 | 1.1 | 0.6 |
| 9.0 | 1.7 | 7.0 | 20.1 | 0.4 | 2.8 | 14.1 | 9.9 | 10.4 | 10.7 | 30.0 | 3.6 |
| 5.6 | 30.8 | 13.3 | 4.2 | 25.5 | 3.4 | 11.9 | 21.5 | 27.6 | 36.4 | 2.7 | 64.0 |
| 1.5 | 2.5 | 27.4 | 1.0 | 27.1 | 20.2 | 16.8 | 5.3 | 9.7 | 27.5 | 2.5 | 27.0 |

Table 4. MLEs for Wheaton River data (standard errors in parentheses)

| Distribution | BPD | BCD | GWD | LWD |
| ---: | ---: | ---: | ---: | ---: |
| Parameter estimates | $\hat{\alpha}=7.6954$ | $\hat{\alpha}=317.0256$ | $\hat{\mu}=-0.6548$ | $\hat{\alpha}=0.1449$ |
|  | $\hat{b}=85.75$ | $(312.5864)$ | $(1.1214)$ | $(0.0472)$ |
|  | $\hat{\theta}=0.1$ | $\hat{b}=1.4584$ | $\hat{\sigma}=3.3672$ | $\hat{\theta}=0.03124$ |
|  | $\hat{k}=0.0208$ | $(0.4899)$ | $(0.7295)$ | $(0.0383)$ |
|  |  | $\hat{\theta}=-0.0482$ | $\hat{\alpha}=1.4848$ | $\hat{k}=1.6396$ |
|  |  | $(1.2301)$ | $(0.3665)$ | $(0.2842)$ |
|  |  | $\hat{\lambda}=0.09617$ | $\hat{\lambda}=8.0323$ | $\hat{\lambda}=6.3766$ |
|  |  | $(0.0688)$ | $(2.8206)$ | $(2.1724)$ |
| Log Likelihood | -272.1280 | -260.4813 | -247.8373 | -247.4916 |
| AIC | 552.256 | 528.952 | 503.700 | 503.000 |
| K-S | 0.1625 | 0.1219 | 0.0662 | 0.0587 |
| $(\boldsymbol{p}$-value) | $(0.0446)$ | $(0.2350)$ | $(0.9101)$ | $(0.9652)$ |



Figure 7. The histogram and the PDFs of the Wheaton River data

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

## Strengths of 1.5 cm Glass Fibers Data

The second application represents fitting the LWD to the strength of 1.5 cm glass data set given in Table 5. The data set is "sample 1" of Smith and Naylor (1987) and deals with the breaking strength of 63 glass fibers of length 1.5 cm , originally obtained by workers at the UK National Physical Laboratory.

Barreto-Souza et al. (2010) applied the beta generalized exponential distribution (BGED) to fit the data and Barreto-Souza, Cordeiro, and Simas (2011) fitted beta Fréchet distribution (BFD) to the data. Recently, Alzaghal, Famoye, and Lee (2013) used the data in an application of the exponentiated Weibull-exponential distribution (EWED).

Table 5. Strength of 1.5 cm glass fibers data

| 0.55 | 0.74 | 0.77 | 0.81 | 0.84 | 0.93 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 1.04 | 1.11 | 1.13 | 1.24 | 1.25 | 1.27 |
| 1.28 | 1.29 | 1.30 | 1.36 | 1.39 | 1.42 |
| 1.48 | 1.48 | 1.49 | 1.49 | 1.50 | 1.50 |
| 1.51 | 1.52 | 1.53 | 1.54 | 1.55 | 1.55 |
| 1.58 | 1.59 | 1.60 | 1.61 | 1.61 | 1.61 |
| 1.61 | 1.62 | 1.62 | 1.63 | 1.64 | 1.66 |
| 1.66 | 1.66 | 1.67 | 1.68 | 1.68 | 1.69 |
| 2.00 | 2.01 | 2.24 | 1.76 | 1.76 | 1.77 |
| 1.70 | 1.70 | 1.73 | 1.84 | 1.84 | 1.89 |
| 1.78 | 1.81 | 1.82 |  |  |  |

Table 6. MLEs for the strength of 1.5 cm glass fibers data (standard errors in parentheses)

| Distribution | BFD | BGE | EWED | LWD |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
| Parameter estimates | $\hat{\alpha}=0.396$ | $\hat{\alpha}=0.4125$ | $\hat{\alpha}=23.614$ | $\hat{\alpha}=1.1907$ |  |
|  | $(0.174)$ | $(0.3020)$ | $(3.954)$ | $(0.7232)$ |  |
|  | $\hat{b}=225.720$ | $\hat{b}=93.4655$ | $\hat{\gamma}=7.249$ | $\hat{\theta}$ | $=21.9641$ |
|  | $(164.476)$ | $(120.0850)$ | $(0.994)$ | $(9.4167)$ |  |
|  | $\hat{\lambda}=1.302$ | $\hat{\alpha}=22.6124$ | $\hat{c}=0.0033$ | $\hat{k}=2.9842$ |  |
|  | $(0.270)$ | $(21.925)$ | $(0.0030)$ | $(1.2329)$ |  |
|  | $\hat{\sigma}=6.863$ | $\hat{\lambda}=0.9227$ |  | $\hat{\lambda}=1.0889$ |  |
|  | $(1.992)$ | $(0.5010)$ |  | $(0.3105)$ |  |
| Log Likelihood | -19.5900 | -15.5995 | -14.3300 | -11.9905 |  |
| AIC | 47.200 | 39.199 | 34.700 | 32.000 |  |
| K-S | 0.2140 | 0.1673 | 0.1370 | 0.1013 |  |
| $(\boldsymbol{p}$-value) | $(0.0060)$ | $(0.0588)$ | $(0.1950)$ | $(0.5373)$ |  |

## ALMHEIDAT ET AL.



Figure 8. The histogram and the PDFs for the glass fibers data

The LWD is fitted to the data and the estimation results and goodness of fit statistics are presented in Table 6. From Table 6, the BGE, EWED, and LWD provide an adequate fit to the data with the LWD providing the best fit among all distributions in Table 6 based on every criterion. The distribution of the data is skewed to the left (skewness $=-0.95$ and kurtosis $=1.10$ ). This suggests that the LWD performs well in modeling left skewed data. Contained in Figure 8 are the histogram of the data and the PDFs of the fitted distributions.

## Australian Athletes Data

In this example, a data set reported by Cook and Weisberg (1994) about Australian Athletes is considered. It contains 13 variables on 102 male and 100 female athletes collected at the Australian Institute of Sport. Jamalizadeh, Arabpour, and Balakrishnan (2011) used the heights for the 100 female athletes and the hemoglobin concentration levels for the 202 athletes to illustrate the application of a generalized skew two-piece skew-normal distribution. Choudhury and Abdul Matin (2011) also used percentage of the hemoglobin blood cell for the male athletes to illustrate the application of an extended skew generalized normal distribution. In this example we consider the percentage of body fat (\%Bfat) variable for the 202 athletes.

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

Table 7. MLEs for the \%Bfat data (standard error in parentheses)

| Distribution | WD | BND | LND | LWD |
| ---: | ---: | ---: | ---: | ---: |
| Parameter estimates | $\hat{k}=2.354$ | $\hat{\alpha}=0.1896$ | $\hat{\lambda}=0.3000$ | $\hat{\alpha}=0.2650$ |
|  | $(0.125)$ | $(0.0549)$ | $(0.0235)$ | $(0.0448)$ |
|  | $\hat{\lambda}=15.313$ | $\hat{\beta}=0.2513$ | $\hat{\mu}=14.632$ | $\hat{\theta}=0.0065$ |
|  | $(0.4852)$ | $(0.0241)$ | $(0.369)$ | $(0.0062)$ |
|  |  | $\hat{\mu}=15.289$ | $\hat{\sigma}=2.5330$ | $\hat{k}=5.626$ |
|  |  | $(1.286)$ | $(0.0682)$ | $(0.635)$ |
|  |  | $\hat{\sigma}=2.495$ |  | $\hat{\lambda}=18.538$ |
|  |  | $(0.165)$ |  | $(1.136)$ |
| Log Likelihood | -642.416 | -649.471 | -644.047 | -623.427 |
| AIC | 1288.8 | 1306.9 | 1294.1 | 1254.9 |
| K-S | 0.1091 | 0.1425 | 0.1599 | 0.0468 |
| $(\boldsymbol{p}$-value) | $(0.0163)$ | $\left(5.4400 \times 10^{-4}\right)$ | $\left(6.4700 \times 10^{-5}\right)$ | $(0.7676)$ |



Figure 9. The histogram and the PDFs for \%Bfat data

The LWD, the beta-normal distribution (BND) defined by Eugene et al. (2002), the logistic-normal\{logistic\} distribution (LND) defined by Alzaatreh et al. (2014b), and the Weibull distribution (WD) are applied to fit the data set. Table 7 contains the estimates, standard errors of the estimates, log-likelihood values, AIC, K-S test statistic, and the corresponding $p$-values.

ALMHEIDAT ET AL.

The histogram and the densities of the fitted distributions are provided in Figure 9. From Figure 9, the distribution of this data appeared to be bimodal and skewed to the right (skewness $=0.759$, kurtosis $=2.827$ ).

From Table 7, LWD has the smallest AIC and K-S statistics and the largest log-likelihood value, which indicates that LWD seems to be superior to the other distributions in fitting the data. Even though the BND has the ability to fit bimodal data, it could not capture the bimodality property in fitting the data. On the other hand, the LND capture the bimodality property but with poor fit to the data. This application suggests that LWD has the ability to adequately fit bimodal data.

## Conclusion

A four-parameter LWD was proposed as an extension of the Weibull distribution and a member of $T$-Weibull $\{Y\}$ family defined by Almheidat et al. (2015). The LWD is found to be unimodal or bimodal and reduces to some existing distributions that are known in the literature. Various properties of the LWD are investigated, including the hazard function, the quantile function, and the regions of unimodality and bimodality. Expressions for the moments, the Shannon's entropy, and the mean deviations are derived. The parameters are estimated by the method of maximum likelihood.

The LWD is fitted to three real data sets to illustrate the application of the distribution. The first data set is the exceedances of flood peaks of Wheaton River, the second is the strength of 1.5 cm glass fibers, and the third is the percentage of the body fat of 202 Australian Athletes. In fitting these data sets, different distributions are compared with the LWD based on goodness of fit statistics. The two most competitive distributions to the LWD are the GWD (used in the flood data set) and the EWED (used in the glass fibers data set). The results show that the LWD outperformed these two distributions in fitting both data sets. LWD has an advantage over several other distributions due to the flexibility of this distribution and its ability to model different shapes in real life data sets, including unimodal and bimodal cases.

## References

Abramowitz, M., \& Stegun, I. (1972). Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover Publications.

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

Akinsete, A., Famoye, F., \& Lee, C. (2008). The beta-Pareto distribution. Statistics, 42(6), 547-563. doi: 10.1080/02331880801983876

Al-Aqtash, R., Famoye, F., \& Lee, C. (2014). On generating a new family of distributions using the logit function. Journal of Probability and Statistical Science, 13(1), 135-152.

Aljarrah, M. A., Lee, C., \& Famoye, F. (2014). On generating T-X family of distributions using quantile functions. Journal of Statistical Distributions and Applications, 1(2), 1-17. doi: 10.1186/2195-5832-1-2

Almheidat, M., Famoye, F., \& Lee, C. (2015). Some generalized families of Weibull distribution: Properties and applications. International Journal of Statistics and Probability, 4(3), 18-35. doi: 10.5539/ijsp.v4n3p18

Alshawarbeh, E., Lee, C., \& Famoye, F. (2012). The beta-Cauchy distribution. Journal of Probability and Statistical Science, 10(1), 41-57.

Alzaatreh, A., Famoye, F., \& Lee, C. (2012). Gamma-Pareto distribution and its applications. Journal of Modern Applied Statistical Methods, 11(1), 78-94. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol11/iss1/7/

Alzaatreh, A., Famoye, F., \& Lee, C. (2013). Weibull-Pareto distribution and its applications. Communications in Statistics - Theory and Methods, 42(9), 1673-1691.

Alzaatreh, A., Lee, C., \& Famoye, F. (2013). A new method for generating families of continuous distributions. METRON, 71(1), 63-79. doi:
10.1007/s40300-013-0007-y

Alzaatreh, A., Famoye, F., \& Lee, C. (2014a). The gamma-normal distribution: Properties and applications. Journal of Computational Statistics \& Data Analysis, 69(1), 67-80. doi: 10.1016/j.csda.2013.07.035

Alzaatreh, A., Famoye, F., \& Lee, C. (2014b). T-normal family of distributions: A new approach to generalize the normal distribution. Journal of Statistical Distributions and Applications, 1(16), 1-18. doi: 10.1186/2195-5832-116

Alzaghal, A., Famoye, F., \& Lee, C. (2013). Exponentiated $T-X$ family of distributions with some applications. International Journal of Statistics and Probability, 2(3), 31-49. doi: 10.5539/ijsp.v2n3p31

Barreto-Souza, W., Cordeiro, G. M., \& Simas, A. B. (2011). Some results for beta Fréchet distribution. Communication in Statistics - Theory and Methods, 40(5), 798-811. doi: 10.1080/03610920903366149

## ALMHEIDAT ET AL.

Barreto-Souza, W., Santos, A. H. S., \& Cordeiro, G. M. (2010). The beta generalized exponential distribution. Journal of Statistical Computation and Simulation, 80(2), 159-172. doi: 10.1080/00949650802552402

Bebbington, M., Lai, C.-L., \& Zitikis, R. (2007). A flexible Weibull extension. Reliability Engineering \& System Safety, 92(6), 719-726. doi: 10.1016/j.ress.2006.03.004

Choudhury, K., \& Abdul Matin, M. (2011). Extended skew generalized normal distribution. METRON, 69(3), 265-278. doi: 10.1007/BF03263561

Choulakian, V., \& Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. Technometrics, 43(4), 478-484. doi: 10.1198/00401700152672573

Cook, R. D., \& Weisberg, S. (1994). An introduction to regression graphics. New York: John Wiley and Sons, Inc.

Eugene, N., Lee, C., \& Famoye, F. (2002). Beta-normal distribution and its applications. Communications in Statistics - Theory and Methods, 31(4), 497512. doi: 10.1081/STA-120003130

Famoye, F., Lee, C., \& Eugene, N. (2004). Beta-normal distribution: Bimodality properties and applications. Journal of Modern Applied Statistical Methods, 3(1), 85-103. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol3/iss1/10/

Famoye, F., Lee, C., \& Olumolade, O. (2005). The beta-Weibull distribution. Journal of Statistical Theory and Applications, 4(2), 121-136.

Galton, F. (1883). Inquiries into human faculty and its development. London: Macmillan and Co.

Jamalizadeh, A., Arabpour, A. R., \& Balakrishnan, N. (2011). A generalized skew two-piece skew-normal distribution. Statistical Papers, 52(2), 431-446. doi: 10.1007/s00362-009-0240-x

Johnson, N. L., Kotz, S., \& Balakrishnan, N. (1994). Continuous univariate distributions (Vol. 1) (2nd ed.). New York: John Wiley and Sons, Inc.

Jones, M. C. (2004). Families of distributions arising from distributions of order statistics. Test, 13(1), 1-43. doi: 10.1007/BF02602999

Lai, C. D. (2014). Generalized Weibull distributions. Berlin: Springer.
Lee, C., Famoye, F., \& Alzaatreh, A. (2013). Methods for generating families of univariate continuous distributions in the recent decades. Wiley Interdisciplinary Reviews: Computational Statistics, 5(3), 219-238. doi: 10.1002/wics. 1255

## A GENERALIZATION OF THE WEIBULL DISTRIBUTION

Mahmoudi, E. (2011). The beta generalized Pareto distribution with application to lifetime data. Mathematics and Computers in Simulation, 81(11), 2414-2430. doi: 10.1016/j.matcom.2011.03.006

Moors, J. J. (1988). A quantile alternative for kurtosis. Journal of the Royal Statistical Society. Series $D$ (The Statistician), 37(1), 25-32. doi:
10.2307/2348376

Mudholkar, G. S., \& Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Transactions on Reliability, 42(2), 299-302. doi: 10.1109/24.229504

Nadarajah, S., \& Kotz, S. (2004). The beta Gumbel distribution. Mathematical Problems in Engineering, 2004(4), 323-332. doi: 10.1155/S1024123X04403068

Nadarajah, S., \& Kotz, S. (2006). The beta exponential distribution.
Reliability Engineering \& System Safety, 91(6), 689-697. doi:
10.1016/j.ress.2005.05.008

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-432.

Smith, R. L., \& Naylor, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. Journal of the Royal Statistical Society. Series C (Applied Statistics), 36(3), 358-369. doi: 10.2307/2347795

Steinbrecher, G., \& Shaw, W. T. (2008). Quantile mechanics. European Journal of Applied Mathematics, 19(2), 87-112. doi: 10.1017/S0956792508007341

Weibull, W. (1939). Statistical theory of the strength of materials. Ingeniörs Vetenskaps Akademiens, Handlingar, 151, 1-45.

Weibull, W. (1951). Statistical distribution functions of wide applicability. Journal of Applied Mechanics, 18(3), 293-297.

Xie, M., \& Lai, C. (1995). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. Reliability Engineering \& System Safety, 52(1), 87-93. doi: 10.1016/0951-8320(95)00149-2

Xie, M., Tang, Y., \& Goh, T. (2002). A modified Weibull extension with bathtub failure rate function. Reliability Engineering \& System Safety, 76(3), 279285. doi: 10.1016/S0951-8320(02)00022-4

## Letters To The Editor

# Reflections Concerning Recent Ban on NHST and Confidence Intervals 

Grayson L. Baird<br>Lifespan Hospital System<br>Providence, RI

Sunny R. Duerr<br>SUNY New Paltz<br>New Paltz, NY

This letter addresses some of the immediate consequences of Basic and Applied Social Psychology's (BASP) ban on null hypothesis significance testing (NHST) and confidence intervals. The letter concludes with three suggestions to improve research in general.

Keywords: NHST, NHSTP, BASP, basic and applied social psychology, ban on NHST, confidence intervals, null hypothesis significance testing

The editorial board of Basic and Applied Social Psychology (BASP) made a bold and unequivocal move by outright banning the use of Null Hypothesis Significance Testing (NHST) and confidence intervals, along with giving Bayesian methods at best conditional consideration (see Trafimow \& Marks, 2015). BASP's reasoning behind said ban is based on common concerns of Frequentist statistics in particular, though concerns of Bayesian statistics were also considered. The reasons for said ban are not of interest here, but rather BASP's particular solution. They stated:
...BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem. (Trafimow \& Marks, 2015, p. 1)

Although BASP's intentions to improve the quality of research are commendable, what is not immediately evident is how the use of strong descriptive

Grayson L. Baird is a Research Statistician with the Lifespan Biostatistics Core. Email him at: grayson_baird@brown.edu.

## BAIRD \& DUERR

statistics and larger sample sizes constitute a framework by which inference on a population may be made. By relying on descriptive statistics alone, BASP removed the notion of probability from their statistical methodology, save for the occasionally sanctioned Bayesian analysis. As a consequence, the scope of BASP's scientific inquiry is therefore limited to the description of samples rather than inference to populations. The danger here is this limitation will not stop some readers from making inferences to populations, but will instead only remove the theoretical basis for doing so-thus blurring the distinction between interpretations of inferential and descriptive statistics.

What is especially curious about BASP's aforementioned stance is their notion of sample size and its curative effects over the stability of descriptive statistics and the size of the sampling error (which are the foundational elements of the confidence interval, simply removing probability). Though descriptive statistics can become more stable and sampling error can decrease as sample size increases, this is only true in part.

The point can be illustrated by use of M\&M's $\mathbb{C}$. Assume there is a single 42 oz "party-size bag" of M\&M's and 20 bags of the regular 1.69 oz store-size bags (totaling only 33.8 oz ), randomly sampled from different stores. Which would produce the better estimate of the color proportions from the factory machine settings: the proportions of the 42 oz bag or the mean of the proportions of the 20 1.69 oz bags? Granted, there is probably only a small difference between the two estimates, by design due to quality control. But if one machine goes out of control and fills a bag with too many green M\&M's, the 42 oz bag will be both large and largely biased, while the 1.69 oz bag will be just one sample out of many.

The issue is clear: sample size alone cannot ensure better estimates of a population, the sampling methods by which a sample is procured are of the upmost importance. To quote former American Statistical Association president Peter Lachenbruch "A large $n$ means nothing if the sampling is biased" (Cochran, 2015, p. 17). The nicety of the M\&M's example is presumably the factory settings (population) are known and we could randomly sample the 1.69 oz bags of M\&M's, if we so desired. Unfortunately, this is difficult in the social and behavioral sciences, all the more reason why Morrison and Henkel (1969) asserted:
...for statistical inference to be possible one must first specify the population and then probability sample from that population. The notions of sampling distribution and sampling error have no meaning in statistical inference apart from the assumption of randomness in the sample selection procedure-

## REFLECTIONS CONCERNING RECENT BAN

randomness being a central feature incorporated in all probability sampling designs. (p. 133)

Indeed, sampling error consists of two components: random and systematic. By increasing the sample size, only the random component of sampling error becomes less of a problem, while the systematic part remains unchanged. As for the stability of descriptive statistics, the law of large numbers ensures statistical consistency of estimates as sample size increases, but again, this property is predicated on the sampling being random; thus, biased sampling using large sample sizes can result in consistent and biased samples.

The point here is not to condemn research done without random samplingrandom sampling is difficult if not prohibitive for many studies, more especially those in the behavioral, educational, and medical sciences; rather, it is to illustrate how BASP's newly adopted methodology is not somehow more resilient to the aforementioned issues, relative to Frequentist and Bayesian methods. Ironically, by requiring authors to use large sample sizes and report descriptive statistics, BASP's prescripts would only help these inferential frameworks, if they were allowed. Instead of removing inference from their methodology, BASP could improve the quality of research by requiring authors to do the three following things:

1. Clearly state the population of interest; not only does this help readers understand the scope of the research, it also provides useful information for conducting meta-analyses and replication studies.
2. Use random sampling methods, when possible. When random sampling is not possible, authors should be required to report what sampling methods were used, why random sampling could not be used, and likely sources of bias in the existing sample, including reporting detailed demographic statistics. This allows readers to evaluate the quality of the study sample, in reference to the population of interest.
3. Instead of relying on one all-or-nothing sample, authors should be required to collect multiple samples when possible (e.g., multiple schools, hospitals, etc.). In addition, $B A S P$ should promote publication of replication studies from existing research.

Implementation of these requirements could certainly be considered state of the art.

## BAIRD \& DUERR

## References

Cochran, J. (2015). [ASA Leaders Reminisce interview with Peter (Tony) Lachenbruch]. Amstat News, 452, 16-17. Retrieved from http://magazine.amstat.org/blog/2015/02/01/peterlachenbruch_feb2015/

Morrison, D. E., \& Henkel, R. E. (1969). Significance tests reconsidered. The American Sociologist, 4(2), 131-140. Available from http://www.jstor.org/stable/27701482

Trafimow, D., \& Marks, M. (2015). Editorial. Basic and Applied Social Psychology, 37(1), 1-2. doi: 10.1080/01973533.2015.1012991

## Instructions for Authors

Authors wishing to submit to $J M A S M$ may do so using the submission form at the journal's website, http://digitalcommons.wayne.edu/jmasm. Three areas are appropriate for JMASM:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in Regular Articles, Brief Reports, and Emerging Scholars is externally peer reviewed, with input from the Editorial Board; work appearing in Statistical Software Applications and Review and JMASM Algorithms and Code is internally reviewed by the Editorial Board. $J M A S M$ charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. JMASM uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) only (Wordperfect and .rtf formats may be acceptable - please inquire). Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are NOT acceptable for manuscript submission.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
6. The submission form requires an Abstract with a 50 word maximum, and a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left justified, indent optional.
7. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
8. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "\&" instead of "and" in multiple author listings.
9. Suggestions for style: Instead of "I drew a sample of 40 " write "A sample of 40 was selected". Use "although" instead of "while," unless the meaning is "at the same time." Use "because" instead of "since," unless the meaning is "after." Instead of "Smith (1990) notes" write "Smith (1990) noted." Do not strike the spacebar twice after a period.

## Journal of Modern Applied Statistical Methods

ISSN: 1538-9472
http://digitalcommons.wayne.edu/jmasm
PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

## Copyrights, Attribution and Usage Policies

Copyright ©2015 JMASM, Inc. JMASM retains the copyright for this work for the entire usual period, but grants assignors the right, after one year from the date of publication, to republish the work in whole or in part anywhere and in any format, provided reference is given to the original publication in $J M A S M$ (see website for further details). Readers may freely access journal content at http://digitalcommons.wayne.edu/jmasm.

## To Advertisers

Advertisements are accepted at the discretion of the editor. Send requests for advertising information to ea@jmasm.com.

Wayne StatE<br>UNVERSITY<br>LIBRARY SYSTEM


[^0]:    Dr. Rao is the Eberly Professor Emeritus of Statistics and the Director of Center for Multivariate Analysis. Email him at crrl @psu.edu. Dr. Lovric is a Professor in the Faculty of Economics. Email him at mlovric@kg.ac.rs.

[^1]:    Dr. Sawilowsky is a Professor in the College of Education and the founding editor of this journal. Email him at professorshlomo @ gmail.com.

[^2]:    Dr. Zumbo is the Paragon UBC Professor of Psychometrics and Measurement and an editor of this journal. Email him at bruno.zumbo @ubc.ca. Dr. Kroc is a post-doctoral Research Fellow in the Measurement, Evaluation, and Research Methodology Program.

[^3]:    Rand Wilcox is an Professor in the Department of Psychology. Email him at: rwilcox@usc.edu. Timothy Hayes is a doctoral candidate in the Psychology Department.

[^4]:    Dr. Stephens is a Faculty member in the Education and Psychology departments. Email him at j.a.stephens@csuohio.edu.

[^5]:    Dr. Al-Omari is in the Department of Mathematics. Email him at: amerialomari @aabu.edu.jo. Dr. Bouza is in the Department of Matemática y Computación. Email him at: cbouza2002 @ yahoo.es. Dr. Covarrubias is in the Unidad Académica de Matemática. Email him at: dcova@uagro.mx. Email Dr. Pal at: rpal2011@gmail.com.

[^6]:    T. N. Sindhu is in the Department of Statistics. Email at: sindhuqau@gmail.com. N. Feroze is in the Departments of Statistics. Email at: navidferoz@hotmail.com. M. Aslam is in the Department of Statistics. Email at: aslamsdqu @yahoo.com.

[^7]:    Dr. Lantz is an Associate Professor of Industrial Engineering in the Department of Technology Management and Economics. Email him at: bjorn.lantz@chalmers.se.

[^8]:    A. K. Sharma is in the Department of Applied Mathematics. Email at: aksharma.ism@gmail.com. G. N. Singh is in the Department of Applied Mathematics. Email at: gnsingh_ism@yahoo.com .

[^9]:    *Note: "*" indicate $\mu_{h(\mathrm{opt})}$ does not exist.

[^10]:    Reba Maji is in the Department of Applied Mathematics. Email at: rebamaji09@gmail.com.

[^11]:    A. K. Sharma is in the Department of Applied Mathematics. Email at: aksharma.ism@gmail.com. Alok Kumar Singh is a Junior Research Fellow in the Department of Applied Mathematics. Email at: aksingh.ism@gmail.com. G. N. Singh is in the Department of Applied Mathematics. Email at: gnsingh_ism@yahoo.com.

[^12]:    Note: "*" indicates $\mu_{0}$ does not exists and "--" indicates no gain.

[^13]:    Dr. Kadilar is an Associate Professor in the Department of Statistics. Email her at: gamzeozl@hacettepe.edu.tr.

[^14]:    Dr. Rahardja is a Statistician. Email her at: rahardja @gmail.com, Dewi.G.Rahardja.civ@mail.mil.

[^15]:    Paul Meyvisch is a Biostatistics Director at Janssen Pharmaceutica NV, an affiliate of Johnson \& Johnson, and a doctoral student at I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium. Email him at pmeyvisc@its.jnj.com.

[^16]:    Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

[^17]:    Note: $\beta$ : skewness; $\gamma$ : kurtosis; CP: Critical Point

[^18]:    Dr. Vossoughi is an Assistant Professor in the Department of Dental Public Health. Email him at: vossoughim@sums.ac.ir.

[^19]:    Dr. Papageorgiou is an Assistant Professor in the Department of Statistics. Email her at ioulia@aueb.gr.

[^20]:    Dr. Olubusoye is a Senior Lecturer in the Department of Statistics. Email him at: oe.olubusoye@ui.edu.ng. Dr. Yaya is a Lecturer in the Department of Statistics. Email at: os.yaya@ui.edu.ng. Mr. Ojo is PhD student in the Department of Statistics at the University of Ibadan and a Lecturer at Federal University of Technology. Email at: daruu208075@yahoo.com.

[^21]:    Dr. Dorugade is an Assistant Professor of Statistics. Email him at: adorugade@rediffmail.com.

[^22]:    Dr. Nwogu is an Associate Professor and Head of the Department of Statistics. Email them at: nwoguec@yahoo.com.

[^23]:    Source: Iwueze and Nwogu (2014).

[^24]:    Dr. Udomboso is a Lecturer in the Depatment of Statistics. Email him at:
    cg.udomboso@gmail.com.Professor Amahia is in the Department of Statistics. Email at: go.amahia@ui.edu.ng. Professor Dontwi is in the Department of Mathematics. Email him at: ikedontwi@hotmail.com.

[^25]:    Dr. Kantam is an Professor of Statistics. Email him at: kantam.rrl@gmail.com. M. S. Ravikumar is a Lecturer and Biostatistician. Email at: msrk.raama@gmail.com.

[^26]:    Dr. Walker is an Associate Professor of Quantitative Methods in the Atkinson Graduate School of Management. Email him at: rwalker@willamette.edu.

[^27]:    B. S. Trivedi is an Associate Professor in the H. L. Institute of Commerce. Email her at: bhaktida.trivedi@ahduni.edu.in. M. N. Patel is in the Department of Statistics. Email them at: mnpatel.stat@gmail.com.

[^28]:    Iman Makhdoom is in the Department of Statistics. Email at: makhdoom@pnu.ac.ir. Parviz Nasiri is in the Department of Statistics. Abbas Pak is in the Department of Computer Sciences. Email at: abbas.pak1982 @ gmail.com.

[^29]:    Dr. Hashemizadeh is an Assistant Professor in the Department of Mathematics. Email her at: hashemizadeh@kiau.ac.ir. Dr. Rahbar is a Professor in the Department of Mathematics.

[^30]:    Dr. Rao is a Visiting Professor in the Department of Statistics. Email him at: arunaraomu@gmail.com. D'Cunha is a Research Scholar student in the Department of Statistics. Email her at: gratiajuliet@gmail.com.

[^31]:    Ms. Rodrigo is a graduate student in the Department of Mathematics and Statistics. Email her at: sarasepa@mail.usf.edu. Mr. Tsokos is a Professor in the Department of Mathematics and Statistics. Email him at: ctsokos@usf.edu. Sharaf is an Assistant Professor of Mathematics and Statistics. Email him at: taysseer.sharaf@sru.edu.

[^32]:    Dr. Dyer is a Professor of Information Systems. Email him at: jdyer@georgiasouthern.edu.

[^33]:    ' 5. Delete Previous Estimates and Run-Length Outputs LastRow = ActiveSheet.UsedRange.Rows.Count: 'Count Last Row of Worksheet LastCol = ActiveSheet.UsedRange.Columns.Count 'Count Last Column of Worksheet
    Range(Cells(2, 6), Cells(LastRow, LastCol)).Select: Selection.ClearContents 'DeLete previous Estimates and RL outputs

[^34]:    Dr. Qin is an Associate Research Scientist of Medicine (Cardiology) in the School of Public Health. Email her at liqin2004@gmail.com.

[^35]:    D. R. Barot is an Assistant Professor in the H. L. Institute of Commerce, Amrut Mody School of Management. Email him at: dinesh.barot@ahduni.edu.in.

[^36]:    A. Hossianzadeh is in the Department of Statistics. Email them at: a.hossainzadeh @yahoo.com. K. Zare is in the Department of Statistics.

[^37]:    Dr. Krstić is a Toxicologist and Human Health Risk Assessment Specialist. Email him at: goran.krstic@fraserhealth.ca. Mr. N. Krstić is a student in the Department of Statistics. Dr. Zambrano-Bigiarini is a Professor of Hydrology.

[^38]:    Dr. Rose is an Adjunct Instructor of Education Evaluation and Research. Email her at: ak1734@wayne.edu. Dr. Markman is a Professor of Educational Psychology and Education Evaluation and Research. Email him at: barry.markman @ wayne.edu.

[^39]:    Dr. Wan Muhamad Amir W Ahmad is an Associate Professor. Email him at: wmamir@usm.my. Mohamad Arif Bin Awang Nawi is a postgraduate student in the Faculty Informatics and Computing. Email him at: anz_wsh@yahoo.com.

[^40]:    > Library(MASS)

[^41]:    Dr. Almheidat is an Assistant Professor in the Department of Basic Sciences. Email her at: malmheidat@uop.edu.jo. Dr. Lee is a Professor in the Department of Mathematics. Email him at: carl.lee@cmich.edu. Dr. Famoye is a Professor in the Department of Mathematics. Email him at: felix.famoye@cmich.edu.

