


5-1-2014

## Vol. 13, No. 1 (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 (2014) "Vol. 13, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 1 , Article 34.

DOI: 10.22237/jmasm/1398918780

Available at: <http://digitalcommons.wayne.edu/jmasm/vol13/iss1/34>

This Full Issue is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

```
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

Vol. 13, No. 1 • May, 2014

# Journal of Modern Applied Statistical Methods

---

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

Jack Sawilowsky  
EDITOR  
Reason Statistical Consulting

Harvey Keselman  
ASSOCIATE EDITOR EMERITUS  
Department of Psychology  
University of Manitoba

Alan Klockars  
ASSISTANT EDITOR EMERITUS  
Educational Psychology  
University of Washington

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

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

Todd C. Headrick  
ASSISTANT EDITOR  
Educational Psychology  
& Special Education  
So. Illinois University–  
Carbondale

Julie M. Smith, PhD  
EDITORIAL ASSISTANT

Joshua Neds-Fox  
EDITORIAL ASSISTANCE

---

*JMASM* (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 computer-intensive 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](mailto:ea@jmasm.com). See back matter for instructions for authors.

# Journal of Modern Applied Statistical Methods

Vol. 13, No. 1

✎ May 2014 ✎

## Table of Contents

### *Regular Articles*

2 – 22	Y. LIU B. D. ZUMBO A. D. WU	Relative Importance of Predictors in Multilevel Modeling
23 – 35	F. GEORGE	A Comparison of Shape and Scale Estimators of the Two-Parameter Weibull Distribution
36 – 54	M. BHANDARY K. FUJIWARA	An Alternative Test for the Equality of Intraclass Correlation Coefficients under Unequal Family Sizes for Several Populations
55 – 70	A. ALZAATREH I. GHOSH H. SAID	On the Gamma-Logistic Distribution
71 – 90	E. N. COMAN E. IORDACHE L. DIERKER J. FIFIELD J. J. SCHENSUL S. SUGGS R. BARBOUR	Statistical Power of Alternative Structural Models for Comparative Effectiveness Research: Advantages of Modeling Unreliability
91 – 109	M. L. LESSER M. B. AKERMAN	An Exploratory Graphical Method for Identifying Associations in $r \times c$ Contingency Tables
110 – 139	L. F. LEACH R. K. HENSON	Bias and Precision of the Squared Canonical Correlation Coefficient Under Nonnormal Data Conditions



140 – 156	<b>T. SHARAF</b> <b>C. P. TSOKOS</b>	Predicting Survival Time of Localized Melanoma Patients Using Discrete Survival Time Method
157 – 173	<b>A. A. SMADI</b> <b>J. J. JABER</b> <b>A. G. AL-ZU'BI</b>	Robustness of Several Estimators of the ACF of AR(1) Process With Non-Gaussian Errors
174 – 186	<b>R. R. L. KANTAM</b> <b>M. C. PRIYA</b> <b>M. S. RAVIKUMAR</b>	Likelihood Ratio Type Test for Linear Failure Rate Distribution vs. Exponential Distribution
187 – 198	<b>S. KUMAR</b> <b>M. VISWANATHAIAH</b>	Population Mean Estimation with Sub Sampling the Non-Respondents Using Two Phase Sampling
199 – 222	<b>J. SUBRAMANI</b> <b>G. PRABAVATHY</b>	Two Parameter Modified Ratio Estimators with Two Auxiliary Variables for Estimation of Finite Population Mean with Known Skewness, Kurtosis and Correlation Coefficient
223 – 233	<b>R. TAILOR</b> <b>H. A. LONE</b>	Separate Ratio-type Estimators of Population Mean in Stratified Random Sampling
234 – 254	<b>J. SUBRAMANI</b> <b>G. PRABAVATHY</b>	Median Based Modified Ratio Estimators with Known Quartiles of an Auxiliary Variable
255 – 266	<b>NURWIANI</b> <b>S. SUNARYO</b> <b>SETIAWAN</b> <b>B. W. OTOK</b>	Ridge Regression in Calibration Models with Symmetric Padding Extension-Daubechies Wavelet Transform Preprocessing
267 – 277	<b>R. R. L. KANTAM</b> <b>V. RAMAKRISHNA</b> <b>M. S. RAVIKUMAR</b>	Estimation and Testing in Type-II Generalized Half Logistic Distribution
278 – 286	<b>A. RASHID</b> <b>T. R. JAN</b>	A Compound of Geeta Distribution with Generalized Beta Distribution
287 – 304	<b>A. PAK</b> <b>G. A. PARHAM</b> <b>M. SARAJ</b>	Inference for the Rayleigh Distribution Based on Progressive Type II Fuzzy Censored Data

305 – 328	<b>S. PUNDIR</b> <b>R. AMALA</b>	Evaluation of Area Under the Constant Shape Bi-Weibull ROC Curve
329 – 338	<b>A. M. ÇILINGIRTÜRK</b> <b>Ö. ERGÜT</b>	Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure
339 – 353	<b>G. PRAKASH</b>	Change Point Estimation for Pareto Type-II Model
354 – 366	<b>A. BHATTACHARJEE</b>	Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis
367 – 379	<b>G. SRINIVASA RAO</b>	Estimation of Reliability in Multicomponent Stress-Strength Based on Generalized Rayleigh Distribution
380 – 396	<b>S. SINGH</b> <b>S. A. SEDORY</b>	Stochastic Randomized Response Model for a Quantitative Sensitive Random Variable
397 – 409	<b>Y. KAWASAKI</b> <b>E. MIYAOKA</b>	Comparison of Three Calculation Methods for a Bayesian Inference of Two Poisson Parameters
410 – 430	<b>O. S. YAYA</b> <b>O. I. SHITTU</b>	Specifying Asymmetric STAR models with Linear and Nonlinear GARCH Innovations: Monte Carlo Approach
431 – 445	<b>O. I. SHITTU</b> <b>K. A. ADEPOJU</b>	On the Exponentiated Weibull Distribution for Modeling Wind Speed in South Western Nigeria
446 – 462	<b>S. S. GANGULY</b>	Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising in Medical Research

### *JMASM Algorithms and Code*

---

463 – 483	<b>Y. PAN</b> <b>M. T. McBEE</b>	A Flexible Method for Conducting Power Analysis for Two- and Three-Level Hierarchical Linear Models in R
-----------	-------------------------------------	--

484 – 513	<b>M. LI</b> <b>J. R. HARRING</b> <b>G. B. MACREADY</b>	Investigating the Feasibility of Using <i>Mplus</i> in the Estimation of Growth Mixture Models
514 – 518	<b>S. MAGGIO</b> <b>S. SAWILOWSKY</b>	JMASM 33: A Two Dependent Samples Maximum Test Calculator: Excel

## **Regular Articles:** **Relative Importance of Predictors in Multilevel Modeling**

**Yan Liu**  
Harvard University  
Cambridge, MA

**Bruno D. Zumbo**  
Univ. of British Columbia  
Vancouver, BC, CAN

**Amery D. Wu**  
Univ. of British Columbia  
Vancouver, BC, CAN

---

The Pratt index is a useful and practical strategy for day-to-day researchers when ordering predictors in a multiple regression analysis. The purposes of this study are to introduce and demonstrate the use of the Pratt index to assess the relative importance of predictors for a random intercept multilevel model.

*Keywords:* Random Intercept model, multilevel model, *Mplus*, Structural equation modeling, Pratt Index

---

### **Introduction**

Multiple regression analysis is a widely used statistical method in many fields. Once predictors in a regression model are selected, it is a common practice for researchers to investigate which predictors explains more variance than others, or to identify a sub-set of predictors that explain most of the variation in the outcome variable. Hence, how to measure the relative importance of explanatory variables has been widely discussed in the regression literature (e.g., Budescu, 1993; Darlington, 1968; Green, Carroll, & Desarbo, 1978; Kruskal, 1987; Pratt, 1987; Thomas, Hughes, & Zumbo, 1998). As is commonly noted in the literature, the relative importance of a predictor reflects how much it contributes to the explanation/prediction of an outcome variable, in the presence of the other correlated predictors.

The Pratt index, a R-square based statistic, has been shown to be a useful and practical strategy when ordering predictors in terms of importance in a

---

*Dr. Liu has a Ph.D. in Psychometrics. Email at: [yan\\_liu@hms.harvard.edu](mailto:yan_liu@hms.harvard.edu). Dr. Zumbo is Professor of Measurement and Associate Editor of JMASM. Email him at: [bruno.zumbo@ubc.ca](mailto:bruno.zumbo@ubc.ca). Dr. Wu is Assistant Professor of Measurement. Email her at: [ameryw@yahoo.com](mailto:ameryw@yahoo.com).*

multiple regression analysis. However, to date, this technique has not been adapted for multilevel model or hierarchical linear model (HLM) analysis because (a) there is no natural R-square measure for a multilevel regression model akin to one in multiple regression that can be partitioned additively, and (b) the within- and between-level correlation matrices are not readily available – both of which are key elements in R-square based methods for variable ordering. However, recent advances on multilevel modeling within a structural equation modeling (SEM) framework provides these two key elements (e.g., [Asparouhov & Muthén, 2006](#); [Muthén, 1994](#)) and hence allows one to apply the Pratt index to multilevel regression models.

As Raudenbush and Bryk (2002) note, the random intercept model is widely used, especially when the clustering is a nuisance factor or one is interested in how the level-2 predictors affect the means of the outcome variable (e.g., [Bryk & Driscoll, 1988](#); [Englert, et al., 1988](#); [Judge, Scott, & Ilies, 2006](#); [Muijs, 2003](#)), and hence the ordering of predictors has practical significance and value. The purpose of this study is to demonstrate how to order the relative importance of predictors in a multilevel regression analysis with a random intercept using the Pratt index ([Pratt, 1987](#); [Thomas, Hughes, & Zumbo, 1998](#); [Zumbo, 2007](#)). The article is organized as follows. First, the Pratt index is briefly described. "Next, the additive property of R-square measures and estimated covariance matrices at within- and between-levels are described. Finally, it will be demonstrated how to use the Pratt index in multilevel regression models using *Mplus* with two examples: (a) a random intercept only multilevel regression analysis, and (b) a random intercept only with a new multilevel regression approach-- latent covariate.

## Pratt Index

Herein a very brief sketch of Pratt's variable ordering measure is provided, similar to the one described in [Zumbo \(2007\)](#). The interested reader is referred to [Pratt \(1987\)](#) and [Thomas, Hughes, & Zumbo \(1998\)](#) for details. Pratt considered a linear regression of the form

$$y = b_0 + b_1x_1 + \dots + b_px_p + \varepsilon, \quad (1)$$

where residual term  $\varepsilon$  is uncorrelated with  $x_1, x_2, \dots, x_p$  and is distributed with mean zero and variance  $\sigma^2$ . The total standardized variance ( $R^2$ ) in a population explained by the model in equation (1) can be written as

$$R^2 = \sum_j b_j \rho_j \quad (2)$$

where  $b_j$  is the standardized regression coefficient corresponding to  $x_j$ , and  $\rho_j$  is the simple correlation (i.e., zero-order correlation) between  $y$  and  $x_j$ . Pratt justified the rule whereby relative importance of a predictor is equated to variance explained, provided that the explained variance attributed to  $x_j$  is  $b_j \rho_j$ , a definition which is widely used in the applied literature (e.g., Green, Carroll & De Sarbo, 1978).

An additional feature of Pratt's measure is that it allows the importance of a subset of variables to be defined additively, as the sum of their individual importance irrespective of the correlation among the predictors. Other commonly used measures (e.g., the standardized beta-weights, the  $t$ -values, zero-order correlations, semi-partial correlations) do not allow for an additive definition and may be problematic with correlated predictor variables.

Thomas, Hughes, and Zumbo (1998) provide a sample interpretation of Pratt's measure based on the geometry of least squares. They considered a sample regression equation,

$$\hat{y} = \hat{b}_1 x_1 + \dots + \hat{b}_p x_p \quad (3)$$

where the  $\hat{b}_j$ s are estimates of the population regression coefficients,  $j = 1, \dots, p$ . They defined the partition of  $R^2$  of  $x_j$ ,  $j=1, \dots, p$ , to be the signed length of the orthogonal projection of  $\hat{b}_j x_j$  onto  $\hat{y}$ , to the length of  $\hat{y}$ . By definition, this ratio represents the proportion of  $R^2$  and sums to 1.0. Furthermore, the partitioning is additive, so that one could, for example, compute the proportion of  $R^2$  attributable to various subsets of the explanatory variables, irrespective of the correlations among the explanatory variables.

One then can partition the resulting  $R^2$  by computing the Pratt index,  $d_j$ ,

$$d_j = \frac{\hat{b}_j \times r_j}{R^2}, \quad (4)$$

where, as above,  $\hat{b}_j$  is the  $j$ th standardized regression coefficient (the "beta"),  $r_j$  is the simple Pearson correlation, also called zero-order correlation, between the response variable and  $j$ th explanatory variable in equations (1) and (3) in samples.

The sum of  $d_j$ , computed from equation (4), over all predictors is one, and the relative importance of predictors can be ordered by  $d_j$ , that is, the larger the value of  $d_j$  the more important the predictor, as per Pratt (1987). Thomas (1992) suggested that as a general rule, if  $d_j < 1/(2p)$  (where  $p$  is the number of predictors), namely half the average importance, then the predictor can be regarded as unimportant.

A variety of strategies have been used in practice in the literature, such as standardized regression coefficients (i.e., beta-weights), zero-order correlations, and the  $t$ -tests and its  $p$ -values for the regression coefficients, but they can give inconsistent results when the predictors are correlated because they do not have the additive property as indicated above. In addition to the Pratt index, two other methods have also been recommended in the literature, dominance analysis (Budescu, 1993) and proportional marginal variance decomposition (i.e., a modified version of dominance analysis) (Feldman, 2005). However, these two methods are computationally intensive with even a modest number of predictors, whereas the Pratt index requires simple computation and is easy to understand and interpret.

## Additive R-squares and Correlations Using SEM

R-square is a widely used global effect size in multiple regression analysis, which is used to quantify the variance in an outcome variable explained by the model (i.e., by all the explanatory variables). However, R-square in a multilevel analysis is not straightforward. Several R-square or effect size measures were suggested in the literature, but none of them is equivalent to the one used in a multiple regression and the calculation of R-square for a random slope model is more complex due to the covariance of residuals between the intercept and slope(s) (Gelman & Pardoe, 2006; Hox, 2010; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Roberts & Monaco, 2006; Singer & Willett, 2003; Snijders & Bosker, 1999).

Based on a SEM framework, the recent advances in multilevel modeling have made it possible for us to use the Pratt index in a multilevel regression analysis. Unlike the conventional multilevel modeling approach, the observed covariance matrix can be decomposed into within- and between-levels orthogonally using the SEM framework. Cronbach and Webb (1979) proposed to decompose the observed individual variables into within- and between- group components, which can be written as  $Y_{\text{tot}} = Y_w + Y_b$ , and the components  $Y_w$  and  $Y_b$  are orthogonal and additive. This decomposition can be used for the partition

of population covariance matrix to  $\Sigma_w$  (within-level covariance matrix) and  $\Sigma_b$  (between-level covariance matrix). Muthén (1989, 1990, 1994) showed that the sample covariance matrices can be used to estimate the multilevel population covariance matrices. In addition, Muthén (1994) showed that the pooled within-level covariance matrix is an unbiased estimate of the population within-level covariance matrix  $\Sigma_w$ , which is given by

$$S_w = \frac{\sum_j \sum_i (Y_{ij} - \bar{Y}_j)(Y_{ij} - \bar{Y}_j)'}{N - G} = \hat{\Sigma}_w$$

where  $i$  denotes individuals,  $j$  denotes groups,  $N$  is the total sample size,  $G$  is the total number of groups,  $Y_{ij}$  denotes individual observations of all observed variables,  $\bar{Y}_j$  denotes the group means of all observed variables, and the symbol prime denotes transpose. Muthén further showed that the sample between-level covariance matrix is an estimate of the composite  $S_b = \hat{\Sigma}_w + c \hat{\Sigma}_b$ , where  $c$  is a scaling factor

$$c = \frac{N^2 - \sum_j n_j^2}{N(G-1)}$$

and  $S_b$  is given by

$$S_b = \frac{\sum_j n(\bar{Y} - \bar{Y}_j)(\bar{Y} - \bar{Y}_j)'}{G-1}.$$

The maximum likelihood estimate of  $\hat{\Sigma}_b$  is  $c^{-1}(S_b - S_w)$ .

The estimated within- and between-level covariance matrices allow us to obtain two key components that are needed for the calculation of the Pratt index: the correlations of the outcome variable with the predictors and the variances of the outcome variable explained by the model at within- and between-levels, respectively. The correlations can be obtained from covariance matrices as the correlation matrices are simply the standardized covariance matrices. The additive



property of estimated variance-covariance matrices at the within- and between-levels makes it possible to obtain the R-square which is conceptually equivalent to the one used in a multiple regression analysis and is always positive – a property that is not always guaranteed by other “R-square” measures discussed in the literature. The total variance of the outcome variable at both levels can be obtained directly from the covariance matrices, the residual variances at both levels can be obtained from a multilevel model analysis, and R-square can be computed from the equation  $R^2 = 1 - \frac{\sigma_e^2}{\sigma_{tot}^2}$ , which applies to both within- and

between-levels. It should be noted that R-square arising from this method is akin to the R-square in regression and hence can be partitioned using Equation (4). However, it should be noted that the additive property of R-square described here only applies to a random intercept regression model and the problems raised by a random slope model are still not yet solved. This limitation is also true for other methods of ordering in multilevel models such as those based on dominance analysis (Luo & Azen, 2013).

The *Mplus* software program has currently made those parameter estimates available in the output file. The covariance as well as correlation matrices at both within- and between-levels can be obtained by requesting “SAMPSTAT” under the “OUTPUT” command. The request for “STANDARDIZED” under the “OUTPUT” command will give R-squares for within- and between-levels, respectively, and the standardized beta-weights (i.e., beta-weights in the section of “STDYX Standardization” in the output). Researchers can also calculate R-square using the variance of outcome variable and the residual variance obtained from the *Mplus* output. Examples of *Mplus* syntax and *Mplus* output can be found in [Appendices A and B](#).

## Two Demonstrations

In this section, the use of the Pratt index with two real data examples is demonstrated. The first is a demonstration of a commonly used model in conventional HLM practice and involves what is often referred to as a random intercept model with predictors at both within- and between-levels. The second is a demonstration of a model that is referred to as a latent covariate approach, wherein the observed predictors are decomposed into two latent components rather than the common practice of aggregating individual observations to form a group level predictor.

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

### Data sources

The data were retrieved from Trends in the International Mathematics and Science Study (*TIMSS*) 2007. TIMSS 2007 Grade-8 mathematics ability scores, plausible values, estimated by item response theory, were used as the outcome variable. For the purpose of demonstration, one of five plausible values for the analyses was chosen. Six predictors (either measured variables or derived indices by TIMSS) were chosen from the students' questionnaire as within-level predictors. These within-level predictors included sex, use of calculator (Calculator), availability of computer (Computer), students' positive affect toward mathematics (Affect), students' valuing of mathematics (Valuing), and students' perception about being safe at school (Safety). Three variables were chosen from the school principal's questionnaire as between-level predictors — good school attendance (Attendance), principals' perception of school climate (Climate), and percentage of students at economic disadvantage in the school (SES).

Among those predictors, Calculator and Computer are on 4-point Likert scale (never, some lessons, half the lessons, & every or almost every lesson); Affect, Valuing, Safety, Attendance, and Climate are on 3-point Likert scale (low, medium, & high); and SES are on 4-point Likert scale (0-10%, 11-25%, 26-50%, & more than 50%). A detailed description of these variables can be found in the TIMSS 2007 User Guide (Foy & Olson, 2009). A total of 120 schools, 3470 students from Hong Kong were included in the analysis and 50.4% of students are girls. It should be noted that the same data set will be used for both demonstrations.

### Demonstration One

**Data analysis.** Please see Raudenbush and Bryk's (2002) case study one for a description of the random intercept model in their notation. In the case herein, the random intercept multilevel regression model was estimated using *Mplus* 6.02 to address how students' mathematics ability was affected by the between-level (school level) factors as well as the within-level (student level) factors. The Pratt indices were computed to answer the question—which predictors are more important when accounting for the variance in the outcome variable (mathematics ability). The two-level random intercept model in the *Mplus* formulation can be described as:

$$\begin{aligned} \text{Within-model: } Y_{ij} = & \beta_{0j} + \beta_1 \text{ Gender} + \beta_2 \text{ Valuing} + \beta_3 \text{ Computer} \\ & + \beta_4 \text{ Affect} + \beta_5 \text{ Calculator} + r_{ij}, \end{aligned} \quad (5)$$

$$\text{Between-model: } \beta_{0j} = \gamma_{00} + \gamma_{01} \text{ Attendance} + \gamma_{02} \text{ Climate} + \gamma_{03} \text{ SES} + u_{0j},$$

where  $i$  denotes the number of students;  $j$  denotes the number of schools;  $\beta_{0j}$  is the random intercept; other  $\beta$ s are the fixed slopes of within-model predictors;  $\gamma_{00}$  is the model grand mean; other  $\gamma$ s are the slopes for the between model predictors;  $r_{ij}$  is the within-level residual; and  $u_{0j}$  is the between-level residual for the random intercept.

**Results** Table 1 shows the results of the multilevel regression analysis with a random intercept and the Pratt index for each predictor. The second column is the standardized regression coefficients (the ‘beta’-weights); the following columns present  $t$ -tests, the corresponding  $p$ -values, zero-order correlations, and Pratt indices. The upper and lower parts of the table contain the results of the student-level (within) and school-level (between) models, respectively. The *Mplus* syntax and output of Demonstrate One can be found in [Appendix A](#).

Using the common practices described above, would make contradictory conclusions about the relative importance of within-level predictors, *sex*, *calculator*, *computer* and *valuing math*, if using different strategies. For example, one would consider *computer* more important than *sex*, *calculator* and *valuing math* if relying on the beta-weights. However, one would regard *sex* more important if relying on  $t$ -tests or the corresponding  $p$ -values or regard *valuing math* more important than *computer*, *sex*, or *calculator* if using simple correlations. These strategies are problematic as they do not have the additive property mentioned earlier.

However, due to its additive property, the Pratt index orthogonally partitions the R-square and sums to one, which can provide us a criterion of how much each predictor contributes to the explained variance in the outcome variable orthogonally. Using the Pratt indices, *calculator* is shown to be more important than *sex*, *computer*, and *valuing math*, but all of them have made trivial contributions to the model relative to *affect*, which accounted for 73.8% of the R-square ( $R^2=0.135$ ).

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

**Table 1.** A Random Intercept Multilevel Regression Analysis and Corresponding Pratt Indices

Within Level	Beta-weight	t-test	p-value	Correlation	Pratt
<i>valuing math</i>	0.048	2.610	0.009	0.158	0.056
<i>computer</i>	-0.104	-5.031	<0.001	-0.077	0.059
<i>sex</i>	0.090	5.108	<0.001	0.094	0.063
<i>calculator</i>	0.090	4.659	<0.001	0.121	0.081
<i>Affect</i>	0.300	16.381	<0.001	0.332	0.738
R-square	0.135				SUM=1.0
Between Level	Beta-weight	t-test	p-value	Correlation	Pratt
<i>school climate</i>	0.249	3.831	<0.001	0.354	0.258
<i>low SES</i>	-0.259	-3.258	0.001	-0.430	0.326
<i>school attendance</i>	0.319	4.196	<0.001	0.447	0.417
R-square	0.342				SUM=1.0

**Note.** The sum of Pratt index of all predictors in either within- or between-levels is not exactly one due to rounding errors from parameter estimates.

For the between-level model, *school attendance* was shown as the most important predictor among the three school variables. The order of importance would also be different, depending on whether beta-weights, correlation, or *t*-tests are used as criterion. The Pratt indices showed that *school attendance* is the most important predictor, which accounted for 41.7% of the explained variance ( $R^2=0.342$ ). The next important predictors are *low SES* and *school climate*, which accounted for 32.6% and 25.8% of the explained variance, R-square, respectively. Using Thomas' (1992) criterion, all the values of Pratt indices are greater than 0.167, so that those between-level predictors could be considered as important predictors.

### Demonstration 2

**Data Analysis** In the second example, a new approach is demonstrated that allows us to examine a predictor at both levels though it is collected at the individual level. In some situations, data was collected from individuals, but it was also desired to investigate them at an aggregate level. For example, imagine students' socioeconomic status (SES) was collected from individual students, but also were interested in the effects of school SES. Rather than aggregating SES by taking an average from within-level, a new approach that decomposes SES variable into two latent components ( $SES_{within}$  and  $SES_{between}$ ) in the multilevel

regression analysis can be used, which would reduce the measurement error arising from aggregating the data as is typical within SEM.

In general, a manifest covariate  $X_{ij}$  can be decomposed into two latent components  $X_{ij} = X_{wij} + X_{bj}$  where  $X_{wij}$  and  $X_{bj}$  are latent covariates. The multilevel equations are defined as  $Y_{ij} = \beta_{0j} + \beta_{1j}X_{wij} + r_{ij}$  and  $\beta_{0j} = \gamma_{00} + \gamma_{01}X_{bj} + u_{0j}$  where all the notation is defined the same as in Equation (3). A detailed description can be found in Asparouhov and Muthen (2006) and Ludtke, et al. (2008). This approach has also been adopted in Preacher, Zyphur, and Zhang's (2010) multilevel mediational models.

This latent variable decomposition approach is used in the second demonstration. Building on the model in the first demonstration, one variable *safety* (How safe students feel at schools) was added, which was collected from individual students, but the effects of safety at both student and school levels can be examined using this latent variable decomposition approach. The 2-level model is presented as follows:

$$\begin{aligned} \text{Within-model: } Y_{ij} &= \beta_{0j} + \beta_1 \text{ Gender} + \beta_2 \text{ Valuing} + \beta_3 \text{ Computer} \\ &\quad + \beta_4 \text{ Affect} + \beta_5 \text{ Calculator} + \beta_6 \text{ Safety}_{\text{within}} + r_{ij}, \\ \text{Between-model: } \beta_{0j} &= \gamma_{00} + \gamma_{01} \text{ Attendance} + \gamma_{02} \text{ Climate} + \gamma_{03} \text{ SES} \\ &\quad + \gamma_{04} \text{ Safety}_{\text{between}} + u_{0j}, \end{aligned} \quad (6)$$

where all parameters are defined as in the Equation (3) except that the variable *safety* was added into the within-level and between-level models in Equation (4);  $\beta_6$  is the slope for *safety* in level-1 model;  $\gamma_{04}$  is the slope for *safety* in level-2 model.

**Results** Table 2 presents the parameter estimates obtained from the random intercept multilevel regression analysis based on a latent variable decomposition approach. The columns 2-5 are the standardized beta-weights, t-tests, corresponding *p*-values, and correlations, respectively. The Pratt indices are calculated and shown in the last column. The *Mplus* syntax and output of Demonstrate Two can be found in Appendix B.

The interesting findings of this analysis are that although it was collected at the individual level *safety* was a trivial predictor at the student level, beta-weight=0.008, *t*=0.478, *p*=0.632, and the corresponding Pratt index=0.001, but it

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

became a salient predictor at the school level, beta-weight=0.306,  $t=2.925$ ,  $p=0.004$ , and the corresponding Pratt index=0.306.

**Table 2.** A Random Intercept Multilevel Regression Analysis Based on a Latent Variable Decomposition Approach and the Corresponding Pratt Indices

Within Level	Beta-weight	t-test	p-value	Correlation	Pratt
<b>Safety</b>	<b>0.008</b>	<b>0.478</b>	<b>0.632</b>	<b>0.005</b>	<b>0.001</b>
<i>valuing math</i>	0.048	2.657	0.008	0.158	0.056
<i>computer</i>	-0.102	-4.920	<0.001	-0.075	0.056
<i>sex</i>	0.093	5.168	<0.001	0.097	0.066
<i>calculator</i>	0.093	4.763	<0.001	0.124	0.085
<i>Affect</i>	0.300	16.365	<0.001	0.333	0.735
R-square	0.136				SUM=1.0
Between Level	Beta-weight	t-test	p-value	Correlation	Pratt
<i>school climate</i>	0.228	3.840	0.005	0.355	0.185
<i>low SES</i>	-0.239	-3.042	0.002	-0.430	0.235
<i>school attendance</i>	0.274	3.755	<0.001	0.447	0.280
<b>Safety</b>	<b>0.306</b>	<b>2.925</b>	<b>0.004</b>	<b>0.427</b>	<b>0.306</b>
R-square	0.436				SUM=1.0

**Note.** The sum of Pratt index of all predictors in either within- or between-levels is not exactly one due to rounding errors from parameter estimates.

Again, the relative importance of predictors would be ordered differently, depending on which criterion, beta-weights, t-tests, or correlations, were used for the judgment. For example, the effect of *safety* at the between level would not be considered as the most important predictor if the judgment is based on correlations or t-tests. However, using Pratt indices, *safety* was regarded as the most important predictor, which accounted for 30.6% of the R-square. The importance of the other between-level predictors is ranked in the following the order, *school attendance* (28%), *low SES* (23.5%), and *school climate* (18.5%). The order of relative importance for the within-level predictors was similar to that of Demonstration One except the inclusion of *safety*, which should not be regarded as an important predictor based on 1/(2p) criterion as it accounted for less than 8.3% of the R-square.

## Concluding Remarks

Ordering the relative importance of predictors has been a common practice in multiple regression analysis, but the methods developed in multiple regression

have not been used in multilevel regression analysis due to several statistical challenges. This study demonstrated how to order the relative importance of predictors in a multilevel regression analysis using the Pratt index. The Pratt index has not been used in multilevel regression analyses mainly because the within- and between-level variance-covariance could not be partitioned orthogonally and thus an R-square measure equivalent to the one used in multiple regression analysis cannot be obtained. The recent advances in multilevel modeling using SEM framework made the R-square available for researchers to compute the Pratt index when conducting a random intercept multilevel regression analysis.

The Pratt index provides a useful tool to day-to-day researchers. As indicated in the introductory section, the Pratt index can be used with random intercept models when one wants to eliminate a nuisance factor arising from clustering or when one is only interested in the relationship between the level-2 predictors and the average scores of outcome.

It should be noted that the Pratt index can currently only apply to a random intercept regression model. The problems of obtaining a R-square with a random slope described above have not been solved yet, such as the R-square values can be negative and the magnitude of global R-square measure depends on the scale of predictors included in the model. The residual covariances of the intercept and slopes give rise to the complexity of partitioning the within- and between-level variances-covariances. Luo and Azen (2013) also discussed this issue in their extension of dominance analysis to multilevel models. They pointed out that the random slope model is problematic when conducting dominance analysis and hence suggested readers to use the random intercept model when using dominance analysis.

It is worth noting that the Pratt index is not used as a strategy to select variables, but a tool for ordering the relative importance of variables once predictors have been chosen. Selection of the variables should be based on the data as well as the theories and literature surrounding the dependent and predictor variables. Moreover, Pratt index can only tell us the statistical importance of variables, but in practice researchers also need to consider the practical/substantive importance of variables.

## References

Asparouhov, T., & Muthen, B. (2006). *Constructing covariates in multilevel regression (Mplus Web Notes No. 11)*. Retrieved from <http://www.statmodel.com>.

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

Bring, J. (1996). A geometric approach to compare variables in a regression model. *American Statistician*, 50, 57-62.

Bryk, A. S. , & Driscoll, M. E. (1988). *An empirical investigation of school as a community*. Madison, WI: University of Wisconsin Research Center on Effective Secondary Schools.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114, 542 - 551.

Cronbach, L. J., & Webb, N. (1979). Between class and within class effects in a reported aptitude  $\times$  treatment interaction: A reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, 67, 717-724.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.

Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., Fear, K. L., & Gregg, S. L. (1988). *A case for writing intervention: Strategies for writing informational text*. East Lansing, MI: Michigan State University, Institute for Research On Teaching.

Feldman, B. (2005). *Relative Importance and Value*, unpublished manuscript (Version 1.1, March 19, 2005). Available online at <http://www.prismanalytics.com/docs/RelativeImportance050319.pdf>.

Foy, P. & Olson, J.F. (Eds.). (2009). *TIMSS 2007 International database and user guide (Trends in International Mathematics and Science Study)*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48, 241-251.

Green, P. E., Carroll, J. D., & Desarbo, W. S. (1978). New Measure of Predictor Variable Importance in Multiple-Regression. *Journal of Marketing Research*, 15, 356-360.

Hoffman, P. J. (1962). Assessment of the independent contributions of predictors. *Psychological Bulletin*, 59, 77-80.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Judge, T., Scott, B., & Ilies, R. (2006). Hostility, job attitudes and workplace deviance: Test of a multilevel model. *Journal of Applied Psychology*, 91, 126 – 138.



- Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.
- Kruskal, W. (1987). Relative importance by averaging over orderings. *American Statistician*, 41, 6-10.
- Ludtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203–229.
- Luo, W., & Azen, R. (2013). Determining predictors importance in hierarchical linear models using dominance analysis. *Journal of Educational and Behavioral Statistics*, 38, 3-31.
- Muijs, D. (2003). The effectiveness of learning support assistants in improving the mathematics in primary school. *Educational Research*, 45, 219–230.
- Muthén, B. (1989). Latent variable modeling in heterogeneous populations. Presidential address to the Psychometric Society, July, 1989. *Psychometrika*, 54, 557-585.
- Muthén, B. (1990). Mean and covariance structure analysis of hierarchical data. Paper presented at the Psychometric Society meeting in Princeton, NJ, June 1990. *UCLA Statistics Series* 62.
- Muthén, B. (1994). Multilevel covariance structure analysis. In J. Hox, & I. Kreft (Eds.), *Multilevel Modeling, a special issue of Sociological Methods & Research*, 22, 376-398.
- Pratt, J. W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained. In T. Pukkila and S. Puntanen (eds.), *Proceedings of the Second International Conference in Statistics* (pp. 245-260). Tampere, Finland: University of Tampere.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209-233.
- Raudenbush, S. W. Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. 2nd edition. Newbury Park , CA : Sage.
- Roberts, J. K., & Monaco, J. P. (April, 2006). *Effect size measures for the two- level linear multilevel model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.

Snijders, T. A. B., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22, 324-363.

Southwell, B. (2005). Between messages and people: A multilevel model of memory for television content. *Communication Research*, 32, 11 -140.

Thomas, D. R. (1992). Interpreting discriminant functions: A data analytic approach. *Multivariate Behavioural Research*, 27, 335-362.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45, 253-275.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

## Appendix A: *Mplus* Syntax and Output for Demonstration One

### *Mplus* Syntax

```

TITLE: this is an example of a two-level regression analysis for a continuous
       dependent variable with a random intercept and observed covariates
DATA:  FILE = HK_example2.dat
       FORMAT ARE 136.F8;
VARIABLE:
  NAMES = idsch idstd y1-y5 sex calculator computer affect valuing
         paredu safty attendan climate ses;
  USEVARIABLES = y1 sex calculator computer affect valuing
                attendan climate ses;
  MISSING = blank;
  WITHIN = sex calculator computer affect valuing
  BETWEEN = attendan climate ses;
  CLUSTER = idsch;
  CENTERING = GRANDMEAN (sex calculat computer affect valuing);
ANALYSIS:  TYPE = TWOLEVEL;
MODEL:
  %WITHIN%
    y1 ON sex calculat computer affect valuing;
  %BETWEEN%
    y1 ON attendan climate ses;
OUTPUT:    SAMPSTAT STANDARDIZED;

```

### *Mplus* Output

#### SAMPLE STATISTICS

#### ESTIMATED SAMPLE STATISTICS FOR WITHIN

##### Covariances

	Y1	SEX	CALCULAT	COMPUTER	AFFECT	VALUING
Y1	3259.376					
SEX	2.687	0.250				
CALCULAT	5.823	-0.037	0.706			
COMPUTER	-3.661	0.051	0.030	0.700		
AFFECT	16.487	0.035	0.093	0.025	0.757	

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

VALUING	5.847	0.007	0.062	0.024	0.192	0.420
Correlations						
	Y1	SEX	CALCULAT	COMPUTER	AFFECT	VALUING
Y1	1					
SEX	0.094	1				
CALCULAT	0.121	-0.089	1			
COMPUTER	-0.077	0.122	0.042	1		
AFFECT	0.332	0.081	0.127	0.035	1	
VALUING	0.158	0.022	0.114	0.045	0.34	1

### ESTIMATED SAMPLE STATISTICS FOR BETWEEN Covariances

	Y1	ATTENDAN	CLIMATE	SES
Y1	5091.540			
ATTENDAN	19.053	0.357		
CLIMATE	14.287	0.049	0.319	
SES	-32.546	-0.225	-0.138	1.127

### Correlations

	Y1	ATTENDAN	CLIMATE	SES
Y1	1			
ATTENDAN	0.447	1		
CLIMATE	0.354	0.144	1	
SES	-0.43	-0.355	-0.23	1

### STANDARADIZED MODEL RESULTS

#### STDYX Standardization

##### Within Level

	Estimate	S.E.	Est./S.E.	P-Value
Y1 ON				
SEX	0.090	0.018	5.108	0.000
CALCULAT	0.090	0.019	4.659	0.000
COMPUTER	-0.104	0.021	-5.031	0.000
AFFECT	0.300	0.018	16.381	0.000
VALUING	0.048	0.018	2.610	0.009

##### Residual Variances

Y1	0.865	0.012	69.355	0.000
----	-------	-------	--------	-------

##### Between Level

# LIU ET AL

Y1 ON

ATTENDAN	0.319	0.076	4.196	0.000
CLIMATE	0.249	0.065	3.831	0.000
SES	-0.259	0.080	-3.258	0.001

Intercepts

Y1	7.939	0.554	14.319	0.000
----	-------	-------	--------	-------

Residual Variances

Y1	0.658	0.077	8.557	0.0000
----	-------	-------	-------	--------

R-SQUARE

Within Level

Observed

Variable	Estimate	S.E.	Est./S.E.	P-Value
Y1	0.135	0.012	10.786	0.000

Between Level

Y1	0.342	0.077	4.448	0.000
----	-------	-------	-------	-------

## Appendix B: Mplus Syntax and Output for Demonstration Two

### Mplus Syntax

```

TITLE: this is an example of a two-level regression analysis for a continuous
       dependent variable with a random intercept - latent variable
       decomposition
DATA:  FILE = HK_example2.dat
       FORMAT ARE 136.F8;
VARIABLE:
       NAMES = idsch idstd y1-y5 sex calculator computer affect valuing
              paredu safty attendan climate ses;
       USEVARIABLES ARE y1 sex calculator computer affect valuing
              safty attendan climate ses;
       MISSING = blank;
       WITHIN = sex calculator computer affect valuing
       BETWEEN = attendan climate ses;
       CLUSTER = idsch;
       CENTERING = GRANDMEAN (sex calculat computer affect valuing
                              safty attendan climate ses);
ANALYSIS:      TYPE = TWOLEVEL;
MODEL:
       %WITHIN%
       y1 ON sex calculat computer affect valuing safty;
       %BETWEEN%
       y1 ON attendan climate ses safty;
OUTPUT:        SAMPSTAT STANDARDIZED;

```

### Mplus Output

#### SAMPLE STATISTICS

#### ESTIMATED SAMPLE STATISTICS FOR WITHIN

##### Covariances

	Y1	SEX	CALCULAT	COMPUTER	AFFECT	VALUING	SAFTY
Y1	3255.204						
SEX	2.757	0.250					
CALCULAT	5.944	-0.038	0.706				

# LIU ET AL

COMPUTER	-3.555	0.052	0.030	0.699			
AFFECT	16.522	0.036	0.093	0.025	0.757		
VALUING	5.853	0.007	0.062	0.024	0.191	0.420	
SAFTY	0.188	-0.040	-0.001	-0.024	0.009	-0.002	0.415

## Correlations

	Y1	SEX	CALCULAT	COMPUTER	AFFECT	VALUING	SAFTY
Y1	1						
SEX	0.097	1					
CALCULAT	0.124	-0.091	1				
COMPUTER	-0.075	0.124	0.043	1			
AFFECT	0.333	0.084	0.127	0.034	1		
VALUING	0.158	0.023	0.114	0.044	0.339	1	
SAFTY	0.005	-0.126	-0.001	-0.045	0.016	-0.004	1

## ESTIMATED SAMPLE STATISTICS FOR BETWEEN

### Covariances

	Y1	SAFTY	ATTENDAN	CLIMATE	SES
Y1	5093.476				
SAFTY	4.285	0.020			
ATTENDAN	19.068	0.015	0.357		
CLIMATE	14.302	0.009	0.049	0.319	
SES	-32.546	-0.020	-0.225	-0.138	1.127

### Correlations

	VALUING	SAFTY	ATTENDAN	CLIMATE	SES
VALUING	1				
SAFTY	0.427	1			
ATTENDAN	0.447	0.183	1		
CLIMATE	0.355	0.108	0.144	1	
SES	-0.430	-0.137	-0.355	-0.230	1

## STANDARDIZED MODEL RESULTS

### STDYX Standardization

#### Within Level

	Estimate	S.E.	Est./S.E.	P-Value
Y1 ON				
SEX	0.093	0.018	5.168	0.000
CALCULAT	0.093	0.020	4.763	0.000
COMPUTER	-0.102	0.021	-4.920	0.000

## RELATIVE IMPORTANCE OF PREDICTORS IN MULTILEVEL MODELING

AFFECT	0.300	0.018	16.365	0.000
VALUING	0.048	0.018	2.657	0.008
SAFTY	0.008	0.016	0.478	0.632
Residual Variances				
Y1	0.864	0.013	67.907	0.000

### Between Level

#### Y1 ON

ATTENDAN	0.274	0.073	3.755	0.000
CLIMATE	0.228	0.059	3.843	0.000
SES	-0.239	0.079	-3.042	0.002
SAFTY	0.306	0.105	2.925	0.003

### Intercepts

Y1	7.972	0.546	14.601	0.000
----	-------	-------	--------	-------

### Residual Variances

Y1	67	03	96	00
----	----	----	----	----

## R-SQUARE

### Within Level

#### Observed

Variable	Estimate	S.E.	Est./S.E.	P-Value
Y1	0.136	0.013	10.660	0.000

### Between Level

Y1	0.433	0.103	4.197	0.000
----	-------	-------	-------	-------



# A Comparison of Shape and Scale Estimators of the Two-Parameter Weibull Distribution

**Florence George**

Florida International University  
Miami, FL

---

Weibull distributions are widely used in reliability and survival analysis. In this paper, different methods to estimate the shape and scale parameters of the two-parameter Weibull distribution have been reviewed and compared, based on the bias, mean square error and variance. Because a theoretical comparison is not possible, an extensive simulation study has been conducted to compare the performance of different estimators. Based on the simulation study it was observed that MLE consistently performs better than other methods.

*Keywords:* Two-parameter Weibull distribution, scale parameters, shape parameters

---

## Introduction

The Weibull distribution is a commonly used model in reliability, life time and environmental data analysis. A considerable literature discussing the methods of estimation of Weibull parameters exists (Sharoon, et al., 2012; Saralees et al., 2011; Saralees et al., 2008) because of its applications in different fields. Kantar and Senoglu (2008) did a simulation comparison of different estimators for scale parameter when shape is known. Balakrishnan and Kateri (2008) showed the existence and uniqueness of maximum likelihood estimates (MLE) of Weibull distribution. Dubey (1967) derived the percentile estimators (Percentile 1) which uses 4 different percentiles to estimate the shape and scale parameters. Seki and Yokoyama (1993) proposed a simple and robust method that uses only two percentiles, 31<sup>st</sup> and 63<sup>rd</sup> percentile (Percentile 2) to estimate both parameters. Moment estimators (MOM) and median rank regression estimators (MRRS) are also commonly used in literature (Kantar and Senoglu, 2008) because of their

---

*Dr. George is Assistant Professor in the Department of Mathematics and Statistics. Email her at: fgeorge@fiu.edu.*

## COMPARISON OF ESTIMATORS OF THE WEIBULL DISTRIBUTION

easiness in computation. Existing methods (namely MLE, MOM, MRRS, Percentile 1, and Percentile 2) for estimating both shape and scale parameters of two-parameter Weibull distribution are here reviewed and compared. A simulation study has been conducted to compare the performance of these methods under same simulation conditions.

### Statistical Methodology

The Weibull distribution has the probability density function,

$f(x) = \alpha \beta^{-\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}$  for  $x \geq 0, \alpha > 0, \beta > 0$ , where  $\alpha$  is the shape parameter and  $\beta$  is the scale parameter. The cumulative distribution function is given by

$$F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^\alpha} \text{ for } x \geq 0.$$

The distribution is reversed J-shaped when  $\alpha < 1$ , exponential when  $\alpha = 1$  and bell-shaped when  $\alpha > 1$  (Kantar and Senoglu, 2008). Because of its wide-variety of shapes it is used extensively in practice for modeling real life data in different fields.

### Maximum Likelihood Estimators (MLE)

The log-likelihood function of a random sample from the two-parameter Weibull distribution is given by

$$\ln L = n \ln \alpha - n \alpha \ln \beta + (\alpha - 1) \sum \ln x - \sum \left(\frac{x}{\beta}\right)^\alpha.$$

This will yield the following two score equations

$$\begin{aligned} \frac{\partial \ln L}{\partial \beta} &= -\frac{n\alpha}{\beta} + \frac{\alpha}{\beta^{\alpha+1}} \sum x^\alpha = 0 \text{ and} \\ \frac{\partial \ln L}{\partial \alpha} &= -\frac{n}{\alpha} - n \ln \beta + \sum \ln x + \frac{\ln \beta}{\beta^\alpha} \sum x^\alpha - \frac{1}{\beta^\alpha} \sum x^\alpha \ln x = 0 \end{aligned}$$

The above two equations can be solved numerically to obtain MLEs.

### Moment Estimators (MOM)

The moment estimators are obtained by equating the population moments to the corresponding sample moments. The first and second moments of Weibull distribution are respectively

$$\mu_1' = \beta \Gamma(1 + \alpha^{-1}) \text{ and}$$

$$\mu_2' = \beta^2 \Gamma(1 + 2\alpha^{-1})$$

The first two moments from the sample are  $m_1' = \frac{1}{n} \sum x$  and  $m_2' = \frac{1}{n} \sum x^2$ .

The moment estimates are obtained by solving the following two equations

$$m_1' = \beta \Gamma(1 + \alpha^{-1})$$

$$m_2' = \beta^2 \Gamma(1 + 2\alpha^{-1})$$

### Median Rank Regression Estimators (MRRS)

MRR is a procedure for estimating the Weibull parameters by fitting a least squares regression line through the points on a probability plot. Thus,

$$\log(1 - F(x)) = -\left(\frac{x}{\beta}\right)^\alpha \text{ and hence}$$

$$\log(\log(1 - F(x))) = \alpha \log x = \alpha \log \beta.$$

This is now a linear model and method of least squares can be used to estimate  $\alpha$  and  $\beta$ . The sample data are first sorted in ascending order and then following Abernethy (2006), the distribution function,  $F(x_i)$  is approximated for each point  $(x_i)$  in the sorted sample as  $F(x_i) = \frac{i - 0.3}{n + 0.4}$ , where  $I$  is the ascending rank of the data point  $x_i$ .

### Percentile Estimators (Percentile 1)

Percentile estimators for both shape and scale parameters were derived by Dubey (1967). He proposed an estimator based on 17<sup>th</sup> and 97<sup>th</sup> percentiles for shape parameter and one based on 40<sup>th</sup> and 82<sup>nd</sup> percentile for scale parameter. The formulae for the shape and scale percentile estimators are presented here; for details refer to Dubey (1967). Let  $p_1 = 0.1673$  and  $p_2 = 0.9737$ . Define  $k_1 = \log(-\log(1 - p_1)) - \log(-\log(1 - p_2))$ . Let  $y_1$  and  $y_2$  represent the 100 $p_1$ th percentile from the data. Then

$$\hat{\alpha} = \frac{-k_1}{\log(y_1) - \log(y_2)}$$

Similarly to estimate  $\beta$ , define  $p_3 = 0.3978$  and  $p_4 = 0.8211$ . Let  $k_2 = \log(-\log(1 - p_3)) - \log(-\log(1 - p_4))$ ;  $k_3 = -\log(1 - p_3)$  and  $w = 1 - \frac{\log(k_3)}{k_2}$ . Let  $y_3$  and  $y_4$  represent the 100 $p_3$ th and 100 $p_4$ th percentile from the data. Then

$$\beta = \exp(w \log(y_3) + (1 - w) \log(y_4)).$$

### Improved Percentile Estimators (Percentile 2)

Seki and Yokoyama (1993) proposed this simple and robust method that uses only two percentiles, 31<sup>st</sup> and 63<sup>rd</sup> percentile to estimate  $\alpha$  and  $\beta$ . The Weibull cumulative distribution function is given by

$$F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^\alpha} \text{ for } x \geq 0.$$

Hence the 100 $p$ th percentile of the Weibull distribution can be written as  $x_p = \beta(-\log(1 - p))^{1/\beta}$ . Then the 100(1 -  $e^{-1}$ ) = 63.2<sup>th</sup> percentile is  $x_{0.632} = \beta$  for any Weibull distribution. This can be used to compute  $\hat{\beta}$ . Therefore, the estimate of the shape parameter can be obtained as  $\hat{\alpha} = \frac{\log(-\log(1 - p))}{\log\left(\frac{x_p}{x_{0.632}}\right)}$ . Seki and

Yokoyama (1993) approximated the numerator of this estimator as  $-1$  and then obtained  $p = 0.31$ , approximately, to obtain  $\hat{\alpha}$ .

## Simulation Study

A simulation study has been conducted to explore the performances of the different methods discussed in this article.

### Simulation Technique

The main objective of this study is to compare the performance of five different methods to estimate the shape and scale parameters of two-parameter Weibull distribution. Weibull distribution with parameters scale = 10 and shape = 0.5, 1, 1.5, 2, 3 and 4 were used to generate 5,000 samples of sizes  $n = 5, 10, 20, 30, 50$  and 100. The estimates are compared using the values of average bias, mean squared error (MSE) and variance. The simulation was done using statistical software R version 2.15.2.

### Results and Discussion

The results of the simulation are shown in [Tables 1 to 3](#). The bias and MSEs from Weibull (10, 0.5) and Weibull (10, 3) are also presented in [Figures 1 to 4](#). From [Tables 1 to 3](#), it can be observed that as sample size increases, bias, MSE and variance decrease. For small sample size, the performance of methods differs significantly. For all methods, absolute bias, MSE and variance decrease as sample size increases. It can be observed from [Tables 1 to 3](#) and [Figures 1 to 4](#), in almost all cases MLE performed better than the other 4 methods and percentile method-1 performed the worst. In some situations, MRRS also performs well, especially for shape estimates. It can also be observed that both percentile estimators perform poorly in estimation of shape. There is no consistency in the performance of estimates by the method of moments. Because MLE is performing consistently better than the other 4 methods practitioners are encouraged to use MLE whenever possible.

# COMPARISON OF ESTIMATORS OF THE WEIBULL DISTRIBUTION

**Table 1.** Bias, Variance and MSE of both Scale and Shape estimates  $\alpha=10$  and  $\beta=0.5$ ; 1

$\alpha, \beta$	$n$		Scale					Shape				
			MLE	MOM	MRRS	Perc1e1	Perc1e2	MLE	MOM	MRRS	Perc1e1	Perc1e2
10, 0.5	5	Bias	2.929	6.216	5.361	5.340	4.394	0.219	0.351	0.025	0.661	0.497
		Vars	156.930	224.302	219.866	282.068	197.209	0.147	0.092	0.085	5.231	0.313
		MSEs	165.511	262.939	248.610	310.581	216.516	0.195	0.216	0.085	5.668	0.560
	10	Bias	1.525	4.606	2.998	2.334	1.773	0.085	0.218	-0.015	0.283	0.193
		Vars	65.147	96.815	84.821	110.433	73.834	0.031	0.033	0.026	0.349	0.050
		MSEs	67.473	118.026	93.811	115.879	76.977	0.038	0.081	0.026	0.429	0.087
	20	Bias	0.705	3.217	1.643	0.979	0.656	0.036	0.136	-0.021	0.112	0.079
		Vars	25.634	40.337	32.389	43.623	30.013	0.010	0.016	0.012	0.069	0.015
		MSEs	26.131	50.689	35.088	44.582	30.444	0.012	0.035	0.013	0.082	0.021
	30	Bias	0.556	2.621	1.258	0.707	0.442	0.026	0.107	-0.017	0.072	0.054
		Vars	17.111	25.216	20.965	26.613	19.722	0.006	0.013	0.008	0.036	0.009
		MSEs	17.420	32.084	22.548	27.113	19.917	0.007	0.024	0.008	0.041	0.012
	50	Bias	0.283	1.912	0.783	0.405	0.207	0.014	0.076	-0.016	0.039	0.034
		Vars	9.585	15.077	11.172	15.220	11.347	0.003	0.008	0.005	0.017	0.005
		MSEs	9.665	18.731	11.785	15.384	11.390	0.004	0.014	0.005	0.019	0.006
	100	Bias	0.174	1.257	0.487	0.236	0.158	0.008	0.048	-0.011	0.021	0.018
		Vars	4.664	8.153	5.320	7.373	5.727	0.002	0.005	0.003	0.007	0.002
		MSEs	4.694	9.734	5.557	7.429	5.752	0.002	0.008	0.003	0.008	0.003
	5	Bias	0.284	0.312	1.232	0.485	0.304	0.443	0.353	0.049	1.297	0.926
		Vars	23.380	23.097	27.835	29.087	24.408	0.621	0.411	0.347	10.932	1.147
		MSEs	23.461	23.194	29.353	29.322	24.501	0.817	0.535	0.349	12.614	2.005
	10	Bias	0.190	0.222	0.840	0.223	0.060	0.173	0.173	-0.032	0.507	0.371
		Vars	11.082	11.192	12.944	15.822	12.013	0.127	0.108	0.106	0.991	0.199
		MSEs	11.118	11.241	13.650	15.872	12.016	0.157	0.137	0.107	1.247	0.336
	20	Bias	0.072	0.100	0.486	0.067	-0.047	0.074	0.087	-0.042	0.225	0.158
		Vars	5.576	5.707	6.335	8.386	6.348	0.043	0.046	0.048	0.281	0.063
		MSEs	5.581	5.717	6.571	8.390	6.350	0.049	0.053	0.050	0.331	0.088
	30	Bias	0.075	0.098	0.412	0.049	-0.006	0.049	0.062	-0.040	0.141	0.108
		Vars	3.671	3.796	4.182	5.412	4.340	0.025	0.029	0.032	0.134	0.037
		MSEs	3.676	3.805	4.351	5.414	4.340	0.027	0.033	0.033	0.154	0.048
	50	Bias	0.050	0.062	0.281	0.028	-0.010	0.029	0.039	-0.031	0.087	0.069
		Vars	2.173	2.269	2.426	3.332	2.635	0.014	0.018	0.020	0.072	0.020
		MSEs	2.175	2.273	2.504	3.332	2.635	0.015	0.019	0.021	0.080	0.025
	100	Bias	0.032	0.034	0.179	0.012	0.002	0.016	0.021	-0.021	0.045	0.037
		Vars	1.117	1.178	1.232	1.698	1.349	0.007	0.010	0.010	0.031	0.010
		MSEs	1.118	1.179	1.264	1.698	1.349	0.007	0.010	0.011	0.033	0.011

**Table 2.** Bias, Variance and MSE of both Scale and Shape estimates  $\alpha=10$  and  $\beta=1.5$ ; 2

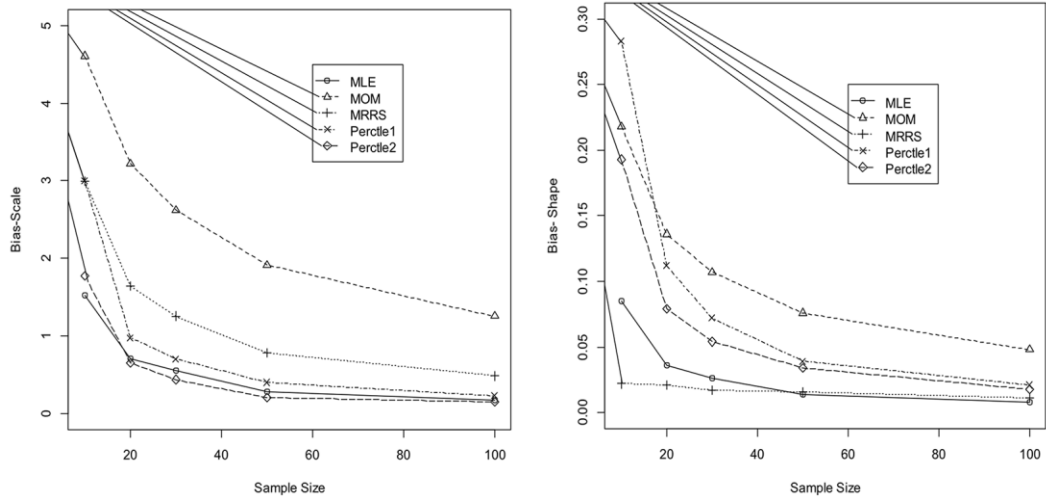
$\alpha, \beta$	$n$		Scale					Shape				
			MLE	MOM	MRRS	Perctle1	Perctle2	MLE	MOM	MRRS	Perctle1	Perctle2
10, 1.5	5	Bias	-0.081	-0.149	0.513	-0.149	-0.226	0.650	0.398	0.059	2.045	1.298
		Vars	9.333	9.331	10.477	11.069	9.546	1.374	0.947	0.713	47.203	2.365
		MSEs	9.340	9.353	10.740	11.091	9.597	1.797	1.106	0.717	51.386	4.050
	10	Bias	0.038	-0.002	0.451	-0.016	-0.092	0.257	0.171	-0.048	0.755	0.537
		Vars	4.948	4.986	5.542	6.870	5.493	0.275	0.231	0.232	2.331	0.427
		MSEs	4.950	4.986	5.746	6.870	5.501	0.342	0.260	0.235	2.901	0.715
	20	Bias	0.039	0.016	0.317	0.016	-0.046	0.114	0.081	-0.065	0.339	0.248
		Vars	2.508	2.517	2.763	3.802	2.851	0.097	0.090	0.109	0.606	0.144
		MSEs	2.509	2.517	2.863	3.802	2.853	0.110	0.096	0.113	0.721	0.205
	30	Bias	0.014	-0.003	0.227	-0.006	-0.064	0.075	0.054	-0.054	0.209	0.167
		Vars	1.669	1.677	1.816	2.439	1.947	0.055	0.054	0.074	0.324	0.083
		MSEs	1.669	1.677	1.867	2.439	1.952	0.061	0.057	0.077	0.368	0.111
	50	Bias	-0.008	-0.019	0.145	-0.021	-0.050	0.041	0.030	-0.050	0.121	0.099
		Vars	0.973	0.979	1.048	1.486	1.166	0.031	0.032	0.045	0.155	0.048
		MSEs	0.973	0.979	1.069	1.486	1.168	0.033	0.033	0.048	0.170	0.058
	100	Bias	0.003	-0.004	0.102	-0.020	-0.027	0.022	0.016	-0.032	0.067	0.054
		Vars	0.493	0.495	0.533	0.748	0.591	0.014	0.016	0.023	0.070	0.021
		MSEs	0.493	0.495	0.543	0.749	0.592	0.015	0.016	0.024	0.074	0.024
	5	Bias	-0.102	-0.123	0.339	-0.200	-0.263	0.868	0.510	0.089	2.499	1.673
		Vars	5.328	5.376	5.763	6.306	5.441	2.301	1.833	1.321	43.582	3.716
		MSEs	5.339	5.391	5.878	6.345	5.511	3.055	2.093	1.329	49.828	6.516
	10	Bias	-0.049	-0.065	0.254	-0.132	-0.189	0.338	0.191	-0.060	1.024	0.699
		Vars	2.797	2.830	3.090	3.811	3.062	0.459	0.390	0.394	4.718	0.699
		MSEs	2.799	2.834	3.154	3.829	3.098	0.573	0.427	0.398	5.767	1.187
	20	Bias	-0.041	-0.051	0.157	-0.089	-0.136	0.151	0.087	-0.080	0.473	0.331
		Vars	1.359	1.368	1.461	2.068	1.560	0.171	0.159	0.197	1.296	0.262
		MSEs	1.361	1.371	1.486	2.076	1.578	0.194	0.166	0.203	1.520	0.371
10, 2	30	Bias	-0.004	-0.009	0.157	-0.023	-0.059	0.096	0.054	-0.081	0.275	0.222
		Vars	0.911	0.915	0.991	1.347	1.059	0.096	0.090	0.125	0.540	0.142
		MSEs	0.911	0.915	1.016	1.348	1.062	0.105	0.093	0.131	0.615	0.191
	50	Bias	-0.007	-0.010	0.110	-0.028	-0.049	0.059	0.035	-0.063	0.166	0.141
		Vars	0.543	0.546	0.596	0.829	0.654	0.055	0.054	0.080	0.269	0.084
		MSEs	0.543	0.546	0.608	0.830	0.656	0.059	0.055	0.084	0.297	0.104
	100	Bias	0.001	-0.002	0.071	-0.020	-0.022	0.031	0.019	-0.040	0.088	0.073
		Vars	0.276	0.276	0.296	0.427	0.332	0.025	0.025	0.041	0.120	0.037
		MSEs	0.276	0.276	0.301	0.427	0.333	0.026	0.026	0.042	0.128	0.043

# COMPARISON OF ESTIMATORS OF THE WEIBULL DISTRIBUTION

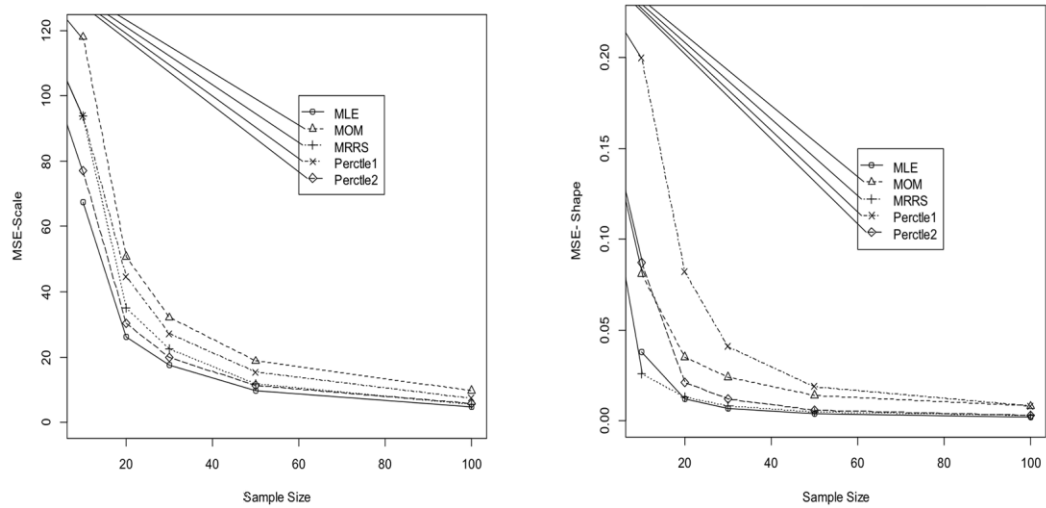
**Table 3.** Bias, Variance and MSE of both Scale and Shape estimates  $\alpha=10$  and  $\beta=3; 4$

$\alpha, \beta$	$n$		Scale					Shape				
			MLE	MOM	MRRS	Perctle1	Perctle2	MLE	MOM	MRRS	Perctle1	Perctle2
10, 3	5	Bias	-0.166	-0.138	0.122	-0.275	-0.320	1.295	0.767	0.131	3.899	2.417
		Vars	2.392	2.439	2.515	2.865	2.489	5.768	5.041	3.285	243.739	8.976
		MSEs	2.419	2.458	2.530	2.941	2.591	7.446	5.629	3.302	258.938	14.818
	10	Bias	-0.053	-0.037	0.147	-0.129	-0.152	0.505	0.276	-0.102	1.461	1.048
		Vars	1.221	1.234	1.290	1.659	1.352	1.100	0.998	0.911	9.896	1.722
		MSEs	1.224	1.235	1.312	1.676	1.374	1.354	1.074	0.922	12.032	2.820
	20	Bias	-0.017	-0.009	0.120	-0.060	-0.070	0.218	0.113	-0.134	0.689	0.487
		Vars	0.602	0.605	0.649	0.896	0.690	0.374	0.359	0.428	2.594	0.572
		MSEs	0.602	0.605	0.663	0.900	0.695	0.422	0.372	0.445	3.070	0.810
	30	Bias	-0.022	-0.017	0.082	-0.049	-0.057	0.153	0.084	-0.110	0.452	0.341
		Vars	0.415	0.417	0.442	0.618	0.487	0.230	0.227	0.294	1.387	0.343
		MSEs	0.416	0.417	0.448	0.620	0.491	0.253	0.234	0.306	1.592	0.459
	50	Bias	-0.020	-0.017	0.058	-0.039	-0.044	0.083	0.044	-0.101	0.254	0.209
		Vars	0.243	0.243	0.259	0.358	0.288	0.121	0.123	0.185	0.643	0.183
		MSEs	0.243	0.243	0.262	0.360	0.290	0.128	0.125	0.196	0.707	0.226
	100	Bias	0.000	0.001	0.047	-0.003	-0.014	0.035	0.014	-0.075	0.098	0.096
		Vars	0.118	0.118	0.128	0.186	0.144	0.058	0.059	0.092	0.265	0.087
		MSEs	0.118	0.118	0.130	0.186	0.144	0.059	0.059	0.098	0.274	0.096
	5	Bias	-0.125	-0.081	0.091	-0.230	-0.258	1.716	1.051	0.164	4.996	3.156
		Vars	1.388	1.411	1.442	1.683	1.471	9.001	7.677	5.003	141.435	13.999
		MSEs	1.404	1.418	1.451	1.736	1.537	11.945	8.782	5.029	166.393	23.962
	10	Bias	-0.078	-0.054	0.076	-0.140	-0.156	0.686	0.401	-0.131	2.017	1.402
		Vars	0.670	0.675	0.709	0.913	0.747	2.063	1.999	1.723	16.360	3.181
		MSEs	0.676	0.678	0.714	0.932	0.772	2.535	2.159	1.740	20.429	5.146
	20	Bias	-0.033	-0.021	0.070	-0.056	-0.076	0.305	0.166	-0.174	0.875	0.659
		Vars	0.362	0.364	0.391	0.545	0.420	0.691	0.695	0.774	4.648	1.023
		MSEs	0.363	0.365	0.396	0.548	0.426	0.783	0.723	0.804	5.414	1.457
	30	Bias	-0.024	-0.016	0.055	-0.038	-0.055	0.194	0.107	-0.156	0.549	0.444
		Vars	0.229	0.230	0.245	0.339	0.271	0.388	0.411	0.523	2.300	0.567
		MSEs	0.229	0.230	0.248	0.341	0.274	0.425	0.423	0.547	2.601	0.764
	50	Bias	-0.016	-0.011	0.042	-0.025	-0.035	0.112	0.060	-0.132	0.322	0.271
		Vars	0.138	0.139	0.149	0.209	0.163	0.224	0.237	0.323	1.087	0.342
		MSEs	0.138	0.139	0.151	0.209	0.165	0.237	0.241	0.341	1.190	0.415
	100	Bias	-0.013	-0.011	0.023	-0.021	-0.025	0.060	0.036	-0.087	0.156	0.147
		Vars	0.066	0.067	0.071	0.104	0.082	0.108	0.116	0.174	0.478	0.157
		MSEs	0.066	0.067	0.072	0.105	0.083	0.111	0.118	0.181	0.503	0.178



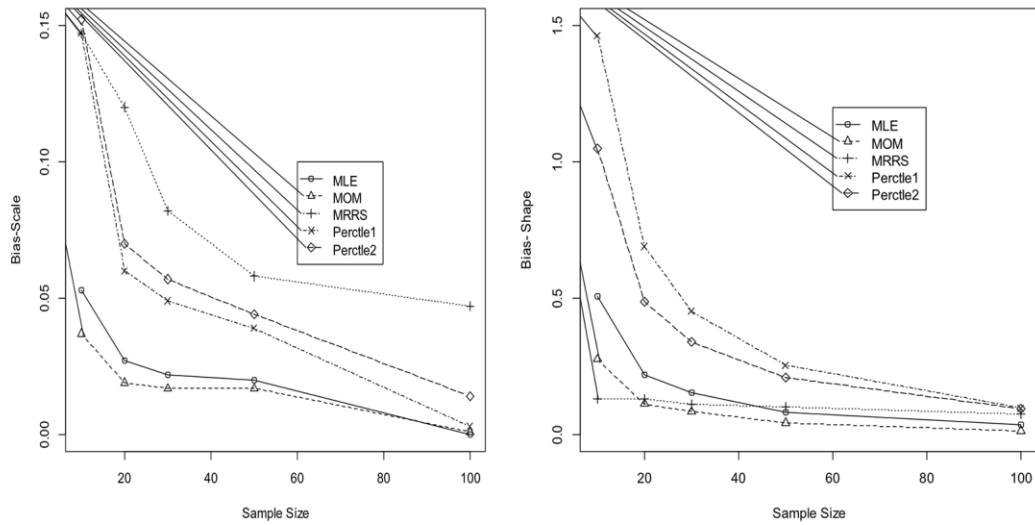


**Figure 1.** Absolute Bias of Scale parameter estimate (left), Shape parameter estimate (right) vs. Sample size from Weibull (10, 0.5)

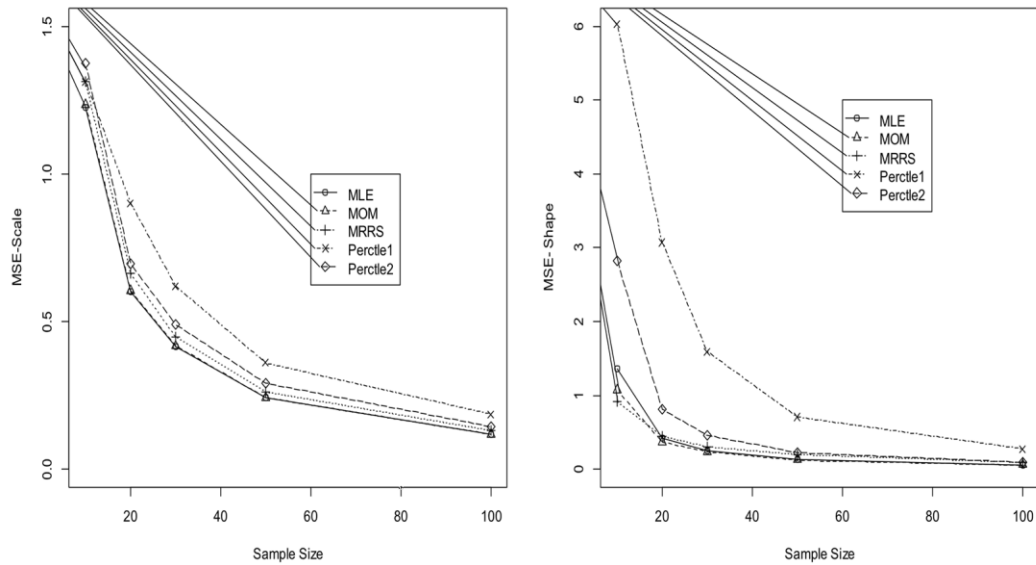


**Figure 2.** MSE of Scale parameter estimate (left), Shape parameter estimate (right) vs. Sample size from Weibull (10, 0.5)

## COMPARISON OF ESTIMATORS OF THE WEIBULL DISTRIBUTION



**Figure 3.** Absolute Bias of Scale parameter estimate (left), Shape parameter estimate (right) vs. Sample size from Weibull (10, 3)



**Figure 4.** MSE of Scale parameter estimate (left), Shape parameter estimate (right) vs. Sample size from Weibull (10, 3)

### Example

Many researchers modeled wind data using the Weibull distribution (Dorvlo, 2002; Weisser, 2003; Celik, 2003). The five methods with an example in Battacharya and Bhattacharjee (2010) will be discussed next. This example provides the average monthly wind speed (m/s) of Kolkata from 1<sup>st</sup> March 2009 to 31<sup>st</sup> March 2009. Table 4 presents the data set.

The estimates of the two-parameter Weibull distribution obtained by fitting to the data using the methods discussed in the article are given in the Table 5. It look like the Percentile 1 and Percentile 2 estimates are at the extreme ends and the MLE estimates lie somewhat between the values of other estimates.

**Table 4.** Average daily wind speed in Kolkata during March 2009.

Date	Speed (m/s)	Date	Speed (m/s)
1	0.56	17	0.28
2	0.28	18	0.83
3	0.56	19	1.39
4	0.56	20	1.11
5	1.11	21	1.11
6	0.83	22	0.83
7	1.11	23	0.56
8	1.94	24	0.83
9	1.11	25	1.67
10	0.83	26	1.94
11	1.11	27	1.39
12	1.39	28	0.83
13	0.28	29	2.22
14	0.56	30	1.67
15	0.28	31	2.22
16	0.28		

**Table 5.** Estimates of two-parameter Weibull by different methods.

Method	Scale	Shape
MLE	1.1550	1.9081
MOM	1.1501	1.8456
MRRS	1.1636	1.8031
Percentile 1	1.1100	1.8055
Percentile 2	1.1816	2.1704

### Conclusion

Five different methods for the joint estimation of both scale and shape parameters of two-parameter Weibull distribution were reviewed in this article. A simulation study was conducted to compare the five methods based on bias, mean square error and variance of estimates. From simulation results, it was observed that MLE performs consistently better than MOM, MRRS, percentile method and

## COMPARISON OF ESTIMATORS OF THE WEIBULL DISTRIBUTION

improved percentile method and therefore MLE estimates are recommended to the practitioners.

### References

- Abernethy, R. B. (2006). (5<sup>th</sup> Ed.). *The new Weibull handbook : reliability & statistical analysis for predicting life, safety, risk, support costs, failures, and forecasting warranty claims, substantiation and accelerated testing, using Weibull, Log normal, crow-AMSAA, probit, and Kaplan-Meier models*. North Palm Beach, FL: R. B. Abernethy.
- Abernethy, R. B., Breneman, J. E., Medlin, C. H., & Reinman, G. L. (1983). *Weibull Analysis Handbook*. Air Force Wright Aeronautical Laboratories Technical Report AFWAL-TR- 83-2079. Available at <http://handle.dtic.mil/100.2/ADA143100>.
- Balakrishanan, N. & Kateri, M. (2008). On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data. *Statistics & Probability Letters*, 78(17): 2971–2975.
- Battacharya, P., & Bhattacharjee, R. (2010). A study on Weibull distribution for estimating the parameters. *Journal of Applied Quantitative Methods*, 5(2): 234-241.
- Celik, A. N. (2003). Energy output estimation for small-scale wind power generators using Weibull-representative wind data. *Journal of Wind Engineering and Industrial Aerodynamics*, 91: 693-707. doi: 10.1016/S0167-6105(02)00471-3
- Dorvlo, A. S. (2002). Estimating wind speed distribution. *Energy Conservation and Management*, 43: 2311-2318.
- Dubey, S. Y. D. (1967). Normal and Weibull distributions. *Naval Research Logistics Quarterly*, 14(1): 69–79. doi: 10.1002/nav.3800140107
- Kantar, Y. M. & Senoglu, B. (2008). A Comparative Study for the Location and Scale Parameters of the Weibull Distribution with a given Shape Parameter. *Computers & Geosciences*, 34: 1900-1909.
- Saralees, N., & Firoozeh, H. (2011). An extension of the exponential distribution. *Statistics - A Journal of Theoretical and Applied Statistics*, 45(6): 543-558.
- Saralees, N., & Kotz, S. (2008). Strength modeling using Weibull distributions. *Journal of Mechanical Science and Technology*, 22: 1247-1254.

Seki, T., & Yokoyama, S. (1993). Simple and robust estimation of the Weibull parameters. *Microelectronics Reliability*, 33(1): 45-52.

Sharoon, H., Muhammad, Q. S., Muhammad, M., & Kibria, B. M. G. (2012). A Note on Beta Inverse - Weibull Distribution. *Communication in Statistics - Theory and Methods*, 42(2): 320-335.

Weisser, D. (2003). A wind energy analysis of Grenada: an estimation using the 'Weibull' density function. *Renewable Energy*; 28(11): 1803-1812.

# An Alternative Test for the Equality of Intraclass Correlation Coefficients under Unequal Family Sizes for Several Populations

**Madhusudan Bhandary**  
Columbus State University  
Columbus, GA

**Koji Fujiwara**  
North Dakota State University  
Fargo, ND

---

An alternative test for the equality of several intraclass correlation coefficients under unequal family sizes based on several independent multinormal samples is proposed. It was found that the alternative test consistently and reliably produced results superior to those of Likelihood ratio test (LRT) proposed by Bhandary and Alam (2000) and  $F_{\max}$  test proposed by Bhandary and Fujiwara (2006) in terms of power for various combinations of intraclass correlation coefficient values and also the alternative test stays closer to the significance level under null hypothesis compared to the Likelihood ratio test and  $F_{\max}$  test. This alternative test is computationally very simple and also can be used for both small sample and large sample situations. An example with real life data is presented.

**Keywords:** Likelihood ratio test,  $F_{\max}$ -test, Alternative test, intraclass correlation coefficient

---

## Introduction

It is sometimes necessary to estimate the correlation coefficient between blood pressures of children on the basis of measurements taken on  $p$  children in each of  $n$  families. The  $p$  measurements on a family provide  $p(p - 1)$  pairs of observations  $(x, y)$ ,  $x$  being the blood pressure of one child and  $y$  that of another. From the  $n$  families a total of  $np(p - 1)$  pairs are generated from which a correlation coefficient is computed in the ordinary way.

---

*Dr. Bhandary is a Professor in the Department of Mathematics. Email him at: [bhandary\\_madhusudan@colstate.edu](mailto:bhandary_madhusudan@colstate.edu). Koji Fujiwara is a graduate student in the Department of Statistics. Email him at: [koji.fujiwara@ndsu.edu](mailto:koji.fujiwara@ndsu.edu).*

The correlation coefficient thus computed is called intraclass correlation coefficient. It is important to have statistical inference concerning intraclass correlation, because it provides information regarding blood pressure, cholesterol etc. in a family within some race in the world.

The intraclass correlation coefficient  $\rho$  has a wide variety of applications. It can be used to measure the degree of intra-family resemblance with respect to characteristics such as blood pressure, cholesterol, weight, height, stature, lung capacity, etc.

Statistical inference concerning  $\rho$  based on a single multinormal sample has been studied by several authors (Scheffe, 1959; Rao, 1973; Rosner, et al., 1977, 1979; Donner and Bull, 1983; Srivastava, 1984; Konishi, 1985; Gokhale and SenGupta, 1986; SenGupta, 1988; Velu and Rao, 1990).

For a two sample problem, Donner and Bull (1983) discussed the likelihood ratio test for testing the equality of two intraclass correlation coefficients based on two independent multinormal samples under equal family sizes. Konishi and Gupta (1987) proposed a modified likelihood ratio test and derived its asymptotic null distribution. They also discussed another test procedure based on a modification of Fisher's Z-transformation following Konishi (1985).

For a several sample problem, Huang and Sinha (1993) considered an optimum invariant test for the equality of intraclass correlation coefficients under equal family sizes for more than two intraclass correlation coefficients based on independent samples from several multinormal distributions.

For unequal family sizes, Young and Bhandary (1998) proposed Likelihood ratio test, large sample Z-test and large sample  $Z^*$ -test for the equality of two intraclass correlation coefficients based on two independent multinormal samples.

For several populations and unequal family sizes, Bhandary and Alam (2000) proposed Likelihood ratio test and large sample ANOVA test for the equality of several intraclass correlation coefficients based on several independent multinormal samples. Bhandary and Fujiwara (2006) proposed  $F_{\max}$  test for the equality of several intraclass correlation coefficients under unequal family sizes. Donner and Zou (2002) proposed asymptotic test for the equality of dependent intraclass correlation coefficients under unequal family sizes.

An alternative test for the equality of several intraclass correlation coefficients is considered based on several independent multinormal samples under unequal family sizes.

A conditional analysis is carried out here, assuming family sizes fixed though unequal.

## A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

It could be of interest to see whether blood pressure or cholesterol or lung capacity, etc., among families in Caucasian, Asian, Hispanic or African races, etc., differ or not; therefore a small sample test for the equality of intraclass correlation coefficients under unequal family sizes has been developed.

Also, an alternative test is proposed for the equality of intraclass correlation coefficients under unequal family sizes, which is computationally very simple. A brief discussion of likelihood ratio test proposed by Bhandary and Alam (2000) and  $F_{\max}$  test proposed by Bhandary and Fujiwara (2006) are provided.

These tests are compared in the section titled *Simulation Results*, using simulation technique. It is found on the basis of simulation study that the alternative test consistently and reliably produced results superior to those of Likelihood ratio test and  $F_{\max}$  test in terms of power for various combination of intraclass correlation coefficient values and also the alternative test stays closer to the significance level under null hypothesis compared to the Likelihood ratio test and  $F_{\max}$  test.

An example with real life data is given in the section titled *Example With Real Life Data*.

### Tests of $H_0 : \rho_1 = \rho_2 = \rho_3$ Versus $H_1 : \text{NOT } H_0$

#### Likelihood Ratio Test

Let  $\underline{X}_i = (x_{i1}, x_{i2}, \dots, x_{ip_i})'$  be a  $p_i \times 1$  vector of observations from the  $i^{th}$  family;  $i = 1, 2, \dots, k$ . The structure of mean vector and the covariance matrix for the familial data is given by the following (Rao, 1973):

$$\underline{\mu}_i = \mu \underline{1}_i \text{ and } \sum_{p_i \times p_i} \sigma^2 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{pmatrix} \quad (1)$$

where  $\underline{1}_i$  is a  $p_i \times 1$  vector of 1's,  $\mu (-\infty < \mu < \infty)$  is the common mean and  $\sigma^2 (\sigma^2 > 0)$  is the common variance of members of the family and  $\rho$ , which is called the intraclass correlation coefficient, is the coefficient of correlation among the members of the family and  $\max_{1 \leq i \leq k} \left( -\frac{1}{p_i - 1} \right) \leq \rho \leq 1$ .



It is assumed that  $\tilde{x}_i \sim N_{p_i}(\tilde{\mu}_i, \tilde{\Sigma}_i); i=1, \dots, k$ , where  $N_{p_i}$  represents  $p_i$ -variate normal distribution and  $\tilde{\mu}_i, \tilde{\Sigma}_i$ 's are defined in (1).

$$\text{Let } \tilde{u}_i = (\tilde{u}_{i1}, \tilde{u}_{i2}, \dots, \tilde{u}_{ip_i})' = \tilde{Q} \tilde{X}_i \quad (2)$$

where  $\tilde{Q}$  is an orthogonal matrix.

Under the orthogonal transformation (2), it can be seen that  $\tilde{u}_i \sim N_{p_i}(\tilde{\mu}_i^*, \tilde{\Sigma}_i^*); i=1, \dots, k$  where  $\tilde{\mu}_i^* = (\mu, 0, 0, \dots, 0)'$  and  $\tilde{\Sigma}_i^* = \sigma^2 \begin{pmatrix} \eta_i & 0 & \dots & 0 \\ 0 & 1-\rho & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1-\rho \end{pmatrix}$  and  $\eta_i = p_i^{-1} \{1 + (p_i - 1)\rho\}$ .

The transformation used on the data from  $\tilde{x}$  to  $\tilde{u}$  above is independent of  $\rho$ . One can use Helmert's orthogonal transformation.

Srivastava (1984) gives estimator of  $\rho$  and  $\sigma^2$  under unequal family sizes which are good substitute for the maximum likelihood estimator and are given by the following:

$$\begin{aligned} \hat{\rho} &= 1 - \frac{\hat{\gamma}^2}{\hat{\sigma}^2} \\ \hat{\sigma}^2 &= (k-1)^{-1} \sum_{i=1}^k (u_{i1} - \hat{\mu})^2 + k^{-1} \hat{\gamma}^2 \left( \sum_{i=1}^k a_i \right) \\ \text{where } \hat{\gamma}^2 &= \frac{\sum_{i=1}^k \sum_{r=2}^{p_i} u_{ir}^2}{\sum_{i=1}^k (p_i - 1)} \\ \hat{\mu} &= k^{-1} \sum_{i=1}^k u_{i1} \\ \text{and } a_i &= 1 - p_i^{-1}. \end{aligned} \quad (3)$$

Now, consider the three sample problem with  $k_1$ ,  $k_2$  and  $k_3$  families from each population.

## A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

Let  $\underset{\sim}{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip_i})'$  be a  $p_i \times 1$  vector of observations from  $i^{th}$  family;  
 $i = 1, \dots, k_1$

$$\text{and } \underset{\sim}{x}_i \sim N_{p_i}(\underset{\sim}{\mu}_{1i}, \underset{\sim}{\Sigma}_{1i}), \text{ where } \underset{\sim}{\mu}_{1i} = \mu_1 \underset{\sim}{1}_i, \underset{\sim}{\Sigma}_{1i} = \sigma_1^2 \begin{pmatrix} 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \dots & \rho_1 \\ \dots & \dots & \dots & \dots \\ \rho_1 & \rho_1 & \dots & 1 \end{pmatrix} \quad (4)$$

$$\text{and } \max_{1 \leq i \leq k_1} \left( -\frac{1}{p_i - 1} \right) \leq \rho_1 \leq 1$$

Let  $\underset{\sim}{y}_j = (y_{j1}, y_{j2}, \dots, y_{jq_j})'$  be a  $q_j \times 1$  vector of observations from  $j^{th}$  family in the  
 second population;  $j = 1, \dots, k_2$

$$\text{and } \underset{\sim}{y}_j \sim N_{q_j}(\underset{\sim}{\mu}_{2j}, \underset{\sim}{\Sigma}_{2j})$$

$$\text{where } \underset{\sim}{\mu}_{2j} = \mu_2 \underset{\sim}{1}_j, \underset{\sim}{\Sigma}_{2j} = \sigma_2^2 \begin{pmatrix} 1 & \rho_2 & \dots & \rho_2 \\ \rho_2 & 1 & \dots & \rho_2 \\ \dots & \dots & \dots & \dots \\ \rho_2 & \rho_2 & \dots & 1 \end{pmatrix} \quad (5)$$

$$\text{and } \max_{1 \leq j \leq k_2} \left( -\frac{1}{q_j - 1} \right) \leq \rho_2 \leq 1$$

Let  $\underset{\sim}{z}_l = (z_{l1}, z_{l2}, \dots, z_{lp_l})'$  be a  $p_l \times 1$  vector of observations from  $l^{th}$  family;  
 $l = 1, 2, \dots, k_3$

$$\text{and } \underset{\sim}{z}_l \sim N_{p_l}(\underset{\sim}{\mu}_{3l}, \underset{\sim}{\Sigma}_{3l}), \text{ where } \underset{\sim}{\mu}_{3l} = \mu_3 \underset{\sim}{1}_l, \underset{\sim}{\Sigma}_{3l} = \sigma_3^2 \begin{pmatrix} 1 & \rho_3 & \dots & \rho_3 \\ \rho_3 & 1 & \dots & \rho_3 \\ \dots & \dots & \dots & \dots \\ \rho_3 & \rho_3 & \dots & 1 \end{pmatrix} \quad (6)$$

$$\text{and } \max_{1 \leq l \leq k_3} \left( -\frac{1}{r_l - 1} \right) \leq \rho_3 \leq 1.$$

Using orthogonal transformation, the data vector can be transformed from  $\tilde{x}_i$  to  $\tilde{u}_i$ ,  $\tilde{y}_j$  to  $\tilde{v}_j$  and  $\tilde{z}_l$  to  $\tilde{w}_l$  as follows:

$$\begin{aligned} \tilde{u}_i &= (u_{i1}, u_{i2}, \dots, u_{ip_i})' \sim N_{p_i}(\mu_{1i}^*, \Sigma_{1i}^*); i = 1, \dots, k_1 \\ \text{and } \tilde{v}_j &= (v_{j1}, v_{j2}, \dots, v_{jq_j})' \sim N_{q_j}(\mu_{2j}^*, \Sigma_{2j}^*); j = 1, \dots, k_2 \\ \text{and } \tilde{w}_l &= (w_{l1}, w_{l2}, \dots, w_{lr_l})' \sim N_{r_l}(\mu_{3l}^*, \Sigma_{3l}^*); l = 1, \dots, k_3 \end{aligned}$$

where,  $\tilde{\mu}_{1i}^* = (\mu_1, 0, 0, \dots, 0)'$ ,  $\tilde{\Sigma}_{1i}^* = \sigma_1^2 \begin{pmatrix} \eta_i & 0 & \dots & 0 \\ 0 & 1 - \rho_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 - \rho_1 \end{pmatrix}$

$$\eta_i = p_i^{-1} \{1 + (p_i - 1)\rho_1\} \tag{7}$$

$$\tilde{\mu}_{2j}^* = (\mu_2, 0, 0, \dots, 0)'$$
,  $\tilde{\Sigma}_{2j}^* = \sigma_2^2 \begin{pmatrix} \xi_j & 0 & \dots & 0 \\ 0 & 1 - \rho_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 - \rho_2 \end{pmatrix}$ 

$$\xi_j = q_j^{-1} \{1 + (q_j - 1)\rho_2\}$$

$$\tilde{\mu}_{3l}^* = (\mu_3, 0, 0, \dots, 0)'$$
,  $\tilde{\Sigma}_{3l}^* = \sigma_3^2 \begin{pmatrix} \varsigma_l & 0 & \dots & 0 \\ 0 & 1 - \rho_3 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 - \rho_3 \end{pmatrix}$ 

and  $\varsigma_l = r_l^{-1} \{1 + (r_l - 1)\rho_3\}$

The transformations used on the data above from  $\tilde{x}$  to  $\tilde{u}$ ,  $\tilde{y}$  to  $\tilde{v}$  and  $\tilde{z}$  to  $\tilde{w}$  are independent of  $\rho_1$ ,  $\rho_2$  and  $\rho_3$ . It is assumed that  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$ .

## A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

Under the above setup, Bhandary and Alam (2000) derived likelihood ratio test statistic for testing  $H_0 : \rho_1 = \rho_2 = \rho_3$  Vs.  $H_1 : NOT H_0$  which is given by the following:

$$\begin{aligned}
 -2 \log \Lambda &= \sum_{i=1}^{k_1} \log \left[ p_i^{-1} \{1 + (p_i - 1) \hat{\rho}\} \right] + \sum_{i=1}^{k_1} (p_i - 1) \log(1 - \hat{\rho}) \\
 &+ \sum_{j=1}^{k_2} \log \left[ q_j^{-1} \{1 + (q_j - 1) \hat{\rho}\} \right] + \sum_{j=1}^{k_2} (q_j - 1) \log(1 - \hat{\rho}) \\
 &+ \sum_{l=1}^{k_3} \log \left[ r_l^{-1} \{1 + (r_l - 1) \hat{\rho}\} \right] + \sum_{l=1}^{k_3} (r_l - 1) \log(1 - \hat{\rho}) \\
 &+ \frac{1}{\hat{\sigma}^2} \left[ \sum_{i=1}^{k_1} \left\{ p_i (u_{i1} - \hat{\mu}_1)^2 / [1 + (p_i - 1) \hat{\rho}] \right\} + \sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 / (1 - \hat{\rho}) \right. \\
 &+ \sum_{j=1}^{k_2} \left\{ q_j (v_{j1} - \hat{\mu}_2)^2 / [1 + (q_j - 1) \hat{\rho}] \right\} + \sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 / (1 - \hat{\rho}) \\
 &\left. + \sum_{l=1}^{k_3} \left\{ r_l (w_{l1} - \hat{\mu}_3)^2 / [1 + (r_l - 1) \hat{\rho}] \right\} + \sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 / (1 - \hat{\rho}) \right] \\
 &- \sum_{i=1}^{k_1} \log \left[ p_i^{-1} \{1 + (p_i - 1) \hat{\rho}_1\} \right] - \sum_{i=1}^{k_1} (p_i - 1) \log(1 - \hat{\rho}_1) \\
 &- \sum_{j=1}^{k_2} \log \left[ q_j^{-1} \{1 + (q_j - 1) \hat{\rho}_2\} \right] - \sum_{j=1}^{k_2} (q_j - 1) \log(1 - \hat{\rho}_2) \\
 &- \sum_{l=1}^{k_3} \log \left[ r_l^{-1} \{1 + (r_l - 1) \hat{\rho}_3\} \right] - \sum_{l=1}^{k_3} (r_l - 1) \log(1 - \hat{\rho}_3) \\
 &- \frac{1}{\hat{\sigma}^2} \left[ \sum_{i=1}^{k_1} \left\{ p_i (u_{i1} - \hat{\mu}_1)^2 / [1 + (p_i - 1) \hat{\rho}_1] \right\} + \sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 / (1 - \hat{\rho}_1) \right. \\
 &+ \sum_{j=1}^{k_2} \left\{ q_j (v_{j1} - \hat{\mu}_2)^2 / [1 + (q_j - 1) \hat{\rho}_2] \right\} + \sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 / (1 - \hat{\rho}_2) \\
 &\left. + \sum_{l=1}^{k_3} \left\{ r_l (w_{l1} - \hat{\mu}_3)^2 / [1 + (r_l - 1) \hat{\rho}_3] \right\} + \sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 / (1 - \hat{\rho}_3) \right] \tag{8}
 \end{aligned}$$

where,  $\Lambda$  = likelihood ratio test statistic,

$\hat{\rho}$  = estimate of common intraclass correlation coefficients under  $H_0$ ,

$\hat{\rho}_1$  = estimate of intraclass correlation coefficient from first sample  
under  $H_1$ ,

$\hat{\rho}_2$  = estimate of intraclass correlation coefficient from second sample  
under  $H_1$ ,

$\hat{\rho}_3$  = estimate of intraclass correlation coefficient from third sample  
under  $H_1$ ,

$\hat{\sigma}^2$  = estimate of  $\sigma^2$

and  $\hat{\mu}_1, \hat{\mu}_2$  and  $\hat{\mu}_3$  are estimates of means from first ,second and third samples respectively.

The estimators  $\hat{\rho}, \hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\sigma}^2, \hat{\mu}_1, \hat{\mu}_2$  and  $\hat{\mu}_3$  can be obtained from Srivastava's estimator given by (3).

It is known from asymptotic theory that  $-2\log \Lambda$  has an asymptotic chi-square distribution with 2 degrees of freedom.

Bhandary and Alam (2000) also suggested large sample ANOVA test and showed through simulation that likelihood ratio test given by (8) consistently produced results superior to those of the large sample ANOVA test.

The likelihood ratio test given by (8) is computationally complex, and used asymptotically – that is, when family sizes are large (at least 30). But situations may also call for a small sample case. An alternative test is here proposed, which is computationally very simple and can be used for both small sample and large sample situations.

### **$F_{\max}$ test**

The  $F_{\max}$  test is described as follows:

$$F_{\max} = \max \{F_1, F_2, F_3, F_4, F_5, F_6\} \quad (9)$$

# A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

$$\text{where } F_1 = \frac{\sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 / \left\{ \sum_{i=1}^{k_1} (p_i - 1) \right\}}{\sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 / \left\{ \sum_{j=1}^{k_2} (q_j - 1) \right\}} \quad (10)$$

$$F_2 = \frac{\sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 / \left\{ \sum_{i=1}^{k_1} (p_i - 1) \right\}}{\sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 / \left\{ \sum_{l=1}^{k_3} (r_l - 1) \right\}} \quad (11)$$

$$F_3 = \frac{\sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 / \left\{ \sum_{j=1}^{k_2} (q_j - 1) \right\}}{\sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 / \left\{ \sum_{l=1}^{k_3} (r_l - 1) \right\}} \quad (12)$$

$$F_4 = 1 / F_1, F_5 = 1 / F_2, \text{ and } F_6 = 1 / F_3 \quad (13)$$

It can be shown using (7) that

$$\frac{\sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2}{\sigma^2(1 - \rho_1)} \sim \chi_{pp}^2, \frac{\sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2}{\sigma^2(1 - \rho_2)} \sim \chi_{qq}^2 \text{ and } \frac{\sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2}{\sigma^2(1 - \rho_3)} \sim \chi_{rr}^2 \quad (14)$$

where,  $\chi_n^2$  denotes chi-square distribution with n degrees of freedom

and  $pp = \sum_{i=1}^{k_1} (p_i - 1); qq = \sum_{j=1}^{k_2} (q_j - 1); rr = \sum_{l=1}^{k_3} (r_l - 1)$ .

Therefore, using (14) under  $H_0$ , the exact distribution of the  $F_1$  given by (10) is  $F_{pp,qq}$ . (15)

Similarly, using (14) under  $H_0$ , the exact distributions of  $F_2, F_3, F_4, F_5$  and  $F_6$  are  $F_{pp,rr}, F_{qq,rr}, F_{qq,pp}, F_{rr,pp}$  and  $F_{rr,qq}$  respectively, where  $F_{n_1, n_2}$  denotes F-distribution with  $n_1$  and  $n_2$  degrees of freedom respectively. (16)

Hence, using (9), (15) and (16) and using Bonferroni's bound, approximate critical value at  $\alpha$  for testing  $H_0$  Vs.  $H_1$  can be proposed as

$$C = \max \left\{ F_{\frac{\alpha}{6}; pp, qq}, F_{\frac{\alpha}{6}; pp, rr}, F_{\frac{\alpha}{6}; qq, rr}, F_{\frac{\alpha}{6}; qq, pp}, F_{\frac{\alpha}{6}; rr, pp}, F_{\frac{\alpha}{6}; rr, qq} \right\} \quad (17)$$

where,  $F_{\gamma; a, b}$  is the upper  $100\gamma\%$  point of F-distribution with degrees of freedom a and b respectively.

The critical region for testing  $H_0$  Vs.  $H_1$  is proposed as follows:

$$F_{\max} > C \quad (18)$$

where  $F_{\max}$  and  $C$  are given by (9) and (17), respectively.

The test statistic  $F_{\max}$  given by (9) is very simple to compute, and the distributions of  $F_1, F_2, F_3, F_4, F_5$  and  $F_6$  are exact and hence can be used for both small sample and large sample situations.

### Alternative test

For the alternative test, the test statistic is described as follows:

$$F_1 = \frac{\sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 / \left\{ \sum_{i=1}^{k_1} (p_i - 1) \right\}}{\left( \sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 + \sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 \right) / \left\{ \sum_{j=1}^{k_2} (q_j - 1) + \sum_{l=1}^{k_3} (r_l - 1) \right\}}$$

$$F_2 = \frac{\sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 / \left\{ \sum_{j=1}^{k_2} (q_j - 1) \right\}}{\left( \sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 + \sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 \right) / \left\{ \sum_{i=1}^{k_1} (p_i - 1) + \sum_{l=1}^{k_3} (r_l - 1) \right\}}$$

## A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

$$F_3 = \frac{\sum_{l=1}^{k_3} \sum_{t=2}^{r_l} w_{lt}^2 / \left\{ \sum_{l=1}^{k_3} (r_l - 1) \right\}}{\left( \sum_{i=1}^{k_1} \sum_{r=2}^{p_i} u_{ir}^2 + \sum_{j=1}^{k_2} \sum_{s=2}^{q_j} v_{js}^2 \right) / \left\{ \sum_{i=1}^{k_1} (p_i - 1) + \sum_{j=1}^{k_2} (q_j - 1) \right\}}$$

$$F_4 = 1 / F_1, F_5 = 1 / F_2 \text{ and } F_6 = 1 / F_3 \quad (19)$$

Using (14), it can be said that under  $H_0$ , the exact distribution of the  $F_1$  is  $F_{pp,qq+rr}$ .

Similarly, under  $H_0$ , the exact distributions of  $F_2, F_3, F_4, F_5$  and  $F_6$  are  $F_{qq,pp+rr}$ ,  $F_{rr,pp+qq}$ ,  $F_{qq+rr,pp}$ ,  $F_{pp+rr,qq}$  and  $F_{pp+qq,rr}$  respectively, where,  $F_{n_1, n_2}$  denotes F-distribution with  $n_1$  and  $n_2$  degrees of freedom respectively.

Set the P-values to be the right tail probability of the statistics calculated above such that  $P_i = P(X > F_i)$  where  $F_i$ 's are explained in (19).

Sort the P-values obtained as above in an ascending order and denote them by  $P_{(1)}, P_{(2)}, \dots, P_{(6)}$ .

$$\text{Reject } H_0 \text{ if } P_{(i)} < \frac{i}{6} \alpha \text{ for some } i \in \{1, 2, \dots, 6\}. \quad (20)$$

In order that  $H_0$  is insignificant, it is required that  $P_{(1)} \geq \frac{1}{6} \alpha, P_{(2)} \geq \frac{2}{6} \alpha, \dots, P_{(6)} \geq \alpha$ . So, if  $P_{(i)} < \frac{i}{6} \alpha$  then the test corresponding to  $P_{(1)}$  is insignificant, corresponding to  $P_{(2)}$  is insignificant, ..., corresponding to  $P_{(i-1)}$  is insignificant and corresponding to  $P_{(i)}$  is significant and the overall test is significant.

### Simulation Results

Multivariate normal random vectors were generated using R program in order to evaluate the power of the alternative test as compared to  $F_{\max}$  test and the LRT test. Five and thirty vectors of family data were created for each of the three populations. The family size distribution was truncated to maintain the family size



at a minimum of 2 siblings and a maximum of 15 siblings. The previous research in simulating family sizes (Rosner et al., 1977; Srivastava and Keen, 1988) determined the parameter setting for FORTRAN IMSL negative binomial subroutine with a mean = 2.86 and a success probability = 0.483. Here, it is set at a mean = 2.86 and a theta = 41.2552.

All parameters were set the same for each population, except the values of  $\rho_1$ ,  $\rho_2$  and  $\rho_3$  which took various combinations over the range of values from 0.1 to 0.9 at increments of 0.1.

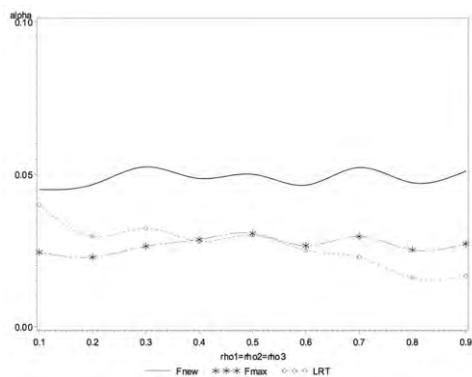
The R program produced estimates of  $\rho_1$ ,  $\rho_2$  and  $\rho_3$  along with  $F_{\max}$  statistic and LRT statistic and the new statistic 10,000 times for each particular combination of population parameters ( $\rho_1$ ,  $\rho_2$  and  $\rho_3$ ).

The frequency of rejection of each test at  $\alpha = 0.05$  was noted and the proportion of rejections are noted for a sample combinations of  $\rho_1$ ,  $\rho_2$  and  $\rho_3$ .

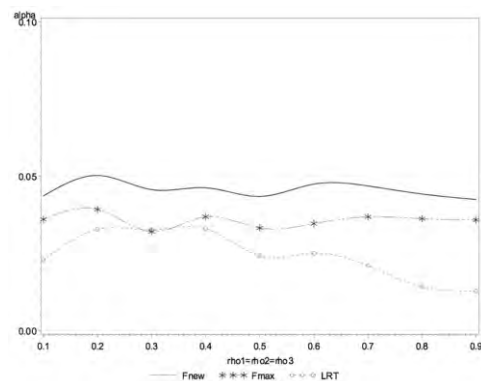
The size comparison for the alternative test,  $F_{\max}$  test and the LRT test for various combinations of  $\rho_1$ ,  $\rho_2$  and  $\rho_3$  is also presented.

A few figures are presented of powers estimates as well as size estimates for these tests. On the basis of this study, it was found that the alternative test showed consistently better results in terms of power as well as in size than LRT and  $F_{\max}$  test. This alternative test is computationally very simple and also can be used for both small sample and large sample situations. The alternative test stays closer to the significance level under null hypothesis compared to the Likelihood ratio test and  $F_{\max}$  test. It is recommend that the alternative test is used in practice.

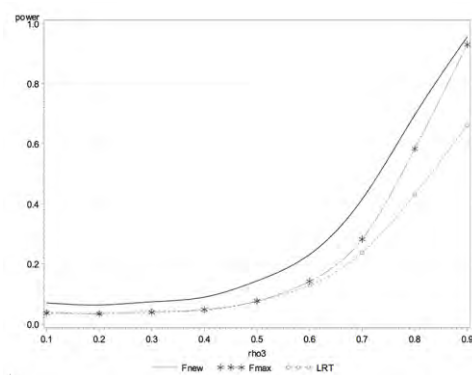
# A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS



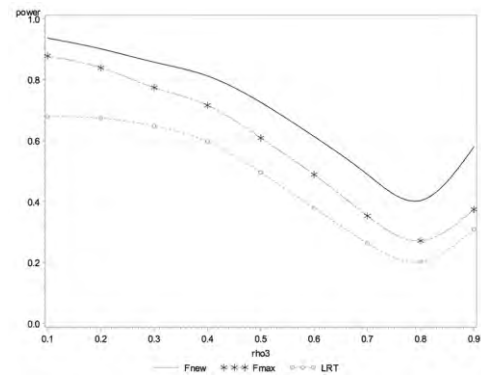
**Figure 1.** Size Estimates ( $\alpha = 0.05$  and  $k = 5$ )



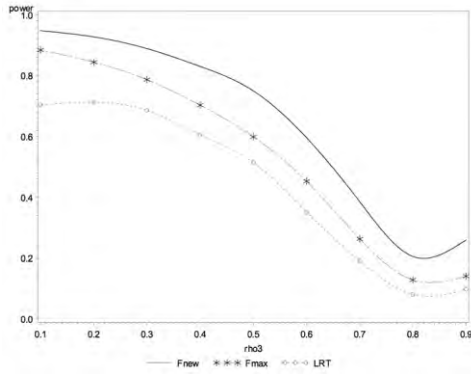
**Figure 2.** Size Estimates ( $\alpha = 0.05$  and  $k = 30$ )



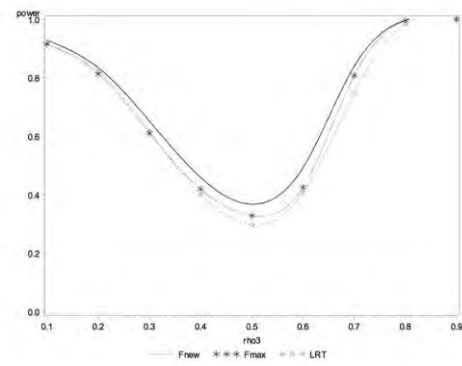
**Figure 3.** Power Estimates ( $\alpha = 0.05$ ,  $k = 5$ ,  $\rho_1 = 0.1$  and  $\rho_2 = 0.3$ )



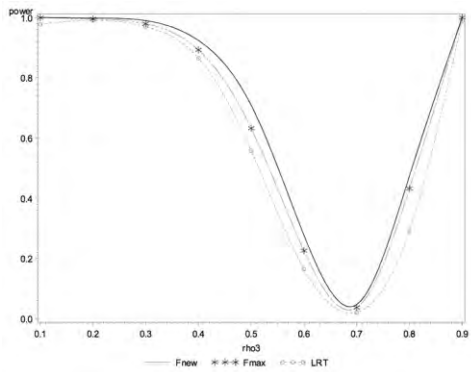
**Figure 4.** Power Estimates ( $\alpha = 0.05$ ,  $k = 5$ ,  $\rho_1 = 0.7$  and  $\rho_2 = 0.9$ )



**Figure 5.** Power Estimates ( $\alpha = 0.05$ ,  $k = 5$ ,  $\rho_1 = 0.9$  and  $\rho_2 = 0.8$ )



**Figure 6.** Power Estimates ( $\alpha = 0.05$ ,  $k = 30$ ,  $\rho_1 = 0.4$  and  $\rho_2 = 0.6$ )



**Figure 7.** Power Estimates ( $\alpha = 0.05$ ,  $k = 30$ ,  $\rho_1 = 0.7$  and  $\rho_2 = 0.7$ )

**Table 1.** Size Estimates ( $\alpha = 0.05$ )

$\rho$	$k = 5$			$k = 30$		
	LRT	$F_{\max}$	$F_{\text{new}}$	LRT	$F_{\max}$	$F_{\text{new}}$
0.1	0.0400	0.0244	0.0450	0.0228	0.0360	0.0436
0.2	0.0296	0.0228	0.0466	0.0328	0.0392	0.0502
0.3	0.0322	0.0264	0.0524	0.0326	0.0320	0.0456
0.4	0.0280	0.0286	0.0486	0.0330	0.0368	0.0462
0.5	0.0300	0.0306	0.0500	0.0242	0.0332	0.0434
0.6	0.0250	0.0266	0.0464	0.0250	0.0346	0.0474
0.7	0.0228	0.0296	0.0522	0.0210	0.0368	0.0468
0.8	0.0160	0.0252	0.0472	0.0142	0.0362	0.0442
0.9	0.0166	0.0272	0.0510	0.0128	0.0358	0.0424

# A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

**Table 2.** Rejection Proportions ( $\alpha = 0.05$ )

$\rho_1$	$\rho_2$	$\rho_3$	$k = 5$			$k = 30$		
			LRT	$F_{\max}$	$F_{\text{new}}$	LRT	$F_{\max}$	$F_{\text{new}}$
0.1	0.3	0.6	0.1290	0.1420	0.2312	0.9182	0.9284	0.9396
0.1	0.3	0.8	0.4310	0.5832	0.6954	0.9736	1.0000	1.0000
0.1	0.5	0.4	0.0794	0.0810	0.1542	0.7454	0.6638	0.7186
0.1	0.5	0.6	0.1472	0.1452	0.2654	0.9316	0.9294	0.9450
0.1	0.5	0.8	0.4134	0.5306	0.6652	0.9764	1.0000	1.0000
0.1	0.6	0.2	0.1392	0.1644	0.2620	0.9318	0.9512	0.9614
0.1	0.6	0.4	0.1354	0.1370	0.2402	0.9164	0.9180	0.9318
0.1	0.6	0.6	0.1752	0.1882	0.3288	0.9604	0.9698	0.9832
0.1	0.6	0.8	0.4122	0.5172	0.6700	0.9780	1.0000	1.0000
0.1	0.7	0.2	0.2500	0.3150	0.4448	0.9682	0.9992	0.9994
0.1	0.7	0.4	0.2386	0.2756	0.4108	0.9690	0.9968	0.9962
0.1	0.7	0.6	0.2518	0.2806	0.4578	0.9716	0.9968	0.9982
0.1	0.7	0.8	0.4558	0.5534	0.7364	0.9774	1.0000	1.0000
0.3	0.3	0.2	0.0290	0.0252	0.0534	0.0846	0.0684	0.0902
0.3	0.3	0.4	0.0336	0.0254	0.0530	0.0904	0.0874	0.1048
0.3	0.3	0.6	0.0896	0.0860	0.1458	0.7282	0.7184	0.7524
0.3	0.3	0.8	0.3786	0.4610	0.5968	0.9988	1.0000	1.0000
0.3	0.5	0.4	0.0428	0.0380	0.0756	0.2446	0.2254	0.2708
0.3	0.5	0.6	0.0818	0.0768	0.1432	0.6008	0.6058	0.6564
0.3	0.5	0.8	0.3206	0.3696	0.4984	0.9978	0.9994	0.9998
0.3	0.6	0.2	0.1052	0.1076	0.1842	0.8546	0.8410	0.8650
0.3	0.6	0.4	0.0770	0.0642	0.1276	0.6194	0.6144	0.6488
0.3	0.6	0.6	0.0968	0.0910	0.1706	0.7276	0.7232	0.7886
0.3	0.6	0.8	0.3090	0.3502	0.4932	0.9974	0.9996	0.9998
0.3	0.7	0.2	0.2158	0.2444	0.3632	0.9866	0.9942	0.9954
0.3	0.7	0.4	0.1566	0.1628	0.2594	0.9426	0.9596	0.9688
0.3	0.7	0.6	0.1554	0.1592	0.2730	0.9226	0.9376	0.9526
0.3	0.7	0.8	0.3392	0.3662	0.5462	0.9984	1.0000	1.0000
0.5	0.3	0.4	0.0424	0.0384	0.0704	0.2348	0.2226	0.2598
0.5	0.3	0.6	0.0774	0.0756	0.1346	0.5986	0.5948	0.6466
0.5	0.3	0.8	0.3306	0.3816	0.5086	0.9986	0.9996	1.0000
0.5	0.5	0.4	0.0350	0.0316	0.0570	0.0936	0.0998	0.1276
0.5	0.5	0.6	0.0340	0.0334	0.0610	0.1144	0.1418	0.1690
0.5	0.5	0.8	0.2116	0.2464	0.3606	0.9730	0.9916	0.9924
0.5	0.6	0.2	0.1132	0.1152	0.1998	0.8388	0.8172	0.8516
0.5	0.6	0.4	0.0478	0.0422	0.0860	0.3062	0.3314	0.3778
0.5	0.6	0.6	0.0386	0.0342	0.0684	0.1150	0.1402	0.1756
0.5	0.6	0.8	0.1694	0.1986	0.2982	0.9350	0.9782	0.9828
0.5	0.7	0.2	0.1996	0.2070	0.3240	0.9750	0.9838	0.9854
0.5	0.7	0.4	0.1068	0.1054	0.1812	0.7638	0.8224	0.8454

Table 2. Continued

$\rho_1$	$\rho_2$	$\rho_3$	$k = 5$			$k = 30$		
			LRT	$F_{\max}$	$F_{\text{new}}$	LRT	$F_{\max}$	$F_{\text{new}}$
0.5	0.7	0.6	0.0580	0.0574	0.1142	0.4240	0.5082	0.5506
0.5	0.7	0.8	0.1648	0.1808	0.3036	0.9100	0.9668	0.9718
0.7	0.3	0.2	0.1966	0.2346	0.3370	0.9868	0.9940	0.9966
0.7	0.3	0.4	0.1582	0.1680	0.2686	0.9534	0.9690	0.9746
0.7	0.3	0.6	0.1536	0.1550	0.2734	0.9302	0.9452	0.9604
0.7	0.3	0.8	0.3616	0.3926	0.5638	0.9980	1.0000	1.0000
0.7	0.5	0.2	0.1924	0.2068	0.3222	0.9770	0.9856	0.9878
0.7	0.5	0.4	0.1104	0.1088	0.1890	0.7676	0.8230	0.8476
0.7	0.5	0.6	0.0594	0.0644	0.1152	0.4272	0.5164	0.5638
0.7	0.5	0.8	0.1578	0.1838	0.3014	0.9238	0.9696	0.9750
0.7	0.6	0.2	0.2146	0.2156	0.3686	0.9800	0.9860	0.9888
0.7	0.6	0.4	0.1088	0.1116	0.1944	0.7580	0.8078	0.8348
0.7	0.6	0.6	0.0396	0.0398	0.0732	0.1454	0.2096	0.2416
0.7	0.6	0.8	0.0852	0.0976	0.1700	0.6738	0.8080	0.8364
0.7	0.7	0.2	0.2650	0.2754	0.4550	0.9914	0.9952	0.9980
0.7	0.7	0.4	0.1354	0.1346	0.2554	0.8650	0.8924	0.9240
0.7	0.7	0.6	0.0358	0.0394	0.0742	0.1648	0.2256	0.2772
0.7	0.7	0.8	0.0386	0.0520	0.0950	0.2916	0.4322	0.4732
0.9	0.3	0.2	0.6918	0.9104	0.9448	0.9978	1.0000	1.0000
0.9	0.3	0.4	0.6914	0.8582	0.9058	0.9996	1.0000	1.0000
0.9	0.3	0.6	0.6584	0.8110	0.8774	0.9996	1.0000	1.0000
0.9	0.3	0.8	0.6780	0.7952	0.8918	0.9998	1.0000	1.0000
0.9	0.5	0.2	0.6888	0.8716	0.9154	0.9962	1.0000	1.0000
0.9	0.5	0.4	0.6394	0.7738	0.8564	1.0000	1.0000	1.0000
0.9	0.5	0.6	0.5408	0.6740	0.7706	1.0000	1.0000	1.0000
0.9	0.5	0.8	0.5152	0.6016	0.7428	0.9998	1.0000	1.0000
0.9	0.6	0.2	0.6806	0.8504	0.9114	0.9972	1.0000	1.0000
0.9	0.6	0.4	0.6200	0.7356	0.8200	1.0000	1.0000	1.0000
0.9	0.6	0.6	0.4552	0.5694	0.6916	0.9996	1.0000	1.0000
0.9	0.6	0.8	0.3686	0.4440	0.5958	0.9964	0.9998	0.9998
0.9	0.7	0.2	0.6782	0.8318	0.8934	0.9974	1.0000	1.0000
0.9	0.7	0.4	0.5848	0.7014	0.8020	1.0000	1.0000	1.0000
0.9	0.7	0.6	0.3790	0.4932	0.6234	0.9980	1.0000	1.0000
0.9	0.7	0.8	0.2014	0.2706	0.3984	0.9666	0.9970	0.9974

### Example with Real Life Data

In this section, three tests using real life data collected from Srivastava and Katapa (1986) are prepared. The data is split randomly into three samples. Table 3 gives the values of pattern intensity on soles of feet in fourteen families, where values for daughters and sons are combined.

**Table 3.** Values of pattern intensity on soles of feet in 14 families

	Family #	Mother	Father	# Siblings	Siblings Values
Sample A	2	2	3	2	2, 3
	7	4	3	7	2, 2, 3, 6, 3, 5, 4
	8	3	7	7	2, 4, 7, 4, 4, 7, 8
	11	5	6	4	5, 3, 4, 4
	14	2	3	3	2, 2, 2
Sample B	1	2	3	2	2, 2
	5	2	3	2	6, 6
	6	4	3	3	4, 3, 3
	9	5	5	2	5, 6
	13	6	3	4	4, 3, 3, 3
Sample C	3	2	3	3	2, 2, 2
	4	2	4	5	2, 2, 2, 2, 2
	10	5	4	3	4, 5, 4
	12	2	4	2	2, 4

First ignoring the father's and mother's values, transform the siblings' values by multiplying each observation vector by Helmert's orthogonal matrix  $Q$

$$\text{where where } Q = \begin{bmatrix} \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \frac{1}{\sqrt{p_i}} & \cdots & \frac{1}{\sqrt{p_i}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \frac{1}{\sqrt{p_i(p_i-1)}} & \cdots & \frac{(p_i-1)}{\sqrt{p_i(p_i-1)}} \end{bmatrix}$$

This gives transformed vectors  $\tilde{u}_i, \tilde{v}_j$  and  $\tilde{w}_l$  respectively for  $i=1,2,\dots,k_1$  ;  $j=1,2,\dots,k_2$  and  $l=1,2,\dots,k_3$ . Here,  $k_1=5, k_2=5$  and  $k_3=4$ .

Srivastava's formula, given by (3), is used to compute intraclass correlation coefficients. The computed values of intraclass correlation coefficients are  $\hat{\rho}_1 = 0.5895$ ,  $\hat{\rho}_2 = 0.9159$  and  $\hat{\rho}_3 = 0.7685$  and  $\hat{\rho} = 0.4923$ .

**Table 4.** Raw Computations

				$i$	$i^* \alpha / 6$	$P(i)$	Col3 < Col2?
F1	9.2874	P1	0.000014	1	0.008333	0.000014	yes
F2	0.1355	P5	0.003139	2	0.016667	0.003139	yes
F3	0.1640	P6	0.003859	3	0.025000	0.003859	yes
F4	0.1077	P3	0.996140	4	0.033333	0.996140	no
F5	7.3798	P2	0.996860	5	0.041667	0.996860	no
F6	6.0994	P4	0.999990	6	0.050000	0.999990	no

**Note.** Conclusion = Reject.

The computed values of LRT statistic and  $F_{\max}$  statistic obtained from formula (8) and (9) respectively are as follows:

**Table 5.** Test Statistics and their Critical Values

	Test Statistic	CV ( $\alpha = 0.01$ )	CV ( $\alpha = 0.05$ )	CV ( $\alpha = 0.10$ )
LRT	7.7820	9.2103	5.9915	4.6052
$F_{\max}$	10.4510	10.1660	6.2630	5.0025
$F_{\text{new}}$	N/A	Reject	Reject	Reject

## References

Bhandary, M. and Alam, M. K. (2000). Test for the equality of intraclass correlation coefficients under unequal family sizes. *Communications in Statistics – Theory and Methods*, 29(4): 755-768.

Bhandary, M. and Fujiwara, K. (2006). A small sample test for the equality of intraclass correlation coefficients under unequal family sizes for several populations. *Communications in Statistics – Simulation and Computation*, 35(3): 765-778.

## A TEST FOR EQUALITY OF CORRELATION COEFFICIENTS

Donner, A. and Bull, S. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics*, 39: 771-775.

Donner, A. and Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *The Statistician*, 51(3): 367-379.

Gokhale, D.V. and SenGupta, A. (1986). Optimal tests for the correlation coefficient in a symmetric multivariate normal population. *J. Statist. Plann. Inference*, 14: 263-268.

Huang, W. and Sinha, B. K. (1993). On optimum invariant tests of equality of intraclass correlation coefficients. *Annals of the Institute of Statistical Mathematics*, 45(3): 579-597.

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, 37: 87-94.

Konishi, S. and Gupta, A. K. (1989). Testing the equality of several intraclass correlation coefficients. *J. Statist. Plann. Inference*, 21: 93-105.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Rosner, B., Donner, A. and Hennekens, C. H. (1977). Estimation of intraclass correlation from familial data. *Applied Statistics*, 26: 179-187.

Rosner, B., Donner, A. and Hennekens, C. H. (1979). Significance testing of interclass correlations from familial data. *Biometrics*, 35: 461-471.

SenGupta, A. (1988). On loss of power under additional information – an example. *Scand. J. Statist.*, 15: 25-31.

Scheffe, H. (1959). *The Analysis of Variance*. New York: Wiley.

Srivastava, M. S. (1984). Estimation of interclass correlations in familial data. *Biometrika*, 71: 177-185.

Srivastava, M. S. and Katapa, R. S. (1986). Comparison of estimators of interclass and intraclass correlations from familial data. *Canadian Journal of Statistics*, 14: 29-42.

Srivastava, M. S. and Keen, K. J. (1988). Estimation of the interclass correlation coefficient. *Biometrika*, 75: 731-739.

Velu, R. and Rao, M. B. (1990). Estimation of parent-offspring correlation. *Biometrika*, 77(3): 557-562.

Young, D. and Bhandary, M. (1998). Test for the equality of intraclass correlation coefficients under unequal family sizes. *Biometrics*, 54(4): 1363-1373.



## On the Gamma-Logistic Distribution

**Ayman Alzaatreh**

Austin Peay State University  
Clarksville, TN

**Indranil Ghosh**

Austin Peay State University  
Clarksville, TN

**Hassan Said**

Austin Peay State University  
Clarksville, TN

---

A new generalization of the logistic distribution is defined and studied, namely, the gamma-logistic distribution. Various properties of the gamma-logistic are obtained. The structural analysis of the distribution includes moments, mode, quantiles, skewness, kurtosis, Shannon's entropy and order statistics. The method of maximum likelihood estimation is proposed for estimating the model parameters. For illustrative purposes, a real data set is analyzed as an application of the gamma-logistic distribution.

*Keywords:* gamma-logistic distribution,  $T-X$  family of distributions, Shannon's entropy, reliability parameter, order statistics

---

### Introduction

The armory of statistical distributions is truly illimitable. New distributions are being unearthed literally on a weekly basis elicited by either theoretical considerations or by pressing practical applications or both. A new class of mixtures of two absolutely continuous distributions is investigated in this article, with the primary objective of exploring its enhanced flexibility in modeling skewed data.

The simplicity of the logistic distribution and its importance as a growth curve has attracted many researchers to study this distribution. Also, the limitation of the shape of the logistic distribution merits further investigation to various other different types of generalized logistic distribution. Many other generalized distributions obtained from the logistic distribution were introduced in the literature to study its skewness properties and to examine its flexibility in modeling skewed data. It is noteworthy to mention that other generalizations of logistic distribution exists in the literature such as Types I, II, III and IV generalized logistic distributions (Johnson et al., 1994). Proposed here is a new

---

*Drs. Alzaatreh and Ghosh are both Assistant Professors in the Department of Mathematics and Statistics. Email Dr. Alzaatreh at: [alzaatreh@apsu.edu](mailto:alzaatreh@apsu.edu). Dr. Hassaan is an Associate Professor in the College of Business.*

distribution, mixing two absolutely continuous distributions, Gamma and logistic, following the idea discussed in Alzaatreh, et al. (2013b).

Let  $F(x)$  be the cumulative distribution function (CDF) of any random variable  $X$  and  $r(t)$  be the probability density function (PDF) of a random variable  $T$  defined on  $[0, \infty)$ . The CDF of the  $T$ - $X$  family of distributions defined by Alzaatreh, et al. (2013b) is given by

$$G(x) = \int_0^{-\log(1-F(x))} r(t) dt. \quad (1)$$

When  $X$  is a continuous random variable, the probability density function of the  $T$ - $X$  family is

$$g(x) = \frac{f(x)}{1-F(x)} r(-\log(1-F(x))) \quad (2)$$

Recently, many generalized distributions have been generated from the  $T$ - $X$  family of distributions such as Weibull-Pareto distribution (Alzaatreh et al., 2013a), gamma-Pareto distribution (Alzaatreh et al., 2012a), gamma-Pareto IV (Alzaatreh and Ghosh, 2013), gamma-half normal distribution (Alzaatreh and Knight, 2013) and the exponentiated-exponential geometric distributions (Alzaatreh et al., 2012b). For more information about methods for generating univariate continuous distributions, one may refer to Lee et al. (2013).

If a random variable  $T$  follows the gamma distribution with parameters  $\alpha$  and  $\beta$ ,  $r(t) = (\beta^\alpha \Gamma(\alpha))^{-1} t^{\alpha-1} e^{-t/\beta}$ ,  $t \geq 0$ , the definition in (2) leads to the gamma- $X$  family with the PDF

$$g(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} f(x) (-\log(1-F(x)))^{\alpha-1} (1-F(x))^{\frac{1}{\beta}-1}. \quad (3)$$

If  $X$  follows a logistic distribution with parameters  $\theta$ ,  $F(x) = 1 - (1 + e^{x/\theta})^{-1}$ ,  $x \in \mathbb{R}$ , then (3) reduces to

$$g(x) = \frac{1}{\theta\beta^\alpha\Gamma(\alpha)} e^{x/\theta} (\log(1 + e^{x/\theta}))^{\alpha-1} (1 + e^{x/\theta})^{-1-1/\beta}, \quad x \in \mathbb{R}; \alpha, \beta, \theta > 0. \quad (4)$$

Note that when  $\alpha = \beta = 1$ , the PDF in (4) reduces to the logistic distribution. When  $\beta = 1$  and  $\alpha = n \in \mathbb{N}$ , the PDF in (4) reduces to the density function of the upper record values,  $X_{U(n)}$ , arising from a sequence  $\{X_i\}_{i=1}^n$  of identically independent logisite random variables (Johnson, et al., 1994, pp. 135). From (4), the CDF of the gamma-logistic distribution can be written as

$$G(x) = \frac{1}{\Gamma(\alpha)} \gamma(\alpha, \beta^{-1} \log(1 + e^{x/\theta})), \quad (5)$$

where  $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$  is the lower incomplete gamma function.

### Some properties of the gamma-logistic distribution

The following lemma provides a characterization of the gamma-logistic distribution which establishes the relation between gamma-logistic and gamma distributions.

#### Lemma 1 (Transformation)

If a random variable  $X$  follows the gamma distribution with parameters  $\alpha$  and  $\beta$ , then  $Y = \theta \log(e^X - 1)$  follows gamma-logistic distribution with parameters  $\alpha$ ,  $\beta$  and  $\theta$ .

**Proof:** The result follows by using the transformation technique.  $\square$

#### Lemma 2 (Mode)

The mode of the gamma-logistic distribution is the solution of the equation  $k(x) = 0$ , where  $k(x) = \log(1 + e^{x/\theta})(1 - \beta^{-1} e^{x/\theta}) + (\alpha - 1)e^{x/\theta}$ .

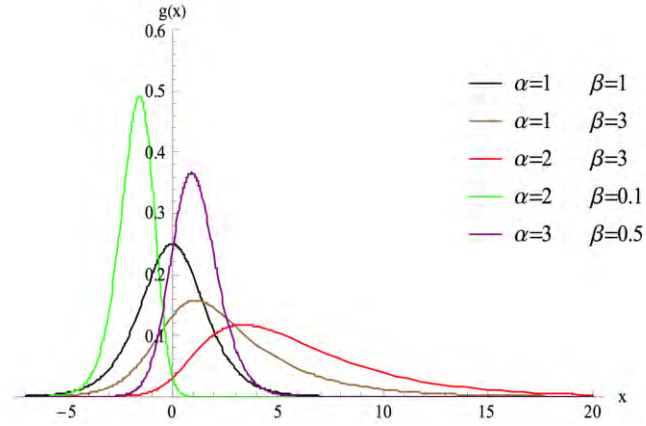
**Proof:** Setting  $g'(x) = 0$  is equivalent to

$$\theta^{-2} \beta^{-\alpha} (\Gamma(\alpha))^{-1} e^{x/\theta} (\log(1 + e^{x/\theta}))^{\alpha-2} (1 + e^{x/\theta})^{-2-1/\beta} k(x) = 0, \text{ where}$$

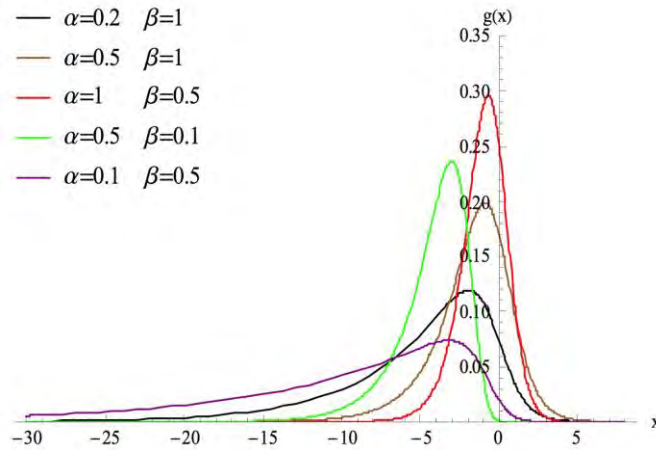
$k(x) = \log(1 + e^{x/\theta})(1 - \beta^{-1} e^{x/\theta}) + (\alpha - 1)e^{x/\theta}$ . Hence, the critical values of  $g(x)$  is the solution of  $k(x) = 0$ .  $\square$

## ON THE GAMMA-LOGISTIC DISTRIBUTION

In Figures 1 and 2, various graphs of  $g(x)$  are provided for different parameter values. The plots indicate that the gamma-logistic distribution is unimodal and can be symmetric, right-skewed or left-skewed.



**Figure 1.** Graphs of the gamma-logistic PDF for various choices of  $\alpha$  and  $\beta$  when  $\theta = 1$ .



**Figure 2.** Graphs of the gamma-logistic PDF for various choices of  $\alpha$  and  $\beta$  when  $\theta = 1$ .

**Lemma 3**

Let  $Q(\lambda)$ ,  $0 < \lambda < 1$ , denotes the quantile function for the gamma-logistic distribution. Then  $Q(\lambda)$  can be written as

$$Q(\lambda) = \theta \log \left( e^{\beta \gamma^{-1}(\alpha, \lambda \Gamma(\alpha))} - 1 \right). \quad (6)$$

**Proof:** The result follows by using  $G(Q(\lambda)) = \lambda$  in (5) and then solving it for  $Q(\lambda)$ .  $\square$

The Shannon's entropy (Shannon, 1948) plays an important role in information theory and it is used as a measure of uncertainty. Shannon's entropy for a random variable  $X$  with PDF  $g(x)$  is defined as  $E[-\log(g(X))]$ . According to Alzaatreh, et al. (2013b), the Shannon's entropy for the gamma- $X$  family can be written as

$$-E \left[ \log \left( f(F^{-1}(1 - e^{-T})) \right) \right] + \alpha(1 - \beta) + \log \beta + \log(\Gamma(\alpha)) + (1 - \alpha)\psi(\alpha), \quad (7)$$

where  $F(\cdot)$  and  $f(\cdot)$  are the CDF and PDF of the Transformer family respectively, and  $T$  follows gamma distribution with parameters  $\alpha$  and  $\beta$ , and  $\psi(\cdot)$  is the digamma function.

The following theorem defines expression for the Shannon's entropy for the gamma-logistic distribution.

**Theorem 1**

The Shannon's entropy for the random variable  $X$  which follows a gamma-logistic distribution is given by

$$\eta_x = \alpha + \log(\beta\theta) + \log \Gamma(\alpha) + (1 - \alpha)\psi(\alpha) + \sum_{k=1}^{\infty} \frac{1}{k(1 + \beta k)^{\alpha}}. \quad (8)$$

**Proof:** In this case,  $F(X) = 1 - (1 + e^{X/\theta})^{-1}$ . So that

$$\log \left\{ f(F^{-1}(1 - e^{-T})) \right\} = \log(e^T - 1) - 2T - \log \theta = \log(1 - e^{-T}) - T - \log \theta, \quad (9)$$

where  $T$  follows the gamma distribution with parameters  $\alpha$  and  $\beta$ . Now, consider  $E(\log(1-e^{-T}))$ . Using the Taylor's series expansion of  $\log(1-e^{-T})$ , one can get

$$\begin{aligned} E(\log(1-e^{-T})) &= -\sum_{k=1}^{\infty} k^{-1} E(e^{-kT}) = -\frac{1}{\beta^{\alpha} \Gamma(\alpha)} \sum_{k=1}^{\infty} k^{-1} \int_0^{\infty} e^{-kt} e^{t/\beta} t^{\alpha-1} dt \\ &= -\sum_{k=1}^{\infty} \frac{1}{k(1+\beta k)^{\alpha}}, \end{aligned}$$

$$\text{Hence, } -E\left[\log\left(f(F^{-1}(1-e^{-T}))\right)\right] = \sum_{k=1}^{\infty} \frac{1}{k(1+\beta k)^{\alpha}} + E(T) + \log \theta. \quad (10)$$

The result in (8) follows by using the fact that  $E(T) = \alpha\beta$  and substituting (10) in (7).  $\square$

## Moments and mean deviations

The moment generating function for the gamma-logistic distribution is given by

$$M_X(t) = E(e^{tX}) = \frac{1}{\beta^{\alpha} \Gamma(\alpha)} \int_{-\infty}^{\infty} e^{x(t+1/\theta)} \left(\log(1+e^{x/\theta})\right)^{\alpha-1} \left(1+e^{x/\theta}\right)^{-1-1/\beta} dx. \quad (11)$$

On using the substitution  $u = \log(1+e^{x/\theta})$ , (11) can be written as

$$M_X(t) = \frac{1}{\beta^{\alpha} \Gamma(\alpha)} \int_0^{\infty} (e^u - 1)^{t\theta} u^{\alpha-1} e^{-u/\beta} du.$$

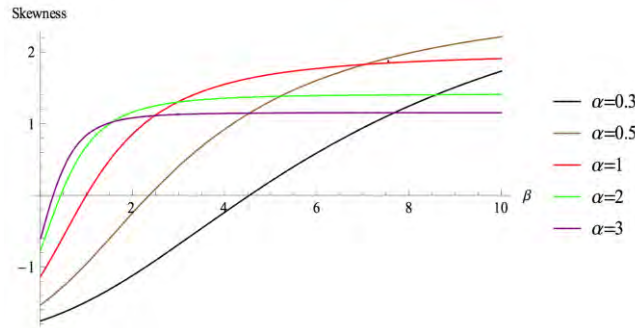
On writing  $(e^u - 1)^{t\theta} = e^{tu\theta} (1 - e^{-u})^{t\theta}$  and using the generalized binomial expression of  $(1 - e^{-u})^{t\theta} = \sum_{k=0}^{\infty} \frac{(-1)^k (t\theta)_{(k)}}{k!} e^{-ku}$ , one can get

$$M_X(t) = \sum_{k=0}^{\infty} \frac{(-1)^k (t\theta)_{(k)}}{k! (\beta(k - t\theta) + 1)^{\alpha}}, \quad (12)$$

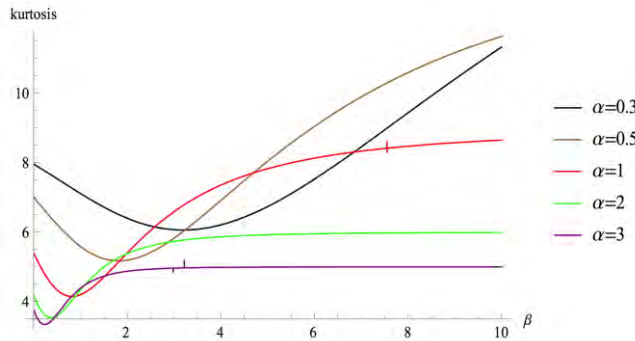
provided that  $t < \theta^{-1}$ , and  $(t\theta)_{(k)} = t\theta(t\theta - 1) \cdots (t\theta - k + 1)$ .

A series expression for the  $r^{\text{th}}$  moments of the gamma-logistic distribution can be obtained by using the fact that  $E(X^r) = \frac{d^r}{dt^r} M_X(t) \big|_{t=0}$ .

Skewness and kurtosis of a distribution can be measured by  $\beta_1 = \mu_3 / \sigma^3$  and  $\beta_2 = \mu_4 / \sigma^4$ , respectively. When the distribution is symmetric,  $\beta_1 = 0$  and when the distribution is right (or left) skew,  $\beta_1 > 0$  (or  $\beta_1 < 0$ ). As  $\beta_2$  increases the tail of the distribution becomes heavier. To investigate the effect of the two shape parameters  $\alpha$  and  $\beta$  on the gamma-logistic distribution,  $\beta_1$  and  $\beta_2$  were computed for different values of  $\alpha$  and  $\beta$ . Figures 3 and 4 display the skewness and the kurtosis for the gamma-logistic distribution when  $\theta = 1$ . From Figure 3, the gamma-logistic distribution can be left skewed, right skewed or symmetric. Also, for fixed value of  $\alpha$ , the skewness is an increasing function of  $\beta$ . As shown in Figure 4, as  $\alpha$  gets smaller, the kurtosis of the gamma-logistic distribution increases rapidly as  $\beta$  increases.



**Figure 3.** Skewness graph for gamma-logistic distribution when  $\theta = 1$ .



**Figure 4.** Kurtosis graph for gamma-logistic distribution when  $\theta = 1$ .

## ON THE GAMMA-LOGISTIC DISTRIBUTION

The deviation from the mean and the deviation from the median are used to measure the dispersion and the spread in a population. If we denote the median by  $M$ , then the mean deviation from the mean,  $D(\mu)$ , and the mean deviation from the median,  $D(M)$ , can be written as

$$D(\mu) = E |X - \mu| = 2\mu G(\mu) - 2 \int_{-\infty}^{\mu} xg(x)dx. \quad (13)$$

$$D(M) = E |X - M| = \mu - 2 \int_{-\infty}^M xg(x)dx. \quad (14)$$

Now, consider

$$\begin{aligned} I_m &= \int_{-\infty}^m xg(x)dx \\ &= \frac{1}{\theta\beta^\alpha\Gamma(\alpha)} \int_{-\infty}^{\infty} xe^{x/\theta} \left(\log(1+e^{x/\theta})\right)^{\alpha-1} \left(1+e^{x/\theta}\right)^{-1-1/\beta} dx \end{aligned} \quad (15)$$

Using the substitution  $u = \log(1+e^{x/\theta})$  in (15) results in

$$I_m = \frac{\theta}{\beta^\alpha\Gamma(\alpha)} \left( \int_0^{\log(1+e^{m/\theta})} u^\alpha e^{-u/\beta} du + \int_0^{\log(1+e^{m/\theta})} \log(1-e^{-u}) u^{\alpha-1} e^{-u/\beta} du \right).$$

Using the Taylor series expansion of  $\log(1-e^{-u})$  results in

$$I_m = \frac{\theta}{\Gamma(\alpha)} \left\{ \gamma\left(\alpha+1, \beta^{-1} \log(1+e^{m/\theta})\right) - \sum_{k=1}^{\infty} \frac{\beta^{k+1}}{k} \gamma\left(\alpha+k, \beta^{-1} \log(1+e^{m/\theta})\right) \right\}. \quad (16)$$

Using equations (5) and (16), the mean deviation from the mean and the mean deviation from the median are

$$D(\mu) = 2\mu \frac{\gamma\left(\alpha, \beta^{-1} \log(1+e^{\mu/\theta})\right)}{\Gamma(\alpha)} - 2I_\mu \text{ and } D(M) = \mu - 2I_M.$$



## Mean residual life function and Reliability parameter

Let  $X$  be a random variable with cumulative distribution function CDF  $F$  such that  $E(X) < \infty$ . The mean residual life (MRL) function  $\xi(x)$  of  $X$  is defined by  $\xi(x) = E(X - x | X > x)$ . The MRL function, also known as expected remaining life function or mean excess function, has been extensively studied in lifetime variables context. It plays a major role in many fields such as industrial reliability, life insurance and biomedical science. For more information about the MRL function, see Kotz and Shanbhag (1980), Hall and Wellner (1979) and Guess and Proschan (1985). The next theorem demonstrates the expression of the MRL function for the Gamma-logistic distribution.

### Theorem 2

Let  $X$  be a random variable which follows the gamma-logistic distribution with parameters  $\alpha$ ,  $\beta$  and  $\theta$ . Then the MRL function is given by

$$\xi(x) = \frac{\theta}{\Gamma(\alpha)} \left\{ \beta \Gamma(\alpha + 1, \beta^{-1} \log(1 + e^{x/\theta})) - \sum_{k=1}^{\infty} \frac{\Gamma(\alpha, (\beta^{-1} + k) \log(1 + e^{x/\theta}))}{k(\beta k + 1)^{\alpha}} \right\} - x, \quad (17)$$

where  $\Gamma(x, a) = \int_x^{\infty} t^{a-1} e^{-t} dt$  is the upper incomplete gamma function.

**Proof:** From (4),

$$E(X | X > x) = \frac{1}{\theta \beta^{\alpha} \Gamma(\alpha)} \int_x^{\infty} t e^{t/\theta} (\log(1 + e^{t/\theta}))^{\alpha-1} (1 + e^{t/\theta})^{-1-1/\beta} dt. \quad (18)$$

On using the substitution  $u = \log(1 + e^{t/\theta})$ , (18) reduces to

$$\begin{aligned} E(X | X > x) &= \frac{\theta}{\beta^{\alpha} \Gamma(\alpha)} \int_{\log(1 + e^{x/\theta})}^{\infty} u^{\alpha-1} e^{-u/\beta} \log(e^u - 1) du \\ &= \frac{\theta}{\beta^{\alpha} \Gamma(\alpha)} \int_{\log(1 + e^{x/\theta})}^{\infty} u^{\alpha-1} e^{-u/\beta} \left( u - \sum_{k=1}^{\infty} e^{-ku} / k \right) du \end{aligned} \quad (19)$$

The result follows from equation (19). □

The reliability parameter  $R$  is defined as  $R = P(X > Y)$ , where  $X$  and  $Y$  are independent random variables. Many applications of the reliability parameter have appeared in the literature such as the area of classical stress-strength model and the breakdown of a system having two components. More applications of the reliability parameter can be found in Hall (1984) and Weerahandi and Johnson (1992). If  $X$  and  $Y$  are two continuous independent random variables with CDFs  $F_1(x)$  and  $F_2(y)$  and their PDFs  $f_1(x)$  and  $f_2(y)$  respectively. Then the reliability parameter  $R$  can be written as  $R = P(X > Y) = \int_{-\infty}^{\infty} F_2(t)f_1(t)dt$ .

The following theorem provides an expression for the reliability parameter  $R$  where the parameter  $\theta$  is fixed.

### Theorem 3

Suppose that  $X$  and  $Y$  are two independent gamma-logistic random variables with parameters  $(\alpha_1, \beta_1, \theta)$  and  $(\alpha_2, \beta_2, \theta)$  respectively. Then

$$P(X > Y) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \sum_{k=0}^{\infty} \left( \frac{\beta_1}{\beta_2} \right)^{\alpha_2+k} \frac{(-1)^k \Gamma(\alpha_1 + \alpha_2 + k)}{k!(k + \alpha_2)}. \quad (20)$$

**Proof:** From (4) and (5),

$$P(X > Y) = \frac{1}{\theta \beta_1^{\alpha_1} \Gamma(\alpha_1) \Gamma(\alpha_2)} \int_{-\infty}^{\infty} \gamma(\alpha_2, \beta_2^{-1} \log(1 + e^{t/\theta})) e^{t/\theta} (1 + e^{t/\theta})^{-1-1/\beta_1} (\log(1 + e^{t/\theta}))^{\alpha_1-1} dt. \quad (21)$$

On using the following series expansion from Nadarajah and Pal (2008),

$$\gamma(\alpha, x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{k+\alpha}}{k!(k + \alpha)} \quad (22)$$

and then the substitution  $w = \log(1 + e^{t/\theta})$ , equation (21) can be written as

$$P(X > Y) = \frac{1}{\beta_1^{\alpha_1} \Gamma(\alpha_1) \Gamma(\alpha_2)} \sum_{k=0}^{\infty} \frac{(-1)^k \beta_2^{-\alpha_2-k}}{k!(k + \alpha_2)} \int_0^{\infty} w^{\alpha_1+\alpha_2+k-1} e^{-w/\beta_1} dw \quad (23)$$

Equation (23) reduces to (20). □

## Order Statistics for the gamma-logistic distribution

Consider the general  $r^{\text{th}}$  order statistic and the asymptotic distributions of the sample minimum and maximum when a random sample of size  $n$  is drawn from the gamma-logistic. From (5), the density function of the  $r^{\text{th}}$  order statistic,  $X_{r:n}$ , is given by

$$\begin{aligned} f_{X_{r:n}}(x) &= \frac{1}{B(r, n-r+1)} (G(x))^{r-1} (1-G(x))^{n-r} g(x) \\ &= \frac{1}{B(r, n-r+1)} g(x) \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \left\{ \frac{\gamma(\alpha, \beta^{-1} \log(1+e^{x/\theta}))}{\Gamma(\alpha)} \right\}^{r+j-1}, \quad x \in \mathbb{R}. \end{aligned}$$

On using (22), the PDF of  $X_{r:n}$  can be written as

$$\begin{aligned} f_{r:n}(x) &= \frac{1}{B(r, n-r+1)} g(x) \sum_{j=0}^{n-r} (-1)^j \binom{n-r}{j} \left\{ \sum_{k=0}^{\infty} \frac{(-1)^k}{\Gamma(\alpha) \beta^{k+\alpha}} \frac{(\log(1+e^{x/\theta}))^{k+\alpha}}{k!(k+\alpha)} \right\}^{r+j-1} \\ &= \frac{1}{B(r, n-r+1)} g(x) \sum_{j=0}^{n-r} \sum_{k_1=0}^{\infty} \cdots \sum_{k_{r+j-1}=0}^{\infty} (-1)^{j+s_k} \binom{n-r}{j} \frac{(\log(1+e^{x/\theta}))^{s_k+(r+j-1)\alpha}}{(\Gamma(\alpha))^{r+j-1} \beta^{s_k+(r+j-1)\alpha} p_k} \\ &= \frac{1}{B(r, n-r+1)} \sum_{j=0}^{n-r} \sum_{k_1=0}^{\infty} \cdots \sum_{k_{r+j-1}=0}^{\infty} (-1)^{j+s_k} \binom{n-r}{j} \frac{\Gamma(s_k+(r+j)\alpha)}{(\Gamma(\alpha))^{r+j} p_k} g(x | s_k+(r+j)\alpha, \beta, \theta), \quad (24) \end{aligned}$$

where  $s_k = \sum_{i=1}^{r+j-1} k_i$  and  $p_k = \prod_{i=1}^{r+j-1} k_i! (k_i + \alpha)$ .

From (24), note that the PDF of the  $r^{\text{th}}$  order statistic  $X_{r:n}$  can be expressed as infinite sums of the gamma-logistic PDFs.

To study the asymptotic distributions of the sample minimum  $X_{1:n}$  and the sample maximum  $X_{n:n}$ , use Theorem 8.3.6 of Arnold et al. (2008) as follows: Since  $G^{-1}(0) = -\infty$ , it follows from the Theorem that the asymptotic distribution of the sample minimum  $X_{1:n}$  will be of the Weibull type with parameter  $\delta > 0$  if

$\lim_{\varepsilon \rightarrow 0^+} \frac{G(\varepsilon x)}{G(\varepsilon)} = x^\delta$ , for all  $x > 0$ . By using the L'Hospital's rule, it can be easily

shown that  $\lim_{\varepsilon \rightarrow 0_+} \frac{G(\varepsilon x)}{G(\varepsilon)} = x \lim_{\varepsilon \rightarrow 0_+} \frac{g(\varepsilon x)}{g(\varepsilon)} = x^\alpha$ . Hence, the asymptotic distribution of  $X_{1:n}$  is of the Weibull type with shape parameter  $\alpha$ . The asymptotic distribution of the sample maximum  $X_{n:n}$  can be viewed as  $G_n(x)$  where  $G_n(x) = 1 - G_1(-x)$ , where  $G_1(x)$  is the CDF of  $X_{1:n}$ .

## Maximum likelihood estimation

Let a random sample of size  $n$  be taken from the gamma-logistic distribution. The log-likelihood function for the gamma-logistic distribution in (4) is given by

$$\begin{aligned} \ell = -n \log \theta - n \log \Gamma(\alpha) - n\alpha \log \beta + \frac{n\bar{X}}{\theta} & - (1 + 1/\beta) \sum_{i=1}^n \log(1 + e^{X_i/\theta}) \\ & + (\alpha - 1) \sum_{i=1}^n \log(\log(1 + e^{X_i/\theta})) \end{aligned} \quad (25)$$

The derivatives of (25) with respect to  $\alpha$ ,  $\beta$  and  $\theta$  are given by

$$\frac{\partial \ell}{\partial \alpha} = -n\psi(\alpha) - n \log \beta + \sum_{i=1}^n \log(\log(1 + e^{X_i/\theta})). \quad (26)$$

$$\frac{\partial \ell}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n \log(1 + e^{X_i/\theta}). \quad (27)$$

$$\frac{\partial \ell}{\partial \theta} = -\frac{n}{\theta} - \frac{n\bar{X}}{\theta^2} - \frac{1}{\theta^2} \sum_{i=1}^n \frac{X_i e^{X_i/\theta}}{(1 + e^{X_i/\theta})} \left( \frac{\alpha - 1}{\log(1 + e^{X_i/\theta})} + \frac{1}{\beta} + 1 \right). \quad (28)$$

The MLE of  $\hat{\alpha}$ ,  $\hat{\beta}$  and  $\hat{\theta}$  are obtained by setting (26), (27) and (28) to zero and solving them iteratively.

## Application

The gamma-logistic is applied to a data set from Brinbaum and Saunders (1969). The data set represents the fatigue life of 6061-T6 aluminum coupons cut parallel

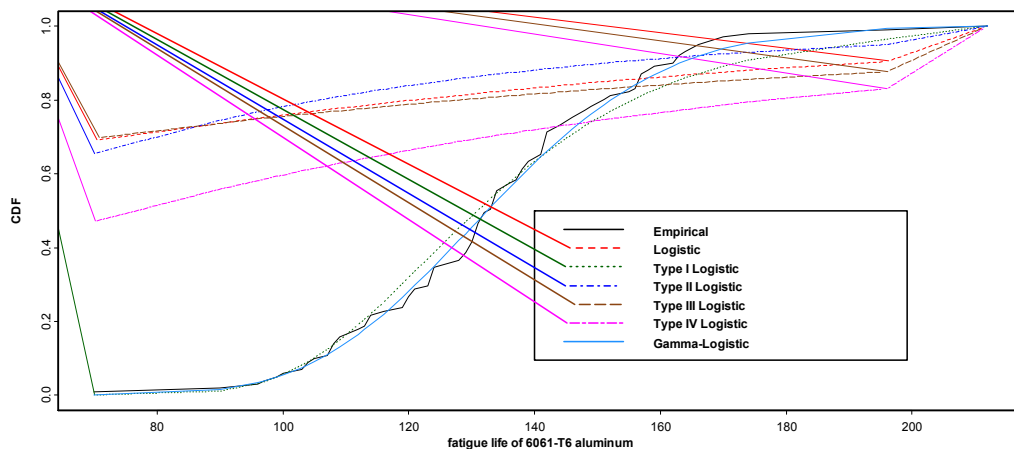
with the direction of rolling and oscillated at 18 cycles per second. The data set was fitted in Table 1 and the result compared with the logistic, Type I logistic, Type II logistic, Type III logistic and Type IV logistic distributions (Johnson et al., 1994). The maximum likelihood estimates, the log-likelihood value, the AIC (Akaike Information Criterion), the Kolmogorov-Smirnov test statistic (K-S) and the p-value for the K-S statistics for the fitted distributions are reported in Table 2. The results from Table 2 indicate that the gamma-logistic distribution provides the best fit among the distributions. Also, the K-S p-values indicate that only gamma-logistic and Type I logistic distributions provide an adequate fit to the data. The empirical and the fitted cumulative distribution functions are displayed in Figure 5. This figure supports the results in Table 2.

**Table 1.** Fatigue Life of 6061-T6 Aluminum

70	90	96	97	99	100	103	104	104	105
107	108	108	108	109	109	112	112	113	114
114	114	116	119	120	120	120	121	121	123
124	124	124	124	124	128	128	129	129	130
130	130	131	131	131	131	131	132	132	132
133	134	134	134	134	134	136	136	137	138
138	138	139	139	141	141	142	142	142	142
142	142	144	144	145	146	148	148	149	151
151	152	155	156	157	157	157	157	158	159
162	163	163	164	166	166	168	170	174	196
212									

**Table 2.** Parameter estimates for the fatigue life of 6061-T6 aluminum coupons data

Distribution	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\theta}$	Log likelihood	AIC	KS	K-S p-value
Logistic	-	-	87.1321	-646.5934	1295.1868	0.7209	0.0000
Logistic I	288.2700	-	21.7088	-462.3355	928.6710	0.0868	0.4323
Logistic II	0.0023	-	0.3021	-595.4801	1194.9602	0.7385	0.0000
Logistic III	0.0000	-	0.0010	-665.4880	1334.9760	0.7189	0.0000
Logistic IV	0.9869	0.0000	0.0022	-595.4802	1196.9604	0.5525	0.0000
Gamma-Logistic	35.6785	1.9360	1.9360	-456.3280	918.6560	0.0530	0.9388



**Figure 5.** CDF for fitted distributions for fatigue life of 6061-T6 aluminum data

## Conclusion

In this article, a special case of the gamma- $X$  family of distributions proposed by Alzaatreh, et al. (2013b), the gamma-logistic distribution is defined and studied. Various properties of the gamma-logistic distribution are studied, including moments, deviations from the mean and median, reliability parameter, Shannon entropy and order statistics. It is observed from figures 3 and 4 that the gamma-logistic distribution exhibits a wide variety of skewness and kurtosis values and the distribution can be symmetric, right skewed or left skewed. A real data set is fitted to the gamma-logistic distribution and compared with other known distributions. The results show that the gamma-logistic distribution provides the best fit among the distributions.

## References

Alzaatreh, A., Famoye, F. & Lee, C. (2013a). Weibull-Pareto distribution and its applications. *Communications in Statistics: Theory & Methods*, 42(9): 1673-1691.

- Alzaatreh, A., Famoye, F. & Lee, C. (2012a). Gamma-Pareto distribution and its applications. *Journal of Modern Applied Statistical Methods*, 11(1): 78-94.
- Alzaatreh, A. & Ghosh, I. (2013). A study of the gamma-Pareto(IV) distribution and its applications. Revised and resubmitted to *Communications in Statistics: Theory & Methods*.
- Alzaatreh, A. & Knight, K. (2013). On the gamma-half normal distribution and its applications. *Journal of Modern Applied Statistical Methods*, 12(1): 103-119.
- Alzaatreh, A., Lee, C. & Famoye, F. (2013b). A new method for generating families of continuous distribution. *Metron: International Journal of Statistics*, 71(1): 63-79.
- Alzaatreh, A., Lee, C. & Famoye, F. (2012b). On the discrete analogues of continuous distributions. *Statistical Methodology*, 9(6): 589-603.
- Arnold, B. C., Balakrishnan, N. & Nagarajah, H. N. (2008): *A first course in Order Statistics*. New York: John Wiley.
- Birirnaum, Z. W. and Saunders, S. C. (1969). A new family of life distributions. *Journal of Applied Probability*, 6: 319-327.
- Guess, F. & Proschan, F. (1985). *Mean residual life: theory and applications (No. FSU-STATISTICS-M702)*. Tallahassee, FL: Florida State University Tallahassee Dept. of Statistics.
- Hall, I. J. (1984). Approximate one-sided tolerance limits for the difference or sum of two independent normal variates. *Journal of Qualitative Technology*, 16: 15-19.
- Hall, I. J., & Wellner, J. A. (1979). *Estimation of mean residual life*. University of Rochester Department of Statistics Technical Report.
- Johnson N. L., Kotz, S., and Balakrishnan N. (1994). *Continuous Univariate Distributions, Volume 2*. New York: John Wiley.
- Kotz, S., & Shanbhag, D. N. (1980). Some new approaches to probability distributions. *Advances in Applied Probability*, 12: 903-921.
- Lee, C., Famoye, F. & Alzaatreh, A. (2013). Methods for generating families of continuous distribution in the recent decades. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5: 219-238.
- Nadarajah, S., & Pal, M. (2008). Explicit expressions for moments of gamma order statistics. *Bulletin of the Brazilian Mathematical Society, New Series*, 39: 45-60.

## ON THE GAMMA-LOGISTIC DISTRIBUTION

Shannon, E. C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27: 379-432.

Weerahandi, S. & Johnson, R. A. (1992). Testing reliability in a stress-strength model when X and Y are normally distributed. *Technometrics*, 38: 83-91.



# Statistical Power of Alternative Structural Models for Comparative Effectiveness Research: Advantages of Modeling Unreliability

**Emil N. Coman**

UConn Health Center  
Farmington, CT

**Eugen Iordache**

Transilvania University  
Brasov, Romania

**Lisa Dierker**

Wesleyan University  
Middletown, CT

**Judith Fifield**

UConn Health Center  
Farmington, CT

**Jean J. Schensul**

Inst. for Community Research  
Hartford, CT

**Suzanne Suggs**

University of Lugano  
Lugano, Switzerland

**Russell Barbour**

Yale University CIRA  
New Haven, CT

---

The advantages of modeling the unreliability of outcomes when evaluating the comparative effectiveness of health interventions is illustrated. Adding an action-research intervention component to a regular summer job program for youth was expected to help in preventing risk behaviors. A series of simple two-group alternative structural equation models are compared to test the effect of the intervention on one key attitudinal outcome in terms of model fit and statistical power with Monte Carlo simulations. Some models presuming parameters equal across the intervention and comparison groups were underpowered to detect the intervention effect, yet modeling the unreliability of the outcome measure increased their statistical power and helped in the detection of the hypothesized effect. Comparative Effectiveness Research (CER) could benefit from flexible multi-group alternative structural models organized in decision trees, and modeling unreliability of measures can be of tremendous help for both the fit of statistical models to the data and their statistical power.

**Keywords:** comparative effectiveness research, quasi-experiment, structural equation modeling, measurement error, internal locus of control, behavioral change

---

---

*Dr. Coman is a Research Associate in the Ethel Donaghue TRIPP Center. Email him at: [comanus@netscape.net](mailto:comanus@netscape.net). Dr. Iordache is Assistant Professor in the Faculty of Silviculture and Forest Engineering. Email him at [i.eugen@unitbv.ro](mailto:i.eugen@unitbv.ro). Dr. Dierker is a Professor in the Psychology Department. Email her at [ldierker@wesleyan.edu](mailto:ldierker@wesleyan.edu). Dr. Fifield is Director of the Ethel Donaghue TRIPP Center. Email her at [fifield@uchc.edu](mailto:fifield@uchc.edu). Dr. Schensul is Founding Director of the Institute for Community Research. Email her at [jean.schensul@icrweb.org](mailto:jean.schensul@icrweb.org). Dr. Suggs is Assistant Professor in the Faculty of Communication Sciences. Email her at [suzanne.suggs@usi.ch](mailto:suzanne.suggs@usi.ch). Dr. Barbour is Associate Director for Research Methods and Analysis. Email him at [russell.barbour@yale.edu](mailto:russell.barbour@yale.edu).*

## Introduction

Assessing intervention effects poses some challenges to researchers, scholars, evaluators, and policy makers, especially when a quasi-experimental design is employed (Judd & Kenny, 1981; Stead, Hastings, & Eadie, 2002). When treatments and interventions move from the trial phase to being implemented on the ground, or Translating Research into Practice (TRIP, Feifer et al., 2004) the question of differential effects is of most concern to practitioners and researchers. Comparative Effectiveness Research (CER, Agency for Healthcare Research and Quality, 2007) is an emerging new approach addressing questions of comparative effects of alternative health interventions implemented in real world settings.

It is particularly difficult to decide on the best comparative results for reporting, when alternative models, accounting for various differences by condition, reach different conclusions. Evaluation challenges posed by health intervention designs in which randomization to conditions is not feasible are illustrated, by comparing alternative Structural Equation Models (SEM, Kline, 2010) testing for comparative intervention effects, in terms of both fit and statistical power. The benefits of modeling unreliability in increasing statistical power to detect true intervention effects are specifically demonstrated.

Evaluating health interventions effects on outcomes in community-based settings involves statistical modeling of non-RCT (Randomized Control Trial) designs, when different comparable groups are contrasted in terms of differential changes or responses to some program. A number of statistical approaches are commonly employed for such tests, among them regression-based linear models testing for the impact of a condition variable (the intervention of interest vs. a comparison condition) on the outcome of interest (Aiken, West, Schwalm, Carroll, & Hsiung, 1998; Bentler, 1991; West, Biesanz, & Pitts, 2000). In real world implementation settings however, the groups always differ in model parameters like baseline means and variances of key outcomes and covariates, as well as in terms of the outcomes change trajectories, or stability.

To accommodate such differences, structural models can be tested in several groups concurrently, like two-group models, thereby accounting for group differences that are commonly overlooked in analyses focused on whole-sample data, like paired t-tests (Macy, Chassin, & Presson, 2013) or analysis of variance (Young, Harrell, Jaganath, Cohen, & Shoptaw, 2013).

Moreover, the very assumptions about various initial and time changing group differences impact how well models fit the data and more importantly the statistical power to detect the effects of interest (Hancock, 2004). These

assumptions need to be flexibly modeled for the estimates of post-test differences or differential changes to be trustworthy (Green & Thompson, 2006). A simple CER model comparison procedure for evaluating true group differences of non-RCT interventions is presented, which specifically tests both the fit to data and the statistical power of alternative SEM models and helps in sorting through competing models, using a decision tree framework. The procedure is repeated for similar models that directly include measurement errors of the measures, and the benefits of modeling unreliability are shown.

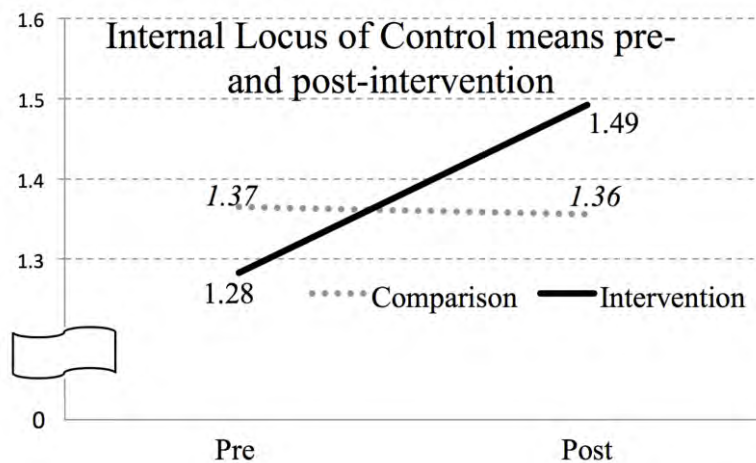
One key outcome was compared between groups of urban minority adolescents from two large cities in the USA, who were enrolled in summer job programs. One youth group was additionally engaged in a youth intervention designed to reduce drug and sexual risk behaviors (Berg, Coman, & Schensul, 2009). Low-income urban youth are often more likely to engage in risky behaviors, like substance use or unprotected sex (Farahmand, Grant, Polo, & Duffy, 2011; Simons-Morton, Crump, Haynie, & Saylor, 1999). A host of factors have been shown to be linked with behaviors that impact youth substance use initiation, like poverty, exposure to violence and drug use in their community (Caldwell, et al., 2004; DeWit, Adlaf, Offord, & Ogborne, 2000; Grant, Stinson, & Harford, 2001; Swahn, et al., 2012). On the other hand, parental support, positive peer influences and social support systems act as protective factors and are often targeted by prevention interventions (Catanzaro & Laurent, 2004; Cleveland, Gibbons, Gerrard, Pomery, & Brody, 2005). Furthermore, youth action and involvement in one's community can reinforce group cohesion and increase individual skills and a sense of self-efficacy and control over their own behaviors (Schensul, Berg, Schensul, & Sydlo, 2004).

YARP (Youth Action Research for Prevention) was a three-year summer and after-school preventive intervention (Berg, Owens, & Schensul, 2002; Reason & Bradbury, 2007). Three youth cohorts were employed and trained over the summer and were instructed to identify a youth-related problem in their community, to develop a research model and an action plan addressing that issue, gather and interpret community data, and actively engage in social action to promote changes in their community. This intervention group was compared to a matched youth group recruited from a comparable summer-job program in a neighboring city with similar economic conditions and ethnic/racial composition.

A primary hypothesis proposed that youth-initiated research for action, along with involvement in multilevel social change activities (or activism) reinforce group cohesion and individual and collective efficacy. As a result, it was expected that among other outcomes, Internal Locus of Control (ILC) would

## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

strengthen in the intervention group compared to the matched comparison group. Figure 1 shows the pre- and post-test sample means of the ILC outcome in the intervention and comparison YARP groups. It is specifically investigated which alternative models testing for intervention effects exhibit both good fit to data and enough statistical power to detect the effects, depending on different model specifications (Hancock, Lawrence, & Nevitt, 2000). The impact of accounting for measurement unreliability in the models, thereby estimating *true* differences of the latent (unobserved) outcome is also explored. The models belong to the Structural Equation Modeling (SEM) framework.



**Figure 1:** Outcome means pre- and post-intervention for the YARP comparison and intervention groups

## Methodology

### Structural equation modeling for intervention effects

A major methodological tool for understanding health intervention processes and assessing comparative outcome effects is the latent linear modeling with multiple simultaneous regression equations, known as Structural Equation Modeling (SEM, Bollen, 1989; Jöreskog, 1973) or covariance structure analysis (Bentler & Dudgeon, 1996). SEM is an enormously flexible technique that can carry out virtually any analysis (Muthén, 2002; Skrondal & Rabe-Hesketh, 2004). Current extensive SEM reviews position it as an integrative general modeling framework,

of which traditional analyses like the t-test, ANOVA, MANOVA, canonical correlation, or discriminant analysis are special cases (Fan, 1997; Graham, 2008; Muthén, 2008; Voelkle, 2007).

A simple SEM setup for testing intervention effects is the common one-group analysis of the effect of a dummy intervention variable on the post-intervention outcome. This approach, called ‘group code’ SEM (Hancock, 1997), tends to overlook however group differences that may need to be modeled, in other words it cannot account for a number of differences between groups, because data from both groups are combined. A more flexible tool is the testing of causal models in multiple groups, which allows for a range of tests of group differences (Bagozzi & Yi, 1989; Kühnel, 1988; Thompson & Green, 2006). Two-group models, like a two-group simple regression, provide parameter estimates for each group (Green & Thompson, 2006), and are more versatile in that they are simultaneously tested in more than one sample, with the options to hold parameters equal or allow them to vary across groups.

The general multiple-group manifest (observed) variable SEM model in multiple groups (indexed by  $g$ ) is of the form:

$$\mathbf{y}_g = \boldsymbol{\tau}_g + \boldsymbol{\Gamma}_g \mathbf{x}_g + \boldsymbol{\zeta}_g \quad (1)$$

where  $\mathbf{y}$  is the  $(q \times 1)$  vector of exogenous and  $\mathbf{x}$  the  $(p \times 1)$  vector of endogenous manifest variables,  $\boldsymbol{\tau}$  is the  $(q \times 1)$  vector of intercepts,  $\boldsymbol{\Gamma}$  represents the  $(q \times p)$  matrix of slopes, and  $\boldsymbol{\zeta}$  the  $(q \times 1)$  vector of residuals (or disturbances). However, when  $m$  latent variables are also modeled, the structure can be expressed separately for the latent variable relationships as:

$$\boldsymbol{\eta}_g = \boldsymbol{\alpha}_g + \mathbf{B}_g \boldsymbol{\eta}_g + \boldsymbol{\Gamma}_g \boldsymbol{\xi}_g + \boldsymbol{\zeta}_g \quad (2)$$

with  $\boldsymbol{\eta}$  being the  $(m \times 1)$  vector of latent endogenous variables,  $\boldsymbol{\alpha}$  the  $(m \times 1)$  vector of factor score means,  $\mathbf{B}$  the  $(m \times m)$  coefficient matrix for the influence of endogenous  $\boldsymbol{\eta}$ 's on  $\boldsymbol{\eta}$ 's,  $\boldsymbol{\Gamma}$  the  $(m \times n)$  coefficient matrix of the effects of the  $n$  exogenous  $\boldsymbol{\xi}$  variables on  $\boldsymbol{\eta}$ 's, and  $\boldsymbol{\zeta}$  is the  $(m \times 1)$  disturbance vector assumed to have an expected value of zero and be uncorrelated with  $\boldsymbol{\xi}$  and  $\boldsymbol{\eta}$ . The model for the measurement part linking the manifest to the latent variables is (Bollen, 1989: 320):

$$\mathbf{y}_g = \boldsymbol{\tau}_{yg} + \boldsymbol{\Lambda}_{yg} \boldsymbol{\eta}_{yg} + \boldsymbol{\varepsilon}_g \quad (3)$$

and

$$\mathbf{x}_g = \tau_{\mathbf{x}g} + \Lambda_{\mathbf{x}g}\eta_{\mathbf{x}g} + \delta_g \quad (4)$$

Model testing in SEM is meant to reproduce the variances, covariances and the means of the observed variables (Bentler & Yuan, 2000; Hancock, 2004). SEM testing requires first the assessment of the fit of the model to the data; the fit is simply the extent to which a model implies means and variances/covariances that are similar to the observed ones. The  $\chi^2$  (chi-squared) fit statistic for instance assesses the closeness between the implied covariance matrix and the sample covariance matrix (Hayduk, 1987). For a multiple-group SEM model, the  $\chi^2$  is obtained as  $(N-1) F_{ML}$  from the fit function  $F_{ML}$ , which is a weighted combination of the  $g$  groups fit functions (Bollen, 1989: 361):

$$F_{gML} = tr(\mathbf{S}_g \Sigma_g^{-1}) + \log |\Sigma_g| - \log |\mathbf{S}_g| - (p + q) \quad (5)$$

where  $\Sigma$  is the population covariance matrix and  $\mathbf{S}$  is the sample covariance matrix.

Lack of  $\chi^2$  fit is generally a function of the constraints imposed on the model (Thompson & Green, 2006). A *two-group* SEM model fits to the extent that it closely reproduces the sample means and covariances in *both* groups, so model misfit can indicate misspecification at the level of both within-group means and covariances (Saris & Satorra, 1993), as well as in the assumptions about cross-group equalities or differences, like the equality of pre-intervention means or variances

However, some specific equality constraints are supported by some data sets and rejected by others (Green & Thompson, 2003), depending on actual community initial conditions, and on differential change processes. For example, the assumption that the path (auto-regressive) coefficients from baseline to post-test outcome are equal in the intervention and comparison groups is rarely true, primarily because the intervention itself is expected to change the stability of the outcome; these assumptions are rarely tested (Bentler, 1991).

To compare groups (like gender, age, or intervention and comparison groups) on the means of the DV (dependent variable, or endogenous) in an SEM framework, researchers evaluate the fit of a structural model of no difference between the focal parameters (i.e. equality of intercepts is imposed) against another model where intercepts differ; if the models fit the data similarly, there is

no difference in intercepts, whereas if the different means model fits significantly better, there is evidence for a systematic group difference.

Acceptable model fit alone however does not ensure that its conclusions are warranted, because alternative well-fitting models may lead researchers to divergent conclusions. This is partly because alternative *well-fitting* models can have different *statistical power* to detect the effects of interest (MacCallum, Lee, & Browne, 2010; Saris & Satorra, 1993), especially for small sample sizes and unequal groups (Hancock, et al., 2000). These models contain different specification errors, and therefore will vary in both fit and testing power. Researchers should then analyze the statistical power of all alternative well-fitting models that can be relied upon for testing the hypothesis of equal post-intervention means.

In summary, there always exists a range of well-fitting models that provide different model-implied estimates of between-group differences, when researchers compare effects of programs across different conditions or settings. For the sake of brevity the focus is on simple models with only one outcome variable measured twice, with the baseline measure affecting the post-test outcome, in two groups, enhanced intervention and comparison, a common quasi-experimental design (Meehl & Waller, 2002). These models can be easily expanded to include covariates and additional intervening factors.

**Analytic steps** Two-group regression models were tested that gradually imposed equality constraints on model parameters across groups, in a hierarchical manner (somewhat similar the SEM decision trees, Brandmaier, von Oertzen, McArdle, et al., 2013), starting with a basic model with all parameters allowed to differ across groups. Specified models with increasingly more parameters were then constrained to be equal across the comparison and intervention groups: baseline means, then baseline variances, then the baseline to post-test regression coefficient, and combinations of them (Mplus syntax outputs are available online at <http://trippcenter.uchc.edu/modeling/files/HEdRes.zip>). The decisions to accept or reject models and equality-constraints are based on chi-square ( $\chi^2$ ) tests and Wald tests. Wald tests are asymptotically equivalent to the chi-square difference tests ( $\Delta\chi^2$ ) and do not require re-specifying the model (Bollen, 1989: 295).

A simple two-group structural model with a baseline outcome causing the post-test outcome yields five model estimated parameters for each group (see for illustration the actual parameters in Figure 2). The model depicts variances as a double headed arrow, or as a covariance of the variable with itself. Such models can specify (or not) equality constraints between some of these parameters, and

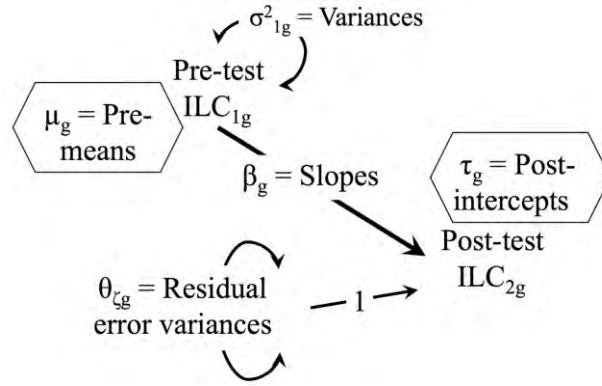


## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

then test the difference between post-intervention intercepts of the outcome. The linear equations can be directly spelled out from the model in Figure 2 as:

$$ILC_{2g} = \tau_g + \beta_g * ILC_{1g} + \zeta_g \quad (6)$$

where  $ILC_{1g}$  and  $ILC_{2g}$  are the baseline and post-test variables,  $\tau_g$  are the intercepts (the values of  $ILC_{2g}$  when  $ILC_{1g}$  are zero),  $\gamma_g$  are the auto-regressive coefficients,  $\zeta_g$  the residual error terms, and  $g$  indexes group (intervention or treatment T, and comparison C). Organizing alternative SEM models using a decision tree that starts with an *all-parameters-different* model, and grows by imposing equality constraints on parameters across groups is proposed.



**Figure 2.** Two group model specification for testing the equality of post-intervention difference  $T_{2C} = T_{2T}$  of the ILC outcome (Note: Hexagons represent means/intercepts; T: treatment group, C: comparison.)

In addition to fit, models differ in statistical power to detect specific effects (Hancock, et al., 2000). The probability of rejecting the hypothesis of equal post-test means, when the means are different in the population, is the statistical power of the test, and should ideally be one. The power of SEM models can be obtained generally by fitting on population data an F (full) model, then an alternative R (restricted) model with an additional constraint of interest (MacCallum, et al., 2010; Satorra & Saris, 1985). Because the population F model fits perfectly, the only worsening (or ‘badness’) of fit of the reduced model R would come from the additional constraint imposed the equality of post-test means in this case. The difference between the two model  $\chi^2$  values represents the noncentrality parameter



for the noncentral distribution with one degree of freedom (Hancock, et al., 2000). Alternatively, the Wald test  $\chi^2$  is an asymptotically equivalent method of estimating power (Buse, 1982).

The statistical power of each alternative model was assessed using Mplus 6 Monte Carlo facility (Muthén & Muthén, 2002), which generates datasets according to an F causal model assumed to be the true in the population, generates simulated sample datasets (in this study, 1,000 simulations), and then can test a constrained model R to each simulated sample dataset. The Mplus output provides descriptives of the percent of times the R replicated models rejected the (assumed false) equality of post-test means, which is the power of the model to detect the effect. Specifically, the power of the model is given by the observed proportion of replication tests for which the Wald test exceeds the critical value of 3.841 (for degree of freedom  $df = 1$ , for the equality of intercepts constraint  $\tau_{C2} = \tau_{I2}$ ). Unreliability was then modeled in both groups statistical power to detect intervention effects was tested for all the new models. (Muthén & Jöreskog, 1983; Thompson & Green, 2006).

**Study setting and data** The research team conducted and evaluated the multi-year YARP project (2002-2005), a youth intervention implemented in Hartford, Connecticut (CT). The Institute for Community Research Institutional Review Board ensured that proper human subjects protocols were followed. The intervention group had  $N_T = 90$  participants who completed all four surveys, recruited from Hartford, CT, of whom 56% were females, 48% Blacks, 37% Latinos, mean age  $M_T = 15.1$  years, while the comparison group had  $N_C = 167$  from a similar inner-city youth in a summer job program in Massachusetts, U.S., with 58% females, 45% Blacks, 44% Latinos, and mean age  $M_C = 15.5$ .

Measures were taken at baseline, 2 month, 6 months, and 1 year in both groups. Internal locus of control was measured with 4 indicators (i.e., ‘I am responsible for accomplishing goals’, ‘Life offers me many choices’, ‘I can do things I set out to do’, and ‘I enjoy having control over own destiny’) from among the Internal subscale items of the Levenson Locus of Control scale (Levenson, 1973) modified for younger ages. For simplicity and because interest lies in long-term and potentially sustainable effects, the focus here is on the difference in changes from baseline to the final fourth measurement time point. A composite of the average items was calculated (rated from strongly disagree = 1, to strongly agree = 4, 4 being greater internality). Basic descriptive, reliabilities, correlations and covariances are shown in Table 1, for each group, and the entire sample. The

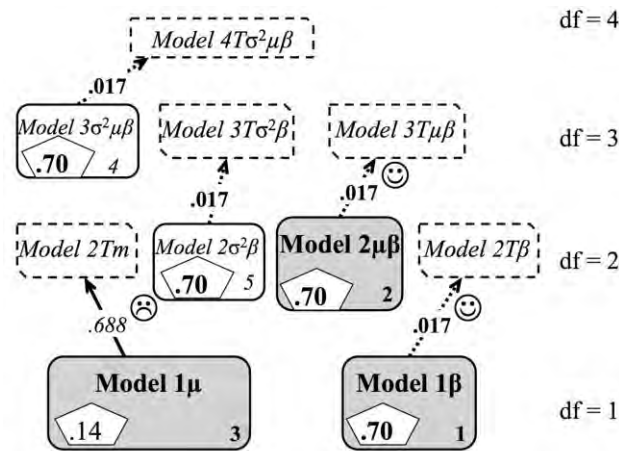
## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

pre- and post-test ILC measures had acceptable reliabilities, Cronbach's alphas between .725 and .871.

**Table 1:** Covariances, correlations, means and Cronbach's  $\alpha$  of the pre- and post-test Internal Locus of Control (ILC) outcome for the two YARP groups and for the whole sample

	Comparison N <sub>C</sub> = 167		Intervention N <sub>T</sub> = 90		Whole sample N = 257		
	ILC1	ILC2	ILC1	ILC2	ILC1	ILC2	Group
ILC1	<i>0.264</i>	<i>0.475*</i>	<i>0.174</i>	<i>0.448*</i>	<i>0.235</i>	<i>0.445*</i>	-0.081 <sup>NS</sup>
ILC2	<b>0.264</b>	<i>0.325</i>	<b>0.174</b>	<i>0.480</i>	<b>0.134</b>	<i>0.385</i>	0.104 <sup>NS</sup>
Group (C/T)	-	-	-	-	<b>-0.019</b>	<b>0.031</b>	<i>0.228</i>
Means $\mu$	1.365	1.356	1.283	1.492	1.337	1.404	0.350
Cronbach's $\alpha$	0.725	0.847	0.726	0.871	.726	.859	-

**Note.** Covariances are shown in bold and below diagonal and correlations above diagonals, variances in italics on the diagonals.



**Figure 3:** Alternative decision-tree SEM modeling for comparing post-intervention observed outcome means in two-group causal models (*Notes:* Shaded models: good chi-square fit; model names indicate which equality constraints are imposed, on:  $\sigma^2$  = variances,  $\mu$  = means;  $\beta$  = autoregressive paths; or T = the test of equality of post-test intercepts; numbers in boxes: in pentagons— power of each model, and lower right - fit ordered from best fitting (1) up; arrows going up show model comparison tests, with p value for significance of Wald test [ $p < .05$  corroborates intervention effect.]

The hypothesis of equal post-intervention ILC means (technically the intercepts  $\tau_{C/T}$ ) was tested with all well-fitting models. The models are shown as a decision tree in Figure 3. The baseline model with  $df = 0$  (the ‘root’) assumes all parameters are different across groups, and each higher layer of nodes adds one more equality constraint, hence estimating one less parameter. When adding the equality constraint between post-test intercepts (the focal parameter) led to a significant worsening of fit, or a significant Wald test statistic, it was concluded that the means were different between groups.

## Results

The results of alternative modeling of the tests of ILC outcome differences are now reported. The three well-fitting models are shown in Table 2, which lists the common SEM measures of fit ordered by descending  $p$  values for  $\chi^2$  larger than .05, and the Wald tests of the post-intervention differences.

**Table 2:** Ordered fit indices, Wald tests, and statistical power for the well-fitting alternative causal models of the YARP intervention effect on Internal Locus of Control

	Model	$\chi^2$	df	$\chi^2 p$	CFI	RMSEA	Wald	Wald $p$	Power
1	$1\beta$ $\beta$ 's equal	1.517	1	0.218	.991	.063	5.685	0.017 ☺	0.70
2	$2\mu\beta$ $\mu$ 's & $\beta$ 's equal	3.436	2	0.179	.976	.075	5.719	0.017 ☺	0.70
3	$1\mu$ $\mu$ 's equal	1.919	1	0.166	.985	.085	0.161	0.688 ☹	0.14

**Note:**  $\mu$  = baseline means;  $\beta$  = auto-regressive path; italics Wald test  $p$  indicate significant intervention effect.

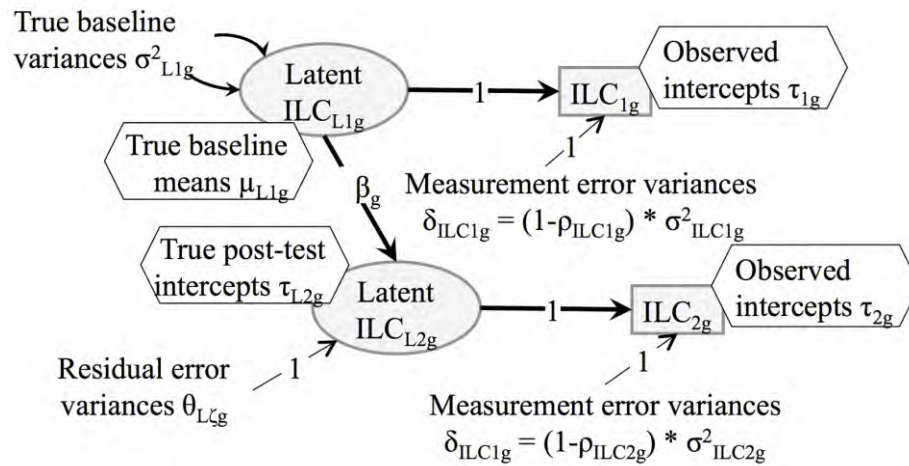
Two well-fitting models,  $1\gamma$ , and  $2\mu\gamma$  indicated that there was indeed a significant intervention effect ( $p = .017$  for the Wald statistic in both), while another well-fitting model,  $1\mu$ , reached another conclusion. Note that the baseline means cannot be deemed statistically different, because the fit of the  $1\mu$  model (baseline means set equal across groups) indicates in fact that the perfectly fitting model with all parameters different (for which  $df = 0$ ) does not worsen significantly when constraining the baseline means to be equal.

The fact that only some models reject the equality of means hypothesis is an indication of differential statistical testing power (Hancock, 2006) linked to model misspecifications (Saris & Satorra, 1993). In other words, some models may have low power to reject the (false) hypothesis of equal post-test means.

## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

In terms of statistical power, the equal baseline means model ( $1\mu$ ) that has initially found no effect yielded a probability to *rightly* reject the (assumed false) equal means hypothesis of  $p = .14$ , while the other two well-fitting models had higher sensitivities of  $p = .70$ . This indicates that for the observed sample sizes of 90 and 167, the models compared here have dramatically different sensitivities to detect the effect of interest. Examination of model fit alone, therefore, without controlling for Type II errors could lead to accepting well fitting models that are not sensitive to detect specific effects (Saris, Satorra, & van der Veld, 2009). In this particular instance, the ‘stress’ induced in this simple linear model by constraining the baseline means to be equal rendered one *well-fitting* model ( $1\mu$ ) seriously *under-powered* to detect the intervention effect. Next it will be shown that this particular model was underpowered because the baseline equality of means assumption was imposed on the unreliable baseline measure.

Informed knowledge of the reliability of an observed variable allows for modeling the true means of latent variables (unattenuated by measurement error). When measurement error is directly specified for composite or single-item variables, each measured variable is in fact subjected to a *mini-factor analysis*, in which a common factor (the true measure) is assumed to be responsible for (acting behind) the observed measure. The reliability of an observed variable is simply the proportion of the observed variance that is true variance, or the squared correlation between the true variable and the observed variable (Raykov, 1997), and a common estimate used in applied research for scale reliability is Cronbach’s alpha coefficient (Raykov & Marcoulides, 2011). Because reliability  $\rho$  is the percentage of variance that is true variance, the complement  $1 - \rho$  is the percentage that is measurement error, hence  $(1-\rho)*\sigma^2_{ILC1}$  is the measurement error variance (MacKinnon, 2008: 189). The measurement error variance for the comparison group  $\delta_{1C}$  for  $ILC_{1C}$  in Figure 4, for example, whose reliability was .73 and variance .26, was fixed at  $(1 - .73) * .26 = .27 * .26 = .070$ .



**Figure 4.** Illustration of two-groups model parameters with measurement errors directly modeled (*Notes:* Hexagons show the means/intercepts;  $\rho$  are reliabilities;  $\sigma^2$  are observed variances;  $g$  indexes group: comparison and intervention.)

When directly modeling the unreliabilities of the baseline and post-intervention ILC outcome in both groups, the power to detect the post-intervention differences in mean ILC of the  $I\mu$  model increases to .716 (from the meager .14 of the manifest ILC model). So when assuming that the *true* (latent) baseline ILC means are equal, the model is better powered to detect the intervention effect unto the reliable (*true*) latent outcome, and the effect emerges as a significant larger increase in the true ILC in the intervention group, Wald test statistic of 6.14 ( $df=1$ ),  $p = .012$ .

## Conclusion

A decision-tree method of comparing alternative models of observed and true outcomes was illustrated (Kaplan, 1990), which tests for post-intervention health outcome differences between community-based groups, based on *both* fit to data and power to detect these effects. This procedure can assist in Comparative Effectiveness Research (CER) by providing the modeling flexibility required by actual data in terms of various group (or community) differences. It is particularly useful when trying to compare effects using summary data from separate studies, when available in the form of means, variances and covariances.

## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

One manifest outcome well-fitting model was under-powered to detect the YARP intervention effect on Internal Locus of Control (ILC), but two other well-fitting models with better statistical power detected a positive effect on ILC in the intervention group. It was found that even small differences in parameters of the unreliable measures create ‘stress’ in the structural models which can render them underpowered to detect the effects of interest. In the illustration, the lack of power of the baseline equal means two-group structural model derived from imposing a plausible equality constraint on the *unreliable* observed ILC measures, rather than on the true (latent) ones.

The structural equation models tested here indicate that the lack of statistical power of the models with unreliable outcomes are due largely to modeling error-in-variable measures (containing measurement errors). The example herein shows the importance of *a priori* specification of alternative models and the utility and relative ease of post-hoc power analysis, and also showed the benefits of directly modeling unreliabilities of outcome measures. The nuanced reporting of the alternative testing and plausibility of competing conclusions is essential for statisticians, prevention and comparative effectiveness researchers, as well as policy makers and community representatives interested in evaluating, replicating or translating successful programs.

Some limitations are worth mentioning. To the extent that one tries out repeated models on the same data, procedure called specification search and available in current SEM software like AMOS (Arbuckle, 2007), the issue of over-fitting the model to the same data (or data dredging, see Brandmaier, et al., 2013) could be a concern (Hayduk, 1987). This procedure is acceptable, if careful planning of model testing under alternative reasonable configurations is undertaken *a priori* (Jöreskog, Bollen, & Long, 1993), being akin to specifying equivalent models before data collection (Hershberger, 1994).

The decision tree modeling approach is useful in identifying and classifying alternative multi-group models according to differential support from multiple-group data in general. It does not of course provide criteria for deciding the true and false nature of the models, but rather their “truth-likeness” or closeness to the truth (Meehl & Waller, 2002). Quasi-experimental designs for instance require the use of covariates to control for additional baseline differences between the groups, and the modeling of selection biases (Muthén & Jöreskog, 1983); however, a basic model was chosen herein for simplicity to illustrate this method.

The method presented here becomes cumbersome when models increase in complexity, e.g. when using multiple indicator measures with numerous possible cross-group constraints, like specific loadings and intercepts (Green & Thompson,

2006). Multiple latent covariates and possibly multiple outcomes with indirect effects complicate the picture even further. Study analyses, however, make clear the benefits of directly modeling unreliability, of careful inspection of alternative models and attending to both model fit measures and statistical power of the models, when comparing the effectiveness of health interventions translated and implemented differently in separate communities.

## References

- Agency for Healthcare Research and Quality. (2007). *Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville, MD. Retrieved from [http://www.effectivehealthcare.ahrq.gov/repFiles/2007\\_10DraftMethodsGuide.pdf](http://www.effectivehealthcare.ahrq.gov/repFiles/2007_10DraftMethodsGuide.pdf)
- Aiken, L., West, S., Schwalm, D., Carroll, J., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, 22(2): 207. doi: 10.1177/0193841X9802200203
- Arbuckle, J. (2007). AMOS 16 User's Guide. Retrieved from <http://support.spss.com/Student/Patches/Amos/Amos5Supplement.pdf>
- Bagozzi, R. P., & Yi, Y. (1989). On the use of structural equation models in experimental designs. *Journal of Marketing Research*, 26(3): 271-284.
- Bentler, P. M. (1991). Modeling of intervention effects. In C. G. Leukefeld & W. J. Bukoski (Eds.), *Drug Abuse Prevention Intervention Research: Methodological Issues* (pp. 159–182). Washington, DC: U.S. Government Printing Office.
- Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, 47(1): 563-592.
- Bentler, P. M., & Yuan, K. H. (2000). On adding a mean structure to a covariance structure model. *Educational and Psychological Measurement*, 60(3): 326. doi: 10.1177/00131640021970574
- Berg, M., Coman, E., & Schensul, J. (2009). Youth Action Research for Prevention: A Multi-level Intervention Designed to Increase Efficacy and Empowerment Among Urban Youth. *American Journal of Community Psychology*, 43(3): 345-359. doi: 10.1007/s10464-009-9231-2



## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

- Berg, M., Owens, D., & Schensul, J. (2002). Participatory action research, service-learning, and community youth development. *CYD Journal: Community Youth Development*, 3(2): 20–25.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley and Sons.
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, 18(1): 71-86. doi: [10.1037/a0030001](https://doi.org/10.1037/a0030001)
- Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *American Statistician*, 36(3): 153-157.
- Caldwell, C. H., Wright, J. C., Zimmerman, M. A., Walsemann, K. M., Williams, D., & Isichei, P. A. C. (2004). Enhancing adolescent health behaviors through strengthening non-resident father-son relationships: a model for intervention with African-American families. *Health Education Research*, 19(6): 644-656. doi: [10.1093/her/cyg078](https://doi.org/10.1093/her/cyg078)
- Catanzaro, S. J., & Laurent, J. (2004). Perceived family support, negative mood regulation expectancies, coping, and adolescent alcohol use: Evidence of mediation and moderation effects. *Addictive Behaviors*, 29(9): 1779-1797. doi: [10.1016/j.addbeh.2004.04.001](https://doi.org/10.1016/j.addbeh.2004.04.001)
- Cleveland, M. J., Gibbons, F. X., Gerrard, M., Pomery, E. A., & Brody, G. H. (2005). The impact of parenting on risk cognitions and risk behavior: A study of mediation and moderation in a panel of African American adolescents. *Child development*, 76(4): 900-916. doi: [10.1111/j.1467-8624.2005.00885.x](https://doi.org/10.1111/j.1467-8624.2005.00885.x)
- DeWit, D. J., Adlaf, E. M., Offord, D. R., & Ogborne, A. C. (2000). Age at first alcohol use: a risk factor for the development of alcohol disorders. *American Journal of Psychiatry*, 157(5): 745-750. doi: [10.1176/appi.ajp.157.5.745](https://doi.org/10.1176/appi.ajp.157.5.745)
- Fan, X. (1997). Canonical Correlation Analysis and Structural Equation Modeling: What Do They Have in Common? *Structural Equation Modeling*, 4(1): 65-79. doi: [10.1080/10705519709540060](https://doi.org/10.1080/10705519709540060)
- Farahmand, F. K., Grant, K. E., Polo, A. J., & Duffy, S. N. (2011). School-Based Mental Health and Behavioral Programs for Low-Income, Urban Youth: A Systematic and Meta-Analytic Review. *Clinical Psychology: Science and Practice*, 18(4): 372-390. doi: [10.1111/j.1468-2850.2011.01265.x](https://doi.org/10.1111/j.1468-2850.2011.01265.x)
- Feifer, C., Ornstein, S., Karson, A. S., Bates, D. W., Jones, K. R., & Vargas, P. A. (2004). From research to daily clinical practice: what are the challenges in"



translation"? *Joint Commission Journal on Quality and Patient Safety*, 30(5): 235-245.

Graham, J. M. (2008). The General Linear Model as Structural Equation Modeling. *Journal of Educational and Behavioral Statistics*, 33(4): 485. doi: [10.3102/1076998607306151](https://doi.org/10.3102/1076998607306151)

Grant, B. F., Stinson, F. S., & Harford, T. C. (2001). Age at onset of alcohol use and DSM-IV alcohol abuse and dependence: a 12-year follow-up. *Journal of Substance Abuse*, 13(4): 493-504. doi: [10.1016/S0899-3289\(01\)00096-7](https://doi.org/10.1016/S0899-3289(01)00096-7)

Green, S., & Thompson, M. (2003). Structural equation modeling in clinical research. In M. C. Roberts & S. S. Illardi (Eds.), *Methods of research in clinical psychology: A handbook* (pp. 138-175). London: Blackwell.

Green, S., & Thompson, M. (2006). Structural equation modeling for conducting tests of differences in multiple means. *Psychosomatic Medicine*, 68(5): 706-717. doi: [10.1097/01.psy.0000237859.06467.ab](https://doi.org/10.1097/01.psy.0000237859.06467.ab)

Hancock, G. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, 30: 91-105.

Hancock, G. (2004). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 317-334). Thousand Oaks, CA: Sage Publications, Inc.

Hancock, G. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69-118). Greenwich, CT: Information Age.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I Error and Power of Latent Mean Methods and MANOVA in Factorially Invariant and Noninvariant Latent Variable Systems. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4): 534 - 556. doi: [10.1207/S15328007SEM0704\\_2](https://doi.org/10.1207/S15328007SEM0704_2)

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Johns Hopkins University Press.

Hershberger, S. (1994). The specification of equivalent models before the collection of data. In A. v. Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 68-108). Thousand Oaks, CA: Sage.

Jöreskog, K. G. (1973). A General Method for Estimating a Linear Structural Equation System. In A. Goldberger & O. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85-112). New York: Seminar Press.

## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

- Jöreskog, K. G., Bollen, K. A., & Long, J. S. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park: Sage.
- Judd, C., & Kenny, D. (1981). *Estimating the effects of social interventions*. Cambridge: Cambridge University Press. Retrieved from [davidakenny.net/doc/JuddKenny1981.pdf](http://davidakenny.net/doc/JuddKenny1981.pdf)
- Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research*, 25(2): 137-155. doi: [10.1207/s15327906mbr2502\\_1](https://doi.org/10.1207/s15327906mbr2502_1)
- Kline, R. (2010). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York, NY: The Guilford Press.
- Kühnel, S. M. (1988). Testing MANOVA Designs with LISREL. *Sociological Methods & Research*, 16(4): 504-523. doi: [10.1177/0049124188016004004](https://doi.org/10.1177/0049124188016004004)
- Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of consulting and clinical psychology*, 41(3): 397-404. doi: [10.1037/h0035357](https://doi.org/10.1037/h0035357)
- MacCallum, R., Lee, T., & Browne, M. (2010). The Issue of Isopower in Power Analysis for Tests of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(1): 23-41. doi: [10.1080/10705510903438906](https://doi.org/10.1080/10705510903438906)
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum Associates.
- Macy, J. T., Chassin, L., & Presson, C. C. (2013). Predictors of health behaviors after the economic downturn: A longitudinal study. *Social Science & Medicine*, 89(0): 8-15. doi: [10.1016/j.socscimed.2013.04.020](https://doi.org/10.1016/j.socscimed.2013.04.020)
- Meehl, P., & Waller, N. (2002). The Path Analysis Controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological Methods*, 7(3): 283-300. doi: [10.1037/1082-989X.7.3.283](https://doi.org/10.1037/1082-989X.7.3.283)
- Muthén, B., & Jöreskog, K. G. (1983). Selectivity Problems in Quasi-Experimental Studies. *Evaluation Review*, 7(2): 139-174. doi: [10.1177/0193841x8300700201](https://doi.org/10.1177/0193841x8300700201)
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29(1): 81-118.

- Muthén, B. O. (2008). Latent variable hybrids: Overview of old and new models. In G. Hancock & K. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4): 599-620. doi: [10.1207/S15328007SEM0904\\_8](https://doi.org/10.1207/S15328007SEM0904_8)
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21: 173-184. doi: [10.1177/01466216970212006](https://doi.org/10.1177/01466216970212006)
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.
- Reason, P., & Bradbury, H. (2007). *The SAGE handbook of action research: Participative inquiry and practice*. Sage Publications Ltd.
- Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181-204). Newbury Park, CA: Sage.
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling*, 16(4): 561-582. doi: [10.1080/10705510903203433](https://doi.org/10.1080/10705510903203433)
- Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50(1): 83-90. doi: [10.1007/BF02294150](https://doi.org/10.1007/BF02294150)
- Schensul, J. J., Berg, M. J., Schensul, D., & Sydlo, S. (2004). Core elements of participatory action research for educational empowerment and risk prevention with urban youth. *Practicing Anthropology*, 26(2): 5-9.
- Simons-Morton, B. G., Crump, A. D., Haynie, D. L., & Saylor, K. E. (1999). Student-school bonding and adolescent problem behavior. *Health Education Research*, 14(1): 99-107. doi: [10.1093/her/14.1.99](https://doi.org/10.1093/her/14.1.99)
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.
- Stead, M., Hastings, G., & Eadie, D. (2002). The challenge of evaluating complex interventions: a framework for evaluating media advocacy. *Health Education Research*, 17(3): 351-364. doi: [10.1093/her/17.3.351](https://doi.org/10.1093/her/17.3.351)
- Swahn, M., Bossarte, R., Choquet, M., Hassler, C., Falissard, B., & Chau, N. (2012). Early substance use initiation and suicide ideation and attempts among

## STATISTICAL POWER OF ALTERNATIVE STRUCTURAL MODELS

students in France and the United States. *International Journal of Public Health*, 57(1): 95-105. doi: [10.1007/s00038-011-0255-7](https://doi.org/10.1007/s00038-011-0255-7)

Thompson, M. S., & Green, S. B. (2006). Evaluating Between-Group Differences in Latent Variable Means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119-170). Greenwich, CT: Information Age.

Voelkle, M. C. (2007). Latent growth curve modeling as an integrative approach to the analysis of change. *Psychology Science*, 49(4): 375. doi: [10.1111/j.1469-8986.2007.00544.x](https://doi.org/10.1111/j.1469-8986.2007.00544.x)

West, S., Biesanz, J., & Pitts, S. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84): Cambridge University Press.

Young, S. D., Harrell, L., Jaganath, D., Cohen, A. C., & Shoptaw, S. (2013). Feasibility of recruiting peer educators for an online social networking-based health intervention. *Health Education Journal*, 72(3): 276-282. doi: [10.1177/0017896912440768](https://doi.org/10.1177/0017896912440768)

# An Exploratory Graphical Method for Identifying Associations in $r \times c$ Contingency Tables

**Martin L. Lesser**

Feinstein Institute for Medical Research  
Manhasset, NY

**Meredith B. Akerman**

Feinstein Institute for Medical Research  
Manhasset, NY

---

On finding a significant association between rows and columns of an  $r \times c$  contingency table, the next step is to study the nature of the association in more detail. The use of a scree plot to visualize the largest contributions to  $X^2$  among all cells in the table in order to determine the nature of the association in more detail is proposed.

*Keywords:* contingency table; graphical method; exploratory analysis; scree plot; contribution to chi-square

---

## Introduction

A graphical method is proposed for exploring associations between rows and columns in an  $r \times c$  contingency table. Typically, the Pearson chi-square test (or alternatively, the Fisher exact test) is used to test for independence of two categorical variables arranged in an  $r \times c$  contingency table. (When one or both categories are ordinal, other procedures more suited to test for ordinal associations are available but the method being proposed here can be applied to both ordinal and non-ordinal data.)

On finding a significant association between rows and columns of an  $r \times c$  table, the next step is to study the nature of the association (i.e., lack of independence) in more detail. One approach is to partition the  $r \times c$  table and to use principles of chi-square partitioning to compare various groupings of rows and columns in order to make sense of the association (Agresti, 1990). Another method is to “collapse” the  $r \times c$  table into some meaningful  $2 \times 2$  table, the results for which are much easier to interpret (Feinstein, 2002). The advantage of

---

*Dr. Lesser is Director, Investigator, and Professor of Biostatistics in the Biostatistics Unit. Email him at: [MLesser@nshs.edu](mailto:MLesser@nshs.edu). Meredith Akerman is a Statistician in the Biostatistics Unit. Email her at: [MAkerman@nshs.edu](mailto:MAkerman@nshs.edu).*

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

the first approach is that it is truly inferential, but the choice of how to partition the table may be impractical for very large  $r \times c$  tables. The second method, while appealing due to its simplicity, may result in combining categories that have no appropriate justification or interpretation with respect to the subject matter being studied.

Consider the situation where the data analyst is interested more in exploration of the association rather than formal inference, in which case an exploratory graphical approach might be appropriate. There is the method known as Correspondence Analysis (CA) with applications in areas of social science, psychology, market research, and, to some extent, biomedical research ([Greenacre, 1984](#); [Greenacre, 1992](#)). This graphical approach is based on linear algebraic techniques, which project the rows and columns of a data matrix in points onto a graph in Euclidean space, from which a better understanding of the data may be derived.

A simpler, yet intuitive method is proposed: exploratory graphical approach based on a method suggested by Snedecor and Cochran ([1989](#)), in which the data analyst identifies the cell entries providing the largest percentage contributions to  $X^2$  because those will suggest departure from the null hypothesis of independence, and will be row-column combinations of interest. Some drawbacks of this approach are that searching an  $r \times c$  table for the “largest” contributions to  $X^2$  can be tedious (especially for large tables), inefficient, and prone to error (i.e., failing to identify all the cells that are “large” contributors). Given these potential problems, a graphical approach to summarizing these contributions would be helpful, especially when there are many cells to analyze.

The graphical approach used herein is to use an adaptation of the scree plot to visualize the largest contributions to  $X^2$  among all of the cells of the  $r \times c$  table. (The scree plot is commonly used in principal components analysis to help choose the most important principal components [[Khattree and Naik, 2000](#)]).

As an example, [Table 1](#) (hypothetical data for illustrative purposes) is a 6 x 5 cross-tabulation of a patient’s primary hospital admitting diagnosis according to the patient’s race. There is a highly significant association between diagnosis and race ( $X^2 = 326.4$ ,  $p < 0.0001$ ). The common interpretation of this significance is that diagnosis is not independent of race or, alternatively, that there are at least two races for which the distributions of diagnosis differ. Which two (or more) columns differ from one another?

**Table 1.** Cross-tabulation of a patient's race according to patient's primary hospital admitting diagnosis

	Primary hospital admitting diagnosis						Total
	DM	Chest pain	CVA	Fever	GI distress	Other	
White	39 23.49	18 10.84	51 30.72	22 13.25	16 9.64	20 12.05	166
Black	11 6.25	15 8.52	8 4.55	2 1.14	92 52.27	48 27.27	176
Hispanic	90 40.54	56 25.23	19 8.56	15 6.76	13 5.86	29 13.06	222
Asian	13 26.53	0 0	14 28.57	7 14.29	15 30.61	0 0	49
Other	44 46.32	18 18.95	10 10.53	11 11.58	9 9.47	3 3.16	95
Total	197	107	102	57	145	100	708

**Note.** The top entry in each cell is the frequency count; the lower entry is the "row percent," which is the percentage based on the row total.

To answer that question, two methods are commonly used. The first is simply to inspect the many so-called "column proportions" and informally, based on subjective visualization, make a judgment as to which columns differ. The second is to more formally perform all 10 pairwise comparisons of the columns using a  $X^2$  test with 5 degrees of freedom and to declare two columns as different if the associated p-value is less than some critical value that is appropriately adjusted for multiple comparisons. (In general there would be  $c!/(2!(c-2)!)$  each with  $r-1$  degrees of freedom.)

The first method is deficient because it is highly subjective and requires simultaneous visual processing of all of the column percentages. The second method has the advantage of being truly inferential, but, in finding two columns that differ, it fails to identify the row locations of those differences.

The graphical method proposed is computationally objective and reproducible and can be easily programmed in most statistical software packages, including SAS<sup>®</sup> for which a publically available macro has been written.

## Methodology

Suppose data are arranged in an  $r \times c$  contingency table. The individual entries in the  $r \times c$  table represent the frequency, or, number, of observations of a given row-column combination (e.g. race and diagnosis as in Table 1.)

Using standard statistical notation, let  $O_{ij}$  represent the observed entry in row  $i$ , column  $j$ ,  $O_{i.}$  the total of all entries in row  $i$ ,  $O_{.j}$  the total of all entries in column  $j$ , and  $E_{ij}$  the expected entry in row  $i$ , column  $j$ . Letting  $n$  denote the sum total of all frequencies entered in the table, the expected frequency of row  $i$ , column  $j$ ,  $E_{ij}$ , is calculated as the product of the total frequency in row  $i$  multiplied by the total frequency in column  $j$ , divided by  $n$  (i.e.,  $E_{ij} = (O_{i.} \times O_{.j}) / n$ ).

Using this notation, the standard Pearson  $X^2$  statistic is calculated as

$$X^2 = \sum_i \sum_j \left[ (O_{ij} - E_{ij})^2 / E_{ij} \right],$$

where the summations correspond to  $i = 1, 2, \dots, r$  and  $j = 1, 2, \dots, c$ . Snedecor and Cochran (1989) denote the contribution of the  $ij^{\text{th}}$  entry to the  $X^2$  statistic as

$$X^2_{ij} = (O_{ij} - E_{ij})^2 / E_{ij}$$

Compute all values of  $X^2_{ij}$  for  $i=1, 2, \dots, r$  and  $j=1, 2, \dots, c$ . Then compute  $P_{ij} = 100 * X^2_{ij} / X^2 =$  percentage of overall  $X^2$  contributed by the  $ij^{\text{th}}$  entry. Snedecor and Cochran (1989) propose that the entries providing the largest percentage contributions to  $X^2$  are those that will suggest departure from the null hypothesis of independence. Note that “contribution to  $X^2$ ” is sometimes referred to as the square of the “standardized residuals” (Agresti, 1990).

The general idea of the proposed graphical method is to compute each table entry's  $P_{ij}$ , order the  $P_{ij}$ s from largest to smallest, and to find the first  $P_{ij}$  for which the remaining ordered  $P_{ij}$ s remain relatively constant. This ordering can be visually displayed in a graph, known as a “scree plot”. The algorithm for constructing the scree plot is given in the following steps:

### Step 1

Order the values of  $P_{ij}$  from largest to smallest and denote the ordered values (i.e. “order statistics”) as  $P_{(1)} \geq P_{(2)} \geq \dots \geq P_{(rc)}$ .



**Step 2**

Plot  $P_{(i)}$  against  $i$  to form a scree plot, analogous to what is done with eigenvalues in principal components analysis (PCA) (Khattree and Naik, 2000).

**Step 3**

Find the cells in the  $r \times c$  table that significantly contribute to the departures from independence. This can be done using any of the following three criteria.

***Cumulative Percent Method*** Find the left-most point on the horizontal axis that corresponds to a cumulative sum of percent contributions to chi-square that totals as close to, but does not exceed some pre-specified percentage,  $\pi$ . For example,  $\pi$  might be set to 50%. It should be noted that  $\pi$  is often chosen arbitrarily with no formal justification of its utility. Using  $\pi = 50\%$  is “middle of the road”. Increasing  $\pi$  would result in a more “liberal” rule, allowing more cells to be implicated in the departure from independence, possibly increasing the false positive rate with respect to identifying the number of such cells. Decreasing  $\pi$  would restrict the number of cells, possibly increasing the false negative rate. (Note that in PCA,  $\pi$ , which would be the cumulative variance explained, is often set to 90% [Khattree and Naik, 2000])

***Subjective Elbow Method*** Find the “bend of the elbow” or “turning point” of the scree plot to determine which cells in the  $r \times c$  table contribute substantially to the  $X^2$  statistic. Typically, the bend in the elbow would be defined as the point on the plot for which all points to the left of it will have a much steeper downward slope than those to the right. The idea behind this choice of a bending point is that the number of cells to be selected is such that the differences between consecutive contributions to chi-square are becoming increasing smaller (Khattree and Naik, 2000). This subjective method is based only on visual inspection of the scree plot. This approach may be useful when there is a fairly clear elbow. The primary shortcoming is that this method is subjective and may not be reproducible between data analysts.

***Objective Elbow Method*** Because the determination of the bend in the elbow using the Subjective Elbow Method is not necessarily reproducible, it is proposed to systematize the identification of the elbow by finding the ordered pair  $(i, P_{(i)})$  which is closest to the origin  $(0,0)$ . This can be done by computing the squared-Euclidean distances of each point on the scree plot,

$(i-0)^2 + (P_{(i)}-0)^2 = i^2 + P_{(i)}^2$  and finding the ordered pair,  $(i^*, P^*)$ , corresponding to the minimum value of those distances (i.e.  $(i^*, P^*)$  is the point closest to the origin). All cells that are represented on the plot with  $i \leq i^*$  would then be implicated in the departure from independence. In the context of a scree plot, which is a plot of a non-increasing concave function, the “ideal” elbow would be two straight line segments connected at a “pivot” point forming an angle of  $90^\circ$  to less than  $180^\circ$  between the segments. For such a function, the bend of the elbow would correspond to the point with minimum distance to the origin. An example of an ideal elbow would be a perfect “L” shape curve with its vertical and horizontal components parallel to the vertical and horizontal axes of the scree plot, respectively.

It should be emphasized that while the proposed method relies on the use of the chi-square statistic, as an exploratory tool, it can be used even when the r x c table does not meet the criteria for the use of the Pearson chi-square test and a Fisher’s exact test would be more appropriate.

For this manuscript, the authors used the PROC FREQ procedure in SAS Version 9.3 (SAS Institute, Cary, NC).

## Results and Examples

The proposed method is illustrated using data from the Asia-Pacific Quality of Life Study (APQOL) in Lung Cancer. (The data are provided courtesy of Drs. Richard Gralla and Patricia Hollen [Gralla, 2013; Thongprassert, 2013]). This data consists of, among other variables, country of diagnosis (China, Korea, Thailand, Taiwan), Karnofsky Performance Status at diagnosis (KPS=50, 60, 70, 80, 90, 100), lung cancer T stage (T0, T1, T2, T3, T4, and TX), node status (N0, N1, N2, N3, NX), and metastasis (M0, M1, MX). [The so-called “TNM staging system” for cancer classifies cancers according to tumor size (T), lymph node involvement (N), and presence or absence of metastatic disease (M). The KPS is a measure of a patient’s general well-being and activities of daily life.] Analyses investigated whether there was any association between any of these variables and country of diagnosis. Standard Pearson chi-square analysis for r x c contingency tables was carried out. Four examples were chosen to illustrate variation in the way that the location of the elbow might be visually and subjectively judged.

### Example 1

Table 2a is the contingency table of Country vs. KPS and displays, respectively, each cell's frequency, deviation from expected ( $O_{ij}-E_{ij}$ ), cell chi-square ( $X^2_{ij}=[O_{ij}-E_{ij}]^2/E_{ij}$ ), and row percent (frequency relative to the row total). As shown in the footnote to Table 2a,  $X^2 = 97.72$ ,  $df = 15$ ,  $p < 0.0001$  and the Fisher exact test yields  $p < 0.0001$ .

**Table 2a.** Country vs. KPS, including frequency, deviation, cell chi-square and row percent.

	50	60	70	80	90	100	Total
China	0	0	8.0000	24.0000	52.0000	15.0000	99
	-0.1920	-0.1920	0.5174	-2.2850	0.7733	1.3779	
	0.1919	0.1919	0.0358	0.1986	0.0117	0.1394	
	0	0	8.0800	24.2400	52.5300	15.1500	
Korea	0	0	8.0000	51.0000	111.0000	8.0000	178
	-0.3450	-0.3450	-5.4530	3.7403	18.8950	-16.4900	
	0.3450	0.3450	2.2106	0.2960	3.8764	11.1050	
	0	0	4.4900	28.6500	62.3600	4.4900	
Thailand	1.0000	0	19.0000	48.0000	41.0000	9.0000	118
	0.7713	-0.2290	10.0810	16.6710	-20.0600	-7.2360	
	2.6016	0.2287	11.3960	8.8705	6.5893	3.2252	
	0.8500	0	16.1000	40.6800	34.7500	7.6300	
Taiwan	0	1.0000	4.0000	14.0000	63.0000	39.0000	121
	-0.2340	0.7655	-5.1450	-18.1300	0.3895	22.3510	
	0.2345	2.4990	2.8949	10.2270	0.0024	30.0050	
	0	0.8300	3.3100	11.5700	52.0700	32.2300	
<b>Total</b>	1	1	39	137	267	71	516

**Note.**  $X^2=97.72$ ,  $df=15$ ,  $p<0.0001$  and Fisher exact test  $p<0.0001$ . The top entry in each cell is the frequency count; the second entry is the cell deviation ( $O-E$ ); the third entry is the cell contribution to chi-square  $[(O-E)^2/E]$ ; the last entry is the "row percent," which is the cell percentage based on the row total.

Table 2b contains the same information as Table 2a (in a list format), where the percent contribution to chi-square of each cell has been computed ( $P_{ij}=100*X^2_{ij}/X^2$ ), the table has been sorted by decreasing  $P_{ij}$ , and the cumulative percent contributions have been computed.

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

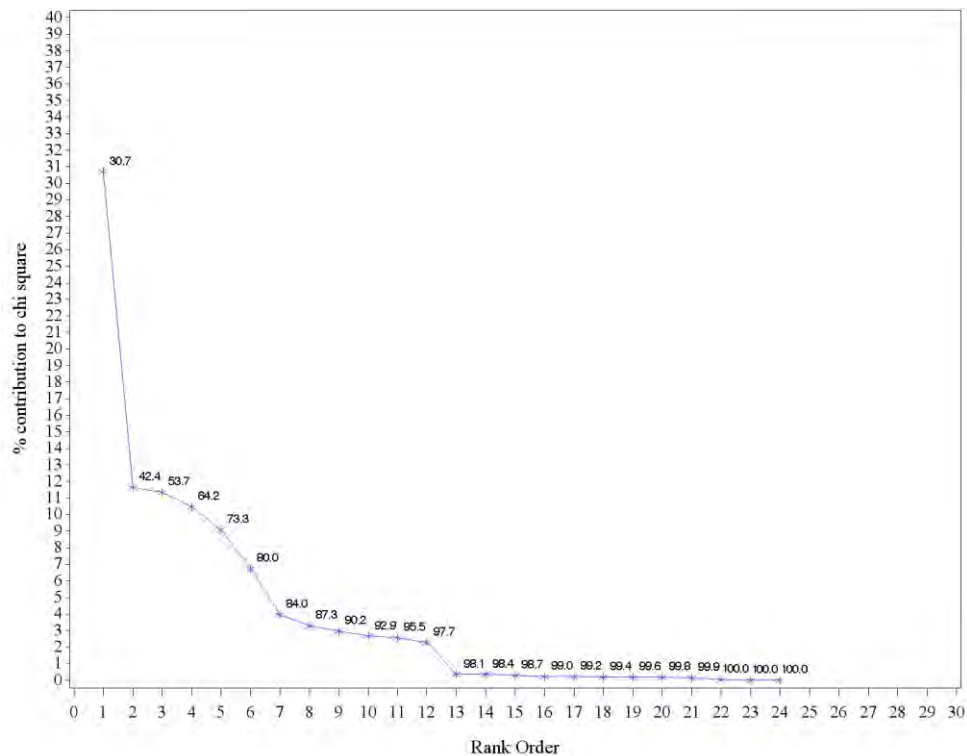
**Table 2b.** Country vs. KPS, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing  $P_{ij}$

Rank	Country	KPS	Cell Chi-Square	Deviation (O-E)	% Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Taiwan	100	30.0048	22.3508	32.2314	30.7046	30.7046
2	Thailand	70	11.3958	10.0814	16.1017	11.6616	42.3661
3	Korea	100	11.1053	-16.4922	4.4944	11.3643	53.7305
4	Taiwan	80	10.2270	-18.1260	11.5702	10.4655	64.1959
5	Thailand	80	8.8705	16.6705	40.6780	9.0773	73.2733
6	Thailand	90	6.5893	-20.0581	34.7458	6.7429	80.0162
7	Korea	90	3.8764	18.8953	62.3596	3.9668	83.9830
8	Thailand	100	3.2252	-7.2364	7.6271	3.3004	87.2834
9	Taiwan	70	2.8949	-5.1453	3.3058	2.9624	90.2458
10	Thailand	50	2.6016	0.7713	0.8475	2.6622	92.9081
11	Taiwan	60	2.4990	0.7655	0.8264	2.5572	95.4653
12	Korea	70	2.2106	-5.4535	4.4944	2.2622	97.7275
13	Korea	50	0.3450	-0.3450	0	0.3530	98.0805
14	Korea	60	0.3450	-0.3450	0	0.3530	98.4335
15	Korea	80	0.2960	3.7403	28.6517	0.3029	98.7364
16	Taiwan	50	0.2345	-0.2345	0	0.2400	98.9764
17	Thailand	60	0.2287	-0.2287	0	0.2340	99.2104
18	China	80	0.1986	-2.2849	24.2424	0.2033	99.4136
19	China	50	0.1919	-0.1919	0	0.1963	99.6100
20	China	60	0.1919	-0.1919	0	0.1963	99.8063
21	China	100	0.1394	1.3779	15.1515	0.1426	99.9489
22	China	70	0.0358	0.5174	8.0808	0.0366	99.9856
23	China	90	0.0117	0.7733	52.5253	0.0119	99.9975
24	Taiwan	90	0.0024	0.3895	52.0661	0.0025	100.0000

Figure 1 displays the corresponding scree plot where each  $P_{(i)}$  is plotted on the vertical axis against its rank order and the plot is further annotated with the respective cumulative cell percentages. Visual inspection of the scree plot (Figure 1) does not reveal a clear cut turning point. Depending on the observer's perspective, rank 2, 7, or 13 could be considered the turning point. Based on the more objective Euclidean distance method, the turning point corresponds to rank 7. (The calculation of each cell's Euclidean distance was deliberately omitted from each table in order to let the reader better appreciate the shortcomings of the visual process of finding the elbow, without being biased by knowing the corresponding distances. For the record, the squared distances for the first 10

ordered cells were 943.8, 140.0, 138.1, 125.5, 107.4, 81.5, 64.7, 74.9, 89.8, and 107.1, with the minimum (64.7) occurring at rank 7.)

Referring back to Table 2b one can examine the ranks of the cells corresponding to ranks 1 through 7 to identify those cells in the table that deviate the most from their expected values, as well as the direction of their deviation under the null hypothesis of independence, in order to better understand the nature of the association. Taiwan appears to have an overrepresentation of patients with KPS 100, while Korea's frequency is less than expected. Patients with KPS 80 tend to be underrepresented in Taiwan, but overrepresented in Thailand. Patients with KPS 90 tend to be underrepresented in Thailand and overrepresented in Korea. Finally, patients with KPS 70 tend to be overrepresented in Thailand.



**Figure 1.** Scree plot of Country vs. KPS data in Table 2.  $P_{(i)}$  is plotted on the vertical axis against its rank order; the plot is annotated with the respective cumulative cell percentages (rounded up).

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

### Example 2

Tables 3a and 3b show the relevant calculations for the association between Country and T stage. In this example, the association is not significant ( $X^2=22.29$ ,  $df=15$ ,  $p=0.10$ , and the Fisher exact test yields  $p=0.085$ .) Although not significant and the general shape of the curve is similar to that in Figure 1, consider this example to show that it may still be of interest to apply the proposed method to discover patterns in the data.

**Table 3a.** Country vs. Tumor stage, including frequency, deviation, cell chi-square and row percent.

	T0	T1	T2	T3	T4	TX	Total
China	0	3	26	19	43	9	100
	-1.758	-1.297	-0.563	-1.508	1.3984	3.7266	
	1.7578	0.3914	0.0119	0.1109	0.047	2.6334	
	0	3	26	19	43	9	
Korea	4	9	47	31	73	8	172
	0.9766	1.6094	1.3125	-4.273	1.4453	-1.07	
	0.3154	0.3505	0.0377	0.5177	0.0292	0.1263	
	2.33	5.23	27.33	18.02	42.44	4.65	
Thailand	5	6	28	22	48	9	118
	2.9258	0.9297	-3.344	-2.199	-1.09	2.7773	
	4.1269	0.1705	0.3567	0.1999	0.0242	1.2396	
	4.24	5.08	23.73	18.64	40.68	7.63	
Taiwan	0	4	35	33	49	1	122
	-2.145	-1.242	2.5938	7.9805	-1.754	-5.434	
	2.1445	0.2943	0.2076	2.5455	0.0606	4.589	
	0	3.28	28.69	27.05	40.16	0.82	
Total	9	22	136	105	213	27	512

**Note.**  $X^2=22.29$ ,  $df=15$ ,  $p=0.10$ ; Fisher exact test  $p=0.085$ . The top entry in each cell is the frequency count; the second entry is the cell deviation ( $O-E$ ); the third entry is the cell contribution to chi-square  $[(O-E)^2 / E]$ ; the last entry is the "row percent," which is the cell percentage based on the row total.

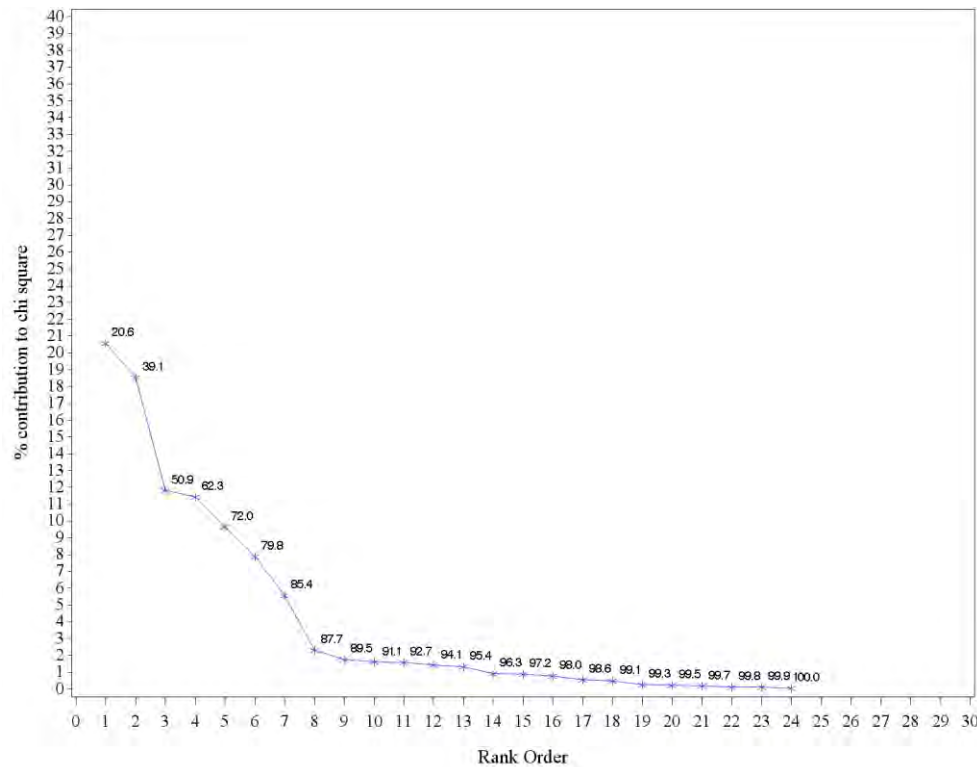
# LESSER & AKERMAN

**Table 3b.** Country vs. Tumor stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing  $P_{ij}$

Rank	Country	Tumor	Cell Chi-Square	Deviation (O-E)	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Taiwan	TX	4.58903	-5.43359	0.8197	20.5890	20.589
2	Thailand	T0	4.12695	2.92578	4.2373	18.5159	39.105
3	China	TX	2.63344	3.72656	9.0000	11.8151	50.920
4	Taiwan	T3	2.54553	7.98047	27.0492	11.4207	62.341
5	Taiwan	T0	2.14453	-2.14453	0	9.6216	71.962
6	China	T0	1.75781	-1.75781	0	7.8866	79.849
7	Thailand	TX	1.23961	2.77734	7.6271	5.5616	85.411
8	Korea	T3	0.51773	-4.27344	18.0233	2.3229	87.733
9	China	T1	0.39142	-1.29688	3.0000	1.7561	89.489
10	Thailand	T2	0.35671	-3.34375	23.7288	1.6004	91.090
11	Korea	T1	0.35046	1.60938	5.2326	1.5723	92.662
12	Korea	T0	0.31543	0.97656	2.3256	1.4152	94.077
13	Taiwan	T1	0.29435	-1.24219	3.2787	1.3206	95.398
14	Taiwan	T2	0.20760	2.59375	28.6885	0.9314	96.329
15	Thailand	T3	0.19986	-2.19922	18.6441	0.8967	97.226
16	Thailand	T1	0.17047	0.92969	5.0847	0.7648	97.991
17	Korea	TX	0.12630	-1.07031	4.6512	0.5666	98.558
18	China	T3	0.11086	-1.50781	19.0000	0.4974	99.055
19	Taiwan	T4	0.06061	-1.75391	40.1639	0.2719	99.327
20	China	T4	0.04701	1.39844	43.0000	0.2109	99.538
21	Korea	T2	0.03771	1.31250	27.3256	0.1692	99.707
22	Korea	T4	0.02919	1.44531	42.4419	0.1310	99.838
23	Thailand	T4	0.02420	-1.08984	40.6780	0.1086	99.947
24	China	T2	0.01191	-0.56250	26.0000	0.0534	100.000

In the scree plot for this example (Figure 2), the bend in the elbow is more obvious than in Figure 1 and appears to be at rank 8. This is confirmed using the Euclidean distance method.

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS



**Figure 2.** Scree plot of Country vs. Tumor stage data in Table 3.

Referring back to Table 3b, it appears that the departures are explained by the frequency distribution of unclassified (TX) and *in situ* (T0) tumors primarily among China, Thailand, and Taiwan. Furthermore, the direction of the deviation from each country can be seen in the column labeled Deviation. China and Thailand appear to have more TX tumors than expected, while Taiwan's frequency is decreased. T0 tumors tend to be underrepresented in China and Taiwan, but overrepresented in Thailand.

Even though the observed association between Country and T stage was not significant (Fisher's  $p=0.085$ ), the observed pattern may still be of clinical interest.

### Example 3

Tables 4a and 4b show the relevant calculations for the association between Country and N stage. In this example, the association is significant ( $X^2=33.96$ ,  $df=12$ ,  $p=0.0007$ .)



# LESSER & AKERMAN

**Table 4a.** Country vs. Node stage, including frequency, deviation, cell chi-square and row percent.

	N0	N1	N2	N3	NX	Total
China	10.0000	8.0000	29.0000	43.0000	10.0000	100
	-3.0860	-0.5940	0.6797	4.5234	-1.5230	
	0.7277	0.0410	0.0163	0.5318	0.2014	
	10.0000	8.0000	29.0000	43.0000	10.0000	
Korea	23.0000	22.0000	44.0000	74.0000	9.0000	172
	0.4922	7.2188	-4.7110	7.8203	-10.8200	
	0.0108	3.5254	0.4556	0.9241	5.9070	
	13.3700	12.7900	25.5800	43.0200	5.2300	
Thailand	19.0000	6.0000	25.0000	43.0000	25.0000	118
	3.5586	-4.1410	-8.4180	-2.4020	11.4020	
	0.8201	1.6907	2.1205	0.1271	9.5615	
	16.1000	5.0800	21.1900	36.4400	21.1900	
Taiwan	15.0000	8.0000	47.0000	37.0000	15.0000	122
	-0.9650	-2.4840	12.4490	-9.9410	0.9414	
	0.0583	0.5887	4.4857	2.1054	0.0630	
	12.3000	6.5600	38.5200	30.3300	12.3000	
Total	67	44	145	197	59	512

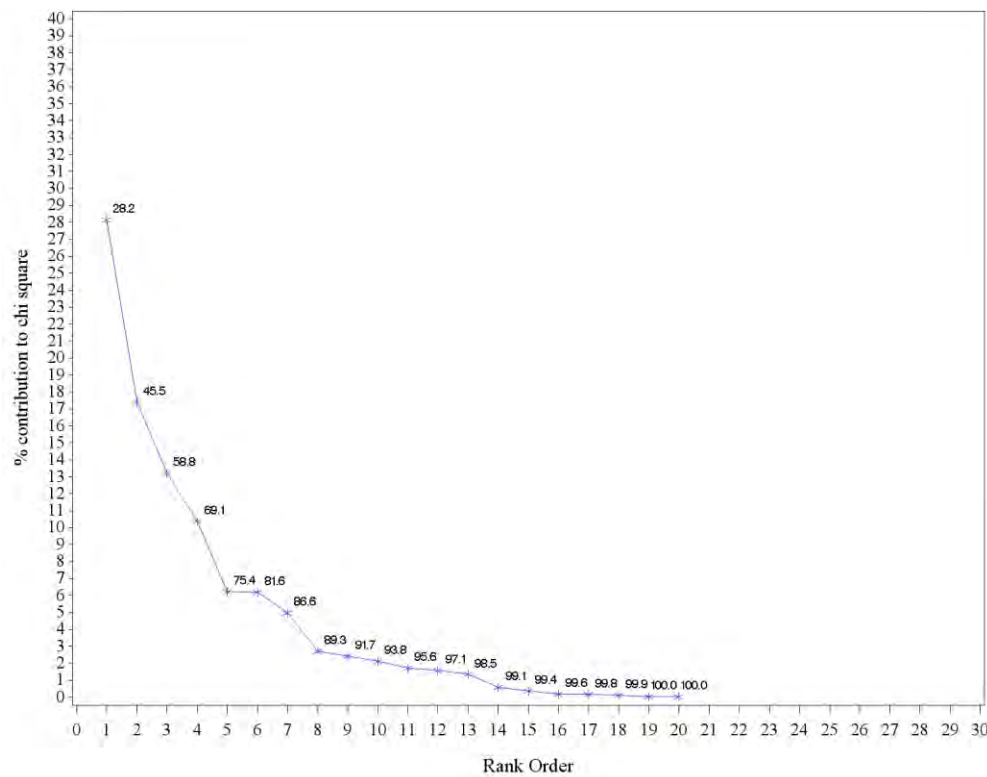
**Note.**  $X^2=33.96$ ,  $df=12$ ,  $p=0.0007$ . The top entry in each cell is the frequency count; the second entry is the cell deviation ( $O-E$ ); the third entry is the cell contribution to chi-square  $[(O-E)^2 / E]$ ; the last entry is the "row percent," which is the cell percentage based on the row total.

**Table 4b.** Country vs. Node stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing  $P_{ij}$

Rank	Country	Nodes	Cell Chi-Square	Deviation (O-E)	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Thailand	NX	9.56146	11.4023	21.1864	28.1532	28.1532
2	Korea	NX	5.90703	-10.8203	5.2326	17.3930	45.5462
3	Taiwan	N2	4.48566	12.4492	38.5246	13.2078	58.7540
4	Korea	N1	3.52544	7.2188	12.7907	10.3805	69.1345
5	Thailand	N2	2.12048	-8.4180	21.1864	6.2437	75.3781
6	Taiwan	N3	2.10542	-9.9414	30.3279	6.1993	81.5774
7	Thailand	N1	1.69070	-4.1406	5.0847	4.9782	86.5556
8	Korea	N3	0.92411	7.8203	43.0233	2.7210	89.2766
9	Thailand	N0	0.82011	3.5586	16.1017	2.4148	91.6914
10	China	N0	0.72773	-3.0859	10.0000	2.1428	93.8341
11	Taiwan	N1	0.58870	-2.4844	6.5574	1.7334	95.5675
12	China	N3	0.53179	4.5234	43.0000	1.5658	97.1334
13	Korea	N2	0.45560	-4.7109	25.5814	1.3415	98.4749
14	China	NX	0.20140	-1.5234	10.0000	0.5930	99.0679
15	Thailand	N3	0.12711	-2.4023	36.4407	0.3743	99.4422
16	Taiwan	NX	0.06304	0.9414	12.2951	0.1856	99.6278
17	Taiwan	N0	0.05831	-0.9648	12.2951	0.1717	99.7995
18	China	N1	0.04102	-0.5938	8.0000	0.1208	99.9203
19	China	N2	0.01631	0.6797	29.0000	0.0480	99.9683
20	Korea	N0	0.01076	0.4922	13.3721	0.0317	100.0000

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

Visual inspection of the scree plot (Figure 3) reveals a much smoother curve than those shown in Figures 1 and 2 and does not reveal a clear cut bending point. Using the Euclidean distance method, the turning point corresponds to rank 5. Referring back to Table 4b, it appears that the departures are explained by the frequency distribution of unclassified (NX) and N2 nodes primarily among Korea, Thailand, and Taiwan. Thailand appears to have an excess of NX nodes, while Korea's frequency is decreased. N2 nodes tend to be underrepresented in Thailand, but overrepresented in Taiwan.



**Figure 3.** Scree plot of Country vs. Node stage data in Table 4.

### Example 4

Tables 5a and 5b show the relevant calculations for the association between Country and M stage. In this example, the association is also significant ( $\chi^2=30.64$ ,  $df=6$ ,  $p<0.0001$ .)

# LESSER & AKERMAN

**Table 5a.** Country vs. Metastasis stage, including frequency, deviation, cell chi-square and row percent.

	M0	M1	MX	Total
China	26.0000	71.0000	3.0000	100
	6.4688	-2.6330	-3.8360	
	2.1425	0.0941	2.1525	
	26.0000	71.0000	3.0000	
Korea	19.0000	147.0000	6.0000	172
	-14.5900	20.3520	-5.7580	
	6.3398	3.2704	2.8196	
	11.0500	85.4700	3.4900	
Thailand	31.0000	71.0000	16.0000	118
	7.9531	-15.8900	7.9336	
	2.7445	2.9048	7.8030	
	26.2700	60.1700	13.5600	
Taiwan	24.0000	88.0000	10.0000	122
	0.1719	-1.8320	1.6602	
	0.0012	0.0374	0.3305	
	19.6700	72.1300	8.2000	
Total	100	377	35	512

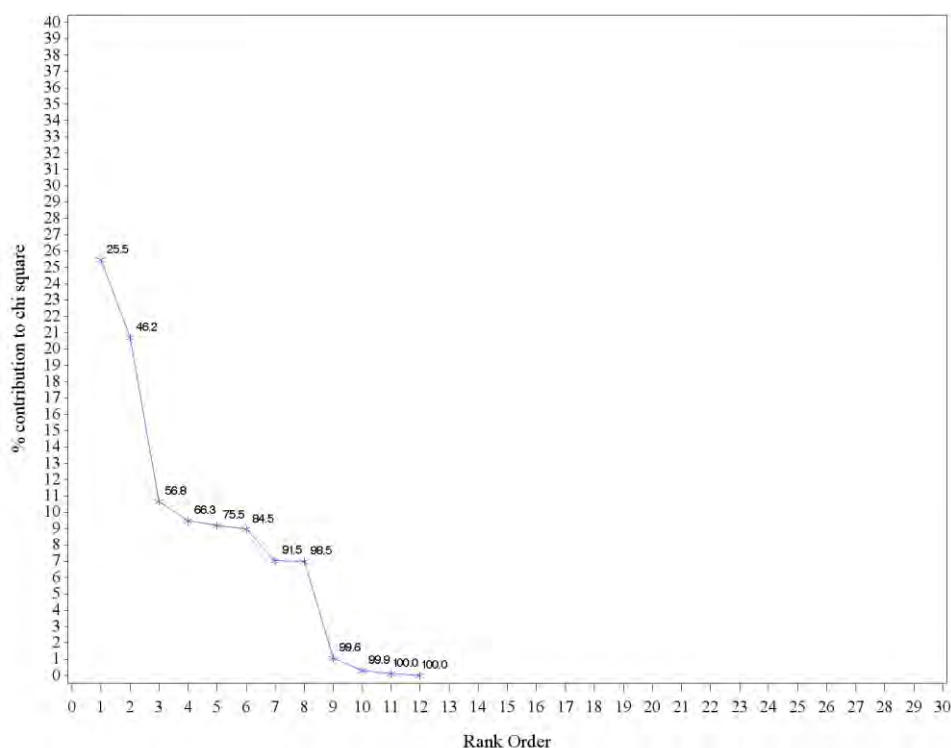
**Note.**  $X^2=30.64$ ,  $df=6$ ,  $p=0.0001$ . The top entry in each cell is the frequency count; the second entry is the cell deviation ( $O-E$ ); the third entry is the cell contribution to chi-square  $[(O-E)^2 / E]$ ; the last entry is the "row percent," which is the cell percentage based on the row total.

**Table 5b.** Country vs. Metastasis stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing  $P_{ij}$

Rank	Country	Metastasis	Cell Chi-Square	Deviation ( $O-E$ )	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Thailand	MX	7.80297	7.9336	13.5593	25.4664	25.466
2	Korea	M0	6.33980	-14.5938	11.0465	20.6911	46.158
3	Korea	M1	3.27036	20.3516	85.4651	10.6734	56.831
4	Thailand	M1	2.90479	-15.8867	60.1695	9.4803	66.311
5	Korea	MX	2.81961	-5.7578	3.4884	9.2023	75.514
6	Thailand	M0	2.74450	7.9531	26.2712	8.9572	84.471
7	China	MX	2.15251	-3.8359	3.0000	7.0251	91.496
8	China	M0	2.14245	6.4688	26.0000	6.9923	98.488
9	Taiwan	MX	0.33048	1.6602	8.1967	1.0786	99.567
10	China	M1	0.09414	-2.6328	71.0000	0.3072	99.874
11	Taiwan	M1	0.03736	-1.8320	72.1311	0.1219	99.996
12	Taiwan	M0	0.00124	0.1719	19.6721	0.0040	100.000

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

The scree plot (Figure 4) does not reveal a clear cut bending point. Either rank 3 or rank 9 could be judged as the turning points. However, using the Euclidean distance method, the turning point corresponds to rank 9. Thailand and Taiwan appear to have an excess of unknown metastases (MX), while Korea and China's frequencies are decreased. Patients with no distant metastases (M0) tend to be underrepresented in Korea, but overrepresented in Thailand and China. Patients with metastases to distant organs (M1) tend to be overrepresented in Korea but underrepresented in Thailand.



**Figure 4.** Scree plot of Country vs. Metastasis stage data in Table 5.

## Conclusion

In statistical problems involving the cross-classification of frequency counts, it is common to test for an association between one variable and another using the well-known Pearson chi-square test (or, alternatively, the Fisher exact test,

particularly for sparse tables). Upon finding a significant association, it is of interest to identify the cells in the table that are “responsible” for the lack of independence. As the dimension of the table gets larger (i.e., the number of rows and/or columns grows larger), it becomes more difficult to identify these row-column combinations.

An exploratory, graphical method of discovering those cells that account for the observed association was proposed. This method is computationally objective and completely reproducible.

The method is based on two frequently used techniques: assessment of contribution to chi-square in contingency tables and construction of scree plots as in principal components analysis. All of the computations required for applying this method are available in virtually all commonly used statistical software packages.

Several examples of  $r \times c$  tables were provided that exemplify the use of this method both when the observed associations are statistically significant and when they are not. The examples illustrate how the use of a cutoff point for the cumulative percent contribution to chi-square (“Cumulative Percent Method” as described above) is purely arbitrary. Of course, most statistical procedures include some elements of arbitrariness – most notably the use of “ $p < 0.05$ ” or “95%” for constructing confidence intervals. The examples further show that visual appraisal of the scree plot (“Subjective Elbow Method”) can be highly subjective and might, therefore, vary from one observer to another.

In order to address these shortcomings, it has been shown how the proposed Objective Elbow Method for exploring contingency tables parallels the currently accepted approach to identifying important principal components in PCA with the addition of an objective and reproducible calculation (Euclidean distance) that identifies the bend in the scree plot that constitutes the “elbow”.

As discussed in the introduction, Correspondence Analysis has been used in the current  $r \times c$  setting. While CA is a useful and powerful method, it requires somewhat specialized, albeit, readily available software (e.g., PROC CORRESP in SAS, CORRESPONDENCE module in SPSS). The proposed method, while not providing the level of detail contained in CA, is much simpler to execute, intuitively appealing to the non-statistician, and requires no more than the ability to perform standard contingency table analysis.

The use of graphical methodology as a complement to inferential analysis is widespread in statistical practice – even in the absence of statistical significance. Common examples include the already cited scree plots in PCA, scatterplots, side-by-side boxplots, receiver operating characteristic (ROC) curves, survival

## EXPLORING R X C CONTINGENCY TABLES WITH SCREE PLOTS

and hazard function curves, ANOVA interaction plots, heat maps in genetics problems, to name only a few.

This method could be readily adopted by investigators in many fields of research involving  $r \times c$  contingency tables because the ability to perform these calculations is readily available in commonly used statistical software packages.

For this manuscript, the PROC FREQ procedure in SAS Version 9.3 (SAS Institute, Cary, NC) was used. The following list shows the availability of the components of the proposed calculation in various software packages.

- SAS (SAS Institute, Cary, NC): PROC FREQ, “cellchi2” TABLE option.
- JMP (SAS Institute, Cary, NC): Contingency Table, choose the drop down labeled “Cell Chi Square”.
- Minitab (Minitab, Inc., State College, PA), Stat: Tables: Cross Tabulation and Chi-Square, check the box labeled “Each cell’s contribution to the Chi-Square statistic”
- Stata (StataCorp LP, College Station, TX): “tabulate” with the cchi2 option
- R (R Foundation for Statistical Computing, r-project.org): chisq.detail
- Excel (Microsoft Corp., Redmond, WA): programmed and calculated by user
- SPSS (IBM Inc., Armonk, NY): Crosstabs, Cells subcommand, check the box labeled “Standardized” under Residuals; contribution to cell chi-square must be programmed and calculated from these Residuals by the user

Finally, it is not proposed that the Objective Elbow Method be rigidly obeyed. This method simply provides a reproducible guidance as to which cells may be responsible for the observed association. Upon finding  $i^*$ , corresponding to the point closest to the origin, the data analyst might also want to consider points to the right of  $i^*$  but very close to it, as other potential cells of interest. Based on study results, the proposed method is believed to be potentially useful to data analysts using large  $r \times c$  tables.

## References

- Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley and Sons.
- Feinstein, A. R. (2002). *Principles of medical statistics*. Boca Raton, FL: Chapman and Hall/CRC.
- Gralla, R. J., Hollen, P., Thongprassert, S., Kim, H. K., Hsia, T. C., Yuankai, S., Kohn, N., & Lesser, M. (2013). *Accurate prediction of survival outcomes in nsccl using a new pro index from the lcsc (lung cancer symptom scale): results of a 622 patient prospective trial*. Presented at ASCO Annual Meeting, Chicago, IL, May 31 – June 4, 2013.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. J. (1992). Correspondence analysis in medical research, *Stat Methods Med Res*, 1(1): 97-117.
- Huber, W., Li, X., & Gentleman, R. (2005). Visualizing Data. In R. Gentleman, V. Carey, et al. (Eds.), *Bioinformatics and computational biology solutions using r and bioconductor* (pp. 161-179). New York: Springer Verlag.
- Khattree, R., & Naik, D. N. (2000). *Multivariate data reduction and discrimination with SAS software*. Cary, NC: SAS Institute Inc.
- Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., & Zhao, Y. (2004). Class Discovery. In *Design and analysis of DNA microarray investigations* (pp. 121-155). New York: Springer Verlag.
- Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods* (Eighth Ed.). Ames, IA: Iowa State University Press.
- Thongprassert, S., Gralla, R. J., Hollen, P., Kim, H. K., Hsia, T. C., Yuankai, S., Kohn, N., & Lesser, M. (2013). *Overcoming barriers in incorporating evaluation of quality of life (QL) and symptoms by using the EPRO version of the LCSS (ELCSS-QL) in a large-scale multinational NSCLC trial (AP-QL trial)*. Presented at ASCO Annual Meeting, Chicago, IL, May 31 – June 4, 2013.

# Bias and Precision of the Squared Canonical Correlation Coefficient Under Nonnormal Data Conditions

**Lesley F. Leach**  
Tarleton State University  
Stephenville, TX

**Robin K. Henson**  
University of North Texas  
Denton, TX

---

Monte Carlo methods were employed to investigate the effect of nonnormality on the bias associated with the squared canonical correlation coefficient ( $R_c^2$ ). The majority of  $R_c^2$  estimates were found to be extremely biased, but the magnitude of bias was impacted little by the degree of nonnormality.

*Keywords:* canonical correlation coefficient, effect size, simulation, nonnormal, canonical correlation analysis

---

## Introduction

Over the last several decades, the movement towards the use of effect size estimates in determining the importance of research results has intensified. This movement can be seen in the editorial policies of at least 25 educational and psychological journals (Wang & Thompson, 2007) that explicitly require the inclusion of effect sizes with statistical results. The sixth edition of the American Psychological Association Publication Manual (APA, 2001) deemed it “almost always necessary to include some measure of effect size” (p. 34) when reporting results. This shift has come with increased awareness that, when used alone to interpret results (i.e., without effect sizes or other statistics),  $p$ -values derived from null hypothesis significance tests (NHSTs) offer little information about the importance of results or their ability to replicate (Cumming, 2008; Henson & Smith, 2000; Kirk, 1996; Kline, 2004; Thompson, 1996, 1998). Effect size estimates offer “practical significance” information by quantifying the magnitude of a difference or relationship between variables. Consequently, numerous authors

---

*Lesley F. Leach is an Associate Professor and Coordinator of the Educational Leadership Doctoral Program in the Department of Educational Leadership and Policy Studies. Email at: leach@tarleton.edu. Robin K. Henson is a professor of Educational Psychology in the College of Education.*



and institutions have argued that effect sizes should be included with statistical results (e.g., APA, 2009, 2001; Henson, 2006; Thompson, 1996, 1998; Wilkinson & APA Task Force on Statistical Inference, 1999).

Although there are many types of effect sizes from which researchers can choose, most fall into two broad categories: (a) standardized mean difference effects and (b) measures of strength of association (Kirk, 1996; Kline, 2004; Olejnik & Algina, 2000; Onwuegbuzie, Levin, & Leech, 2003), although other statistics such as Huberty's *I* index certainly also qualify (Huberty & Lowman, 2000). Outside of the correlation coefficient, one of the most common effect sizes reported in the literature is  $R^2$ , likely due to the fact that  $R^2$  is routinely provided as part of the regression output in most statistical packages (Kirk, 1996).

There has been some debate as to whether effect sizes should be included with all NHST results, even those that are not statistically significant or only for those tests that are statistically significant (Roberts & Henson, 2002; Robinson & Levin, 1997). Some researchers have gone as far as to recommend that hypothesis tests be banned entirely (e.g. Carver, 1993) and replaced with effect size estimates or other statistics (see also Harlow, Mulaik, Steiger, 1997, for a broader discussion). These views notwithstanding, there seems to be current consensus that effect sizes can add considerable value to research interpretation.

However, effect sizes are not without their limitations and can be "subject to as much abuse and misuse as are tests for statistical significance" (Onwuegbuzie, Levin, & Leech, 2003, p. 38; see also O'Grady, 1982; Robinson & Levin, 1997). Many researchers are unaware that effect size estimates can be criticized on some of the same grounds as NHSTs, including but not limited to the fact that effects can vary according to sample size and variability, and they are often impacted by the shape of the data, including departures from normality (Knapp & Sawilowsky, 2001; Onwuegbuzie & Levin, 2003; Onwuegbuzie, Levin, & Leech, 2003). As Henson (2006) noted, "If we fail to adequately understand what our effect sizes do and do not tell us, then we may fall victim to new misconceptions about our research methods" (p. 610).

### **$R^2$ Effect Size**

For example, studies have shown  $r^2$  (Wang & Thompson, 2007; Zimmerman, Zumbo, & Williams, 2003) and its analog,  $R^2$ , to often overestimate the effect found in the population (Carter, 1979; Fan, 2001; Larson, 1931; Snyder & Lawson, 1993; Thompson, 1999; Yin & Fan, 2001). By design, the ordinary least squares estimation method commonly used in regression analyses seeks to

maximize the correlation between variables resulting in the largest possible effect size. To obtain the greatest possible effect, the analyses capitalize on *all* the variance in a given sample, including the variance attributable to sampling error (Thompson & Kieffer, 2000; Wang & Thompson, 2007). Because the effect size accounts for variability unique to the sample – variance that is unlikely to be found in the population or future samples – the resulting  $R^2$  is often a biased estimate of the effect in the population or in future samples (Roberts & Henson, 2002; Snyder & Lawson, 1993; Yin & Fan, 2001). Similar to the univariate application of  $R^2$ , studies have shown that the multivariate extension of  $R^2$ , the squared canonical correlation coefficient ( $R_c^2$ ) can be positively biased due to the influence of sampling error as well (Thompson, 1990; Thorndike & Weiss, 1973).

### Canonical Correlation Analysis

Like other multivariate methods, CCA has seen increased use in educational and psychological research, presumably due to its ability to limit experimentwise error rates and the fact that, by design, research studies using multivariate methods such as CCA often more accurately reflect the situations to which researchers wish to generalize (Fish, 1988; Henson, 1999; Sherry & Henson, 2005). Its primary purpose is to describe the relationship between synthetic composites of two sets of variables, although CCA can theoretically be extended to more than two variable sets.

Like other parametric methods, CCA applies weights, called standardized canonical function coefficients, to observed variables to create synthetic variables. The measure of effect, or canonical correlation coefficient ( $R_c$ ), is calculated as the simple bivariate correlation between the two synthetic variables (Campbell & Taylor, 1996; Henson, 2000; Sherry & Henson, 2005; Thompson, 1984, 1991). It is important to note that the goal of CCA is to maximize this correlation. It is in this optimization process, however, that sample-specific variation can become problematic because, although it was considered in determining the sample magnitude of effect, the sample-specific variance cannot be expected to exist in a new sample. Thus, one would anticipate a lower magnitude of effect in the population and/or replication with a new sample than  $R_c$  identifies. When squared, the canonical correlation ( $R_c^2$ ) represents the proportion of variance that two synthetic CCA composites linearly share (Henson, 2000; Sherry & Henson, 2005; Thompson, 1984) and, in doing so, signifies the percentage of variability in the criterion variable set that can be explained with knowledge of the predictor variable set.

### Bias in $R_c$ Estimates

Empirical studies investigating bias in  $R_c$  estimates have found mixed results. Thorndike and Weiss (1973) first investigated the impact of sampling error on the canonical correlation using data from two different sources – clients of the Minnesota Division of Vocational Rehabilitation and data from Thorndike et al. (1968) ( $N=789$  and  $505$ , respectively). The analyses were split into two studies, and subjects from both sources were randomly split into two groups each ( $n=418$  and  $371$  and  $n=246$  and  $259$ , respectively) for a total of four subgroups. The canonical correlations were compared to the cross-validated canonical correlations for each of the four subgroups. According to the authors, large differences demonstrated that sample-specific covariation could affect sample results.

Barcikowski and Stevens (1975) also investigated the effect of sampling error on the canonical correlation, but with results that differed from Thorndike and Weiss (1973). They selected 8 correlation matrices from the literature [two of which were from Thorndike and Weiss (1973)] and, using a procedure described by Huberty (1969), generated population matrices with the same properties as the selected datasets. Sample correlation matrices were generated from the population matrices, and canonical correlation analyses were performed. The number of variables ranged from 7 to 41, and the sample sizes ranged from 200-3000 in increments of 200; each sample size was replicated 100 times. The results indicated that the canonical correlations were “very stable under replication” (Barcikowski & Stevens, 1975, p. 362), even in the cases of small sample sizes (e.g., 100-200).

Thompson (1990) investigated bias in the canonical correlation that resulted in findings that conflicted with Barcikowski and Stevens (1975). Unlike the studies by Thorndike and Weiss (1973) and Barcikowski and Stevens (1975) that employed real and modeled data, respectively, Thompson used Monte Carlo methods to simulate data such that it met predetermined properties. A fully-crossed design was employed that varied the following conditions: (a) between-set correlations, (b) within-set correlations, (c) sample-size to variable ratios, and (d) variable sets. Sixty-four (i.e., 4 correlation matrices x 4 sample sizes x 4 variable sets) condition combinations were investigated. For each condition combination, 1,000 random samples were drawn and analyzed.

The ratio of subjects to variables emerged as the best predictor of bias in all six analyses (i.e., one for each of three  $R_c^2$  deviation scores and three  $R_c^2$  standard error deviations). Because the bivariate correlations between these values were positive, Thompson (1990) concluded that a greater number of subjects per

variable could potentially lead to less bias in  $R_c^2$ . Furthermore, Thompson found that, though the estimates of  $R_c$  were somewhat positively biased, the bias was minimal unless a ratio “as small as three to one” was used (p. 27). Finally, Thompson contended that even then the bias could be minimized in some situations if the value of  $R_c$  was moderate to large (e.g., greater than .40).

Thompson (1990) only examined the accuracy of  $R_c^2$  when the multivariate normality assumption of CCA was met. Whereas the normality assumption is formally required only when testing the statistical significance of canonical results (Marascuilo & Levin, 1983; Sherry & Henson, 2005; Thompson, 1984), when normality is not met, distribution shapes must still be reasonably comparable. If not, entries in the matrix of association used to derive canonical estimates may be attenuated, which could compromise the results including the magnitude of the effect (Thompson, 1984). Studies have shown, however, that few educational and psychological datasets are exactly normally distributed (Blair, 1981; Bradley, 1968, 1982; Micceri, 1989; Pearson & Please, 1975) and, as such, there is a need to investigate the performance of CCA under nonnormal data conditions to inform the use of CCA in applied studies.

### **Purpose of the Study**

As a result of the equivocal prior findings and the lack of investigation of nonnormal distributional conditions, this study compared the degree of bias associated with the squared canonical correlation coefficient ( $R_c^2$ ) gained from distributions possessing varying degrees of nonnormality to that found with multivariate normal distributions. Additional study factors were included to explore potential bias in this multivariate effect size across common conditions and to allow comparison with prior studies. Monte Carlo simulation methodology was used to fulfill this purpose.

### **Methodology**

#### **Design**

A fully-crossed design was employed in this study, manipulating the following conditions: (a) distribution shape, (b) variable sets, (c) sample sizes, (d) correlation matrices with varied between- and within-set correlations. See Table 1 for the conditions and their respective levels. Six distribution shapes were investigated, as well as 4 variable sets, 4 sample sizes, and 7 correlation matrices

(manipulating both the between- and within-set correlations), resulting in a total of 672 manipulated conditions. Five-thousand samples were drawn for each condition for a total of 3,360,000 canonical analyses

**Table 1:** Summary of Data Conditions Manipulated in the Study

Data condition	Levels Manipulated		
Distribution shape	$k = -1, 0, 1, 3, 5, 8$		
Variable Sets	$6 + 6$ ( $v=12$ ) $4 + 4$ ( $v=8$ ) $4 + 2$ ( $v=6$ ) $10 + 2$ ( $v=12$ )		
Sample size: variable ratio	3:1, 10:1, 25:1, 40:1		
Correlation matrices	Matrix	Between-set correlation	Within-set correlation
	<b>A</b>	0	0
	<b>B</b>	.1 (small)	.3 (moderate)
	<b>C</b>	.1 (small)	.5 (large)
	<b>D</b>	.3 (moderate)	.3 (moderate)
	<b>E</b>	.3 (moderate)	.5 (large)
	<b>F</b>	.5 (large)	.3 (moderate)
	<b>G</b>	.5 (large)	.5 (large)

**Note.**  $k$  denotes univariate kurtosis. The various variable sets are denoted in the following manner: no. of variables in the predictor set + no. of variables in the criterion set (total number of variables in both sets).

**Multivariate normality** The shapes of the distributions were manipulated to facilitate comparison of results under normal theory to those found under multivariate nonnormal data conditions. Specifically, this study examined the impact of varying levels of kurtosis ( $k$ ) on the squared canonical correlation coefficient. Five multivariate nonnormal datasets were generated such that all marginal distributions in each dataset possessed the following levels of univariate kurtosis: (a) negligible kurtosis ( $k = -1, 0, 1, 3$ ) and (b) moderate kurtosis ( $k = 5, 8$ ). These value ranges are consistent with studies investigating the effect of nonnormality on other sample statistics (e.g., Curran, West, & Finch, 1996; Olsson, Foss, Troye, & Howell, 2000).

It is unrealistic to expect that multivariate datasets seen in practical applications would typically possess equal univariate kurtoses across the marginal

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

distributions (Yuan & Bentler, 1997). But, for the sake of clarity and ease of interpretation, this procedure was used in this study as it has been in past investigations (e.g., Curran, West, & Finch, 1996; Fouladi, 2000; Nevitt & Hancock, 2001; Olsson, Foss, Troye, & Howell, 2000). The results from these nonnormal distributions were compared to those from a multivariate normal distribution. Because tests of variances and covariances (e.g., CCA) in normal distributions have been found to be more affected by kurtosis than skewness (Mardia, Kent, & Bibby, 1979), skewness was held constant at symmetrical (i.e., skewness = 0).

**Variable sets** We incorporated the following variable sets [denoted as the number of variables in the predictor variable set + the number in the dependent set]: (a) 6+6 ( $v=12$ ), (b) 4 + 4 ( $v=8$ ), (c) 4 + 2 ( $v=6$ ), (d) 10 + 2 ( $v=12$ ). These sets replicate the variable sets used by Thompson (1990) and represent sets that one would likely see in behavioral studies.

**Sample size to variable ratios** Sample size to variable ratios of 3, 10, 25, and 40 per variable were chosen to represent those likely seen in behavioral research. They are consistent with other studies investigating the accuracy of canonical correlation results (see, for example, Thompson, 1990).

**Correlation matrices** Six combinations of small, moderate, and large within- and between-set correlations made up the population correlation matrices in addition to a “null” model with all correlations equal to zero. Cohen’s (1988) conventions for values of  $r$  to correspond to his  $d$  benchmarks were used to determine the entries in the correlation matrix ( $r = .1$ ,  $.3$ , and  $.5$  indicating small, medium, and large effects, respectively).

It is important to note that the benchmarks provided by Cohen (1988) were not intended to be used as rigid criteria for determining result importance. Effects should always be considered in the context of the study from which they result as well as the broader literature to determine if they indicate a small, moderate, or large effect. In this article, we use the wording *small*, *moderate*, and *large* only to refer to the various effects; our choice of wording does not indicate that the various magnitudes will always represent small, moderate, and large effects, respectively. Furthermore, Cohen’s effect size rules of thumb were originally presented for use in univariate contexts. In multivariate contexts, one could conceivably expect larger effects as a result of the additional variance made available for prediction by multiple dependent variables. There is little research to

support these guidelines for multivariate outcomes, however, so the univariate approximations were used in this study.

Varying combinations of between- and within-set correlations were used to define the correlation matrices (excluding the null model with  $r_b=0$  and  $r_w=0$ ). Within-set correlations were limited to moderate (.3) and large (.5) correlations because, in a typical CCA analysis, one would often expect the correlations within the variable sets to be moderately, if not highly, correlated. Between-set correlations would likely possess a wider range, and, as such, we chose to use small (.1), medium, and large correlations in this study. The combinations for the various population correlation matrices are presented in Table 1.

### Data Generation and Analysis

Populations of data were randomly generated that mirrored the correlation matrices at the kurtosis levels previously specified. A total of 42 multivariate populations ( $N=100,000$  each) were created (i.e., all paired combinations of the 6 distribution shapes [1 multivariate normal and 5 kurtotic] and the 7 correlation matrices). See Appendix A for information regarding the data generation procedure.

Sample canonical analyses were performed using SAS® (SAS Institute, Inc., Cary, NC, [www.sas.com](http://www.sas.com)) version 9.1.3 syntax. The variance explained ( $R_c^2$ ) for each of the first three canonical functions was computed. The accuracy of  $R_c^2$  was then calculated as the difference between the sample  $R_c^2$  and population  $R_c^2$  values. The average level of accuracy, or bias, of the  $R_c^2$  estimates was calculated as the mean of the accuracy values for each condition combination, and the precision of the  $R_c^2$  estimates was represented by the standard deviation of the respective accuracy values. Bias was considered to be extreme if it exceeded  $\pm.30 R_c^2$ ; bias was considered to be minimal (and thus acceptable) if it was less than or equal to  $\pm.30 R_c^2$ .

Analysis of variance (ANOVA) was used to identify the influence of each condition on the variability of the accuracy values. The accuracy – i.e., the differences between the population  $R_c^2$ s and the sample  $R_c^2$ s – acted as the dependent variable (DV) whereas the four conditions made up the independent variables. Only main effects were considered in this study. Main effects were evaluated based on statistical significance of the  $F$  tests (alpha<.02 - value determined using the Bonferroni correction) as well as from  $\eta^2$  and  $\omega^2$  effect size values ( $\omega^2$  was included as a theoretical adjustment for sampling error).



## Results

### Accuracy and Precision of $R_c^2$

Results from the first three functions were analyzed for each correlation matrix. All function II and III results were found to be extremely biased across all correlation combinations. For this reason and for the sake of brevity, only results from the first functions are reported and discussed in the present article. Second and third function results are available from the authors upon request.

The bias and precision (*SD*) of the sample  $R_c^2$  accuracy values for correlation matrices **A** through **G** are presented in Tables A1 through A7 in Appendix B, respectively. Note that all of the condition combinations for correlation matrices **A** and **C** produced extremely biased  $R_c^2$  accuracy values. Likewise, all but two of the combinations (97.92% of 96 cases) for correlation matrix **B** produced  $R_c^2$  values that were extremely biased.

Correlation matrix **D** produced extremely biased accuracy values in only 22.92% of the 96 condition combinations. All condition combinations for correlation matrix **D** with sample size to variable ratios greater than or equal to 10:1 produced minimal amounts of bias. Conversely, the majority of the bias (91.67% of 21 cases) with a sample size to variable ratio of 3:1 were found to be extreme. Only two of the 3:1  $n:v$  ratio condition combinations produced minimal bias; all other cases met the criteria to be considered extremely biased.

Similar results were found with the correlation matrix **E** results. In this case, 43.75% of the 96 condition combinations produced extreme bias. As a general rule, the condition combinations that were found to possess minimal levels of bias had sample size to variable ratios greater than or equal to 25:1. Unlike the other correlation matrices, results from all condition combinations in correlation matrices **F** and **G** were found to contain minimal bias.

The average bias and precision values by the various condition levels are presented in Table 2. As demonstrated in the table, bias generally decreased as the sample size to variable ratio increased. The most dramatic decrease in bias was seen in the difference in bias between the sample sizes of 3:1 and 10:1 (mean difference of .19). Differences between subsequent sample size to variable ratios were comparatively small. Bias values varied across the other condition combinations.



**Table 2:** Descriptive Statistics for Function I Bias by Sample Size to Variable Ratio, Variable Set, Univariate Kurtosis Level, and Correlation Matrix

Condition	<i>M</i>	<i>SD</i>	<i>n</i>
Sample size: variable ratio ( <i>n:v</i> )			
3:1	0.26	0.17	840,000
10:1	0.07	0.07	840,000
25:1	0.03	0.04	840,000
40:1	0.02	0.03	840,000
Variable Set			
6 + 6	0.10	0.14	840,000
4 + 4	0.10	0.14	840,000
4 + 2	0.09	0.01	840,000
10 + 2	0.09	0.13	840,000
Expected Kurtosis ( <i>k</i> )			
-1	0.09	0.14	560,000
0	0.09	0.13	560,000
1	0.09	0.14	560,000
3	0.09	0.14	560,000
5	0.10	0.14	560,000
8	0.10	0.14	560,000
Correlation Matrix			
<b>A</b> ( $r_b = 0, r_w = 0$ )	0.15	0.16	480,000
<b>B</b> ( $r_b = .1, r_w = .3$ )	0.13	0.16	480,000
<b>C</b> ( $r_b = .1, r_w = .5$ )	0.14	0.16	480,000
<b>D</b> ( $r_b = .3, r_w = .3$ )	0.07	0.11	480,000
<b>E</b> ( $r_b = .3, r_w = .5$ )	0.10	0.13	480,000
<b>F</b> ( $r_b = .5, r_w = .3$ )	0.03	0.06	480,000
<b>G</b> ( $r_b = .5, r_w = .5$ )	0.04	0.08	480,000

The precision of results, or standard deviation of the accuracy values, appeared to increase (i.e., the *SD* value decreased) as the sample size to variable ratio increased. Although decreased standard errors would be expected with

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

increased sample size, a dramatic difference in precision was detected between results with a sample size to variable ratio of 3:1 versus results from an  $n:v$  ratio of 10:1 (difference of .10). The precision of results was varied across variable sets. The (6+6), (4+4), and (10+2) variable sets produced results with roughly equal amounts of precision ( $SD=.14$ , .14, and .13, respectively). But the (4+2) variable set saw extremely precise results overall ( $SD=.01$ ). Normality (or nonnormality) of the distributions seemed to matter little in the precision of results. The results had approximately the same precision regardless of the value of kurtosis (ranged from .13-.14). However, the precision of results varied by correlation matrix. Correlation matrices **F** and **G** saw greater precision (.06 and .08, respectively) than matrices **A**, **B**, **C**, **D**, and **E** (ranged from .11-.16). Because correlation matrices **F** and **G** had higher between- and within-set correlations, these results suggest that higher between- and within-set correlations may influence the precision of  $R_c^2$  estimates. But these results should be taken tentatively because they are based on descriptive analyses alone; further exploration is needed.

### Explanation of Variability in $R_c^2$ Bias

An analysis of variance (ANOVA) was run to determine which of the study factors could account for the variability in the accuracy values. The ANOVA summary table for the function I results can be found in Table 3.

**Table 3:** ANOVA Summary Table for Explanation of the Sources of Variation in Function I  $R_c^2$  Bias

Source of Variation	SS	df	MS	F	p	$\eta^2$	$\omega^2$
Expected univariate kurtosis (k)	24.90	5	4.98	702.0	<.001	<.001	<.001
Sample size: variable ratio (n:v)	31776.88	3	10592.29	1493068.8	<.001	0.51	0.51
Variable set	8.14	3	2.71	382.3	<.001	<.001	<.001
Correlation matrix	7060.75	6	1176.79	165878.3	<.001	0.11	0.11
Error	23836.75	3359982	0.01				
Total	62707.42	3359999					

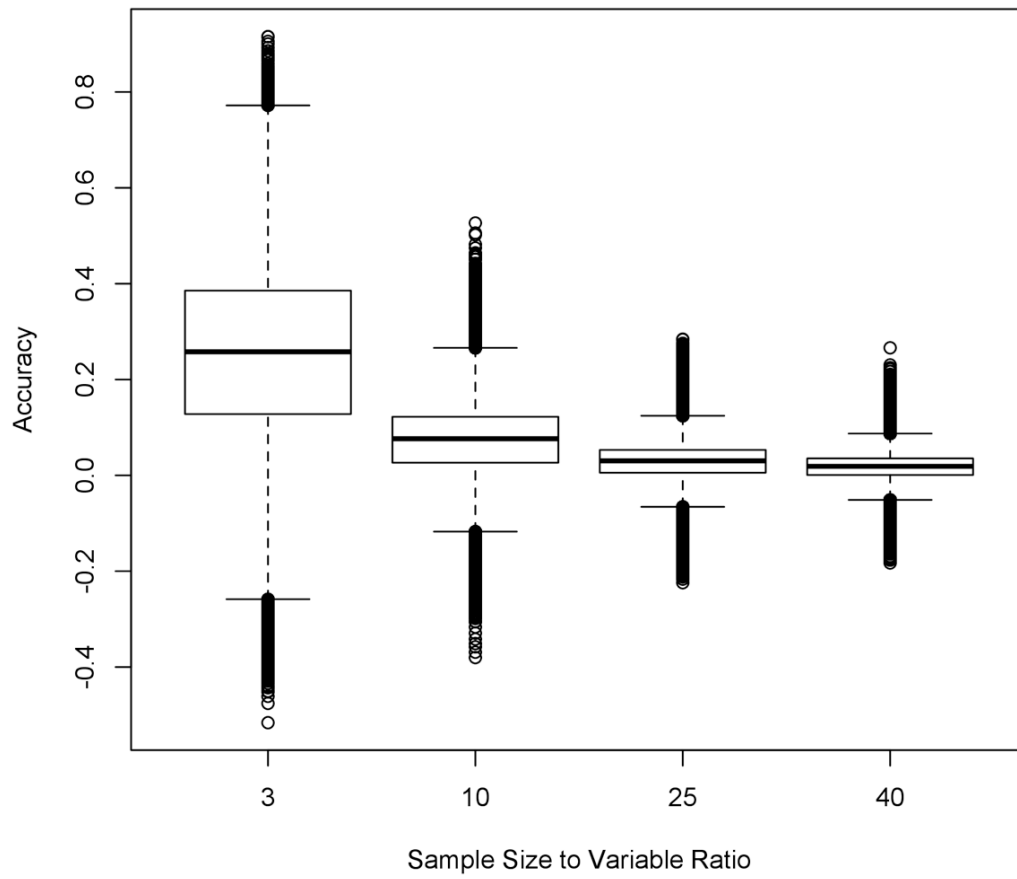
The Levene's test for homogeneity of variance was statistically significant as might be expected given the large number of simulated conditions,  $F(671, 3359328) = 2413.64$ ,  $p < .001$ . Upon visual inspection of the variances by condition, we determined that the variances were roughly homogenous and

therefore most likely met the assumption of homogeneity of variance in this balanced design. Furthermore, the equal variances assumption is primarily related to the Type I error rate involved with the  $F$  tests. Because statistical significance of the ANOVA results was not our primary interest, meeting this assumption was less of a concern for this study.

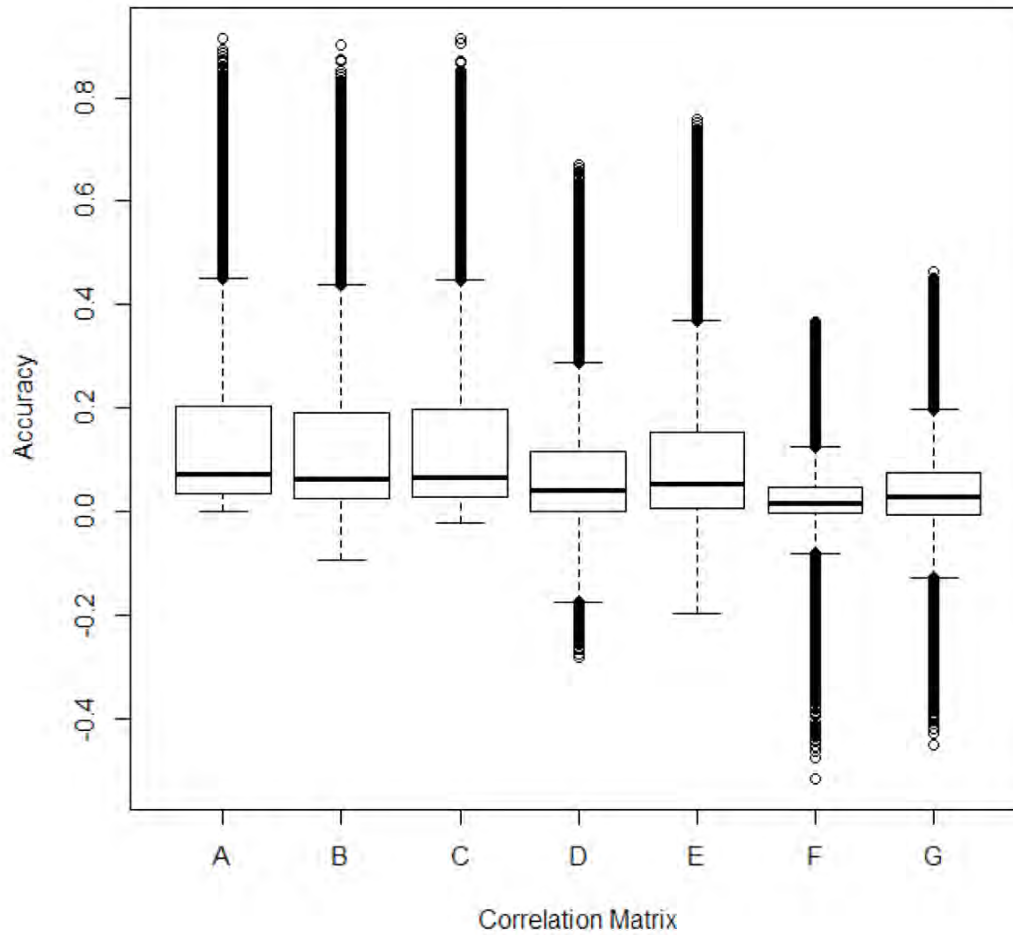
All five conditions produced statistically significant main effects with  $p < .001$  in all cases. It is apparent from examination of the  $\eta^2$  values, however, that not all of the main effects were noteworthy. The sample size to ratio variable explained the greatest amount of variation in function I bias, accounting for 51% ( $\eta^2 = .51$ ;  $\omega^2 = .51$ ) of the variation in the DV. The only other condition that had a somewhat notable effect on the DV was the correlation matrix, and it displayed a considerably weaker relationship with the DV than the sample size to variable ratio. The correlation matrix variable explained 11% ( $\eta^2 = .11$ ;  $\omega^2 = .11$ ) of the variation in function I bias.

Based on these results, it is apparent that, of the five conditions manipulated in this study, the sample size to variable ratio had the largest effect on function I bias (depicted in [Figure 1](#)). And, this effect was considerable given the fact that it could explain approximately half of the function I bias variation. It is worth noting that the 3:1 sample size to variable ratio had, by far, the greatest bias of all the ratios ( $M = .26$ ,  $SD = .17$ ), with less bias for the 10:1, 25:1, and 40:1 ratios ( $M = .07$ ,  $SD = .07$ ;  $M = .03$ ,  $SD = .04$ ; and  $M = .02$ ,  $SD = .03$ ; respectively). Larger sample size to variable ratios seemed to help decrease bias in  $R_c^2$ , particularly when  $n:v \geq 10:1$ . The correlation matrix variable demonstrated a small, but still noteworthy effect in comparison (depicted in [Figure 2](#)). The correlation matrices with larger between- and within-set correlations – correlation matrices **F** and **G** ( $r_b = .43$ ,  $r_w = .50$  and  $r_b = .50$ ,  $r_w = .50$ , respectively) – had less bias than matrices **A**, **B**, **C**, **D**, and **E**, leading to the conclusion that larger correlations may help decrease bias in  $R_c^2$ .

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT



**Figure 1:** Boxplot of function I  $R_c^2$  bias by sample size to variable ratio across all other conditions ( $N=3,360,000$ ;  $n=840,000$ ).



**Figure 2:** Boxplot of function I  $R_c^2$  bias by correlation matrix across all other conditions ( $N=3,360,000$ ;  $n=480,000$ ). Correlation matrix **A** was created with  $r_b = 0$  and  $r_w = 0$ , correlation matrix **B** with  $r_b = .1$  and  $r_w = .3$ , correlation matrix **C** with  $r_b = .1$  and  $r_w = .5$ , correlation matrix **D** with  $r_b = .3$  and  $r_w = .3$ , correlation matrix **E** with  $r_b = .3$  and  $r_w = .5$ , correlation matrix **F** with  $r_b = .5$  and  $r_w = .3$ , and correlation matrix **G** with  $r_b = .5$  and  $r_w = .5$ .

## Conclusion

Overall, a large percentage of the first function results (47.92% of 672 total combinations) across correlation matrices provided minimal amounts of bias. With the exceptions of two minimal mean differences for the normal distribution of correlation matrix **B** data (i.e., the matrix with  $r_b=.10$  and  $r_w=.30$ ), all cases

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

with minimal bias were found with the correlation matrices reporting combinations of between-set and within-set correlations that were each greater than or equal to .30 (i.e., correlation matrices **D**, **E**, **F**, and **G**).

As demonstrated by the ANOVA results, the accuracy of  $R_c^2$  was largely impacted by the sample size to variable ratio. This can be seen in the bias values; as the sample size to variable ratio increased, bias consistently decreased. This finding is not surprising given the impact of both sample size and the number of variables on the theoretical amount of sampling error present. As sample size increases, sampling error theoretically decreases. The number of variables in a model typically has the opposite effect on sampling error; as the number of variables increases, so does the theoretical amount of sampling error present in a sample. It logically follows that a larger sample size to variable ratio would likely help decrease the amount of sampling error (i.e., bias) in canonical results.

The current results speak directly to the sample size needed to obtain reasonable outcomes from CCA analyses. It is apparent from the descriptive statistics in Table 2 that, across all condition combinations, the 3:1 sample size to variable ratio produced substantial bias; bias dramatically decreased when a sample size to variable ratio of at least 10:1 was used. Larger sample sizes (e.g., 25:1 and 40:1) produced even less biased results, further demonstrating the fact that larger sample size to variable ratios are ideal conditions for decreasing bias in  $R_c^2$ . This finding echoes that found by Thompson (1990), who found the sample size to variable ratio to be the best predictor of bias in the squared canonical correlation coefficient. Furthermore, Thompson found that, though the estimates of  $R_c$  were somewhat positively biased, the bias was minimal unless a ratio of “as small as three to one” was used (p. 27). This was also the case in this study. Dramatic decreases in bias were seen between the sample size to variable ratios of 3:1 and 10:1. Thompson pointed out, however, that, with a small sample size to variable ratio (e.g.,  $n:v=3:1$ ), the bias could be minimized in some situations if the value of  $R_c$  was moderate to large. As can be seen across the matrices, this was somewhat true for the data with higher between- and within-set correlations in this study.

None of the other conditions, including the marginal kurtosis level, notably impacted the accuracy of  $R_c^2$ . These results mirror Barcikowski and Stevens (1975) and Thompson (1990) that involved normal distributions. Keep in mind, however, that this study was limited to negligibly and moderately kurtotic distributions; data with more extreme kurtosis could have a differential effect on  $R_c^2$  estimates.

In this study, precision of the  $R_c^2$  values was examined only through descriptive analyses (i.e., standard deviation of the differences between the population and sample  $R_c^2$  values) because there is only one value in each cell for all replications. Although conclusions are limited as a result of the descriptive analyses, some general comments about the precision of  $R_c^2$  can be made. As with the accuracy of the  $R_c^2$  estimates, it appears that the precision of  $R_c^2$  may increase as the  $n:v$  ratio increases. This is logical given the effect of larger samples and fewer variables on sampling error. When sample size is maximized and the number of variables are minimized, a greater  $n:v$  ratio will likely produce more precise  $R_c^2$  results. The pattern of results by the variable sets is somewhat unclear and needs further investigation. More often than not, the (6 + 6) and (10 + 2) produced the most precise results. And, for correlation matrices **A**, **B**, **C**, **D**, and **E**, the precision values generally remained the same for the various marginal kurtosis levels. Matrices **F** and **G** with higher between- and within-set correlations saw greater precision by comparison. But, because these results are based on descriptive analyses alone, these results should be taken tentatively and should likely only be used to inform future studies.

### Recommendations for Practice

Based on the results of the study, several recommendations are warranted in the use of canonical correlation analyses in educational and psychological research. First, it is recommended that a sample size to variable ratio of at least 10:1 be used in CCA analyses to lessen the bias that may affect  $R_c^2$  results. As was seen in the descriptive statistics presented in Table 2, under these study conditions, using an  $n:v$  ratio of 10:1 versus 3:1 led to dramatic reductions in bias. It would not be unlikely to expect similar results in applied studies under similar conditions.

Greater sample size to variable ratios may also provide more precise results as well. Because larger sample size to variable ratios reduce bias even more, however, researchers are encouraged to use the largest sample that is available to them and the fewest variables that will adequately represent their model. Maximizing the sample size and minimizing the number of study variables will help to increase the  $n:v$  ratio and subsequently likely reduce bias and increase precision in the results.

Second, because the univariate kurtosis level was shown to not substantially impact results, researchers can be relatively confident that, when  $k$  is homogenous across variables and within the range of -1 to 8,  $R_c^2$  bias is not likely to be greater or less than that that would be found with results from a normal distribution.

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

Therefore, it is reasonable that multivariate distributions that consist of moderately kurtotic univariate distributions can be treated in the same manner as normal distributions. These results may be encouraging to applied researchers given the fact that in practice, educational and psychological distributions are rarely exactly normally distributed (Blair, 1981; Bradley, 1968, 1982; Micceri, 1989; Pearson & Please, 1975). Resulting  $R_c^2$  values are likely to be accurate in cases even with data that are moderately kurtotic.

Recommendations should be heeded with the limitations of the study in mind, however. Because the data were simulated, we were not able to model every conceivable condition that could impact the squared canonical correlation coefficient. Further research could seek to extend this study with a larger range of population effect sizes, sample sizes, distributional shapes, and numbers of variables.

Despite its limitations, the findings from this study revealed important conditions to consider in the use of the squared canonical correlation coefficient, particularly under nonnormal data conditions. These findings and recommendations are meant to impact research practice and provide more accurate applications of canonical correlation analysis, particularly as regards the use of the squared canonical correlation coefficient.

## References

- American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.
- American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Barcikowski, R. S., & Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. *Multivariate Behavioral Research, 10*, 353-364.
- Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research, 51*, 499-507.
- Bradley, J. V. (1968). *Distribution free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.



- Bradley, J. V. (1982). The insidious L-shaped distribution. *Bulletin of the Psychonomic Society*, 20: 85-88.
- Campbell, K. T., & Taylor, D. L. (1996). Canonical correlation analysis as a general linear model: A heuristic lesson for teachers and students. *The Journal of Experimental Education*, 64: 157-172.
- Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating the population multiple correlation coefficients. *Educational and Psychological Measurement*, 39: 261-266.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education*, 61: 287-292.
- Cohen, J. (1988). *Statistical power analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3: 286-300.
- Curran, P. J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1: 16-29.
- Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, 94: 275-282.
- Fan, X., Felsövályi, Á., Sivo, S.A., Keenan, S.C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.
- Fish, L. J. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development*, 21: 130-137.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43: 521-532.
- Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, 7: 356-410.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.
- Henson, R. K. (1999). Multivariate normality: What is it and how is it assessed? In B. Thompson (Ed.), *Advances in Social Science Methodology*, Vol. 5 (pp.193-211). Stamford, CT: JAI Press.

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

Henson, R. K. (2000). Demystifying parametric analyses: Illustrating canonical correlation analysis as the multivariate general linear model. *Multiple Linear Regression Viewpoints*, 26: 11-19.

Henson, R. K. (2006). Effect size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, 34: 601-629.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the APA Task Force Report and current trends. *Journal of Research and Development in Education*, 33: 285-296.

Huberty, C. J. (1969). *An empirical comparison of selected classification rules in multiple discriminant analysis*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, 60: 543-563.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 27: 179-182.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56: 746-759.

Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, 70: 65-79.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, D.C.: American Psychological Association.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22: 45-55.

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105: 156-166.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8: 353-377.

- O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, 92: 766-777.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25: 241-286.
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7: 557-595.
- Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods*, 2: 133-151.
- Onwuegbuzie, A. J., Levin, J. R., & Leech, N. L. (2003). Do effect-size measures measure up?: A brief assessment. *Learning Disabilities: A Contemporary Journal*, 1(1): 37-40.
- Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, 62: 223-241.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, 62: 241-253.
- Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, 26(5): 21-26.
- Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, 84: 37-48.
- Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, 61: 334-349.
- Thompson, B. (1984). *Canonical correlation analyses: Uses and interpretation*. Thousand Oaks, CA: Sage.
- Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. *Educational and Psychological Measurement*, 50: 15-31.

## BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. *Measurement and Evaluation in Counseling and Development*, 24(2): 80-96.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55: 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25(2): 26-30.

Thompson, B. (1998, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, 9: 165-181.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education*, 70: 80-93.

Thompson, B., & Kieffer, K. M. (2000). Interpreting statistical significant test results: A proposed new "what if" method. *Research in the Schools*, 7(2): 3-10.

Thorndike, R. M., & Weiss, D. J. (1973). A study of the stability of canonical correlations and canonical components. *Educational and Psychological Measurement*, 36: 861-878.

Thorndike, R. M., Weiss, D. J., & Davis, R. V. (1968). Multivariate relationships between a measure of vocational interests and a measure of vocational needs. *Journal of Applied Psychology*, 52: 491-496.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, 51: 473-481.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48, 465-471.

Wang, Z., & Thompson, B. (2007). Is the Pearson  $r^2$  biased, and, if so, what is the best correction formula? *The Journal of Experimental Education*, 75: 109-125.

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54: 594-604.

Yin, P., & Fan, X. (2001). Estimating  $R^2$  shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education*, 69: 203-224.

Yuan, K. H., & Bentler, P. M. (1997). Generating multivariate distributions with specified marginal skewness and kurtosis. In W. Bandilla and F. Faulbaum (Eds.), *SoftStat' 97-Advances in Statistical Software 6* (pp. 385-391). Stuttgart, Germany: Lucius and Lucius.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica*, 24: 133-158.

## Appendix A

Procedures presented by Fleishman (1978), Kaiser and Dickman (1962), Vale and Maurelli (1983) were used to generate the multivariate random distributions in this study. More extreme values of kurtosis (e.g.,  $k = 15, 25$ ) were considered, but rejected because the data generation procedure could not produce distributions that contained the desired levels of nonnormality.

Forty-two populations (i.e., one for each combination of the 6 kurtosis levels and 7 correlation matrices) were generated, and sample canonical analyses were performed using SAS<sup>®</sup> (SAS Institute, Inc., Cary, NC, [www.sas.com](http://www.sas.com)) version 9.1.3 syntax. It is important to note that correlation matrix **F** was intended to have  $r_b=.3$  and  $r_w=.5$ , but the resulting matrix was not of full rank. For that reason, we generated correlation matrix **F** to have  $r_b=.43$  and  $r_w=.5$ , the correlations that were the closest to the intended values that would generate a matrix of full rank. The syntax was written by the authors using the reference by Fan, Felsővályi, Sivo, and Keenan (2002). For the sake of brevity, the syntax was not included in this article; copies can be obtained from the authors.

Several checks were incorporated into the SAS<sup>®</sup> code to insure its accuracy. First, the variables in each of the populations were checked to make sure that they truly approximated the pre-specified correlations and kurtosis levels. Second, values of the condition variables were saved for each of the 3,360,000 canonical analyses so that they could be compared with the expected values. Third, calculations for randomly selected cases were manually checked to verify their accuracy.

## Appendix B

**Table A1:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **A**

Var. Set	n: v	Expected k					
		-1	0	1	3	5	8
v=12 (6+6)	3.1	<b>.45<sup>a</sup></b> (.09) <sup>b</sup>	<b>.45</b> (.09)	<b>.45</b> (.09)	<b>.45</b> (.09)	<b>.45</b> (.09)	<b>.45</b> (.10)
	10.1	<b>.14</b> (.04)	<b>.14</b> (.03)	<b>.14</b> (.04)	<b>.14</b> (.04)	<b>.14</b> (.04)	<b>.14</b> (.04)
	25.1	<b>.06</b> (.02)	<b>.06</b> (.02)	<b>.06</b> (.02)	<b>.06</b> (.02)	<b>.06</b> (.02)	<b>.06</b> (.02)
	40.1	<b>.04</b> (.01)	<b>.03</b> (.01)	<b>.04</b> (.01)	<b>.04</b> (.01)	<b>.04</b> (.01)	<b>.04</b> (.01)
v=8 (4+4)	3.1	<b>.41</b> (.12)	<b>.41</b> (.12)	<b>.41</b> (.12)	<b>.41</b> (.12)	<b>.41</b> (.12)	<b>.41</b> (.13)
	10.1	<b>.13</b> (.05)	<b>.13</b> (.05)	<b>.13</b> (.05)	<b>.13</b> (.05)	<b>.13</b> (.05)	<b>.13</b> (.05)
	25.1	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)
	40.1	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)
v=6 (4+2)	3.1	<b>.36</b> (.14)	<b>.36</b> (.15)	<b>.36</b> (.14)	<b>.36</b> (.14)	<b>.36</b> (.15)	<b>.36</b> (.15)
	10.1	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.11</b> (.06)
	25.1	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)
	40.1	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)
v=12 (10+2)	3.1	<b>.38</b> (.10)	<b>.38</b> (.10)	<b>.38</b> (.10)	<b>.38</b> (.10)	<b>.38</b> (.10)	<b>.38</b> (.11)
	10.1	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.12</b> (.04)	<b>.12</b> (.04)	<b>.12</b> (.04)
	25.1	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)
	40.1	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)	<b>.03</b> (.01)

**Note.** Correlation matrix **A** was created with  $r_w=0$  and  $r_b=0$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm .30 R_c^2$  criterion for extreme bias. <sup>a</sup>The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup>The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

# BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

**Table A2:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **B**

Var. Set	n: v	Expected k					
		-1	0	1	3	5	8
v=12 (6+6)	3.1	<b>.41</b> <sup>a</sup> (.09) <sup>b</sup>	<b>.41</b> (.09)	<b>.41</b> (.09)	<b>.41</b> (.09)	<b>.41</b> (.09)	<b>.41</b> (.10)
	10.1	<b>.11</b> (.05)	<b>.11</b> (.04)	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.12</b> (.05)	<b>.12</b> (.05)
	25.1	<b>.04</b> (.03)	<b>.04</b> (.04)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)
v=8 (4+4)	3.1	<b>.39</b> (.12)	<b>.38</b> (.12)	<b>.38</b> (.12)	<b>.38</b> (.12)	<b>.39</b> (.13)	<b>.39</b> (.13)
	10.1	<b>.11</b> (.05)	<b>.11</b> (.05)	<b>.11</b> (.06)	<b>.11</b> (.06)	<b>.11</b> (.06)	<b>.11</b> (.06)
	25.1	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.03)
v=6 (4+2)	3.1	<b>.34</b> (.14)	<b>.34</b> (.15)	<b>.34</b> (.15)	<b>.34</b> (.15)	<b>.34</b> (.15)	<b>.34</b> (.16)
	10.1	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.07)
	25.1	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.04)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.03)
v=12 (10+2)	3.1	<b>.36</b> (.10)	<b>.30</b> (.10)	<b>.36</b> (.10)	<b>.36</b> (.10)	<b>.36</b> (.11)	<b>.36</b> (.11)
	10.1	<b>.10</b> (.05)	<b>.08</b> (.06)	<b>.10</b> (.05)	<b>.10</b> (.05)	<b>.10</b> (.05)	<b>.10</b> (.05)
	25.1	<b>.04</b> (.03)	<b>.03</b> (.04)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.03)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)

**Note.** Correlation matrix **B** was created with  $r_w=.3$  and  $r_b=.1$   $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm .30 R_c^2$  criterion for extreme bias. <sup>a</sup>The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup>The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.



**Table A3:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **C**

Var. Set	$n: v$	Expected $k$					
		-1	0	1	3	5	8
$v=12$ (6+6)	3.1	<b>.43<sup>a</sup></b> (.09) <sup>b</sup>	<b>.43</b> (.09)	<b>.42</b> (.09)	<b>.43</b> (.09)	<b>.43</b> (.10)	<b>.44</b> (.10)
	10.1	<b>.13</b> (.04)	<b>.13</b> (.04)	<b>.13</b> (.04)	<b>.13</b> (.04)	<b>.13</b> (.04)	<b>.13</b> (.05)
	25.1	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.04</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)	<b>.05</b> (.02)
	40.1	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)
$v=8$ (4+4)	3.1	<b>.40</b> (.12)	<b>.40</b> (.12)	<b>.40</b> (.12)	<b>.40</b> (.12)	<b>.40</b> (.12)	<b>.40</b> (.13)
	10.1	<b>.12</b> (.05)	<b>.12</b> (.05)	<b>.11</b> (.05)	<b>.12</b> (.05)	<b>.12</b> (.06)	<b>.12</b> (.05)
	25.1	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)
	40.1	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.03</b> (.02)	<b>.02</b> (.02)	<b>.03</b> (.02)
$v=6$ (4+2)	3.1	<b>.35</b> (.14)	<b>.35</b> (.14)	<b>.35</b> (.15)	<b>.35</b> (.15)	<b>.35</b> (.15)	<b>.35</b> (.15)
	10.1	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)	<b>.10</b> (.06)
	25.1	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)	<b>.04</b> (.03)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)
$v=12$ (10+2)	3.1	<b>.37</b> (.10)	<b>.37</b> (.10)	<b>.37</b> (.10)	<b>.37</b> (.10)	<b>.37</b> (.10)	<b>.37</b> (.11)
	10.1	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.11</b> (.04)	<b>.11</b> (.05)
	25.1	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)	<b>.04</b> (.02)
	40.1	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)	<b>.02</b> (.02)

**Note.** Correlation matrix **C** was created with  $r_a=.5$  and  $r_b=.1$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm .30R_c^2$  criterion for extreme bias. <sup>a</sup>The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup>The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

# BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

**Table A4:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **D**

Var. Set	n: v	Expected $k$					
		-1	0	1	3	5	8
v=12 (6+6)	3.1	.15 <sup>a</sup> (.09) <sup>b</sup>	<b>.16</b> (.09)	.15 (.09)	<b>.16</b> (.09)	<b>.16</b> (.09)	<b>.17</b> (.09)
	10.1	.04 (.06)	.04 (.06)	.04 (.06)	.04 (.06)	.05 (.06)	.05 (.06)
	25.1	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)
	40.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)
v=8 (4+4)	3.1	<b>.19</b> (.12)	<b>.19</b> (.12)	<b>.19</b> (.12)	<b>.20</b> (.13)	<b>.20</b> (.13)	<b>.21</b> (.13)
	10.1	.05 (.08)	.05 (.08)	.05 (.08)	.05 (.08)	.05 (.09)	.06 (.09)
	25.1	.02 (.05)	.02 (.05)	.02 (.05)	.02 (.05)	.02 (.06)	.02 (.06)
	40.1	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.05)
v=6 (4+2)	3.1	<b>.22</b> (.15)	<b>.22</b> (.15)	<b>.22</b> (.15)	<b>.22</b> (.16)	<b>.23</b> (.16)	<b>.23</b> (.16)
	10.1	.06 (.10)	.06 (.09)	.06 (.10)	.06 (.10)	.06 (.10)	.07 (.11)
	25.1	.02 (.06)	.02 (.06)	.02 (.06)	.02 (.06)	.02 (.06)	.03 (.07)
	40.1	.01 (.05)	.01 (.05)	.01 (.05)	.01 (.05)	.02 (.05)	.02 (.05)
v=12 (10+2)	3.1	<b>.20</b> (.11)	<b>.20</b> (.11)	<b>.20</b> (.11)	<b>.20</b> (.11)	<b>.20</b> (.11)	<b>.22</b> (.11)
	10.1	.05 (.07)	.05 (.07)	.05 (.07)	.06 (.07)	.06 (.07)	.07 (.07)
	25.1	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.05)	.03 (.05)
	40.1	.01 (.04)	.01 (.03)	.01 (.03)	.01 (.03)	.02 (.04)	.02 (.04)

**Note.** Correlation matrix **D** was created with  $r_w=.3$  and  $r_b=.3$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm .30 R_c^2$  criterion for extreme bias. <sup>a</sup> The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup> The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

**Table A5:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **E**

Var. Set	n: v	Expected $k$					
		-1	0	1	3	5	8
$v=12$ (6+6)	3.1	<b>.28<sup>a</sup></b> (.10) <sup>b</sup>	<b>.28</b> (.10)	<b>.28</b> (.09)	<b>.28</b> (.10)	<b>.28</b> (.10)	<b>.30</b> (.11)
	10.1	.07 (.07)	.07 (.07)	.07 (.07)	.07 (.07)	.07 (.07)	<b>.08</b> (.07)
	25.1	.03 (.04)	.03 (.04)	.03 (.04)	.03 (.04)	.03 (.05)	.03 (.05)
	40.1	.02 (.03)	.02 (.03)	.02 (.03)	.02 (.04)	.02 (.04)	.02 (.04)
$v=8$ (4+4)	3.1	<b>.28</b> (.12)	<b>.27</b> (.12)	<b>.27</b> (.13)	<b>.28</b> (.13)	<b>.28</b> (.13)	<b>.30</b> (.14)
	10.1	.07 (.08)	<b>.07</b> (.08)	<b>.07</b> (.08)	<b>.07</b> (.08)	<b>.08</b> (.08)	<b>.08</b> (.09)
	25.1	.03 (.05)	.03 (.05)	.03 (.05)	.03 (.05)	.03 (.05)	.03 (.06)
	40.1	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.04)
$v=6$ (4+2)	3.1	<b>.27</b> (.16)	<b>.26</b> (.15)	<b>.26</b> (.15)	<b>.27</b> (.16)	<b>.27</b> (.16)	<b>.28</b> (.17)
	10.1	<b>.07</b> (.09)	<b>.07</b> (.09)	<b>.06</b> (.09)	<b>.07</b> (.09)	<b>.07</b> (.10)	<b>.08</b> (.10)
	25.1	.02 (.06)	.03 (.06)	.02 (.06)	.03 (.06)	.03 (.05)	.03 (.06)
	40.1	.02 (.05)	.01 (.05)	.02 (.05)	.02 (.05)	.02 (.05)	.02 (.05)
$v=12$ (10+2)	3.1	<b>.27</b> (.11)	<b>.27</b> (.11)	<b>.27</b> (.11)	<b>.28</b> (.11)	<b>.27</b> (.11)	<b>.29</b> (.12)
	10.1	<b>.07</b> (.07)	<b>.07</b> (.07)	<b>.07</b> (.07)	<b>.08</b> (.07)	<b>.08</b> (.07)	<b>.08</b> (.07)
	25.1	.03 (.04)	.03 (.04)	.03 (.04)	.03 (.04)	.03 (.04)	.03 (.05)
	40.1	.02 (.03)	.02 (.03)	.02 (.03)	.02 (.03)	.02 (.03)	.02 (.04)

**Note.** Correlation matrix **E** was created with  $r_w=.5$  and  $r_b=.3$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm.30 R_c^2$  criterion for extreme bias. <sup>a</sup> The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup> The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

# BIAS AND PRECISION OF THE SQUARED CANONICAL COEFFICIENT

**Table A6:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **F**

Var. Set	n: v	Expected k					
		-1	0	1	3	5	8
v=12 (6+6)	3.1	.03 <sup>a</sup> (.02) <sup>b</sup>	.03 (.02)	.02 (.02)	.03 (.02)	.03 (.02)	.04 (.02)
	10.1	.01 (.01)	.01 (.02)	.01 (.01)	.01 (.02)	.01 (.02)	.01 (.02)
	25.1	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)	.01 (.01)
	40.1	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)	<.01 (.01)
v=8 (4+4)	3.1	.07 (.07)	.06 (.07)	.06 (.07)	.07 (.07)	.07 (.07)	.08 (.07)
	10.1	.02 (.04)	.02 (.04)	.02 (.04)	.02 (.05)	.02 (.05)	.03 (.05)
	25.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)
	40.1	<.01 (.02)	<.01 (.02)	<.01 (.02)	.01 (.02)	.01 (.02)	.01 (.03)
v=6 (4+2)	3.1	.10 (.13)	.10 (.12)	.10 (.12)	.11 (.12)	.11 (.12)	.13 (.12)
	10.1	.03 (.08)	.03 (.08)	.02 (.07)	.03 (.08)	.03 (.08)	.04 (.08)
	25.1	.01 (.05)	.01 (.05)	.01 (.05)	.01 (.05)	.01 (.05)	.02 (.05)
	40.1	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)
v=12 (10+2)	3.1	.09 (.07)	.08 (.07)	.08 (.06)	.09 (.07)	.09 (.07)	.10 (.07)
	10.1	.02 (.04)	.02 (.04)	.02 (.04)	.03 (.04)	.03 (.04)	.04 (.05)
	25.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.02 (.03)
	40.1	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)

**Note.** Correlation matrix **F** was created to have  $r_w=.3$  and  $r_b=.5$ , but limitations with the data generation procedures required us to create a correlation matrix with  $r_w=.43$  and  $r_b=.5$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm.30 R_c^2$  criterion for extreme bias. <sup>a</sup> The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup> The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

**Table A7:** Bias and Precision of Function I Sample  $R_c^2$  Values for Correlation Matrix **G**

Var. Set	n: v	Expected $k$					
		-1	0	1	3	5	8
v=12 (6+6)	3.1	.08 <sup>a</sup> (.06 <sup>b</sup> )	.08 <sup>a</sup> (.06)	.08 (.06)	.08 (.06)	.08 (.06)	.09 (.06)
	10.1	.02 (.04)	.02 (.04)	.02 (.04)	.03 (.04)	.03 (.04)	.03 (.04)
	25.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)
	40.1	.01 (.02)	.01 (.02)	<.01 (.02)	.01 (.02)	.01 (.02)	.01 (.02)
v=8 (4+4)	3.1	.11 (.10)	.10 (.10)	.10 (.10)	.11 (.10)	.11 (.10)	.13 (.10)
	10.1	.03 (.06)	.03 (.06)	.03 (.06)	.03 (.06)	.04 (.07)	.04 (.07)
	25.1	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.02 (.04)	.02 (.05)
	40.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.04)
v=6 (4+2)	3.1	.13 (.14)	.12 (.13)	.12 (.13)	.13 (.14)	.14 (.14)	<b>.16</b> (.15)
	10.1	.03 (.09)	.03 (.08)	.03 (.09)	.04 (.09)	.04 (.09)	.05 (.09)
	25.1	.01 (.06)	.01 (.06)	.01 (.06)	.02 (.06)	.02 (.06)	.02 (.06)
	40.1	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.04)	.01 (.05)	.01 (.05)
v=12 (10+2)	3.1	.12 (.08)	.11 (.09)	.12 (.08)	.12 (.09)	.13 (.09)	.14 (.09)
	10.1	.04 (.05)	.03 (.05)	.03 (.05)	.04 (.06)	.04 (.06)	.05 (.06)
	25.1	.01 (.04)	.01 (.03)	.01 (.03)	.02 (.04)	.02 (.04)	.02 (.04)
	40.1	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)	.01 (.03)

**Note.** Correlation matrix **G** was created with  $r_w=.5$  and  $r_b=.5$ .  $k$  denotes univariate kurtosis.  $n=5,000$  per cell. Bolded entries represent values that exceeded the  $\pm .30 R_c^2$  criterion for extreme bias. <sup>a</sup> The bias of the  $R_c^2$  values is denoted as the mean difference between the known population value of  $R_c^2$  and the average sample value of  $R_c^2$  across 5,000 sample replications. <sup>b</sup> The precision of the  $R_c^2$  values is denoted as the standard deviation of the accuracy values.

# Predicting Survival Time of Localized Melanoma Patients Using Discrete Survival Time Method

**Taysseer Sharaf**  
University of South Florida  
Tampa, FL

**Chris P. Tsokos**  
University of South Florida  
Tampa, FL

---

Melanoma is the most fatal type of skin cancer. It is ranked first in death of skin cancer diseases. This study establishes a statistical model that can predict the survival time of localized melanoma patients, as a function of age at diagnosis, tumor thickness, and extension of the tumor (tumor invasion). The discrete time survival method was used to build the statistical model. The patients involved in the current study were observed from the SEER database. Patients were divided into nine groups according to age at diagnosis. Variation in survival time was found to be significant among some of the age groups.

*Keywords:* melanoma, survival time, discrete survival time, skin cancer, SEER, localized melanoma

---

## Introduction

Melanoma is a malignant tumor associated with skin cancer. If melanoma is detected at a late stage, it can spread to other parts of the body and that's what makes it a lethal form of cancer. More general information about melanoma can be found in ([www.melanoma.org](http://www.melanoma.org)), (Markovic, et al., 2007) and (Mackle, et al., 2009). Over the last decades, the incidence of melanoma has been rapidly increasing in the United States. It appears more in white populations than other races. According to clinical studies, risk factors of melanoma are but not limited to, ultraviolet light exposure, moles, light hair, freckling and family history of melanoma. Some of the statistical analyses done on the risk factors are shown in (Gandini, et al., 2005c; Naldi, et al., 2000; Cho, et al., 2005).

---

*Mr. Sharaf is a Graduate Teaching Assistant in the Department of Mathematics and Statistics. Email him at <mailto:tsharaf@mail.usf.edu>. Dr. Tsokos is Distinguished University Professor in the Department of Mathematics and Statistics. Email him at [chris.tsokos@gmail.com](mailto:chris.tsokos@gmail.com).*

Consider the survival time for melanoma patients. The primary objective is the time between when the patient is diagnosed with melanoma and when death occurs. The study includes the effect of three risk factors that drives the survival time for a patient. Those risk factors are Age at diagnosis, tumor thickness and extension of the tumor (invasion of tumor through the body). Other factors include gender and sequence number (a number that indicates how many tumors the patient had prior to being diagnosed with melanoma). The main concern will be in estimating the survival time of melanoma patients diagnosed at stage one (localized Melanoma). More information regarding staging is discussed in the Methodology; for updates on the staging of melanoma visit [www.cancer.gov](http://www.cancer.gov). Soong, et al. (2010) developed an electronic prediction tool based on the American Joint Committee on Cancer (AJCC) melanoma staging database, to predict survival outcome of localized melanoma. Other predictive models of survival for localized melanoma have been developed in the United States and other countries (Clark, et al., 1989; MacKie, et al., 1995; Barnhill, et al., 1996; Schuchter, et al., 1996; Sahin, et al., 1997; Soong, et al., 2003). Soong, et al. (2010) used the Cox survival function model, which considers the survival time as a continuous random variable, where most survival times are recorded in discrete form as a number of months or years. They used same three risk factors in their analysis beside the primary melanoma site and primary tumor ulceration. As shown in Table 3, there exist 10 primary melanoma sites, and in order to reduce the variation in the model (biological variation between humans is a lurking variable), only one site of the ten was studied. Allison (1982) mentioned that in continuous survival time, maximum likelihood method ignores the discrete character of the data.

Xie, Mchugo, Drake, and Sengupta (2003), summarized the advantages of using discrete-time survival analysis. These were initially suggested by Singer & Willett (1993) as primarily useful for many longitudinal studies in clinical settings where data are often collected at discrete time periods. Secondly, the analysis facilitates the examination of the shape of the hazard function. Third, the analysis is simple and convenient to use, because it is a modification of the logistic regression model. Lastly and most important, time-varying covariates can easily be included in the model. After Cox presented the discrete time survival model, two basic versions of logistic models were introduced: the ordinal version and the dichotomous version. The dichotomous version (Allison, 1982; Singer & Willett, 1993; Xie, et al., 2003) represents each survival time as a set of indicators of whether or not an individual failed at each time point, until a person either experiences the event or is censored.

## Methodology

Allison (1982) and Singer and Willett (1993) proposed a discrete survival time method. The method starts by dividing the continuous time into an infinite sequence of contiguous time  $(0, t_1), (t_1, t_2), \dots, (t_{k-1}, t_k), \dots$ , and so on. Let  $k$  represent the number of time intervals. In this case, time is recorded in months, where time is divided into 20 intervals each consists of 12 months:  $(1,12), (12,24), \dots, (228,240)$ . If a patient's survival time is 7 months, then this patient's event is classified as happening during the 1<sup>st</sup> time interval; if another patient's survival time is 50 months, then this is classified as happening during the 5<sup>th</sup> time interval.

To estimate the survival function, start with the discrete-time hazard model (Allison, 1982)

$$h_{ik} = \frac{1}{1 + EXP\left\{-\left(\alpha_1 T_{1ik} + \alpha_2 T_{2ik} + \dots + \alpha_J T_{Jik}\right) - \left(\beta_1 X_{1ik} + \beta_2 X_{2ik} + \dots + \beta_p X_{pik}\right)\right\}} \quad (1)$$

where  $[T_{1ik}, T_{2ik}, \dots, \alpha_J T_{Jik}]$  are a sequence of dummy variables, with values  $[t_{1ik}, t_{2ik}, \dots, t_{Jik}]$  indexing time periods, where  $J$  refers to the last time period observed for any individual in the sample. If individual  $i$  was observed (experienced the event or censored) in the fourth period, then  $J = 4$ , and the time period's dummy variables are defined identically for each individual;  $t_{1ik} = 1$  when  $j = 1$  and 0 when  $j$  takes any other value.

The coefficients  $(\alpha_1 \alpha_2 \dots \alpha_J)$  act as the intercept parameters for the baseline hazard in each time period, and the coefficients  $(\beta_1 \beta_2 \dots \beta_p)$  describe the effect of the predictors on the baseline hazard in the logit scale. Singer and Willett (1993) discussed briefly the procedures to construct the likelihood function (in terms of the discrete hazard function) used to estimates the latter intercepts and slope parameters.

The likelihood function presented by Singer and Willett (1993) is given by

$$L = \prod_{i=1}^n \prod_{k=1}^{j_i} h_{ik}^{y_{ik}} (1 - h_{ik})^{(1-y_{ik})} \quad (2)$$



where  $y_{ik}$  is a sequence of dummy variables that records the event history for patient  $i$ , whose values are defined as:

$$y_{ik} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ patient experienced the event in period } k \\ 0 & \text{if the } i^{\text{th}} \text{ patient did not experience the event in period } k \end{cases}$$

The likelihood function is identical to the likelihood function to a sequence of  $N = (k_1 + k_2 + \dots + k_n)$  independent Bernoulli trials with parameters  $h_{ik}$ .

Using results by Allison (1982), the  $y_{ik}$  values can be considered as the outcome variable in a logistic regression analysis, which provides a simple model to obtain the maximum likelihood estimate rather than finding the solution by maximizing equation (2).

For discrete event history data, each record consists of the information for one patient like survival time, Age and whether or not the patient time is censored or not. In order to apply the logistic model discussed previously, the data need to be converted into new person-period data, in which each patient will have multiple records, one per time period of observation. As shown by Singer and Willett (1993), the new person-period data will contain the information about the  $k^{\text{th}}$  time period as follows

- **The time indicators.** The set of dummy variables  $[T_{1ik}, T_{2ik}, \dots, \alpha_j T_{jik}]$ .
- **The predictors.** Covariates under study, where the ability exists to use the time-varying covariates that have values differs from time period to time period.
- **The event indicator (response variable in the logistic model).** This variable records whether the event of interest occurred in period  $j$  or not. The variable takes value 1 if the event occurred, takes 0 if did not.

In this study, the survival time of melanoma patients diagnosed in the period of 1988 to 2008 was considered. The survival time was recorded up to the nearest month. The time period of the study was divided into 20 intervals, one year each. Besides the covariates age at diagnosis, tumor size and extension of the tumor, there were 20 dummy variables representing the 20 time periods as shown in Tables 1 and 2. In Table 1, there is a record of 3 patients as extracted from Surveillance Epidemiology and End Results database (SEER) database. In Table

## PREDICTING SURVIVAL TIME OF LOCALIZED MELANOMA PATIENTS

2, there is a representation of the conversion of the data to fit the new person-period data to be used for the logistic model.

**Table 1.** A record of 3 patients from SEER database.

Patient ID	Survival Time (mos)	Age Diagnosis	Tumor Size	Extension of Tumor
7013574	49	84	120	10
8862253	3	66	230	30
8869492	86	61	134	30

Table 1 shows that the first patient survival time is 49 months. The first patient lived for four years and died in the first month of the fifth year, which means that the event took place during the fifth time period. In the new data setting the first patient will have five records, one record corresponding to every time period (from the first to the fifth). The event indicator variable will take 0 for the first four records and 1 in the fifth record where the event took place. Table 2 shows this conversion for the three patients in table one.

**Table 2.** A sample of three patients from the new person-period data.

Indc.	ID	ST	A	TS	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>	D <sub>6</sub>	D <sub>7</sub>	D <sub>8</sub>	...	D <sub>20</sub>
0	1	49	84	120	1	0	0	0	0	0	0	0	...	0
0	1	49	84	120	0	1	0	0	0	0	0	0	...	0
0	1	49	84	120	0	0	1	0	0	0	0	0	...	0
0	1	49	84	120	0	0	0	1	0	0	0	0	...	0
1	1	49	84	120	0	0	0	0	1	0	0	0	...	0
1	2	3	66	230	1	0	0	0	0	0	0	0	...	0
0	3	86	61	134	1	0	0	0	0	0	0	0	...	0
0	3	86	61	134	0	1	0	0	0	0	0	0	...	0
0	3	86	61	134	0	0	1	0	0	0	0	0	...	0
0	3	86	61	134	0	0	0	1	0	0	0	0	...	0
0	3	86	61	134	0	0	0	0	1	0	0	0	...	0
0	3	86	61	134	0	0	0	0	0	1	0	0	...	0
0	3	86	61	134	0	0	0	0	0	0	1	0	...	0
1	3	86	61	134	0	0	0	0	0	0	0	1	...	0

The variables  $D_1, D_2, \dots, D_{20}$  represent the 20 dummy variables for the time intervals. Each row shows patient  $i$  information during each time interval until the event occurs or he/she is censored. The first column in Table 2 (Indc.) represents the indicator variable, which takes 0 if the event did not take place during the

current time period interval, or 1 if the event occurs during the time period interval. The patient's ID was changed to ID (1, 2 and 3) to optimize the table. The setting of the data shown in Table 2 can allow us to use time varying covariates easily, but because the data used does not support this information the covariates were repeated. Only the tumor size of the patient is known at the time of diagnosis; no follow up information was supported. Age at diagnosis can be changed, but no big difference will appear (as will be shown); ages were grouped into 9 intervals, each interval covering 10 years.

The data set used was collected from the Surveillance Epidemiology and End Results database (SEER) 1973-2008. 208,143 patients were diagnosed in the United States from 1973 through 2008. Taking into account the patients that were confirmed dead because of melanoma cancer, and removing all the missing records in the covariates shown in Table 4, results in studying the patients diagnosed from 1988 through 2008. Melanoma cancer is classified into 4 stages as shown in Table 3, and there exist 10 sites of the skin where the cancer appears. In order to reduce the variation (biological difference between humans) and to get less prediction error in the statistical model, only 'skin of trunk' patients in stage 1 (localized melanoma) are considered in this study. The sample size used in this study is 1,240. Table 3 shows the description of the coding used for the primary site and the staging of the cancer.

The risk factors (affecting the survival time of patients) that were involved in the study are age of the patient, the tumor size and the tumor extension (how far the tumor spread). Around 99 percent of the 1,240 patients were white (due to the fact that melanoma is rare in people with dark skin), which is why race was not considered in the modeling aspects.

Age of patients at Diagnosis was classified into 9 groups, 10 years each, starting from 11 to 20 years in the first group through 91 to 100 years old. Tumor thickness (instead of size as known in other cancer types, the thickness is measured, but size will be referred to for the remainder of this article) was classified into 3 groups, the first group from 1mm to 50mm, the second group from 51mm to 300mm and the last group from 301mm to 992mm. In the current sample the extension of the tumor contains 4 levels which are as stated in SEER EOD-88 3<sup>rd</sup> edition; (10) for papillary dermis (the middle layer of skin) invaded, (20) for papillary-reticular dermal interface invaded, (30) for reticular dermis invaded and finally (40) localized. A summary of the number of patients lying in the groups stated previously is shown in Table 4.

## PREDICTING SURVIVAL TIME OF LOCALIZED MELANOMA PATIENTS

**Table 3.** SEER coding for the stage and primary site for Melanoma.

Variable	Code	Description
SEER Historic stage	0	In situ : A tumor which has not penetrated the basement membrane nor extended beyond the epithelial issue
	1	Localized : An invasive neoplasm confined entirely to the organ of origin
	2	Regional: A neoplasm that has gone beyond the bounds of the organ of origin or into regional lymph nodes.
	4	Distant : A neoplasm that has spread to parts of the body distant from the primary tumor
Primary Site	C440	Skin of lip
	C441	Eyelid
	C442	External Ear
	C443	Skin of other and unspecified parts of face
	C444	Skin of Scalp and neck
	C445	Skin of Trunk
	C446	Skin of upper limb and shoulder
	C447	Skin of lower limb and hip
	C448	Overlapping lesion of skin
	C449	Skin NOS

Table 4 illustrates that around 46.5% of the patients are diagnosed at the third level of tumor extension where the tumor invaded into the reticular dermis, indicating that there is a delay from patients until they figured out that they needed medical attention. It must be stressed that during the current study no treatment effects were added to the statistical model, so study results are considered as if patients did not get any treatment. The different treatment effects and histology effects will be studied in further publications.

Descriptive statistics of the survival time of melanoma patients recorded in months from time of diagnosis till death for each age group are recorded in Table 5. Because the survival time distribution is skewed, it is important to estimate the median survival time for the melanoma patients. The median will be more informative than the mean in this case. The large variance of survival time inside each group can be seen. This assures the presence of independent variables affecting the survival time.

**Table 4.** Distribution of the 1240 patients on the various groups.

Factor		Extension of the Tumor				Total	%
Age	Tumor Size	10	20	30	40		
11-20	1						
	2	1	4			5	0.4838710
	3				1	1	
21-30	1	3	10	2	1	16	
	2	3	11	12	4	30	4.2741935
	3			6	1	7	
31-40	1	8	3	11	3	25	
	2	11	52	50	8	121	13.3870970
	3		5	15		20	
41-50	1	14	9	13	2	38	
	2	10	87	73	12	182	20.3225810
	3		8	20	4	32	
51-60	1	17	9	19	7	52	
	2	17	61	73	14	165	20.3225810
	3		3	27	5	35	
61-70	1	14	3	11	5	33	
	2	14	59	87	18	178	20.4032260
	3		4	33	5	42	
71-80	1	18	7	12	4	41	
	2	8	36	60	15	119	15.2419350
	3		7	19	3	29	
81-90	1	6	2	11	3	22	
	2	3	9	14	5	31	5.0806452
	3		3	5	2	10	
91-100	1	1		1		2	
	2		1	2		3	0.4838710
	3			1		1	
%		11.94000	31.69000	46.53000	9.83871	1240	100%

**Table 5.** Descriptive statistics of survival time.

Age	Survival Time				
	Count	Mean	Median	St. deviation	Skewness
11-20	6	74.50	57.00	42.43	1.39
21-30	53	61.40	46.00	44.10	1.11
31-40	166	66.95	52.50	48.48	1.21
41-50	252	67.83	55.50	45.53	1.15
51-60	252	56.90	43.00	41.17	1.32
61-70	253	53.96	41.00	42.11	1.58
71-80	189	50.38	45.00	32.86	0.89
81-90	63	38.44	34.00	28.05	1.26
91-100	6	32.17	26.00	31.92	1.91

## Results

Three out of the four models proposed models are discussed. The first model is the baseline model, estimating the survival function using the time periods. The second model introduces age at diagnosis as the first covariate with the time periods. The third model uses age at diagnosis and tumor size as covariates with the time periods. The fourth model introduces all three covariates with the time periods.

### Model 1

The baseline model is the starting point of the proposed modeling procedure. The simplest hazard model from equation (1), considering only the 20 dummy variables that represent the time effect, is considered. The baseline hazard model for this case is represented as:

$$\text{logit}(h_i) = (\alpha_1 t_1 + \alpha_2 t_2 + \dots + \alpha_{20} t_{20}) \quad (3)$$

Equation (3) represents the log transform of equation (1), where

$$\text{logit}(h_{ik}) = \log\left(\frac{h_{ik}}{1-h_{ik}}\right) \quad (4)$$

This model will answer the basic question ‘what is the probability of obtaining the event (melanoma patient dies due to the cancer) in each time period?’ In other words what is the probability that a melanoma patient will survive for one, or two years, etc.

The parameters in equation (3) can be converted by exponentiation the right hand side. For example if it is desired to know the estimate of the probability of event occurrence in the fifth interval will be equal to

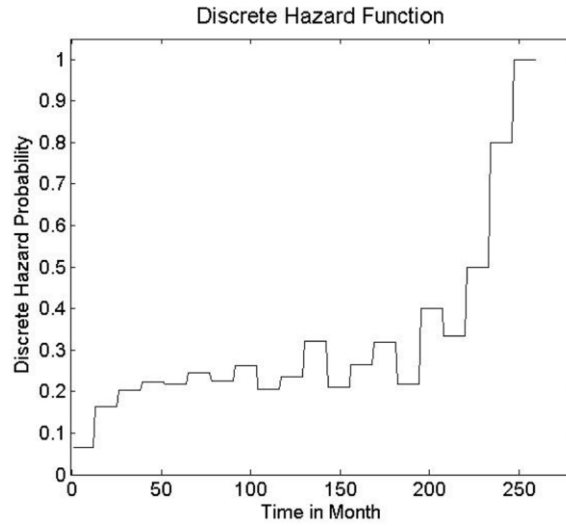
$$\hat{h}_5 = \frac{1}{(1 + e^{-\alpha_5})} \quad (5)$$

The estimates of the baseline hazard are presented in Table 6. After estimating the baseline hazard one can calculate the corresponding survival function using

$$\hat{S}_k = \prod_{j=1}^k (1 - \hat{h}_j) \quad (6)$$

**Table 6.** Estimates of the baseline hazard parameters.

Param.	$\hat{\alpha}$	Param.	$\hat{\alpha}$
D1	-2.687	D11	-0.743
D2	-1.642	D12	-1.326
D3	-1.372	D13	-1.017
D4	-1.242	D14	-0.758
D5	-1.275	D15	-1.273
D6	-1.125	D16	-0.405
D7	-1.233	D17	-0.693
D8	-1.031	D18	0.000
D9	-1.349	D19	1.386
D10	-1.185	D20	21.203

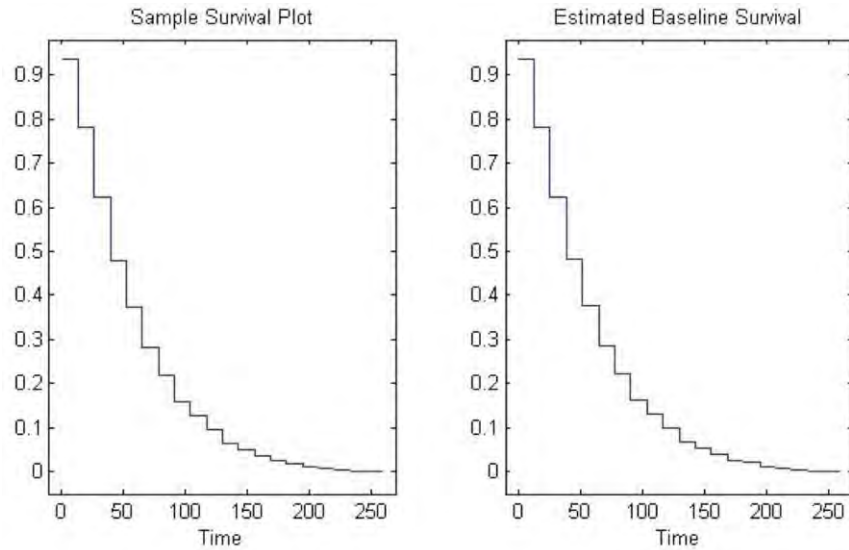
**Figure 1.** Scatter plot for the estimated baseline Hazard function.

Thus to illustrate the output, the hazard probability of first time period is calculated and by substituting the values in equation (3)  $t_2, t_3, \dots, t_{20} = 0$  while

$t_1 = 1, \hat{h}_1 = \frac{1}{(1+e^{2.687})} = 0.06$ , and to get the hazard probability of the second interval  $\hat{h}_2 = \frac{1}{(1+e^{1.642})} = 0.16$ , and so on. The coefficient of the last time period

had a high negative value, but was actually insignificant in the modeling process. This was due to the occurrence of only one event, which had held out till the final period in the sample. Also, removing the last interval from the modeling process did not induce any significant change to the  $-2\log\text{likelihood}$ , which was used to pick the best model from the four models that were tested.

Once calculated for all 20 time periods the baseline survival function can be calculated using equation (6). Graphical representation of the estimated discrete hazard function is shown in Figure 1. And a comparison of the estimated base line survival function by the model and the sample survival function is shown in Figure 2.



**Figure 2.** Sample Survival Function and Estimated Baseline Survival.

In Figure 2 it is shown that the estimated base line model fits the data well. This is also supported by the residual analysis from the logistic model used to estimate the baseline hazard function. The model residuals came to be uncorrelated and with constant variance. Because the distribution of the survival



time is skewed, the median survival time is of great interest, as shown in the first graph in Figure 2: the estimated median survival time when  $S(t) = 0.5$ , which is equal to  $\hat{t}_{0.5} = 48$  months. It is customary to see the discrete survival function as a step down function in graphs, but for technical purposes and comparison issues, connected lines are used in these Figures rather than a step down function.

## Model 2

The second model is to see the effect of the covariates on the survival time. The first covariate will be age at diagnosis. As discussed in the previous section the age of patient at diagnosis is grouped into 9 groups, 10 year interval each. The nine groups will be represented by 8 dummy variables with the first age group as the base. The parameter estimates are represented in Table 7.

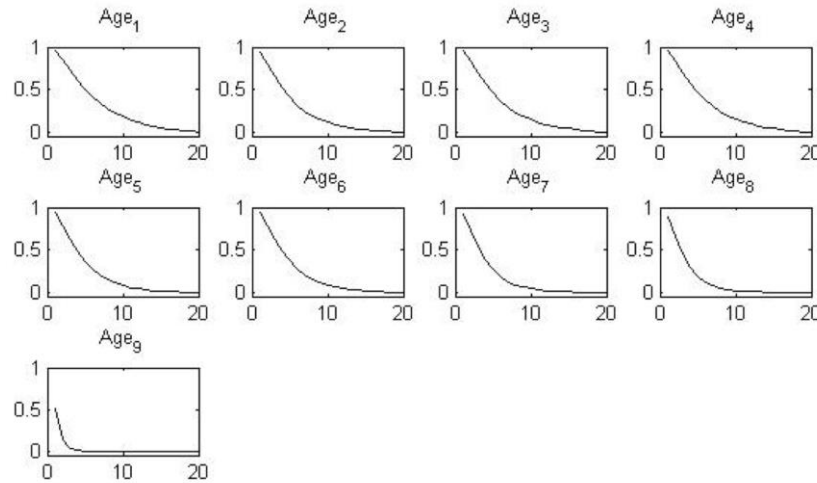
**Table 7.** Parameter estimates for discrete hazard function with Age as covariate

Param.	$\alpha$	Param.	$\alpha$	Param.	$\beta$
D1	-3.13	D11	-1.02	age_1	0.28
D2	-2.04	D12	-1.60	age_2	0.11
D3	-1.75	D13	-1.28	age_3	0.09
D4	-1.60	D14	-1.00	age_4	0.41
D5	-1.62	D15	-1.51	age_5	0.39
D6	-1.46	D16	-0.63	age_6	0.73
D7	-1.55	D17	-0.90	age_7	1.00
D8	-1.33	D18	-0.22	age_8	3.11
D9	-1.64	D19	1.18		
D10	-1.48	D20	21.11		

The estimates for the alpha parameters correspond to the time periods. The first beta estimate 0.28 corresponds to the second age group (Age 21-30); recall that first age group is at base level. For example, the estimated discrete hazard probability for the first age group (Age 11-20) of the first time period is  $\hat{h}_1 = \frac{1}{(1 + e^{-(-3.13)})} = 0.041887$ , the estimate of the survival probability for the same age group in the first time period by equation (4)  $\hat{S}_1 = (1 - \hat{h}_1) = 0.958$ . Similar calculations were followed to get the survival for the 20 time periods for each Age

## PREDICTING SURVIVAL TIME OF LOCALIZED MELANOMA PATIENTS

group. Figure 3 shows the graph illustration for those survival functions, for each age group.



**Figure 3.** Estimated Survival function plot from Model 2, for different Age groups.

Looking at the survival plots in Figure 3, the duration time of the melanoma cancer is same for patients in the age group 5 and 6. The duration time for the second age group is lower than that for Age group 3 and 4, which is closer to the duration time of the first age group. For the last age group (Ages 91-100) the estimated median survival time corresponds to the first time period, which means it is between 1 and 12 months.

### Model 3

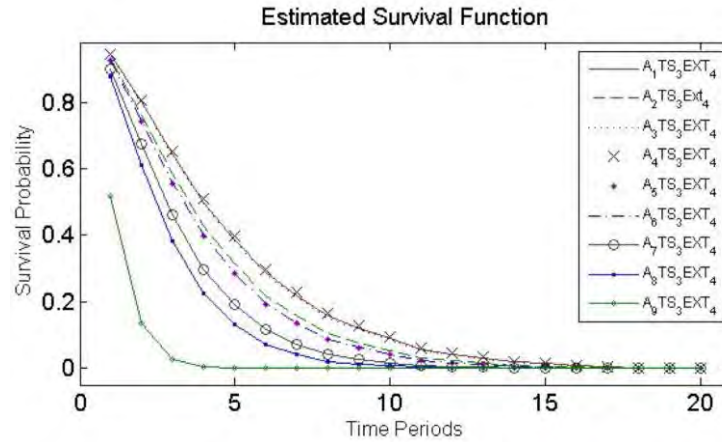
Adding more covariates made the model more significant. Table 8 shows the different models that were applied to the data along with the corresponding  $-2\log\text{likelihood}$  and Cox & Snell R-square. The model that best fits the data is the last one with the three covariates: Age at diagnosis, Tumor Size and Extension of the tumor.

This model with 0.383 Cox and Snell R-square is considered to be significant for the analysis of binary data. This model gave more informative estimates about the behavior of the survival time of melanoma cancer patients, across the 9 age groups. A plot showing the estimated survival time using the last

model based on patients diagnosed with tumor size of group 3 and extension of tumor of group 4 (refer to Table 4) is given in Figure 4.

**Table 8.** -2loglikelihood for various models

Model	-2loglikelihood	Cox & Snell R-square
Base	6088.32	0.369
Base + Age	6016.71	0.376
Base + Age + Tumor Size	5981.91	0.379
Base + Age + Tumor Size + Extension of Tumor	5939.26	0.383



**Figure 4.** Estimated Survival Function for the different age groups using Model 4.

Figure 4 represents the estimated survival probability for the 9 age groups: A1 for first age group, A2 for second age group, ... , A9 for the ninth group. According to the model (Model 4), the following results are found for patients diagnosed with tumor size between (301mm to 992mm) and at the fourth level of tumor extension (No treatment was involved in this model):

- A patient diagnosed at ages 11-20 and 31-50 have the same estimate of median survival time less than 5 years.
- A patient diagnosed at ages 21-30, have an estimate of median survival time less than 4 years.

## PREDICTING SURVIVAL TIME OF LOCALIZED MELANOMA PATIENTS

- Age group 5 (51-60) and 6 (61-70) have the same estimate of median survival time between 3 to 4 years.
- [Figure 4](#) shows that the maximum estimated survival time for all age groups is around 15 years from the time of diagnosis.

### Conclusion

A statistical model was developed to predict the survival time of localized melanoma patients using the discrete survival time method. The discrete survival time method gives better results when applied on follow-up data sets. If the information about the progress a patient's tumor thickness and the time of treatment patients took is available, the results become more accurate and show less prediction error.

Four different statistical models were developed with a recommended model (fourth one) to be the best model for predicting the survival time of a given localized melanoma patient. This model is the one that takes into consideration the patient's age at diagnosis, tumor thickness and extension of the tumor.

In comparison with research by Soong, et al. (2010), the primary melanoma site was not considered as one of covariates in the model. Patients were divided into 10 groups according to the primary site and model each group separately. Results from this study show less error compared with Soong, et al. (2010) because the variation due to the difference in the primary site was removed.

### References

- Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology* 13: 61-98.
- Barnhill, R. L., Fine, J. A., Roush, G. C., & Berwick, M. (1996). Predicting five-year outcome for patients with cutaneous melanoma in population-based study. *Cancer* 78(3): 427-432.
- Bataille, V. & de Vries, E. (2008). Melanoma – Part 1: epidemiology, risk factors, and prevention. *BMJ* 337: 1287-1291.
- Cho, E., Rosner, B. A., & Colditz, G. A. (2005). Risk Factors of Melanoma by Body Site, *Cancer Epidemiology, Biomarkers & Prevention* 14: 1241-44.
- Clark, W. H. Jr., Elder, D. E., Guerry, D. IV, Braitman, L. E., Trock, B. J., Schultz, D., Synnestvedt, M., & Halpern, A. C. (1989). Model predicting survival

in stage I melanoma based upon tumor progression. *J Natl Cancer Inst.* 81(24): 1893-1904.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Abeni, D., Boyle, P., & Melchi, C. F. (2005a). Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi, *European Journal of Cancer*, 41(1): 28-44.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Picconi, O., Boyle, P., & Melchi, C. F. (2005b). Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure, *European Journal of Cancer*, 41(1): 45-60.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Zanetti, R., Masini, C., Boyle, P., & Melchi, C. F. (2005c). Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors, *European Journal of Cancer*, 41(14): 2040-59.

MacKie, R. M., Aitchison, T., Sirel, J. M., McLaren, K., & Watt, D. C. (1995). Prognostic models for subgroups of melanoma patients from the Scottish Melanoma Group database 1979-86, and their subsequent validation. *Br J Cancer* 71(1): 173-176.

Mackie, R. M., Hauschild A., & Eggermont, A. M. M. (2009). Epidemiology of Invasive Cutaneous melanoma, *Annals of Oncology* 20(Suppl 6): vi1-vi7.

Markovic, S. N., et al., (2007). Malignant melanoma in the 21st century, Part 1: epidemiology, risk factors, screening, prevention, and diagnosis. *Mayo Clinic Proceedings*, 82(3): 364-80.

Melanoma Research Foundation. (2011). *Melanoma Statistics & Facts, 2011*, Retrieved October 23, 2012 from <http://www.melanoma.org/learn-more/melanoma-101/melanoma-statistics-facts>

Naldi, L., Lorenzo, I. G., Parazzini, F., Gallus, S., & La Vecchia, C. (2000). Pigmentary traits, modalities of sun reaction, history of sunburns, and melanocytic nevi as risk factors for cutaneous malignant melanoma in the Italian population, *Cancer*, 88(12): 2703-2710

National Cancer Institute. *Staging*, (2011a, January 11). Retrieved October 23, 2012 from <http://www.cancer.gov/cancertopics/wyntk/skin/page9>

National Cancer Institute. *What do you need to know About Melanoma and other Skin Cancers*, (2011b, January 11). Retrieved October 23, 2012 from <http://www.cancer.gov/cancertopics/wyntk/skin/page4>

National Cancer Institute. (2012). *Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009)*,

## PREDICTING SURVIVAL TIME OF LOCALIZED MELANOMA PATIENTS

National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.

Sahin, S., Rao, B., Kopf, A. W., Lee, E., Rigel, D. S., Nossa, R., Rahman, I. J., Wortzel, H., Marghoob, A. A., & Bart, R.S. (1997). Predicting ten-year survival of patients with primary cutaneous melanoma: a corroboration of a prognostic model. *Cancer* 80(8): 1426-1431.

Schuchter, L., Schultz, D. J., Synnestvedt, M., Trock, B. J., Guerry, D., Elder, D. E., Elenitsas, R., Clark, W. H., & Halpern, A. C. (1996). A prognostic model for predicting 10-years survival in patients with primary melanoma: The Pigmented Lesion Group. *Ann Intern Med.* 125(5): 369-375.

Singer, J. D. & Willett, J. B. (1993). It's about Time: Using Discrete- Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics* 18(2): 155-195.

Soong, S.-J., Zhang, Y., & Desmond, R. (2003). Models for predicting outcome. In: Balch, C. M., Houghton, A. N., Sober, A., & Soong, S.-J., Eds. *Cutaneous melanoma* (4th ed., pp. 77-90). St. Louis: Quality Medical Publishing.

Soong, S.-J., Ding, S., Coit, D., Balch, C. M., Gershenwald, J. E., Thompson, J. F., Gimotty, P., & the AJCC Melanoma Task Force. (2010). Predicting the Survival Outcome of Localized Melanoma: An Electronic Prediction Tool Based on the AJCC Melanoma Database, *Annals of Surgical Oncology* 17(8): 2006-2014.

Svetomir, N. M., Erickson, L. A., Rao, R. D., Weenig, R. H., Pockaj, B. A., Bardia, A., Vachon, C. M., Schild, S. E., McWilliams, R. R., Hand, J. L., Laman, S. D., Kottschade, L. A., Maples, W. J., Pittelkow, M. R., Pulido, J. S., Cameron, J. D., Creagan, E. T., & the Melanoma Study Group of the Mayo Clinic Cancer Center. (2007). Malignant Melanoma in the 21st Century, Part 1: Epidemiology, Risk Factors, Screening, Prevention, and Diagnosis. *Mayo Clinic Proceedings*, 82(3): 364-380.

Xie, H., McHugo, G., Drake, R., & Sengupta, A. (2003). Using Discrete-Time Survival Analysis to Examine Patterns of Remission From Substance Use Disorder Among Persons With Severe Mental Illness. *Mental Health Services Research* 5(1): 55-64.

# Robustness of Several Estimators of the ACF of AR(1) Process With Non-Gaussian Errors

**A. A. Smadi**

Yarmouk University  
Irbid, Jordan

**J. J. Jaber**

University of Jordan  
Aqaba, Jordan

**A. G. Al-Zu'bi**

Yarmouk University  
Irbid, Jordan

---

The autocorrelation function (ACF) plays an important role in the context of ARMA modeling, especially for their identification and estimation. This study considers the robust estimation of the ACF of the AR(1) model if the white noise (WN) process is non-Gaussian. Three estimators including the ordinary moment estimator and two other (robust) estimators are considered. The impacts of the deviation from normality of the WN process on those estimators in terms of bias, MSE and distribution via Monte-Carlo simulation are examined. The empirical distribution of those estimators when the errors are normal,  $t$ , Cauchy and exponential are studied. Results show that the moment estimator is more affected by the change of the white noise distribution than other considered estimators.

*Keywords:* autocorrelation function, robust estimation, Monte-Carlo simulation, kernel density estimation

---

## Introduction

A time series (TS) can be defined as a sequence of observations taken sequentially in time. Time series can be observed in different fields; for example, in agriculture, business, engineering and medical studies. The list of areas in which time series is observed, studied and analyzed is endless. A major feature in the development of time series is an assumption of some form of statistical equilibrium, or known as stationarity. There are two types of stationarity; the first is called strict stationarity, and the other type is called weak stationarity. In practice, it is very difficult to examine time series being strictly stationary. Further, a stochastic process  $\{X_t\}$  is weak stationary if its mean is constant and the auto-

---

*Dr. Smadi is in the Department of Statistics. Email him at: [asmadi@yu.edu.jo](mailto:asmadi@yu.edu.jo). Dr. Jaber is in the Department of Risk Management & Insurance. Dr. Al-Zu'bi is in the Department of Statistics.*

## ROBUSTNESS OF ESTIMATORS OF THE ACF OF AR(1) PROCESS

covariance function (ACVF) depends on the time lag only, i.e.,  $Cov(X_t, X_{t+k}) = \gamma_k$  as well as its ACF  $\rho_k = \gamma_k / \gamma_0$ . For more details on the ACVF and ACF of stationary time series and their properties, see Wei (2006, p. 12).

The class of autoregressive moving average (ARMA) models is widely known for modeling stationary time series (Wei, 2006, p. 56–64). The stochastic process  $\{X_t\}$  is said to follow the ARMA( $p, q$ ) model if:

$$X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + u_t - \theta_1 u_{t-1} - \dots - \theta_q u_{t-q}$$

where  $\phi_1, \dots, \phi_p$  and  $\theta_1, \dots, \theta_q$  are the AR and MA parameters, respectively, and  $u_t$  is the white noise (WN) process, assumed iid  $(0, \sigma_u^2)$  and usually assumed normal. A detailed account on ARMA models, their autocorrelation functions and building methodology is found in Box, et al. (1994).

Beside the mixed ARMA model, the ARMA( $p, q$ ) models also include as special cases the pure AR and pure MA models when  $q = 0$  and  $p = 0$ , respectively. In particular, the pure AR(1) model is given by:

$$X_t = c + \phi X_{t-1} + u_t \quad (1)$$

which is stationary if  $|\phi| < 1$  (Wei, 2006).

The ACF plays an important role in the Box and Jenkins methodology for building ARMA models, especially for the identification and estimation of those models (Wei, 2006). In fact, there are other identification tools for the ARMA models, including the inverse ACF method (Cleveland, 1972); Akaike information criterion (AIC) (Akaike, 1974); the R and S array method (Gray, et al., 1978) and the corner method (Beguín, et al., 1980).

Consider the robust estimation of the ACF of AR(1) model if the WN process  $\{u_t\}$  is non-Gaussian. Berkoun, et al. (2003) investigated robust inference for serial correlation in AR(1) process in the presence of a single additive outlier. Assuming that  $\{X_1, \dots, X_n\}$  is a time series following the zero-mean AR(1) model contaminated with a single additive outlier, they investigated three estimators of  $\rho_1$ , namely:



$$r_1 = \frac{\sum_{t=2}^n Z_t Z_{t-1}}{\sum_{t=2}^n Z_{t-1}^2} \quad (2)$$

$$\rho_1^* = Med \left\{ \frac{Z_2}{Z_1}, \frac{Z_3}{Z_2}, \dots, \frac{Z_n}{Z_{n-1}} \right\} \quad (3)$$

and

$$\tilde{\rho}_1 = \frac{Med \{Z_1 Z_2, Z_2 Z_3, \dots, Z_{n-1} Z_n\}}{Med \{Z_1^2, Z_2^2, \dots, Z_{n-1}^2\}} \quad (4)$$

where  $Z_t = X_t - \bar{X}$  are the mean-subtracted data,  $Med(\cdot)$  stands for the median,  $r_1$  is the ordinary moment estimator of  $\rho_1$  whereas  $\rho_1^*$  and  $\tilde{\rho}_1$  are two robust estimators of  $\rho_1$  originally proposed by Hurwicz (1950) and Haddad (2000), respectively. Berkoun, et al. (2003) showed that the inference of  $\rho_1$  based on  $r_1$  is highly sensitive to a single additive outlier.

Smadi, et al. (2009) generalized these estimators for the periodic AR(1) model. They again observed that the counterpart of  $r_1$  is more sensitive to additive outliers than other estimators.

For higher time lags  $k = 1, 2, \dots$  the estimators in (2) – (4) generalize to estimate  $\rho_k$  as follows:

$$r_k = \frac{\sum_{t=k+1}^n Z_t Z_{t-k}}{\sum_{t=2}^n Z_{t-1}^2} \quad (5)$$

$$\rho_k^* = Med \left\{ \frac{Z_{k+1}}{Z_1}, \frac{Z_{k+2}}{Z_2}, \dots, \frac{Z_n}{Z_{n-k}} \right\} \quad (6)$$

and

$$\tilde{\rho}_k = \frac{\text{Med}\{Z_1 Z_{k+1}, Z_2 Z_{k+2}, \dots, Z_{n-k} Z_n\}}{\text{Med}\{Z_1^2, Z_2^2, \dots, Z_{n-k}^2\}} \quad (7)$$

In this research, the main objective is to study the statistical properties; namely the mean, variance as well as the distribution of various estimators of  $\rho_k$ . This study is restricted to the AR(1) model as (1) along various distributions for the WN process. Therefore, it focuses on the robustness of estimators above subject to the distribution of the WN process.

In the literature of time series analysis, the area of robust inference has found considerable attention. Denby and Martin (1979) proposed the generalized M-estimates for autoregressive processes and Bustos and Yohai (1986) took the auto-covariance structure of time series into consideration when robustifying the estimators. Zieliński (1999) investigated the median-unbiased estimation of the stationary AR(1) process. Molinares, et al. (2009) investigated robust estimation in long-memory processes when the data contains additive outliers.

Besides, several articles focused on the estimation of ACF of stationary time series including the work of Berkoun, et al. (2003) mentioned above. Smadi, et al. (2009) and Smadi (2013) generalized the work of Berkoun, et al. (2003) to periodic AR models. Alternatively, Hassani (2010) found that the distributions of a set of sample autocorrelations are neither independent nor identically distributed. This finding implies that the result of diagnostic check and model building based on  $r_k$ , especially in the presence of some suspect data can be quite misleading. Kan and Wang (2010) provide an algorithm for evaluating the exact distribution of the sample autocorrelations.

### Some properties of the ACF of AR(1) model

Let  $\{X_t\}$  be a stationary time series, then the auto-covariance function (ACVF) and the autocorrelation function (ACF) depend on time lag only. Based on a realization  $\{X_1, \dots, X_n\}$ , the moment estimator of  $\rho_k$  is given by  $r_k$  defined in (5) above (Wei, 2006). For large  $n$ ,  $r_k$  is approximately normally distributed with mean  $\rho_k$ . Also, for a stationary Gaussian process, based on Bartlett (1946), Brockwell & Davis (2002) have shown that for  $k > 0$  and  $k + j > 0$ ,

$$\text{Cov}(r_k, r_{k+j}) \cong \frac{1}{n} \sum_{i=1}^{\infty} \{(\rho_{i+k} + \rho_{i-k} - 2\rho_i \rho_k) \times (\rho_{j+k} + \rho_{j-k} - 2\rho_j \rho_k)\} \quad (8)$$

which in turn for  $j = 0$  reduces to

$$Var(r_k) \cong \frac{1}{n} \sum_{i=1}^{\infty} (\rho_{i+k} + \rho_{i-k} - 2\rho_i \rho_k)^2 \quad (9)$$

For more details on properties of  $r_k$  for stationary time series, see Wei (2006).

Now, if  $\{X_t\}$  follows the zero-mean AR(1) Model, as (1) with  $c = 0$  and  $|\phi| < 1$ , then  $\rho_k = \phi^k$ ,  $k = 0, 1, 2, \dots$  and (8) reduces to

$$\begin{aligned} Var(r_k) &\approx \frac{1}{n} \sum_{i=1}^k \phi^{2k} (\phi^{-1} - \phi^i)^2 + \sum_{i=k+1}^{\infty} \phi^{2i} (\phi^{-k} - \phi^k)^2 \\ &\approx \frac{1}{n} \left\{ (1 - \phi^{2k})(1 + \phi^2)(1 - \phi^2)^{-1} - 2k\phi^{2k} \right\}. \end{aligned} \quad (10)$$

A further approximation of (10) gives for  $k = 1$  (Cryer & Chan, 2008)

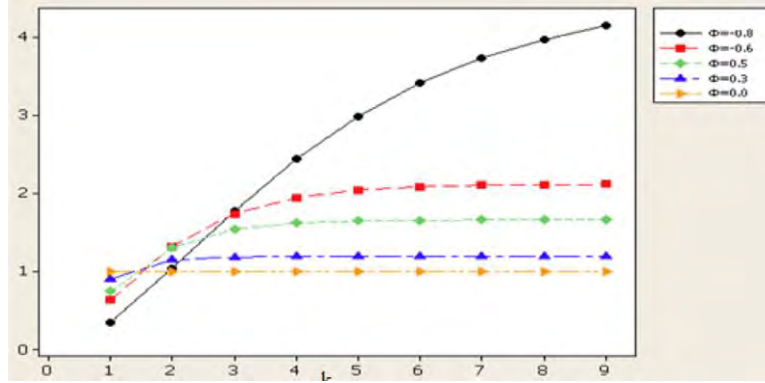
$$Var(r_k) \approx \frac{1 - \phi^2}{n}. \quad (11)$$

So that the closer  $\phi$  is to  $\pm 1$  the more accurate the estimate of  $\rho_1 (= \phi)$  becomes. For large values of  $k$ , the terms in (10) involving  $\phi^k$  could be ignored so that

$$Var(r_k) \approx \frac{1}{n} \left[ \frac{1 + \phi^2}{1 - \phi^2} \right]. \quad (12)$$

In Figure 1,  $nVar(r_k)$  is sketched based on (10) for some selected values of  $\phi$ . Notice in this figure that for  $k = 1$ , as  $|\phi|$  gets larger,  $Var(r_k)$  is decreasing. Therefore, with stronger autocorrelation among the data, the moment estimator  $r_1$  is more accurate. The opposite happens for  $k \geq 1$ , that is as  $|\phi|$  approaches one,  $Var(r_k)$  becomes larger.

## ROBUSTNESS OF ESTIMATORS OF THE ACF OF AR(1) PROCESS



**Figure 1.**  $n\text{Var}(r_k)$  for the AR(1) model for several values of  $\phi$  and  $k$ .

For the AR(1) model, (8) can also be simplified for general  $0 < k < k + j$  as (Cryer & Chan, 2008)

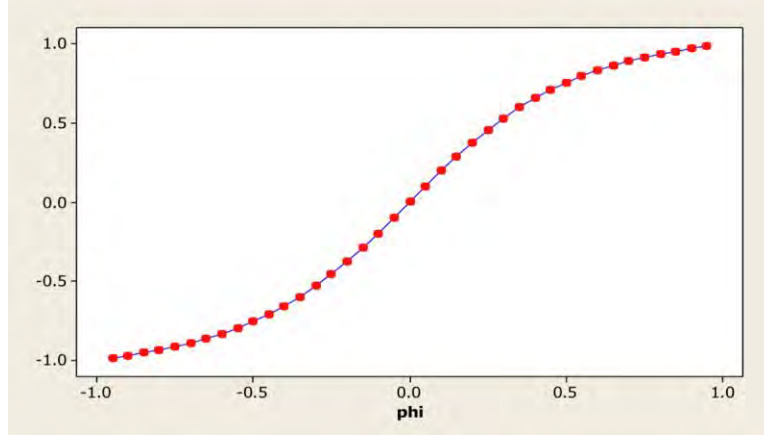
$$\text{Cov}(r_k, r_{k+j}) \cong \frac{1}{n} \frac{(\phi^j - \phi^{2k+j})(1 + \phi^2)}{(1 - \phi^2)} + j\phi^j - (2k + j)\phi^{2k+j}.$$

In particular,

$$\text{Corr}(r_1, r_2) \cong 2\phi \sqrt{\frac{1 - \phi^2}{1 + 2\phi^2 + 3\phi^4}}.$$

Figure 2 shows  $|\text{Corr}(r_1, r_2)|$  for some selected values of  $\phi$ . This figure shows a stronger association between  $r_1$  and  $r_2$  when  $|\phi|$  is closer to one. More precisely, for the AR(1) time series data, when  $\phi \approx 1$  ( $\phi \approx -1$ ) a large positive (negative)  $r_1$  is expected to be followed by a relatively large positive (positive)  $r_2$ . This agrees with the theoretical ACF of AR(1) model,  $\rho_k = \phi^k$ , which is alternating for negative values of  $\phi$ .

Notice from the discussion above, that for stationary time series data,  $r_k$  is asymptotically unbiased. The formulas for the variance and covariances among various sample autocorrelations depend mainly on the theoretical ACF of the model and they are again asymptotic. In the following example, using Monte-Carlo simulation, the bias and MSE of  $r_k$  for the AR(1) model are studied and the accuracy of the asymptotic variance of  $r_k$  given by (11) and (12) is investigated.



**Figure 2.** Asymptotic  $\text{Corr}(r_1, r_2)$  for AR(1) model for several values of  $\phi$ .

### Example 1

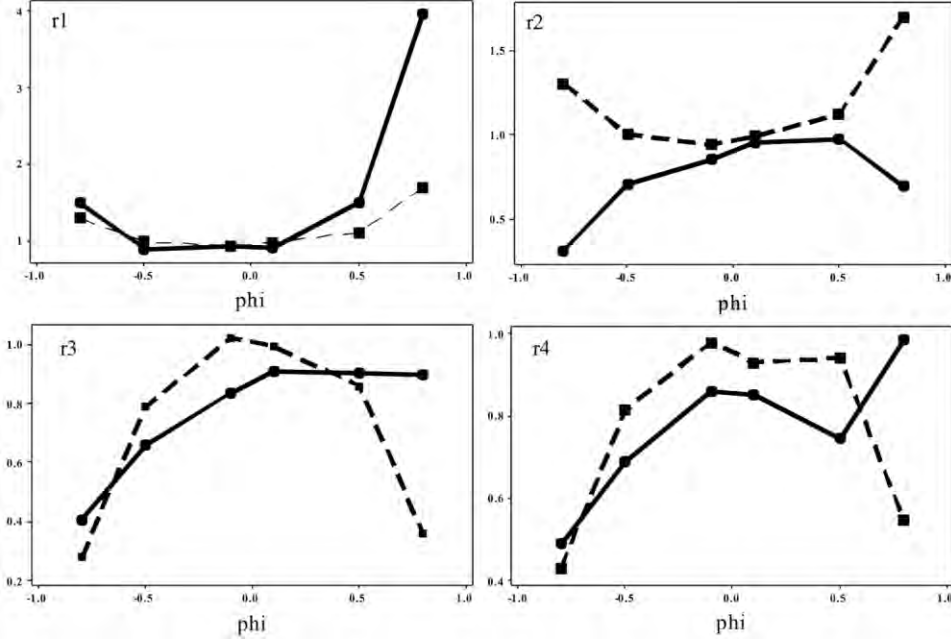
The accuracy of the formulas for  $\text{Var}(r_k)$  given in (11) and (12) are now studied. Assuming the zero-mean AR(1) model with  $\phi = -0.8, -0.5, -0.1, 0.1, 0.5, 0.8$ , one thousand realizations each of length  $n = 30, 100$  are generated from this model assuming that the WN process is iid  $N(0, 1)$ , then the sample ACF for lags  $k = 1, \dots, 5$  is computed, then the sample MSE are computed (in terms of  $\rho_1$  and its estimates  $r_{1(1)}, \dots, r_{1(1000)}$ ) as

$$MSE = \frac{1}{1000} \sum_{i=1}^{1000} \left( \rho_1 - r_{1(i)} \right)^2.$$

finally  $\text{Rel-MSE} = \text{MSE}/\text{Var}(r_k)$  are computed, where (11) and (12) are used for  $\text{Var}(r_k)$ , which in turn are sketched in Figure 3. The simulations are carried out using the R-package through the R-command `sim.ARIMA`

In view of Figure 3, it can be seen that  $\text{Rel-MSE}$  is close to one (which means that the asymptotic formulas in (11) and (12) become more accurate) when  $\phi$  is close to zero and  $n$  is large. In addition, it seems that (11) underestimates the actual variance for  $r_1$ . For  $k \geq 2$ , the asymptotic variance is defined by (12). For  $r_2$  with  $n = 100$  this formula again underestimates the actual variance but, unexpectedly, not true for  $n = 30$ . For larger time lags, it is seen that  $\text{Rel-MSE} < 1$ , so that (12) overestimates the actual variance of  $r_k$ . Therefore, in practice, (11)

and (12) should be used with caution as they may produce poor results depending on the type and strength of autocorrelation among data and the realization length.



**Figure 3.** The Rel-MSE of  $r_k$  for the AR(1) model for various values of  $\phi$  ( $n = 30$ : ~,  $n=100$ : ---)

### The empirical distributions of some robust estimators of $\rho_1$ for the AR(1) model

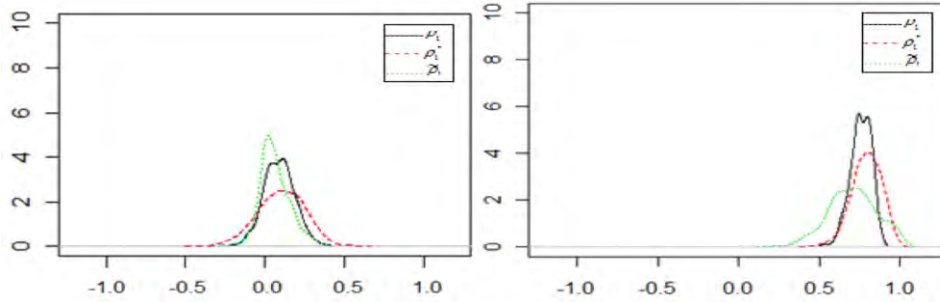
An estimation procedure is said to be robust if it is little influenced by blatant departures from assumptions. Such procedures aim to minimize the influence of outliers or departure from model assumptions while performing at the same time as well as the optimum methods when assumptions hold (Sprenst and Smeeton, 2001).

In (3) and (4), two robust estimators of  $\rho_1$  are defined due to Berkoun, et al. (2003) which have been generalized to higher time lags in (6) and (7). Recall that  $\rho_1$  is of particular importance specially in the AR(1) model for which  $\rho_1 = \phi$ . This value determines the strength and pattern of all remaining autocorrelations. Also, in many routine statistical analyses, as for instance in testing for autocorrelated

errors in regression analysis, only  $\rho_1$  is usually investigated. The following example investigates the distributions of various estimators of  $\rho_1$  for the AR(1) model.

### Example 2

Assuming the zero-mean AR(1) model with WN following  $N(0, \sigma^2)$ , then using Monte-Carlo simulation the distributions of  $r_1, \rho_1^*$  and  $\tilde{\rho}_1$  for  $\phi = 0.1, 0.8, n = 100$  and  $\sigma^2 = 1$  are compared. An r-code is written by the authors to accomplish this job. The empirical distributions of various estimators are obtained based on one thousand repetitions using the r-command (density). Results are presented graphically in Figure 4. This figure shows that the empirical distributions of various estimators are unimodal and nearly symmetric. The distribution of  $\rho_1^*$  seems closer to normality than other distributions. As  $\phi$  is increased from 0.1 to 0.8, the location of various distributions is shifted up towards 0.8. Also, the variability in the distributions of  $r_1$  and  $\rho_1^*$  is decreased more than that of the second robust estimator,  $\tilde{\rho}_1$ .



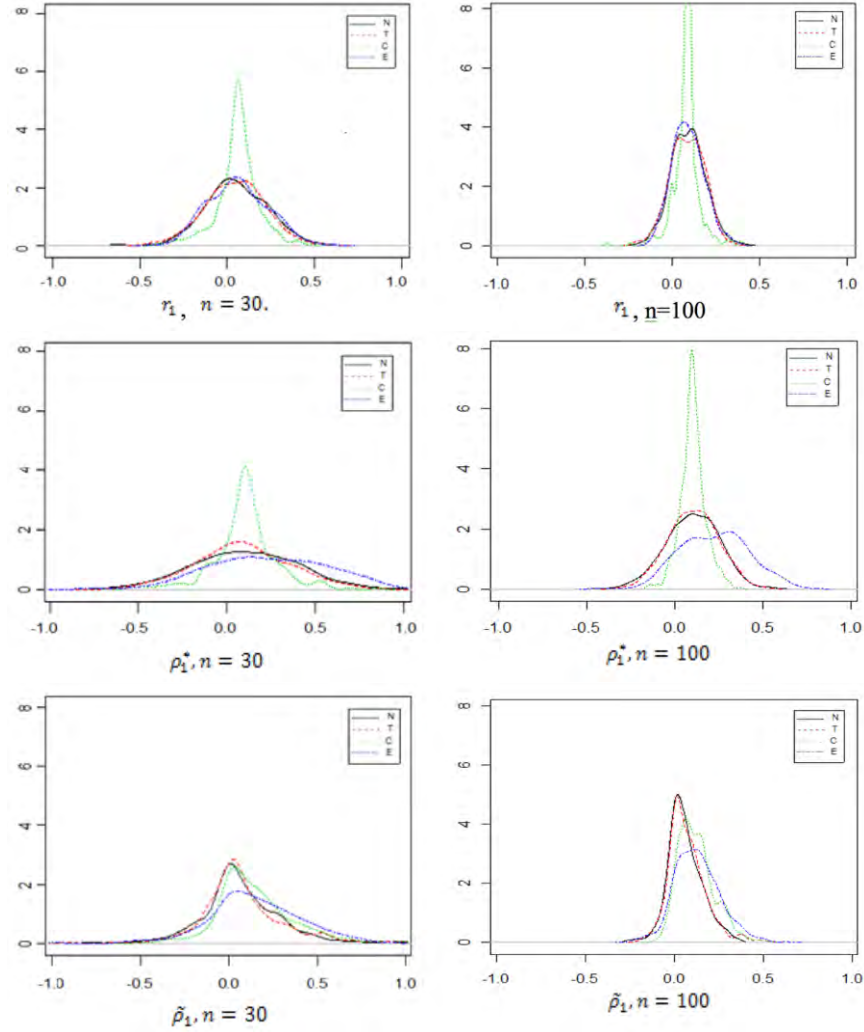
**Figure 4.** The empirical pdf of three estimators of  $\rho_1$  for AR(1) model; for  $\phi = 0.1$  (left) and  $\phi = 0.8$  (right),  $n = 100$  and  $\sigma^2 = 1$

In traditional time series analysis, it is usually assumed that the WN terms in the AR(1) model are iid  $N(0, \sigma^2)$ , as above. Therefore, it is crucial to explore the robustness of various estimators of the ACF if the WN terms are not normal. In the following example the empirical distributions of various estimators for  $\rho_1$  in the AR(1) model are investigated assuming that the WN terms follow the normal, student-t, Cauchy and exponential. The choice of the student-t and Cauchy

## ROBUSTNESS OF ESTIMATORS OF THE ACF OF AR(1) PROCESS

distributions was to study the effect of tail-heaviness of WN distribution whereas the exponential distribution is used to study the effect of skewness of WN distribution.

### Example 3



**Figure 5.** Empirical distributions of  $r_1$ ,  $\rho_1^*$  and  $\tilde{\rho}_1$  of AR(1) model; for  $\phi = 0.1$ , for various error distributions.

Assuming the zero-mean AR(1) model with WN following  $N(0,1)$ ,  $t_5$ , Cauchy (0,1), and (the zero-mean)  $\text{Exp}^*(1)$  (that is, the ordinary exponential distribution



with mean 1 but shifted left by one unit). Again, using Monte-Carlo simulation the distribution of  $r_1, \rho_1^*$  and  $\tilde{\rho}_1$  for  $\phi = 0.1, 0.8$ ,  $n = 30, 100$  are compared. The empirical distributions of various estimators is obtained based on 1,000 repetitions using the r-command (density). The results are summarized in Figure 5. Also, the p-values of two tests of normality for the empirical distributions, namely the Shapiro-Wilk test (SWT) and the Anderson-Darling test (ADT), are presented in Table 1. To perform these tests, the R-commands shapiro.test(X) and ad.test(x) were used which belong, respectively to the stats and nortest R-packages. A detailed account of these tests and other tests of normality is found in Thode (2002).

**Table 1.** Normality tests of the distributions of  $r_1, \rho_1^*$  and  $\tilde{\rho}_1$  of AR(1) model; for  $\phi=0.1, 0.8, n=30, 100$ , along various error distributions.

$\phi$	$n$	N			T		
		$r_1$	$\rho_1^*$	$\tilde{\rho}_1$	$r_1$	$\rho_1^*$	$\tilde{\rho}_1$
0.1	30	0.0700	0.0458	2.50E-11	0.5264	0.9339	5.61E-12
		0.0076	0.0259	2.20E-16	0.5338	0.9008	2.20E-16
	100	0.4270	0.7529	7.76E-11	0.3117	0.7291	2.20E-16
		0.2414	0.7529	1.39E-15	0.6098	0.9912	2.20E-16
0.8	30	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
		2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	100	9.23E-12	5.63E-09	1.96E-07	1.28E-11	1.19E-08	7.80E-07
		6.05E-14	5.12E-07	2.00E-03	1.06E-13	3.48E-11	2.44E-03
$\phi$	$n$	C			E		
		$r_1$	$\rho_1^*$	$\tilde{\rho}_1$	$r_1$	$\rho_1^*$	$\tilde{\rho}_1$
0.1	30	9.54E-16	2.20E-16	2.20E-16	0.01542	1.21E-05	3.35E-07
		2.20E-16	2.20E-16	2.20E-16	0.01491	3.02E-05	1.60E-11
	100	2.20E-16	3.41E-15	2.20E-16	6.92E-05	4.12E-05	9.54E-15
		2.20E-16	2.20E-16	2.20E-16	6.42E-05	7.41E-05	1.43E-15
0.8	30	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
		2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16	2.20E-16
	100	2.20E-16	2.20E-16	2.20E-16	3.29E-10	1.74E-13	3.77E-09
		2.20E-16	2.20E-16	6.57E-15	2.16E-08	3.61E-14	3.47E-06

In view of Table 1, it can be seen that the majority of distributions are far from normality, especially when the WN distribution is far from normality. The

departure from normality is specifically seen for smaller  $n$  as well as larger  $\phi$ . The normality assumption is validated only with  $\phi = 0.1$ , for  $r_1$  and  $\rho_1^*$  along normal and  $t$  WN distributions. In summary, results indicated that the assumption of normality for any of the estimators of  $\rho_1$  considered here is mostly invalid. This result agrees with those of Hassani (2010).

As far as Figure 5 is considered, it can be seen that the empirical distribution of  $r_1$  is not affected by the WN distribution, except the Cauchy case which showed a much higher kurtosis than other distributions. This is true for  $n = 30$  and  $100$ . A nearly similar conclusion is seen for  $\rho_1^*$ . For  $\tilde{\rho}_1$ , no significant differences are seen, especially for  $n = 30$ , among its empirical distributions in terms of WN distribution including the Cauchy distribution, whereas all distributions here show some positive skewness. In overall, it seems that  $\tilde{\rho}_1$  is less affected by the change of WN distribution as compared to other estimators.

### **Bias and MSE for various estimators of ACF of AR(1) with Gaussian errors**

Now, go back to the case of normal WN distribution and again the AR(1) model. The objective is to study the precision and accuracy of various estimators for  $\rho_k$  defined in (5) – (7). Example 1 defined the empirical MSE of  $r_1$ . Similarly, the bias of  $r_1$  is defined as

$$Bias = \frac{1}{1000} \sum_{i=1}^{1000} (\rho_1 - r_{1(i)}).$$

Because interest is in determining the bias and MSE for various time lags and  $\phi$  which in turn change the theoretical autocorrelations, the better comparable measures are computed, namely the relative bias (RB) and relative root MSE (RRMSE) defined as

$$RB(\hat{\theta}) = \frac{B(\hat{\theta})}{\theta}$$

and

$$RRMSE(\hat{\theta}) = \frac{\sqrt{MSE(\hat{\theta})}}{|\theta|}.$$

**Table 2.** The RB and RRMSE for various estimators of the ACF of AR(1) model.

$\phi$	Est	$n = 30$				
		1	2	3	4	5
-0.8	$r_k$	0.079 (0.168)	0.172 (0.340)	0.232 (0.485)	0.313 (0.665)	0.355 (0.827)
	$p_k^*$	0.038 (0.235)	0.056 (0.418)	0.078 (0.626)	0.111 (0.899)	0.153 (1.197)
	$\tilde{p}_k$	0.127 (0.331)	0.176 (0.484)	0.199 (0.651)	0.218 (0.871)	0.244 (1.136)
-0.5	$r_k$	0.062 (0.301)	0.240 (0.792)	0.234 (1.528)	0.565 (3.131)	(0.013) (5.987)
	$p_k^*$	0.028 (0.494)	0.081 (1.219)	0.188 (2.520)	0.203 (5.219)	0.013 (11.136)
	$\tilde{p}_k$	0.201 (0.561)	0.282 (1.061)	0.355 (2.043)	0.394 (4.089)	0.262 (8.723)
-0.1	$r_k$	(0.268) (1.755)	2.610 (17.020)	(20.900) (168.500)	287.000 (1711.700)	(1960.000) (16763.100)
	$p_k^*$	(0.040) (2.888)	(0.950) (27.530)	9.000 (291.900)	137.000 (3108.100)	1080.000 (31686.000)
	$\tilde{p}_k$	0.213 (2.211)	(0.130) (22.270)	6.800 (226.100)	130.000 (2370.700)	300.000 (25059.900)
0.1	$r_k$	0.442 (1.732)	4.180 (17.970)	47.100 (175.500)	295.000 (1702.900)	3280.000 (17406.900)
	$p_k^*$	0.023 (2.827)	0.790 (29.780)	17.000 (302.700)	(147.000) (3028.200)	(470.000) (32542.300)
	$\tilde{p}_k$	0.247 (2.112)	0.450 (22.270)	11.400 (230.000)	(138.000) (2267.200)	(920.000) (24819.300)
0.5	$r_k$	0.209 (0.388)	0.526 (0.932)	0.966 (1.787)	1.634 (3.256)	2.906 (6.336)
	$p_k^*$	0.060 (0.507)	0.076 (1.257)	0.198 (2.747)	0.195 (5.445)	0.221 (11.250)
	$\tilde{p}_k$	0.218 (0.562)	0.270 (1.077)	0.289 (2.172)	0.350 (4.443)	0.346 (9.230)
0.8	$r_k$	0.200 (0.272)	0.395 (0.508)	0.582 (0.721)	0.780 (0.945)	0.986 (1.170)
	$p_k^*$	0.035 (0.225)	0.073 (0.428)	0.094 (0.657)	0.105 (0.891)	0.127 (1.242)
	$\tilde{p}_k$	0.129 (0.329)	0.182 (0.497)	0.197 (0.673)	0.214 (0.863)	0.224 (1.134)

# ROBUSTNESS OF ESTIMATORS OF THE ACF OF AR(1) PROCESS

**Table 2, cont.**

		$n = 100$				
$\phi$	Est	1	2	3	4	5
-0.8	$r_k$	0.026 (0.086)	0.056 (0.176)	0.077 (0.273)	0.104 (0.387)	0.119 (0.514)
	$p_k^*$	0.012 (0.127)	0.024 (0.243)	0.027 (0.363)	0.024 (0.504)	0.011 (0.680)
	$\tilde{p}_k$	0.110 (0.229)	0.183 (0.357)	0.237 (0.465)	0.245 (0.555)	0.251 (0.669)
-0.5	$r_k$	0.013 (0.173)	0.064 (0.458)	0.068 (0.933)	0.190 (1.986)	0.061 (3.945)
	$p_k^*$	0.006 (0.268)	(0.001) (0.655)	0.000 (1.353)	0.061 (2.889)	(0.125) (5.778)
	$\tilde{p}_k$	0.237 (0.381)	0.352 (0.632)	0.414 (0.990)	0.411 (1.893)	0.346 (3.592)
-0.1	$r_k$	(0.037) (0.964)	0.970 (10.200)	(10.800) (100.000)	124.000 (1000.000)	(720.000) (10148.900)
	$p_k^*$	0.029 (1.565)	0.680 (15.750)	(8.800) (154.600)	(17.000) (1486.600)	520.000 (16522.700)
	$\tilde{p}_k$	0.414 (1.058)	0.640 (9.327)	(4.700) (93.270)	(20.000) (866.000)	(10.000) (9434.000)
0.1	$r_k$	0.158 (0.990)	1.360 (10.050)	14.000 (97.470)	102.000 (964.400)	1460.000 (10148.900)
	$p_k^*$	0.046 (1.533)	0.100 (16.210)	(2.700) (156.800)	21.000 (1612.500)	470.000 (16673.300)
	$\tilde{p}_k$	0.438 (1.058)	0.330 (9.434)	(3.700) (96.950)	(6.000) (927.400)	80.000 (9848.900)
0.5	$r_k$	0.053 (0.183)	0.141 (0.478)	0.278 (1.002)	0.549 (1.992)	0.986 (3.880)
	$p_k^*$	0.008 (0.271)	0.009 (0.650)	(0.010) (1.415)	0.122 (2.920)	0.064 (5.831)
	$\tilde{p}_k$	0.242 (0.381)	0.352 (0.620)	0.397 (1.024)	0.406 (1.979)	0.480 (3.677)
0.8	$r_k$	0.052 (0.098)	0.108 (0.200)	0.168 (0.309)	0.230 (0.431)	0.302 (0.574)
	$p_k^*$	0.009 (0.123)	0.010 (0.228)	0.029 (0.348)	0.018 (0.473)	0.014 (0.660)
	$\tilde{p}_k$	0.115 (0.229)	0.190 (0.347)	0.235 (0.451)	0.256 (0.530)	0.275 (0.639)

Therefore, assuming the same model and settings in [Example 1](#), and based on one thousand realizations, the RB and RRMSE for  $r_k$ ,  $p_k^*$  and  $\tilde{p}_k$  are computed and then summarized in [Table 2](#). The main advantage of adopting RB and RRMSE is that they can be used to compare any two cases (cells) within [Table 2](#), regarding the value of  $\phi$ ,  $n$  or estimator.

The first conclusion from Table 2 is that, for fixed  $\phi$  and  $n$  the RB and RRMSE increase for all estimators as the time lag  $k$  is increasing. The RB and RRMSE also increase for all estimators as  $|\phi|$  approaches zero. Thus, it may be concluded that with stronger autocorrelation among data, all estimators perform better than for weaker autocorrelation. It can also be seen that the RB and RRMSE for negative values of  $\phi$  are slightly smaller than their corresponding positive values.

As the sample size increases, it can be seen that the RB and RRMSE are decreasing for  $r_k$  and  $\rho_k^*$ . For  $\rho_k^*$ , the RRMSE is decreasing along  $n$ , but the RB shows no clear pattern.

When  $|\phi|$  is large, no big differences are seen in RB and RRMSE for various estimators, while discrepancies appear as  $|\phi|$  gets closer to zero. In overall, it seems that  $\rho_k^*$  is better than other estimators in terms of RB and  $r_k$  is better than other estimators in terms of RRMSE.

## Conclusions

This study considered the statistical properties of the ACF of the AR(1) model, beginning with instigating some asymptotic formulas for the variances and covariance for the sample ACF ( $r_k$ ) for AR(1) model. It was noticed that some asymptotic formulas for  $Var(r_k)$  are not accurate, especially for strong autocorrelation ( $|\phi|$  closer to one).

Later, the empirical distributions for three estimators of the first lag autocorrelation,  $r_1$ ,  $\rho_1^*$  and  $\tilde{\rho}_1$ , were studied, where the later two estimators are two robust estimators of  $\rho_1$ . These distributions are investigated for various error distributions. It is noticed that the empirical distributions of  $r_1$  and  $\rho_1^*$  are only affected when the error distribution is Cauchy, while the third estimator  $\tilde{\rho}_1$  is found more robust in this regard. Conversely, it is seen that the majority of empirical distributions are far from normality for all estimators.

Earlier the accuracy and precision of higher lags estimators of  $\rho_k$ , were studied, namely  $r_k$ ,  $\rho_k^*$  and  $\tilde{\rho}_k$  for the AR(1) model with normal errors. It is seen that, all estimators were more accurate and precision for  $|\phi|$  closer to one and small time lags. Besides, the RB and RRMSE dramatically increase when  $\phi$  is closer to zero and large time lags. In overall,  $r_k$  perform better than other

## ROBUSTNESS OF ESTIMATORS OF THE ACF OF AR(1) PROCESS

estimators in terms of RRMSE while  $\rho_k^*$  is better than other estimators in terms of RB.

Finally, this study indicates that the moment estimator of the ACF of AR(1) model is an important tool in the identification and estimation of such models. It seems that it behaved well when the error of distributions is non-normal. However, the accuracy and precision of the moment ACF may suffer for weaker autocorrelations among data and higher time lags. It seems that more effort is needed following the current work, either regarding the model type of data or improving the accuracy and precision of the sample ACF of data.

## References

- Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals of the Institute of Statistical Mathematics*, 26: 363-387.
- Bartlett, M. S. (1946). On the theoretical specification of sampling properties of auto-correlated time series. *Journal of the Royal Statistical Society, Series B*, 8: 27-41.
- Beguín, J. M., Gouriéroux, C., & Monfort, A. (1980). Identification of a mixed autoregressive moving average process: The corner method. In O. D. Anderson (Ed.). *Time series* (pp. 423-436). Amsterdam: North-Holland.
- Berkoun, Y., Fellag, H., & Zieliński, R. (2003). Robust Testing serial correlation in AR (1) processes in the presence of a single additive outlier. *Communications in Statistics, Theory and Methods*, 32(8): 1527-1540.
- Box, G., Jenkins, G., & Reinsel, G. (1994). *Time series analysis, forecasting and control*, 3<sup>rd</sup> Edition. New York, NY: Prentice-Hall.
- Brockwell, P., & Davis, R. (2002). *Introduction of time series and forecasting*, 2<sup>nd</sup> Edition. New York, NY: Springer-Verlag.
- Bustos, O. H., & Yohai, V. J. (1986). Robust estimates of ARMA models. *Journal of the American Statistical Association*, 81: 69-155.
- Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics*, 14: 277-293.
- Cryer, J., & Chan, K. (2008). *Time series analysis with applications in R*, 2<sup>nd</sup> Edition. New York, NY: Springer.

- Denby, L., & Martin, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association*, 74: 140-146.
- Gray, H. L., Kelley, G. D., & McIntire, D. D. (1978). A new approach to ARMA modeling, *Communications in Statistics*, 87: 1-77.
- Haddad, J. N. (2000). On robust estimation in the first-order autoregressive processes, *Communications in Statistics, Theory and Methods*, 29(11): 45-54.
- Hassani, H. (2010). A note on the sum of the sample autocorrelation function. *Journal Physica A*: 1601-1606.
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35: 73-101.
- Hurwicz, L. (1950). Least-squares bias in time series. In T. C. Koopmans (Ed.). *Statistical inference in dynamic economic models*, (pp. 365-383). New York, NY: Wiley and Sons.
- Kan, R., & Wang, X. (2010). On the distribution of the sample autocorrelation coefficients. *Econometrics*, 154: 101-121.
- Molinares, F., Reisen, V., & Neto, F. (2009). Robust estimation in long-memory processes under additive outliers. *Journal of Statistical Planning and Inference*, 139: 2511-2525.
- Smadi, A. A. (2013). A Comparison of Some Estimators of the Seasonal ACF for Various PAR Models. Accepted in the *Journal of Applied Probability and Statistics*, January, 2013.
- Smadi, A., Abu-Affouna, N., & Al-Quraan, A. (2009). Robust estimation of the seasonal autocorrelation of the PAR (1) model. *Jordan Journal of Mathematics and Statistics*, 2(2): 105-118.
- Sprent, P., & Smeeton, N. C. (2001). *Applied nonparametric statistical methods*, 3<sup>rd</sup> Edition. Boca Raton, FL: Chapman and Hall/CRC.
- Thode Jr., H. C. (2002). *Testing for normality*. New York, NY: Marcel Dekker.
- Wei, W. W. (2006). *Time series analysis, univariate and multivariate methods*, 2<sup>nd</sup> Edition. Boston, MA: Addison-Wesley.
- Zieliński, R. (1999). A median-unbiased estimator of the AR(1) coefficient. *Journal of Time Series Analysis*, 20(4): 477-481.

# Likelihood Ratio Type Test for Linear Failure Rate Distribution vs. Exponential Distribution

**R. R. L. Kantam**

Acharya Nagarjuna University  
Guntur-Andhra Pradesh, India

**M. C. Priya**

Acharya Nagarjuna University  
Guntur-Andhra Pradesh, India

**M. S. Ravikumar**

Acharya Nagarjuna University  
Guntur-Andhra Pradesh, India

---

The Linear Failure Rate Distribution (LFRD) is considered. The graphs of its probability density function are examined for selected parameter combinations. Some of them are similar to the well-known exponential distribution. Incidentally exponential distribution is one of the two component models of the LFRD model. In view of the simpler form of exponential model as applicable in inference, looking at the frequency curves of LFRD, a test statistic is proposed based on ratio of likelihood functions containing the standard forms of the density functions of both LFRD and Exponential to discriminate between LFRD and exponential models. The critical values and the powers of the test statistic are developed.

*Keywords:* Linear failure rate distribution, likelihood ratio type, test statistic, power

---

## Introduction

In reliability studies, series systems are one of many popular system configurations. If a series system has two components having independently distributed lifetime random variables with failure rate functions  $h_1(x)$  and  $h_2(x)$  then the reliability of the series system is

$$R(x) = \exp \left[ - \int_0^x \{ h_1(t) + h_2(t) \} dt \right] \quad (1)$$

The corresponding cumulative distribution function, failure density function and failure rate function are respectively given by

---

*Dr. Kantam is a Professor in the Department of Statistics. Email him at [kantam.rrl@gmail.com](mailto:kantam.rrl@gmail.com). M. S. Ravikumar is a UGC Research Fellow in the Department of Statistics. Email him at: [msrk.raama@gmail.com](mailto:msrk.raama@gmail.com).*



$$F(x) = 1 - \exp \left[ - \int_0^x \{h_1(t) + h_2(t)\} dt \right] \quad (2)$$

$$f(x) = \frac{d}{dx} F(x) \quad (3)$$

$$h(x) = \frac{f(x)}{R(x)} \quad (4)$$

Taking  $h_1(x), h_2(x)$ , as the failure rates of the exponential and Rayleigh distributions in (1) results in the most commonly used Linear Failure Rate Distribution (LFRD). More specifically, if  $h_1(x) = a$  and  $h_2(x) = bx$  then the failure density function, cumulative distribution function, hazard or failure rate function of LFRD is:

$$f(x) = (a + bx) e^{-\left(ax + \frac{bx^2}{2}\right)}; x > 0, a > 0, b > 0 \quad (5)$$

$$F(x) = 1 - e^{-\left(ax + \frac{bx^2}{2}\right)}; x > 0, a > 0, b > 0 \quad (6)$$

$$h(x) = a + bx \quad (7)$$

Bain (1974) seems to be one of the earliest works that has touched upon LFRD as a model useful for analysis in life testing. Ananda Sen (2005) gave a detailed review along with the distributional characteristics and inferential aspects of LFRD. Some basic features of LFRD are as follows:

Mean:

$$\mu = \sqrt{\frac{2\pi}{b}} e^{\frac{a^2}{2b}} \left( 1 - \phi\left(\frac{a}{\sqrt{b}}\right) \right) \quad (8)$$

where  $\phi$  denotes the cumulative distribution function of a standard normal variate.

## LIKELIHOOD RATIO TYPE TEST FOR LINEAR FAILURE RATE

Variance:

$$\sigma^2 = \frac{2}{b}(1 - a\mu) - \mu^2 \quad (9)$$

Mode:

$$M = \left( \sqrt{\frac{1}{b} - \frac{a}{b}} \right) I(a^2 < b) \quad (10)$$

where  $I(.)$  denotes indicator function.

100 p<sup>th</sup> Percentile:

$$F^{-1}(p) = \sqrt{\left(\frac{a}{b}\right)^2 - \frac{2 \log(1-p)}{b}} - \frac{a}{b} \quad (11)$$

and hence median is

$$M_d = \sqrt{\left(\frac{a}{b}\right)^2 - \frac{2 \log(0.5)}{b}} - \frac{a}{b} \quad (12)$$

In biological sciences this is called 50% survival time denoted by  $t_{50}$ .

Recurrence relation for raw moments is

$$\mu_k' = \frac{a}{k+1} \mu_{k+1}' + \frac{b}{k+2} \mu_{k+2}'; \quad k = 0, 1, 2, \dots \quad (13)$$

The second, third and fourth non-central moments are

$$\mu_2' = \frac{2}{b}(1 - a\mu) \quad (14)$$

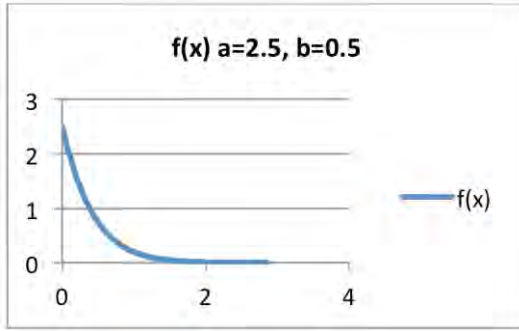
$$\mu_3' = \frac{3}{b} \left( \mu - \frac{a}{b}(1 - a\mu) \right) \quad (15)$$

$$\mu_4' = \frac{8}{b^2} + \frac{4a^2}{b^3} - \mu \left( \frac{12a}{b^2} + \frac{4a^3}{b^3} \right) \quad (16)$$

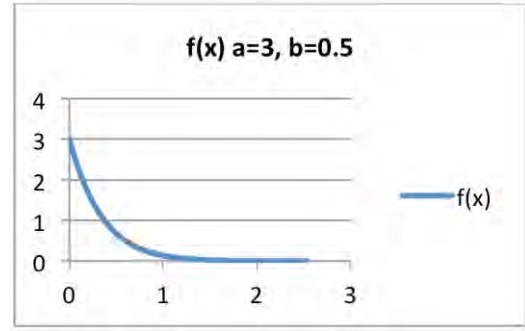
where  $\mu$  is the mean of the distribution given by (8).

It can be seen from (10) that LFRD has a non-zero mode only if its parameters  $a$  and  $b$  satisfy the relation  $a^2 < b$  with  $a > 0, b > 0$ .

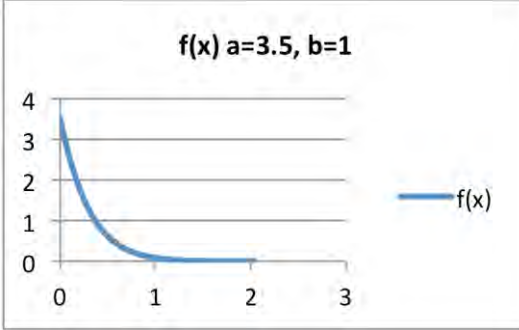
The graphs of LFRD density function for some combinations of the parameters  $a, b$  are shown in the following figures.



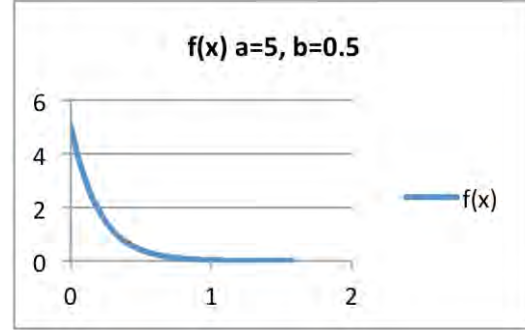
**Figure 1.** LFRD Density function when  $a = 2.5, b = 0.5$



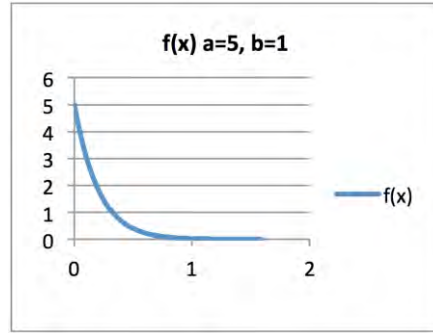
**Figure 2.** LFRD Density function when  $a = 3, b = 0.5$



**Figure 3.** LFRD Density function when  $a = 3.5, b = 1$



**Figure 4.** LFRD Density function when  $a = 5, b = 0.5$



**Figure 5.** LFRD Density function when  $a = 5$ ,  $b = 1$

In Figures 1 – 5, the combinations of  $a$  and  $b$  are bound by  $a^2 > b$ , accordingly the mode is zero and the graphs are similar to that of exponential distribution. These characteristics of LFRD and its component distribution-exponential, motivated us to study the discriminatory aspect between LFRD and exponential through statistical test procedures. Such studies of discriminatory problems between probability models are made by Gupta, et al. (2002), Gupta and Kundu (2003a), Gupta and Kundu (2003b), Kundu and Gupta (2004a, 2004b), Kundu and Manglick (2004), Kundu, et al. (2005), Kundu and Manglick (2005), Kundu (2005), Kundu and Raqab (2007), Arabin and Kundu (2009), Arabin and Kundu (2010), Arabin and Kundu (2012a), Arabin and Kundu (2012b) and the references therein. The rest of the article is organised as follows. The methodology of the proposed LR type criterion for testing is described in the next section. The critical values of the test statistic are presented in following section. The aspects of power of the proposed test statistic are given in the final section, with a comparative study.

## LR Type Methodology

Consider LFRD as a null population for example,  $P_0$ , the exponential model is regarded as an alternative population such as  $P_1$ . Let  $x_1, x_2, \dots, x_n$  be a given random sample of size  $n$ . Let  $L_1$  denote the value of the likelihood function at the sample  $x_1, x_2, \dots, x_n$  with reference to the population  $P_1$ .  $L_1$  is obtained as follows. Considering  $x_1, x_2, \dots, x_n$  as a sample from  $P_1$  with some method of point estimation using the  $P_1$  as the mathematical model, substituting the values of the

estimates so obtained and the sample observations  $x_1, x_2, \dots, x_n$  in  $L_1$  results in a value of  $L_1$  from the sample  $x_1, x_2, \dots, x_n$  with respect to  $P_1$ . Using the sample  $x_1, x_2, \dots, x_n$  with  $P_0$  as the model one can get estimates of the parameters of  $P_0$  thereby getting the value of the likelihood function in relation to  $P_0$  at  $x_1, x_2, \dots, x_n$  the parameters of  $P_0$  as estimated using  $x_1, x_2, \dots, x_n$ .  $L_0$  is thus the value of likelihood function substituting the same sample  $x_1, x_2, \dots, x_n$  and the estimates of  $P_0$ . Thus for the same sample  $x_1, x_2, \dots, x_n$ , two values of likelihood function with respect to  $P_0$  as well as  $P_1$  were obtained.

Generally in likelihood ratio test procedure the MLEs of the parameters in  $L_1$  and  $L_0$  are substituted thereby getting the value of  $L_1/L_0$  at a given samples  $x_1, x_2, \dots, x_n$  with the parameters of  $P_1, P_0$  estimated by ML method using the respective models. Because likelihood is also joint probability of the sample  $x_1, x_2, \dots, x_n$ , had the sample belonged to  $P_0$  the ratio  $L_1/L_0$  tends to be very small. If it is the other way—that is the sample is truly from  $P_1$ —then the ratio  $L_1/L_0$  tends to be very large. Hence the ratio  $L_1/L_0$  can be a criterion to test whether the sample  $x_1, x_2, \dots, x_n$  actually belongs to the population  $P_1$  or  $P_0$ . If  $L_1/L_0$  is very small it may be stated that the sample belongs to  $P_0$ . Thus the ratio  $L_1/L_0$  decides the sample to have belonged to either  $P_1$  or  $P_0$ . It is therefore necessary to get critical values for  $L_1/L_0$  to decide whether a given sample belongs to  $P_1$  or  $P_0$ . In turn this leads to the knowledge of percentiles of the sampling distribution of  $L_1/L_0$ . In the proposed method of testing LFRD vs. exponential, point estimates of the parameters were used in both null and alternative populations using any other point estimation instead of the classical ML method, because MLEs of LFRD parameters are not analytically available. Similar testing processes were adopted by other researchers (Gupta & Kundu, 2003a; Kundu, et al., 2005). The proposed method is named the LR Type Criterion. In the discussion, the methods of point estimation that are considered are Least Squares estimators, Percentiles estimators, and Weighted Least Squares Estimators. The sampling distribution of  $L_1/L_0$  is not mathematically tractable. The percentiles of  $L_1/L_0$  were obtained through Monte-Carlo simulation as described in the following section. For comparison purposes, the following parametric combinations were chosen.

## LIKELIHOOD RATIO TYPE TEST FOR LINEAR FAILURE RATE

**Table 1.** Parametric combinations chosen for the study.

Least Squares Estimators		Percentiles Estimators		Weighted Least Squares Estimators	
$a$	$b$	$a$	$b$	$a$	$b$
0.5	4.0	0.5	4.0	0.5	4.0
2.5	0.5	2.5	0.5	2.5	0.5
3.0	0.5	3.0	0.5	3.0	0.5
3.5	1.0	3.5	1.0	3.5	1.0
5.0	0.5	5.0	0.5	5.0	0.5
5.0	1.0	5.0	1.0	5.0	1.0

### LR Type Test Statistic – Critical Values

A random sample of size  $n$  is generated from LFRD ( $P_0$ ) with parameter combinations as specified in the Table 1. Using that sample the parameters of LFRD are estimated by least square method / percentile method / weighted least square method given method of estimation. The estimates so obtained are substituted in  $P_0$  in the respective places of the parameters along with the sample observations used to get those estimates thus having an estimated value of  $L_0$ . Using the same sample, the parameters appearing in  $P_1$  are estimated by a least square method / percentile method / weighted least square method in succession using the model  $P_1$  method suitable for  $P_1$ . Here because  $P_1$  is an exponential distribution the MLEs of parameters of  $P_1$  were calculated using formulae and expressions suitable for  $P_1$ . The estimates of the parameters of  $P_1$  so obtained are then substituted in  $P_1$  along with the sample observations used to get the estimates. Thus estimated likelihood function  $L_1$  are obtained by three separate methods. The ratio  $L_1/L_0$  for different samples with the same parameter combinations as described in the previous section is calculated for each sample. This procedure was repeated 10,000 times for accuracy and precision. Among these 10,000 values, various specified cut off points (percentiles) would form the critical values of  $L_1/L_0$  useful for testing. These are given below in the following Tables 2 and 3, for only the parameters ( $a=2.5$ ,  $b=0.5$ ), ( $a=3$ ,  $b=0.5$ ). Results of other parameter combinations are available from the authors.

**Table 2a:** Percentiles of  $L_1/L_0$  ::  $P_0$ : LFRD vs  $P_1$ : EXP, Least Square Estimation, ( $a=2.5$ ,  $b=0.5$ )

$n$	Least Square Estimation			
	5	10	15	20
0.00100	0.05555	0.00866	0.00641	0.00495
0.00135	0.05579	0.00980	0.00700	0.00551
0.00270	0.05802	0.01389	0.01311	0.01005
0.00500	0.06338	0.01944	0.01997	0.01839
0.01000	0.07127	0.03196	0.03634	0.04059
0.02500	0.09607	0.07852	0.09572	0.09172
0.05000	0.15049	0.17091	0.17663	0.16909
0.10000	0.27829	0.32933	0.33165	0.32362
0.90000	1.45170	1.30077	1.31607	1.35776
0.95000	2.36966	1.56559	1.55775	1.59782
0.97500	4.97214	2.00069	1.86525	1.85212
0.99000	20.67554	3.27230	2.50671	2.34857
0.99500	89.41741	6.02098	3.90709	3.01258
0.99730	206.88170	10.79545	5.50735	4.79198
0.99865	938.89170	20.64189	19.63486	12.46792
0.99000	1441.98200	40.78289	23.69878	36.68090

**Table 2b:** Percentiles of  $L_1/L_0$  ::  $P_0$ : LFRD vs  $P_1$ : EXP, Weighted Least Square Estimation, ( $a=2.5$ ,  $b=0.5$ )

$n$	Weighted Least Square Estimation			
	5	10	15	20
0.00100	0.05558	0.00865	0.00560	0.00541
0.00135	0.05639	0.00958	0.00619	0.00693
0.00270	0.06081	0.01370	0.01244	0.00971
0.00500	0.06562	0.01851	0.01819	0.01855
0.01000	0.07279	0.03215	0.03719	0.04195
0.02500	0.09342	0.07797	0.09896	0.09637
0.05000	0.14239	0.16794	0.17843	0.18065
0.10000	0.26167	0.32321	0.33875	0.34647
0.90000	1.42631	1.28926	1.36464	1.46510
0.95000	2.39327	1.58677	1.65037	1.76804
0.97500	5.02094	2.18780	2.14986	2.28297
0.99000	19.63531	3.80238	4.02971	3.72403
0.99500	88.76622	9.38806	9.31526	8.65864
0.99730	222.91150	19.79771	28.90935	27.69206
0.99865	825.53910	58.24844	314.41790	122.53770
0.99000	1537.66000	125.22960	826.64140	388.92530

# LIKELIHOOD RATIO TYPE TEST FOR LINEAR FAILURE RATE

**Table 2c:** Percentiles of  $L_1/L_0 :: P_0$ : LFRD vs  $P_1$ : EXP, Percentile Estimation, ( $a=2.5$ ,  $b=0.5$ )

$n$	Percentile Estimation			
	5	10	15	20
0.00100	0.07292	0.01208	0.00611	0.00468
0.00135	0.07628	0.01361	0.00718	0.00559
0.00270	0.08055	0.01710	0.00989	0.00979
0.00500	0.08669	0.02131	0.01592	0.01877
0.01000	0.09456	0.03308	0.03379	0.03679
0.02500	0.12043	0.07695	0.08074	0.07452
0.05000	0.16311	0.15190	0.15107	0.15069
0.10000	0.24330	0.28240	0.28224	0.28175
0.90000	2.08305	1.52860	1.46043	1.44967
0.95000	4.89041	2.23917	1.97528	1.82456
0.97500	14.79908	4.16817	3.02435	2.50522
0.99000	123.33970	19.42763	7.95037	5.62689
0.99500	748.87240	71.67762	31.90508	13.60665
0.99730	2710.38500	246.98620	100.23880	55.98616
0.99865	71595.25000	623.14900	454.89490	233.64480
0.99900	190377.10000	897.07890	952.26130	833.10900

**Table 3a:** Percentiles of  $L_1/L_0 :: P_0$ : LFRD vs  $P_1$ : EXP, Least Square Estimation, ( $a=3.0$ ,  $b=0.5$ )

$n$	Least Square Estimation			
	5	10	15	20
0.00100	0.05603	0.01062	0.00596	0.00513
0.00135	0.05691	0.01139	0.00662	0.00634
0.00270	0.06038	0.01314	0.01177	0.01171
0.00500	0.06443	0.02129	0.02265	0.02324
0.01000	0.06995	0.03725	0.04223	0.04346
0.02500	0.08877	0.09321	0.09890	0.09904
0.05000	0.13725	0.17716	0.17834	0.18288
0.10000	0.26580	0.33410	0.34319	0.33358
0.90000	1.43639	1.33367	1.35457	1.37469
0.95000	2.31841	1.62922	1.59357	1.61963
0.97500	4.98869	2.14252	1.94302	1.93538
0.99000	21.02987	4.00168	2.80106	2.50630
0.99500	80.51004	8.20346	3.78306	3.15834
0.99730	252.88440	20.03408	6.46744	3.97503
0.99865	3116.18000	71.98767	11.33482	6.36183
0.99900	59094.28000	179.53870	17.87834	7.78476



**Table 3b:** Percentiles of  $L_1/L_0$  ::  $P_0$ : LFRD vs  $P_1$ : EXP, Weighted Least Square Estimation, ( $a=3.0$ ,  $b=0.5$ )

$n$	Weighted Least Square Estimation			
	5	10	15	20
0.00100	0.05647	0.01066	0.00576	0.00664
0.00135	0.05761	0.01104	0.00669	0.00795
0.00270	0.06264	0.01395	0.01140	0.01456
0.00500	0.06704	0.02150	0.02225	0.02212
0.01000	0.07299	0.03735	0.04243	0.04817
0.02500	0.08761	0.08531	0.09748	0.10723
0.05000	0.12932	0.17186	0.18351	0.19496
0.10000	0.25456	0.32422	0.34730	0.36366
0.90000	1.42080	1.31768	1.38883	1.49770
0.95000	2.32665	1.65014	1.67716	1.86482
0.97500	4.88276	2.28681	2.14252	2.36452
0.99000	20.92875	4.86321	3.67318	3.89300
0.99500	72.31535	11.28078	6.79731	6.48627
0.99730	281.68090	32.09840	21.00146	24.05159
0.99865	2668.58100	204.66170	82.91345	187.71560
0.99000	60999.62000	313.55800	123.97500	744.18340

**Table 3c:** Percentiles of  $L_1/L_0$  ::  $P_0$ : LFRD vs  $P_1$ : EXP, Percentile Estimation, ( $a=3.0$ ,  $b=0.5$ )

$n$	Percentile Estimation			
	5	10	15	20
0.00100	0.07245	0.01317	0.00645	0.00460
0.00135	0.07337	0.01435	0.00712	0.00557
0.00270	0.08110	0.01997	0.01176	0.01181
0.00500	0.08790	0.02728	0.01912	0.02171
0.01000	0.09718	0.04102	0.03919	0.03662
0.02500	0.11891	0.08040	0.08506	0.08822
0.05000	0.16062	0.15065	0.15994	0.16028
0.10000	0.25323	0.28654	0.30645	0.29008
0.90000	2.06479	1.54506	1.49559	1.46275
0.95000	5.04268	2.27774	1.96774	1.85958
0.97500	14.98131	4.37480	3.03259	2.57786
0.99000	95.64787	17.19165	7.48988	4.85088
0.99500	765.44120	76.76962	19.14950	14.15270
0.99730	4913.02900	229.08730	59.52394	70.34382
0.99865	343286.90000	526.59070	325.89340	280.56010
0.99000	2031568.00000	1125.17300	478.65110	711.98170

## LR Type Test Statistic – Power

The LR type statistic suggested would be meaningful only if it is able to distinguish between the null and alternative populations. As is generally considered, the level of significance was fixed at 0.05. The critical value of  $L_1/L_0$  corresponding to the level of significance 0.05 is (corresponding to the percentile at 0.95) identified from the relevant portion of Tables 2 and 3.

10,000 random samples of size each  $n = 5$  (5) 20, from the alternative population (exponential) are generated. The MLE (reciprocal of sample mean) of the parameter of the alternative population, the individual sample values are substituted in  $L_1$  to get the value of  $L_1$ . Using the same sample the value of  $L_0$  as described in the previous section is also computed in order to get 10,000 values of  $L_1/L_0$  for a given sample size, for a given parametric combination and for a given method of point estimation applied to the parameters of  $P_0$ . The proportion of values of  $L_1/L_0$  that exceeded the critical value ( $c_0$ ) out of 10,000 is computed and is considered as the power of the test statistic at level of significance 0.05.

**Table 4.** Powers of LR Test Criterion at  $\alpha = 0.05$  Parameter Estimates Using P.E., L.S.E., W.L.S.E. Methods

Parameter Combinations	Estimation Method											
	Percentile				Least Squares				Weighted Least Squares			
	$n=5$	$n=10$	$n=15$	$n=20$	$n=5$	$n=10$	$n=15$	$n=20$	$n=5$	$n=10$	$n=15$	$n=20$
$a=2.5, b=0.5$	0.0539	0.0601	0.0606	0.0729	0.0598	0.0692	0.0735	0.0737	0.0587	0.0672	0.0646	0.0697
$a=3, b=0.5$	0.0516	0.0585	0.0608	0.0704	0.0612	0.0619	0.0676	0.0700	0.0608	0.0607	0.0609	0.0599
$a=3.5, b=1$	0.0534	0.0586	0.0613	0.0726	0.0632	0.0714	0.0740	0.0786	0.0621	0.0668	0.0806	0.0678
$a=5, b=0.5$	0.0505	0.0500	0.0533	0.0608	0.0581	0.0570	0.0592	0.0599	0.0589	0.0559	0.0525	0.0543
$a=5, b=1$	0.0505	0.0540	0.0549	0.0606	0.0920	0.0639	0.0645	0.0619	0.0571	0.0613	0.0534	0.0589
$a=0.5, b=4$	0.0517	0.1126	0.3692	0.6813	0.1987	0.4105	0.6137	0.7472	0.2018	0.0613	0.5280	0.6024

A large value of the power shows that the test statistic is able to distinguish between the null and alternative populations. A small value of the power would show the indistinguishability between  $P_1$  and  $P_0$  as decided by LR type test statistic. The powers so obtained are given in Table 4, treated separately for each method of estimation at a specified level of significance 0.05.

The tabulated power values are very poor touching a maximum of 0.092 at  $n=5, a=5, b=1$ . These recorded powers show that the LR type test statistic is not able to discriminate between LFRD and exponential at all the values of  $n$  and the

respective parametric combinations across the methods of estimation, except the last row of each table. It shows that exponential distribution can be used as an alternative for LFRD without much loss whereas the last row of each table shows that LFRD and exponential stand apart from each other for  $a=0.5$ ,  $b=4$ . It is therefore concluded that the simple and powerful inferential tools available for exponential may be used for LFRD also. The discrimination between LFRD and exponential is clear as evident from the last row of each table.

## References

- Arabin K. D., & Kundu, D. (2009). Discriminating among the log-normal, Weibull and generalized exponential distributions. *IEEE Transactions on Reliability*, 58(3): 416-424.
- Arabin K. D., & Kundu, D. (2010). Discriminating between the log-normal and log-logistic distributions. *Communications in Statistics – Theory and Methods*, 39: 280-292.
- Arabin K. D., & Kundu, D. (2012a). Discriminating Between the Weibull and Log-Normal Distributions for Type-II Censored Data. *Statistics*, 46(2): 197-214.
- Arabin K. D., & Kundu, D. (2012b). Discriminating Between Bivariate Generalized Exponential and Bivariate Weibull Distributions. *Chilean Journal of Statistics*, 3(1): 93-110.
- Bain, L. J. (1974). Analysis for the Linear Failure-Rate Life-Testing Distribution. *Technometrics*, 16(4): 551-559.
- Gupta, R. D., & Kundu, D. (2003a). Discriminating Between the Weibull and the GE Distributions. *Computational Statistics and Data Analysis*, 43: 179-196.
- Gupta, R. D., & Kundu, D. (2003b). Closeness of Gamma and Generalized Exponential Distributions. *Communications in Statistics – Theory and Methods*, 32(4): 705-721.
- Gupta, R. D., Kundu, D., & Manglick, A. (2002). Probability of Correct Selection of Gamma Versus GE or Weibull vs. GE Models Based on Likelihood Ratio Test. In Y. P. Chaubey, Ed. *Recent Advances in Statistical Methods* (pp. 147-156). London: World Scientific Publishing Company Inc.
- Kundu, D. (2005). Discriminating between the Normal and Laplace Distributions. In N. Balakrishnan, H. N. Nagaraja and N. Kannan (Eds.).

## LIKELIHOOD RATIO TYPE TEST FOR LINEAR FAILURE RATE

*Advances in Ranking and Selection, Multiple Comparisons, and Reliability* (pp. 65-78). Boston: Birkhauser.

Kundu, D., & Gupta, R. D. (2004a). Discriminating Between the Gamma and Generalized Exponential Distributions. *Journal of Statistical Computation and Simulation*, 74(2): 107-121.

Kundu, D., & Gupta, R. D. (2004b). Discriminating Between the Weibull and Log-Normal Distributions. *Naval Research Logistics*, 51(6): 893-905.

Kundu, D., Gupta, R. D., & Manglick, A. (2005). Discriminating Between the Log-Normal and the Generalized Exponential Distributions. *Journal of Statistical Planning and Inference*, 127: 213-227.

Kundu D., & Manglick, A., (2004). Discriminating between the Weibull and Log-Normal Distributions. *Naval Research Logistics*, 51(6): 893-905.

Kundu, D., & Manglick, A. (2005). *Discriminating Between the gamma and log-normal Distributions*. *Journal of Applied Statistical Sciences*, 14(1-2): 175-187.

Kundu, D., & Raqab, M. Z. (2005). Generalized Rayleigh Distribution: Different Methods of Estimation. *Computational Statistics and Data Analysis*, 49: 187-200.

Kundu, D., & Raqab, M. Z. (2007). Discriminating Between the Log-Normal and Generalized Rayleigh Distributions. *Statistics*, 41(6): 505-515.

Sen, A. (2005). Linear failure rate distribution. In Kotz, Balakrishnan, Read and Vidakovic, Eds. *Encyclopedia of Statistical Sciences*, (Vol. 6), pp. 4212-4217.

# Population Mean Estimation with Sub Sampling the Non-Respondents Using Two Phase Sampling

**Sunil Kumar**  
Alliance University  
Karnataka, India

**M. Viswanathaiah**  
Shantha Group of Institutions  
Karnataka, India

---

The problem of non-response in double (or two phase) sampling is dealt with combined ratio, product and regression estimators. Expressions of bias and MSE for these estimators are obtained. Comparisons of a proposed strategy with a usual unbiased estimator and other estimators are carried out and results obtained are illustrated numerically using an empirical sample.

*Keywords:* Study variable, auxiliary variable, bias, mean squared error, non-response

---

## Introduction

In surveys regarding human populations, it is common for some information to be missing, even after some callbacks. Hansen and Hurwitz (1946) considered the problem of non-response while estimating a population mean by taking a sub sample from the non-respondent group and proposed an estimator by considering the information available from response and non-response groups. In estimating population parameters such as the mean, total or ratio, product and regression, sample survey experts sometimes use auxiliary information to improve the precision of estimates. Using Hansen and Hurwitz's (1946) technique, several authors including Cochran (1977), Rao (1986, 1987), Khare and Srivastava (1993, 1995, 1997), Okafor and Lee (2000), Lundström and Särndal (2001), Särndal and Lundström (2005), Tabasum and Khan (2004, 2006), Singh and Kumar (2008, 2009a, b, 2010), and Singh, et al. (2010) have suggested improvements to the population mean estimation procedure in the presence of non-response using an auxiliary variable.

---

*Sunil Kumar is an Associate Professor. Email at [sunilbhoulgal06@gmail.com](mailto:sunilbhoulgal06@gmail.com). M. Viswanathaiah is a Professor. Email at [vmatam@gmail.com](mailto:vmatam@gmail.com).*

Following Singh and Ruiz Espejo (2007), a class of ratio-product estimators in two phase sampling in the presence of non-response is suggested in this article, and its properties studied. An estimator was studied by using one auxiliary variable for two phase sampling, which is the combined regression with Okafor and Lee's (2000) estimator and Singh and Ruiz Espejo's (2007) estimator for no information case. The conditions for attaining minimum mean squared error of the proposed classes of estimators were obtained. A comparison of the proposed estimator with other estimators was conducted and a numerical illustration is provided to support the proposed estimator.

### Double Sampling Ratio, Product and Regression Estimator

Let  $y$  and  $x$  be the study and auxiliary variables with population means  $\bar{Y}$  and  $\bar{X}$  respectively. The population is divided into  $N_1$  (responding) and  $N_2$  (non-responding) units such that  $N_1 + N_2 = N$ . When the population mean  $\bar{X}$  of the auxiliary variable  $x$  is unknown, it is suggested that a first phase sample of size  $n'$  be selected from the population of size  $N$  using the simple random sampling without replacement (SRSWOR) method, and observing the information on variable  $x$ . From these selected  $n'$  units, a second phase sample size  $n(< n')$  is selected for the study variable  $y$ , and it is observed that  $n_1$  units respond and  $n_2$  units do not respond in the sample of size  $n$ . Further, from  $n_2$  non-responding units, select a sub sample of size  $r\left(\frac{n_2}{k}\right); k > 1$  using SRSWOR. Hence, there are  $n_1 + r$  responding units on  $y$ . Consequently, to estimate  $\bar{Y}$  using the sub sampling scheme suggested by Hansen and Hurwitz (1946),

$$\bar{y}^* = w_1 \bar{y}_1 + w_2 \bar{y}_{2r},$$

where  $w_1 = (n_1/n)$ ,  $w_2 = (n_2/n)$ ;  $\bar{y}_1$  and  $\bar{y}_{2r}$  denotes the sample means of the  $y$  variable based on  $n_1$  and  $r$  units, respectively.

Similarly, to estimate the population mean  $\bar{X}$  of the auxiliary variable  $x$ , the estimator  $\bar{x}^*$ ,

$$\bar{x}^* = w_1 \bar{x}_1 + w_2 \bar{x}_{2r} \quad (1)$$

with variance

$$Var(\bar{x}^*) = \left( \frac{1}{n} - \frac{1}{N} \right) S_x^2 + \frac{W_2(k-1)}{n} S_{x(2)}^2$$

where  $S_x^2$  and  $S_{x(2)}^2$  are the population mean square of the auxiliary variable  $y$  for the entire population and for the non-responding portion of the population.

Khare and Srivastava (1993) proposed ratio and product methods for estimators respectively as:

$$t_{1R} = \bar{y}^* \left( \frac{\bar{x}'}{\bar{x}^*} \right)$$

and

$$t_{1P} = \bar{y}^* \left( \frac{\bar{x}^*}{\bar{x}'} \right).$$

The  $MSE$ 's of the estimators  $t_{1R}$  and  $t_{1P}$  to the first degree of approximation, are

$$MSE(t_{1R}) = \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) \{ S_y^2 + R(R - 2\beta_{yx}) S_x^2 \} + \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k-1)}{n} \{ S_{y(2)}^2 + R(R - 2\beta_{yx(2)}) S_{x(2)}^2 \} \right] \quad (2)$$

$$MSE(t_{1P}) = \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) \{ S_y^2 + R(R + 2\beta_{yx}) S_x^2 \} + \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k-1)}{n} \{ S_{y(2)}^2 + R(R + 2\beta_{yx(2)}) S_{x(2)}^2 \} \right] \quad (3)$$

where  $R = (\bar{Y}/\bar{X})$ ,  $S_{yx} = \rho_{yx} S_y S_x$ ,  $S_{yx(2)} = \rho_{yx(2)} S_{y(2)} S_{x(2)}$ ,  $\beta_{yx} = (S_{yx}/S_x^2)$ ,

$K_{yx(2)} = (S_{yx(2)}/S_{x(2)}^2)$ , and  $\rho_{yx}$  and  $\rho_{yx(2)}$  respectively denote the correlation coefficients between  $x$  and  $y$  for the whole population and for the non-response group of the population.

## POPULATION MEAN ESTIMATION WITH SUB SAMPLING

Okafor and Lee (2000) proposed a double (two phase) sampling regression estimator in the presence of non-response on study as well as auxiliary variables, as

$$t_{3Re} = \bar{y}^* + b(\bar{x}' - \bar{x}^*) \quad (4)$$

where

$$\hat{b} = (s_{xy}^* / s_x^{*2}), \quad s_{xy}^* = \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i + k \sum_{i=1}^r x_i y_i - n \bar{x} \bar{y}^* \right), \quad s_x^{*2} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 + k \sum_{i=1}^r x_i^2 - n \bar{x} \bar{x}^* \right).$$

The MSE of the estimator  $t_{3Re}$  is

$$MSE(t_{3Re}) = \left[ \left( \frac{1}{n} - \frac{1}{n'} \right) (1 - \rho_{yx}^2) S_y^2 + \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \frac{W_2(k-1)}{n} \{ S_{y(2)}^2 + \beta_{yx} (\beta_{yx} - 2\beta_{yx(2)}) S_{x(2)}^2 \} \right] \quad (5)$$

Singh and Ruiz Espejo (2007) defined an estimator in presence of non-response as

$$t_{SR}^* = \bar{y}^* \left\{ \alpha \frac{\bar{x}'}{\bar{x}^*} + (1 - \alpha) \frac{\bar{x}^*}{\bar{x}'} \right\} \quad (6)$$

where  $\alpha$  is any suitably chosen constant.

For  $\alpha = 0, 1$ , the class of estimators  $t_{SR}^*$  reduces to the Khare and Srivastava (1993, 1995) and Tabasum and Khan (2004) product and ratio type estimators, that is,  $t_{1P}$  and  $t_{1R}$ .

The MSE of the estimator  $t_{SR}^*$  to the first degree of approximation is

$$MSE(t_{SR}^*) = \left[ \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \{ S_y^2 + R(2\alpha - 1)(R(2\alpha - 1) - 2\beta_{yx}) S_x^2 \} \right. \\ \left. + \frac{W_2(k-1)}{n} \{ S_{y(2)}^2 + R(2\alpha - 1)(R(2\alpha - 1) - 2\beta_{yx(2)}) S_{x(2)}^2 \} \right] \quad (7)$$

which is the minimum, when

$$\alpha = \frac{1}{2} \left( 1 + \frac{D^*}{RD} \right),$$



where

$$D = \left\{ \left( \frac{1}{n} - \frac{1}{n'} \right) S_x^2 + \frac{W_2(k-1)}{n} S_{x(2)}^2 \right\}, D^* = \left\{ \left( \frac{1}{n} - \frac{1}{n'} \right) K_{yx} S_x^2 + \frac{W_2(k-1)}{n} K_{yx(2)} S_{x(2)}^2 \right\}.$$

Thus, the minimum  $MSE$  of  $t_{SR}^*$  is given by

$$\begin{aligned} MSE(t_{SR(opt)}^*) = & \left[ \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \left\{ S_y^2 + \frac{D^*}{D} \left( \frac{D^*}{D} - 2\beta_{yx} \right) S_x^2 \right\} \right. \\ & \left. + \frac{W_2(k-1)}{n} \left\{ S_{y(2)}^2 + \frac{D^*}{D} \left( \frac{D^*}{D} - 2\beta_{yx(2)} \right) S_{x(2)}^2 \right\} \right]. \end{aligned} \quad (8)$$

## The Proposed Estimator

An estimator was developed using one auxiliary variable for two phase sampling for estimating the population mean  $\bar{Y}$  of a study variable  $y$  in the presence of non-response. Okafor and Lee's (2000) estimator is combined with the estimator  $t_{SR}^*$ . Thus, the proposed estimator is:

$$\begin{aligned} t_{CR}^* &= \left\{ \bar{y}^* + b(\bar{x}' - \bar{x}^*) \right\} \left\{ \varphi \frac{\bar{x}'}{\bar{x}^*} + (1-\varphi) \frac{\bar{x}^*}{\bar{x}'} \right\} \\ &= t_{3Re} \left\{ \varphi \frac{\bar{x}'}{\bar{x}^*} + (1-\varphi) \frac{\bar{x}^*}{\bar{x}'} \right\} \end{aligned} \quad (9)$$

where  $\varphi$  is any suitably chosen constant and  $t_{3Re}$  is defined at (4).

To obtain the bias and mean squared error of  $t_{CR}^*$ ,

$$\bar{y}^* = \bar{Y}(1 + \varepsilon_0), \bar{x}^* = \bar{X}(1 + \varepsilon_1), \bar{x}' = \bar{X}(1 + \varepsilon_2), s_{xy}^* = S_{xy}(1 + \varepsilon_3), s_x^{*2} = S_x^2(1 + \varepsilon_4),$$

such that

$$E(\varepsilon_i) = 0 \quad \forall i = 0 \text{ to } 4;$$

POPULATION MEAN ESTIMATION WITH SUB SAMPLING

$$E(\varepsilon_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \frac{W_2(k-1)}{n} S_{y(2)}^2; \quad E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2 + \frac{W_2(k-1)}{n} S_{x(2)}^2;$$

$$E(\varepsilon_2^2) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_x^2; \quad E(\varepsilon_0 \varepsilon_1) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{yx} + \frac{W_2(k-1)}{n} S_{yx(2)};$$

$$E(\varepsilon_0 \varepsilon_2) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_{yx}; \quad E(\varepsilon_1 \varepsilon_2) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_x^2;$$

$$E(\varepsilon_1 \varepsilon_3) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\mu_{21}}{n \bar{X} S_{xy}} + \frac{W_2(k-1)}{n} \frac{\mu_{21(2)}}{\bar{X} S_{xy}};$$

$$E(\varepsilon_2 \varepsilon_3) = \frac{N(N-n')}{(N-1)(N-2)} \frac{\mu_{21}}{n' \bar{X} S_{xy}};$$

$$E(\varepsilon_1 \varepsilon_4) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\mu_{30}}{n \bar{X} S_x^2} + \frac{W_2(k-1)}{n} \frac{\mu_{30(2)}}{\bar{X} S_x^2};$$

$$E(\varepsilon_2 \varepsilon_4) = \frac{N(N-n')}{(N-1)(N-2)} \frac{\mu_{30}}{n' \bar{X} S_x^2};$$

where

$$\mu_{rs} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^r (y_i - \bar{Y})^s; \quad \mu_{rs(2)} = \frac{1}{N_2} \sum_{i=1}^{N=N_1+N_2} (x_i - \bar{X}_2)^r (y_i - \bar{Y}_2)^s;$$

$$\bar{X}_2 = \frac{1}{N_2} \sum_{i=1}^{N=N_1+N_2} x_i; \quad \bar{Y}_2 = \frac{1}{N_2} \sum_{i=1}^{N=N_1+N_2} y_i; \quad (r, s)$$

being non negative integers.

Expanding  $t_{CR}^*$  in terms of  $\varepsilon$ 's results in

$$t_{CR}^* = \bar{Y} \left\{ 1 + \varepsilon_0 + A_0 (\varepsilon_2 - \varepsilon_1) (1 + \varepsilon_3) (1 + \varepsilon_4)^{-1} \right\} \left\{ \varphi (1 + \varepsilon_2) (1 + \varepsilon_1)^{-1} + (1 + \varphi) (1 + \varepsilon_1) (1 + \varepsilon_2)^{-1} \right\} \quad (10)$$

where  $A_0 = (\beta/R)$ ;  $R = (\bar{Y}/\bar{X})$ .

Assume that  $|\varepsilon_4| < 1$ ,  $|\varepsilon_1| < 1$  and  $|\varepsilon_2| < 1$  so that  $(1 + \varepsilon_4)^{-1}$ ,  $(1 + \varepsilon_1)^{-1}$  and  $(1 + \varepsilon_2)^{-1}$  are expandable in terms of  $\varepsilon$ 's. Expanding the right hand side of (10) in terms of  $\varepsilon$ 's and neglecting terms of  $\varepsilon$ 's with power greater than two results in:

$$(t_{CR}^* - \bar{Y}) = \bar{Y} \left\{ \varepsilon_0 - \varepsilon_2 + \varepsilon_1 + \varepsilon_2^2 - \varepsilon_1\varepsilon_2 - \varepsilon_0\varepsilon_2 + \varepsilon_0\varepsilon_1 + \varphi(2\varepsilon_2 - 2\varepsilon_1 + \varepsilon_1^2 - \varepsilon_2^2) + 2\varphi(\varepsilon_0\varepsilon_2 - \varepsilon_0\varepsilon_1) \right. \\ \left. + 2A_0\varphi(\varepsilon_2^2 + \varepsilon_1^2 - 2\varepsilon_1\varepsilon_2) + A_0(\varepsilon_2 - \varepsilon_1 - \varepsilon_2^2 - \varepsilon_1^2 + \varepsilon_1\varepsilon_2 - \varepsilon_2\varepsilon_4 + \varepsilon_1\varepsilon_4 + \varepsilon_2\varepsilon_3 - \varepsilon_1\varepsilon_3) \right\}. \quad (11)$$

Taking expectations of both sides of (11) results in the bias of  $t_{CR}^*$  to the first degree of approximation, as

$$B(t_{CR}^*) = \left[ \begin{aligned} & \frac{1}{\bar{X}} \left( \frac{1}{n} - \frac{1}{n'} \right) \left\{ (1 - 2\varphi) S_{yx} - (R - \varphi(R + 2\beta_{yx})) S_x^2 \right\} \\ & + \frac{W_2(k-1)}{n\bar{X}} \left\{ (1 - 2\varphi) S_{yx(2)} + \varphi(R + 2\beta_{yx}) S_{x(2)}^2 \right\} \\ & - K_{yx} \left\{ \frac{N^2}{(N-1)(N-2)} \left( \frac{1}{n} - \frac{1}{n'} \right) \left( \frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2} \right) + \frac{W_2(k-1)}{n} \left( \frac{\mu_{21(2)}}{S_{xy}} - \frac{\mu_{30(2)}}{S_x^2} \right) \right\} \end{aligned} \right]. \quad (12)$$

Squaring both sides of (11) and neglecting terms of  $\varepsilon$ 's with power greater than two results in

$$(t_{CR}^* - \bar{Y})^2 = \bar{Y}^2 \left\{ \varepsilon_0 + (2\varphi - 1)(\varepsilon_2 - \varepsilon_1) + A_0(\varepsilon_2 - \varepsilon_1) \right\}^2 \\ = \bar{Y}^2 \left\{ \varepsilon_0^2 + (2\varphi - 1)^2 (\varepsilon_2^2 + \varepsilon_1^2 - 2\varepsilon_1\varepsilon_2) + A_0^2 (\varepsilon_2^2 + \varepsilon_1^2 - 2\varepsilon_1\varepsilon_2) \right. \\ \left. + 2(2\varphi - 1)(\varepsilon_0\varepsilon_2 - \varepsilon_0\varepsilon_1) + 2A_0(2\varphi - 1)(\varepsilon_2^2 + \varepsilon_1^2 - 2\varepsilon_1\varepsilon_2) + 2A_0(\varepsilon_0\varepsilon_2 - \varepsilon_0\varepsilon_1) \right\}. \quad (13)$$

Taking expectations of both sides of (13) results in the *MSE* of  $t_{CR}^*$  to the first degree of approximation as:

$$MSE(t_{CR}^*) = \left[ \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \left\{ S_y^2 + (2\varphi - 1 + A_0)^2 R^2 S_x^2 - 2(2\varphi - 1 + A_0) R S_{yx} \right\} \right. \\ \left. + \frac{W_2(k-1)}{n} \left\{ S_{y(2)}^2 + (2\varphi - 1 + A_0)^2 R^2 S_{x(2)}^2 - 2(2\varphi - 1 + A_0) R S_{yx(2)} \right\} \right] \quad (14)$$

which is the minimum

$$\varphi = \frac{1}{2} \left( 1 - A_0 + \frac{D^*}{RD} \right).$$

The resulting minimum mean squared error of  $t_{CR}^*$  is therefore given by:

$$\begin{aligned} MSE(t_{CR(opt)}^*) &= \left[ \left( \frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left( \frac{1}{n} - \frac{1}{n'} \right) \left\{ S_y^2 + \frac{D^*}{D} \left( \frac{D^*}{D} - 2\beta_{yx} \right) S_x^2 \right\} \right. \\ &\quad \left. + \frac{W_2(k-1)}{n} \left\{ S_{y(2)}^2 + \frac{D^*}{D} \left( \frac{D^*}{D} - 2\beta_{yx(2)} \right) S_{x(2)}^2 \right\} \right] \\ &= MSE(t_{SR(opt)}^*). \end{aligned} \quad (15)$$

From (1), (2), (3), (5), (7) and (14),

$$Var(\bar{y}^*) - MSE(t_{CR}^*) = \left\{ (2\varphi - 1 + A_0)^2 R^2 D - 2R(2\varphi - 1 + A_0) D^* \right\} \quad (16)$$

$$MSE(t_{1R}) - MSE(t_{CR}^*) = \left\{ R^2 D - 2RD^* - (2\varphi - 1 + A_0)^2 R^2 D + 2(2\varphi - 1 + A_0) RD^* \right\} \quad (17)$$

$$MSE(t_{1P}) - MSE(t_{CR}^*) = \left\{ R^2 D + 2RD^* - (2\varphi - 1 + A_0)^2 R^2 D + 2(2\varphi - 1 + A_0) RD^* \right\} \quad (18)$$

$$MSE(t_{3Re}) - MSE(t_{CR}^*) = - \left\{ 4\varphi^2 R^2 D - 4\varphi R(RD - \beta_{yx} D + D^*) + R^2 D \right\} \quad (19)$$

$$MSE(t_{SR}^*) - MSE(t_{CR}^*) = RD \left\{ 2A_0 - A_0^2 - 4\alpha A_0 \right\} + 2A_0 D^* \quad (20)$$

The differences given by (16), (17), (18), (19) and (20) are positive, respectively, if

$$\left. \begin{aligned} & \text{either } 0 < \varphi < \frac{1}{2} \left( 1 - A_0 \right) + \frac{D^*}{RD} \\ & \text{or } \frac{1}{2} \left( 1 - A_0 \right) + \frac{D^*}{RD} < \varphi < 0 \end{aligned} \right\} \quad (21)$$

$$\left. \begin{aligned} & \text{either } \frac{2D^* - \beta_{yx}D}{2RD} < \varphi < \left(1 - \frac{\beta_{yx}}{2R}\right) \\ & \text{or } \left(1 - \frac{\beta_{yx}}{2R}\right) < \varphi < \frac{2D^* - \beta_{yx}D}{2RD} \end{aligned} \right\} \quad (22)$$

$$\left. \begin{aligned} & \text{either } \frac{-\beta_{yx}}{2R} < \varphi < \left(1 + \frac{D^*}{RD} - \frac{\beta_{yx}}{2R}\right) \\ & \text{or } \left(1 + \frac{D^*}{RD} - \frac{\beta_{yx}}{2R}\right) < \varphi < \frac{-\beta_{yx}}{2R} \end{aligned} \right\} \quad (23)$$

$$\left. \begin{aligned} & \text{either } -\frac{1}{2} < \varphi < \left(\frac{1}{2} + \frac{D^*}{RD}\right) \\ & \text{or } \left(\frac{1}{2} + \frac{D^*}{RD}\right) < \varphi < -\frac{1}{2} \end{aligned} \right\} \quad (24)$$

$$\left. \begin{aligned} & \text{either } 0 < \varphi < \frac{1}{2} \left(1 - \frac{A_0}{2}\right) + \frac{D^*}{2RD} \\ & \text{or } \frac{1}{2} \left(1 - \frac{A_0}{2}\right) + \frac{D^*}{2RD} < \varphi < 0 \end{aligned} \right\} \quad (25)$$

The proposed estimator  $t_{CR}^*$  is more robust than estimators  $\bar{y}^*$ ,  $t_{1R}$ ,  $t_{1P}$  and  $t_{SR}^*$  respectively, if (21) – (25) hold true.

## Empirical Study

To examine the robustness of the proposed estimators, consider the following data sets (Khare & Sinha, 2004, p. 53) from a survey on physical growth of an upper socioeconomic group of 95 Varanasi school children under an Indian Council of Medical Research (ICMR) study, Department of Pediatrics, Banaras Hindu University (BHU) during 1983-1984. The first 25% (24 children) were considered non-response units.

## POPULATION MEAN ESTIMATION WITH SUB SAMPLING

The parameter values related to the study variable  $y$  (weight in kg) and the auxiliary variable  $x$  (chest circumference in cm) were:

$$\bar{Y} = 19.4968, \bar{X} = 55.8611w, S_y = 3.0435, S_x = 3.2735, S_{y(2)} = 2.3552, S_{x(2)} = 2.5137, \rho_{yx} = 0.8460, \rho_{yx(2)} = 0.7290, R = 0.3490, \beta_{yx} = 0.7865, \beta_{yx(2)} = 0.6829, W_2 = 0.25, N_2 = 24, N_1 = 71, N = 95, n = 35, n' = 70.$$

The percent relative efficiencies (PREs) of different suggested estimators were computed with respect to a usual unbiased estimator  $\bar{y}^*$  for different values of  $k$ .

**Table 1:** Percent relative efficiency of different  $\bar{Y}$  estimators with respect to  $\bar{y}^*$ .

Estimators	(1/k)			
	(1/5)	(1/4)	(1/3)	(1/2)
$\bar{y}^*$	100.00	100.00	100.00	100.00
$t_{1R}$	165.65	165.35	164.95	164.41
$t_{3Re}$	218.74	220.25	222.29	225.16
$t_{SR(opt)}^* = t_{CR(opt)}^*$	220.00	221.16	222.81	225.18

Table 1 shows that

- (i) the PRE's of the estimators  $t_{3Re}$  and  $t_{CR(opt)}^*$  increase as the value of  $k$  increases, while the PRE's of the estimator  $t_{1R}$  decrease as the value of  $k$  increases.
- (ii) the performance of the proposed estimator  $t_{CR(opt)}^*$  is the best among all other estimators  $\bar{y}^*$ ,  $t_{1R}$  and  $t_{3Re}$  because it has the largest gain in efficiency.

Based on these study results, the proposed estimator  $t_{CR}^*$  is recommended for use in practice.

## References

- Cochran, W. G. (1977). *Sampling Techniques*, 3<sup>rd</sup> Edition. New York, NY: John Wiley and Sons.
- Hansen, M. H., & Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41: 517- 529.
- Khare, B. B., & Sinha, R. R. (2004). Estimation of finite population ratio using two phase sampling scheme in the presence of non-response. *Aligarh Journal of Statistics*, 24: 43-56.
- Khare, B. B., & Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy Science Letters*, 16: 111-114.
- Khare, B. B., & Srivastava, S. (1995). Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response. *Proceedings of the Indian National Science Academy*, 65: 195-203.
- Khare, B. B., & Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of non response. *Communication in Statistics - Theory and Methods*, 26: 1779-179.
- Lundström, S., & Särndal, C. E. (2001). *Estimation in the presence of non response and frame imperfections*. Örebro: SCB-Tryck.
- 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.
- Rao, P. S. R. S. (1986). Ratio estimation with sub sampling the non-respondents. *Survey Methodology*, 12: 217-230.
- Rao, P. S. R. S. (1987). *Ratio and regression estimates with sub sampling the non-respondents*. Paper presented at a special contributed session of the International Statistical Association Meeting, Sept., 2-16, 1987, Tokyo, Japan.
- Särndal, C. E. & Lundström, S. (2005). *Estimation in surveys with non-response*. New York: John Wiley and Sons.
- Singh, H. P., & Ruiz Espejo, M. (2007). Double sampling ratio-product estimator of a finite population mean in sample survey. *Journal of Applied Statistics*, 34(1): 71-85.
- Singh, H. P., & Kumar, S. (2008). A regression approach to the estimation of finite population mean in presence of non-response. *Australian and New Zealand Journal of Statistics*, 50(4): 395-408.

## POPULATION MEAN ESTIMATION WITH SUB SAMPLING

Singh, H. P., & Kumar, S. (2009a). A general class of estimators of the population mean in survey sampling using auxiliary information with sub sampling the non-respondents. *The Korean Journal of Applied Statistics*, 22(2): 387-402.

Singh, H. P., & Kumar, S. (2009b). A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *SORT*, 33(1): 71-84.

Singh, H. P., & Kumar, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers*, 51: 559 -582.

Singh, H. P., Kumar, S., & Kozak, M. (2010). Improved estimation of finite-population mean when sub-sampling is employed to deal with non-response. *Communication in Statistics - Theory and Methods*, 39(5): 791-802.

Tabasum, R., & Khan, I. A. (2004). Double sampling for ratio estimation with non- response. *Journal of the Indian Society of Agricultural Statistics*, 58: 300-306.

Tabasum, R., & Khan, I. A. (2006). Double sampling ratio estimator for the population mean in presence of non-response. *Assam Statistical Review*, 20: 73-83.



# Two Parameter Modified Ratio Estimators with Two Auxiliary Variables for Estimation of Finite Population Mean with Known Skewness, Kurtosis and Correlation Coefficient

**Jambulingam Subramani**  
Pondicherry University  
Puducherry, India

**G. Prabavathy**  
Pondicherry University  
Puducherry, India

---

Consider the two parameter modified ratio estimators for the estimation of finite population mean using the skewness, kurtosis and correlation coefficient of two auxiliary variables. The efficiencies of the proposed modified ratio estimators are assessed with that of the simple random sampling without replacement (SRSWOR) sample mean and some of the existing ratio estimators in terms of mean squared errors. The entire above is explained with the help of certain natural populations available in the literature.

*Keywords:* Mean squared error; natural populations; percentage relative efficiency; simple random sampling

---

## Introduction

In survey sampling, consider the problem of estimating the population mean  $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$  for a finite population  $U = \{U_1, U_2, \dots, U_N\}$  of  $N$  distinct and identifiable units, where the value  $Y_i$  is measured on  $U_i$ ,  $i = 1, 2, 3, \dots, N$ . Normally the population mean is estimated by the sample mean obtained from a random sample of size  $n$  drawn by simple random sampling without replacement (SRSWOR) from a finite population, when there is no auxiliary information available. Suppose that there is an auxiliary variable  $X$  available that is positively correlated with a study variable  $Y$ , in this case, either a ratio estimator or linear regression estimator may be used to improve the efficiency of the SRSWOR

---

*Dr. Subramani is an Associate Professor in the Department of Statistics. Email him at: [drjsubramani@yahoo.co.in](mailto:drjsubramani@yahoo.co.in). G. Prabavathy is a Ph.D. Scholar in the Department of Statistics. Email her at: [praba.gopal.23@gmail.com](mailto:praba.gopal.23@gmail.com).*

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

sample mean under certain conditions (see, Cochran (1977) and Murthy (1967) for example). Further improvements can be achieved on the ratio estimator by using known parameters such as skewness, kurtosis, quartiles and coefficient of variation of the auxiliary variable; the resulting estimators are called modified ratio estimators. For further details on the modified ratio estimators, readers are referred to Kadilar and Cingi (2004, 2009), Singh and Tailor (2003, 2005), Singh (2003), Sisodia and Dwivedi (1981), Subramani (2013), Subramani and Kumarapandiyan (2012a, b, c, 2013), Upadhyaya and Singh (1999), and Yan and Tian (2010).

If two auxiliary variables exist, then several modified ratio estimators have been proposed by linking together ratio estimators, product estimators and regression estimators in order to obtain more efficient estimators. For more detailed discussion about ratio estimators and their modifications using two auxiliary variables readers are referred to: Abu-Dayyeh et al. (2003), Bandyopadhyay (1980), Cochran (1940), Kadilar and Cingi (2004, 2005), Khare et al. (2013), Murthy (1967), Naik and Gupta (1991), Olkin (1958), Perri (2004, 2007), Rao and Mudholkar (1967), Raj (1965), Sahoo and Swain (1980), Singh (2003), Singh (1965, 1967), Singh and Tailor (2003, 2005), Srivenkataramana (1980), Srivenkataramana and Tracy (1981), Tailor et al. (2011), and Tracy et al. (1996).

### Existing Estimators with and without auxiliary variables

If  $(y_1, y_2, \dots, y_n)$  is a random sample of size  $n$  drawn from a population of size  $N$  using SRSWOR, then the population mean  $\bar{Y}$  can be estimated by the sample mean  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ , which is an unbiased estimator, and its variance is given by:

$$V(\bar{y}) = \frac{(1-f)}{n} S_y^2, \text{ where } S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^N (Y_i - \bar{Y})^2, f = \frac{n}{N}. \quad (1)$$

The ratio estimator for estimating the population mean  $\bar{Y}$  of the study variable  $Y$  is defined as

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R} \bar{X}. \quad (2)$$

The mean squared error of the ratio estimator  $\hat{\bar{Y}}_R$  to the first degree of approximation is:

$$MSE(\hat{Y}_R) = \frac{(1-f)}{n} \bar{Y}^2 (C_y^2 + C_X^2 - 2p_{xy} C_x C_y). \quad (3)$$

Singh (2003) suggested a ratio estimator with two auxiliary variables for estimating a population mean:

$$\hat{Y}_1 = \bar{y} \left( \frac{\bar{X}_1}{x_1} \right) \left( \frac{\bar{X}_2}{x_2} \right). \quad (4)$$

The mean squared error of  $\hat{Y}_1$  to the first order of approximation is:

$$MSE(\hat{Y}_1) = \frac{(1-f)}{n} \bar{Y}^2 (C_y^2 + C_{X_1}^2 + C_{X_2}^2 - 2p_{yx_1} C_{x_2} C_y + 2p_{x_1x_2} C_{x_1} C_{x_2}) \quad (5)$$

Singh and Tailor (2005) suggested the following modified ratio cum product estimator with known correlation coefficient between auxiliary variables:

$$\hat{Y}_2 = \bar{y} \left( \frac{\bar{X}_1 + p_{x_1x_2}}{\bar{x}_1 + p_{x_1x_2}} \right) \left( \frac{\bar{x}_2 + p_{x_1x_2}}{\bar{X}_2 + p_{x_1x_2}} \right). \quad (6)$$

The mean squared error of  $\hat{Y}_2$  to the first order of approximation is:

$$MSE(\hat{Y}_2) = \frac{1-f}{n} \bar{Y}^2 \left[ C_y^2 + \mu_1^* C_{x_1}^2 (\mu_1^* - 2K_{yx_1}) + \mu_2^* C_{x_2}^2 (\mu_2^* + 2(K_{yx_2} - \mu_1^* K_{x_1x_2})) \right] \quad (7)$$

where

$$K_{yx_1} = \rho_{yx_1} \frac{C_y}{C_{x_1}}, K_{yx_2} = \rho_{yx_2} \frac{C_y}{C_{x_2}}, K_{x_1x_2} = \rho_{x_1x_2} \frac{C_{x_1}}{C_{x_2}}, \mu_1^* = \frac{\bar{X}_1}{\bar{X}_1 + \rho_{x_1x_2}} \text{ and } \mu_2^* = \frac{\bar{X}_2}{\bar{X}_2 + \rho_{x_1x_2}}$$

and  $\rho_{x_1x_2}$  is the coefficient of correlation between  $X_1$  and  $X_2$ .

Kadilar and Cingi (2005) proposed a new ratio estimator using two auxiliary variables as:

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

$$\hat{\bar{Y}}_3 = \bar{y} \left( \frac{\bar{X}_1}{\bar{x}_1} \right)^{\alpha_1} \left( \frac{\bar{X}_2}{\bar{x}_2} \right)^{\alpha_2} + b_1(\bar{X}_1 - \bar{x}_1) + b_2(\bar{X}_2 - \bar{x}_2). \quad (8)$$

The mean squared error of  $\bar{Y}_3$  to the first order of approximation is:

$$MSE(\hat{\bar{Y}}_3) \cong \frac{1-f}{n} \left\{ \begin{aligned} & S_y^2 + (\alpha_1 R_1 + B_1)^2 S_{x_1}^2 \\ & + (\alpha_2 R_2 + B_2)^2 S_{x_1}^2 \\ & - 2(\alpha_1 R_1 + B_1) S_{yx_1} \\ & - 2(\alpha_2 R_2 + B_2) S_{yx_2} \\ & + 2(\alpha_1 R_1 + B_1)(\alpha_2 R_2 + B_2) S_{x_1 x_2} \end{aligned} \right\} \quad (9)$$

where  $B_1 = \frac{S_{xy}}{S_{x_1}^2}$ ,  $B_2 = \frac{S_{xy}}{S_{x_2}^2}$ ,  $R_1 = \frac{\bar{Y}}{\bar{X}_1}$  and  $R_2 = \frac{\bar{Y}}{\bar{X}_2}$ .

Perri (2007) suggested some modified ratio cum product estimators using two auxiliary variables for estimating the population mean:

$$\hat{\bar{Y}}_4 = \bar{y} \frac{\hat{t}_2}{\hat{t}_1} \frac{\bar{X}_1}{\bar{X}_2}, \quad \hat{\bar{Y}}_5 = \bar{y} \frac{\bar{X}_1}{\hat{t}_1} \frac{\bar{X}_2}{\hat{t}_2} \quad \text{and} \quad \hat{\bar{Y}}_6 = \bar{y} \frac{\hat{t}_1}{\hat{t}_2} \frac{\bar{X}_2}{\bar{X}_1}, \quad (10)$$

where  $\hat{t}_1 = \bar{x}_1 + \alpha_1(\bar{X}_1 - \bar{x}_1)$  and  $\hat{t}_2 = \bar{x}_2 + \alpha_2(\bar{X}_2 - \bar{x}_2)$ .

The mean squared errors of  $\hat{\bar{Y}}_4$ ,  $\hat{\bar{Y}}_5$ ,  $\hat{\bar{Y}}_6$  to the first order of approximation are:

$$MSE(\hat{\bar{Y}}_4) = \frac{1-f}{n} \left[ S_y^2 + \gamma_{x_1}^2 + \gamma_{x_2}^2 - 2(\gamma_{yx_1} - \gamma_{yx_2} + \gamma_{x_1 x_2}) \right] \quad (11)$$

$$MSE(\hat{\bar{Y}}_5) = \frac{1-f}{n} \left[ S_y^2 + \gamma_{x_1}^2 + \gamma_{x_2}^2 - 2(\gamma_{yx_1} + \gamma_{yx_2} - \gamma_{x_1 x_2}) \right] \quad (12)$$

$$MSE(\hat{\bar{Y}}_6) = \frac{1-f}{n} \left[ S_y^2 + \gamma_{x_1}^2 + \gamma_{x_2}^2 + 2(\gamma_{yx_1} - \gamma_{yx_2} - \gamma_{x_1 x_2}) \right] \quad (13)$$

where  $\gamma_{x_1x_2} = (1-\alpha_1)(1-\alpha_2)R_1R_2S_{x_1x_2}$ ,  $\gamma_{x_1} = (1-\alpha_1)R_1S_{x_1}$ ,  $\gamma_{x_2} = (1-\alpha_2)R_2S_{x_2}$ ,  $\gamma_{yx_1} = (1-\alpha_1)R_1S_{yx_1}$  and  $\gamma_{yx_2} = (1-\alpha_2)R_2S_{yx_2}$ .

This article is concerned with estimating the population mean of a study variable  $Y$  by two parameter modified ratio estimators with known correlation coefficient, skewness and kurtosis of two auxiliary variables  $X_1$  and  $X_2$ .

## Proposed Two Parameter Modified Ratio Estimators

Whenever one or two auxiliary variables exist, a number of estimators including ratio, regression, product and chain ratio type estimators and their linear combinations have been proposed in the literature. These estimators are improved by using the known values of parameters such as skewness, kurtosis and coefficient of variation of the auxiliary variables. All of these estimators are functions of the ratio, product, regression estimators and their linear combinations; hence, an attempt is made herein to introduce the weighted average of the ratio estimators whenever there are two auxiliary variables available. As a result, two parameter modified ratio estimators with known correlation coefficient, skewness, kurtosis and their linear combinations of two auxiliary variables are proposed.

When the coefficient of kurtosis  $\beta_2(X_1)$  of the auxiliary variable  $X_1$ , and  $\beta_2(X_2)$  of the auxiliary variable  $X_2$  is known, the following two parameter modified ratio estimator is proposed:

$$\hat{\bar{Y}}_{SP1} = \bar{y} \left( \frac{\alpha_1 [\bar{X}_1 + \beta_2(X_1)] + \alpha_2 [\bar{X}_2 + \beta_2(X_2)]}{\alpha_1 [\bar{x}_1 + \beta_2(X_1)] + \alpha_2 [\bar{x}_2 + \beta_2(X_2)]} \right). \quad (14)$$

Using the linear combinations of coefficient of kurtosis  $\beta_2(X_1)$  of the auxiliary variable  $X_1$ ,  $\beta_2(X_2)$  of the auxiliary variable  $X_2$  and correlation coefficient  $\rho_{x_1x_2}$  between  $X_1$  and  $X_2$ , the following two parameter modified ratio estimators are proposed:

$$\hat{\bar{Y}}_{SP2} = \bar{y} \left( \frac{\alpha_1 [\rho_{x_1x_2} \bar{X}_1 + \beta_2(X_1)] + \alpha_2 [\rho_{x_1x_2} \bar{X}_2 + \beta_2(X_2)]}{\alpha_1 [\rho_{x_1x_2} \bar{x}_1 + \beta_2(X_1)] + \alpha_2 [\rho_{x_1x_2} \bar{x}_2 + \beta_2(X_2)]} \right) \quad (15)$$

and

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

$$\hat{Y}_{SP3} = \bar{y} \left( \frac{\alpha_1 [\beta_2(X_1)\bar{X}_1 + \rho_{x_1x_2}] + \alpha_2 [\beta_2(X_2)\bar{X}_2 + \rho_{x_1x_2}]}{\alpha_1 [\beta_2(X_1)\bar{x}_1 + \rho_{x_1x_2}] + \alpha_2 [\beta_2(X_2)\bar{x}_2 + \rho_{x_1x_2}]} \right). \quad (16)$$

Using the linear combinations of coefficient of skewness  $\beta_1(X_1)$  of the auxiliary variable  $X_1$ ,  $\beta_1(X_2)$  of the auxiliary variable  $X_2$ , coefficient of kurtosis  $\beta_2(X_1)$  of the auxiliary variable  $X_1$  and  $\beta_2(X_2)$  of the auxiliary variable  $X_2$  the following two parameter modified ratio estimators are proposed:

$$\hat{Y}_{SP4} = \bar{y} \left( \frac{\alpha_1 [\beta_2(X_1)\bar{X}_1 + \beta_1(X_1)] + \alpha_2 [\beta_2(X_2)\bar{X}_2 + \beta_1(X_2)]}{\alpha_1 [\beta_2(X_1)\bar{x}_1 + \beta_1(X_1)] + \alpha_2 [\beta_2(X_2)\bar{x}_2 + \beta_1(X_2)]} \right) \quad (17)$$

and

$$\hat{Y}_{SP5} = \bar{y} \left( \frac{\alpha_1 [\beta_1(X_1)\bar{X}_1 + \beta_2(X_1)] + \alpha_2 [\beta_1(X_2)\bar{X}_2 + \beta_2(X_2)]}{\alpha_1 [\beta_1(X_1)\bar{x}_1 + \beta_2(X_1)] + \alpha_2 [\beta_1(X_2)\bar{x}_2 + \beta_2(X_2)]} \right). \quad (18)$$

In general, the estimators proposed in (14) to (18) can be defined as particular cases of the estimator:

$$\hat{Y}_{SPT} = \bar{y} \left( \frac{\alpha_1 [\bar{X}_1 + T_1] + \alpha_2 [\bar{X}_2 + T_2]}{\alpha_1 [\bar{x}_1 + T_1] + \alpha_2 [\bar{x}_2 + T_2]} \right). \quad (19)$$

For suitable choices of  $T_1$  and  $T_2$  in (19), the estimators defined in (14) to (18) are obtained.

Suppose that,

- i. if  $T_1 = \beta_2(X_1)$  and  $T_2 = \beta_2(X_2)$  in (19), then  $\hat{Y}_{SPT}$  becomes  $\hat{Y}_{SP1}$  as defined in (14);
- ii. if  $T_1 = \frac{\beta_2(X_1)}{\rho_{x_1x_2}}$  and  $T_2 = \frac{\beta_2(X_2)}{\rho_{x_1x_2}}$  in (19), then  $\hat{Y}_{SPT}$  becomes  $\hat{Y}_{SP2}$  as defined in (15);
- iii. if  $T_1 = \frac{\rho_{x_1x_2}}{\beta_2(X_1)}$  and  $T_2 = \frac{\rho_{x_1x_2}}{\beta_2(X_2)}$  in (19), then  $\hat{Y}_{SPT}$  becomes  $\hat{Y}_{SP3}$  as defined in (16);

- iv. if  $T_1 = \frac{\beta_1(X_1)}{\beta_2(X_1)}$  and  $T_2 = \frac{\beta_1(X_2)}{\beta_2(X_2)}$  in (19), then  $\hat{Y}_{SPT}$  becomes  $\hat{Y}_{SP4}$  as defined in (17); and
- v. if  $T_1 = \frac{\beta_2(X_1)}{\beta_1(X_1)}$  and  $T_2 = \frac{\beta_2(X_2)}{\beta_1(X_2)}$  in (19), then  $\hat{Y}_{SPT}$  becomes  $\hat{Y}_{SP5}$  as defined in (18).

### Derivation of Mean Squared Error of the proposed estimators

The mean squared error of the proposed estimator  $\hat{Y}_{SPT}$  is derived as follows. If  $e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$ ,  $e_1 = \frac{\bar{x}_1 - \bar{X}_1}{\bar{X}_1}$ , and  $e_2 = \frac{\bar{x}_2 - \bar{X}_2}{\bar{X}_2}$ , then  $\bar{y} = \bar{Y}(1 + e_0)$ ,  $\bar{x}_1 = \bar{X}_1(1 + e_1)$ , and  $\bar{x}_2 = \bar{X}_2(1 + e_2)$ . From the definition of  $e_0$  and  $e_1$ ,  $E[e_0] = E[e_1] = 0$  is obtained where  $E[e_0^2] = \frac{(1-f)}{n} C_y^2$ ,  $E[e_1^2] = \frac{1-f}{n} C_{x_1}^2$ ,  $E[e_2^2] = \frac{1-f}{n} C_{x_2}^2$ ,  $E(e_0 e_1) = \frac{1-f}{n} \rho_{yx_1} C_y C_{x_1}$ ,  $E(e_0 e_2) = \frac{1-f}{n} \rho_{yx_2} C_y C_{x_2}$  and  $E(e_1 e_2) = \frac{1-f}{n} \rho_{x_1 x_2} C_{x_1} C_{x_2}$ .

The proposed estimator  $\hat{Y}_{SPT}$  can be written in terms of  $e_0$ ,  $e_1$  and  $e_2$  as:

$$\begin{aligned} \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) \left( \frac{\alpha_1[\bar{X}_1 + T_1] + \alpha_2[\bar{X}_2 + T_2]}{\alpha_1[\bar{X}_1(1 + e_1) + T_1] + \alpha_2[\bar{X}_2(1 + e_2) + T_2]} \right) \\ \Rightarrow \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) \left( \frac{\alpha_1 \bar{X}_1 + \alpha_2 \bar{X}_2 + \alpha_1 T_1 + \alpha_2 T_2}{\alpha_1 \bar{X}_1 + \alpha_2 \bar{X}_2 + \alpha_1 T_1 + \alpha_2 T_2 + \alpha_1 \bar{X}_1 e_1 + \alpha_2 \bar{X}_2 e_2} \right) \\ \Rightarrow \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) \left( \frac{1}{1 + \theta'_1 e_1 + \theta'_2 e_2} \right), \theta'_1 = \frac{\alpha_1 \bar{X}_1}{\alpha_1[\bar{X}_1 + T_1] + \alpha_2[\bar{X}_2 + T_2]} \text{ and } \theta'_2 = \frac{\alpha_2 \bar{X}_2}{\alpha_1[\bar{X}_1 + T_1] + \alpha_2[\bar{X}_2 + T_2]} \\ \Rightarrow \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) (1 + \theta'_1 e_1 + \theta'_2 e_2)^{-1} \\ \Rightarrow \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) \left( 1 - \theta'_1 e_1 - \theta'_2 e_2 + (\theta'_1 e_1 + \theta'_2 e_2)^2 \right) \\ \Rightarrow \hat{Y}_{SPT} &= \bar{Y}(1 + e_0) \left( 1 - \theta'_1 e_1 - \theta'_2 e_2 + \theta_1'^2 e_1^2 + \theta_2'^2 e_2^2 + 2\theta'_1 e_1 \theta'_2 e_2 \right) \end{aligned}$$

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

Neglecting higher order terms

$$\hat{Y}_{SPT} - Y = \bar{Y}e_0 - \bar{Y}\theta'_1e_1 - \bar{Y}\theta'_2e_2 + \bar{Y}\theta'_1e_1\theta'_2e_2 - \bar{Y}\theta'_1e_0e_1 - \bar{Y}\theta'_2e_0e_2$$

and squaring and taking expectations on both sides results in:

$$MSE(\hat{Y}_{SPT}) = E(\hat{Y}_{SPT} - \bar{Y})^2 = \bar{Y}^2 E(e_0 - \theta'_1e_1 - \theta'_2e_2)^2$$

$$\Rightarrow MSE(\hat{Y}_{SPT}) = \bar{Y}^2 E(e_0^2 + \theta_1'^2 e_1^2 + \theta_2'^2 e_2^2 - 2\theta_1' e_0 e_1 - 2\theta_2' e_0 e_2 + 2\theta_1' \theta_2' e_1 e_2)$$

$$\Rightarrow MSE(\hat{Y}_{SPT}) = \bar{Y}^2 \{E(e_0^2) + \theta_1'^2 E(e_1^2) + \theta_2'^2 E(e_2^2) - 2\theta_1' E(e_0 e_1) - 2\theta_2' E(e_0 e_2) + 2\theta_1' \theta_2' E(e_1 e_2)\}$$

$$MSE(\hat{Y}_{SPT}) = \frac{1-f}{n} \bar{Y}^2 \{C_y^2 + \theta_1'^2 C_{x_1}^2 + \theta_2'^2 C_{x_2}^2 - 2\theta_1' \rho_{yx_1} C_y C_{x_1} - 2\theta_2' \rho_{yx_2} C_y C_{x_2} + 2\theta_1' \theta_2' \rho_{x_1 x_2} C_{x_1} C_{x_2}\} \quad (20)$$

The proposed modified ratio estimator  $\hat{Y}_{SPT}$  can be easily generalized to include several auxiliary variables. If  $X_1, X_2, \dots, X_k$  are  $k$  auxiliary variables that are positively correlated with a study variable  $Y$ , then the generalized modified ratio estimator is defined as

$$\hat{Y}_{GSPT} = \bar{y} \left( \frac{\alpha_1 [\bar{X}_1 + T_1] + \alpha_2 [\bar{X}_2 + T_2] + \alpha_3 [\bar{X}_3 + T_3] + \dots + \alpha_k [\bar{X}_k + T_k]}{\alpha_1 [\bar{X}_1 + T_1] + \alpha_2 [\bar{X}_2 + T_2] + \alpha_3 [\bar{X}_3 + T_3] + \dots + \alpha_k [\bar{X}_k + T_k]} \right)$$

where  $\alpha_1, \alpha_2, \dots, \alpha_k$  are the weights and the  $T_1, T_2, \dots, T_k$  are the known parameters of the auxiliary variables.

### Efficiency Comparisons

The efficiencies of the proposed estimators for estimating the finite population mean are assessed with that of SRSWOR sample mean and other existing estimators, as previously proposed.

From expressions (20) and (1), the proposed estimators  $\hat{Y}_{SPT}$  are more efficient than the SRSWOR sample mean  $\bar{y}_r$ . The derived conditions are:



$$MSE\left(\hat{\bar{Y}}_{SPT}\right) \leq V(\bar{y}_r)$$

if

$$\theta_1^2 C_{x_1}^2 + \theta_2^2 C_{x_2}^2 \leq 2\left(\theta_1' \rho_{yx_1} C_y C_{x_1} + \theta_2' \rho_{yx_2} C_y C_{x_2} - \theta_1' \theta_2' \rho_{x_1 x_2} C_{x_1} C_{x_2}\right) \quad (21)$$

From expressions (20) and (5), the proposed estimators  $\hat{\bar{Y}}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{\bar{Y}}_1$ . The derived conditions are:

$$MSE\left(\hat{\bar{Y}}_{SPT}\right) \leq MSE(\hat{\bar{Y}}_1)$$

if

$$(\theta_1^2 - 1)C_{x_1}^2 + (\theta_2^2 - 1)C_{x_2}^2 \leq 2\left\{(\theta_1' - 1)\rho_{yx_1} C_y C_{x_1} + (\theta_2' - 1)\rho_{yx_2} C_y C_{x_2} - (\theta_1' \theta_2' - 1)\rho_{x_1 x_2} C_{x_1} C_{x_2}\right\} \quad (22)$$

From expressions (20) and (7), the proposed estimators  $\hat{\bar{Y}}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{\bar{Y}}_2$ . The derived conditions are:

$$MSE\left(\hat{\bar{Y}}_{SPT}\right) \leq MSE(\hat{\bar{Y}}_2)$$

if

$$\begin{aligned} &(\theta_1^2 - \mu_1^{*2})C_{x_1}^2 + (\theta_2^2 - \mu_2^{*2})C_{x_2}^2 \leq \\ &2\left\{(\theta_1' - \mu_1^*)\rho_{yx_1} C_y C_{x_1} + (\theta_2' - \mu_2^*)\rho_{yx_2} C_y C_{x_2} - (\theta_1' \theta_2' + \mu_1^* \mu_2^*)\rho_{x_1 x_2} C_{x_1} C_{x_2}\right\} \end{aligned} \quad (23)$$

From expressions (20) and (9), the proposed estimators  $\hat{\bar{Y}}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{\bar{Y}}_3$ . The derived conditions are:

$$MSE\left(\hat{\bar{Y}}_{SPT}\right) \leq MSE(\hat{\bar{Y}}_3)$$

if

$$\begin{aligned} &\alpha_1^2 (R_{sp}^2 - R_1^2) S_{x_1}^2 + \alpha_2^2 (R_{sp}^2 - R_2^2) S_{x_2}^2 + B_1 S_{yx_1} + B_2 S_{yx_2} \leq \\ &2\left\{R_{sp}'(\alpha_1 S_{yx_1} + \alpha_2 S_{yx_2}) - [\alpha_1 \alpha_2 R_{sp}'^2 - (\alpha_1 R_1 + B_1)(\alpha_2 R_2 + B_2)] S_{x_1 x_2}\right\} \end{aligned} \quad (24)$$

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

From expressions (20) and (11), the proposed estimators  $\hat{Y}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{Y}_4$ . The derived conditions are:

$$MSE(\hat{Y}_{SPT}) \leq MSE(\hat{Y}_4)$$

if

$$\begin{aligned} & \left[ \alpha_1^2 R_{sp}'^2 - (1 - \alpha_1)^2 R_1^2 \right] S_{x_1}^2 + \left[ \alpha_2^2 R_{sp}'^2 - (1 - \alpha_2)^2 R_2^2 \right] S_{x_2}^2 \leq \\ & 2 \left\{ S_{yx_1} \left[ \alpha_1 R_{sp}' - (1 - \alpha_1) R_1 \right] + S_{yx_2} \left[ \alpha_2 R_{sp}' + (1 - \alpha_2) R_2 \right] \right. \\ & \quad \left. - S_{x_1 x_2} \left[ \alpha_1 \alpha_2 R_{sp}'^2 + R_1 R_2 (1 - \alpha_1)(1 - \alpha_2) \right] \right\} \end{aligned} \quad (25)$$

From expressions (20) and (12), the proposed estimators  $\hat{Y}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{Y}_5$ . The derived conditions are:

$$MSE(\hat{Y}_{SPT}) \leq MSE(\hat{Y}_5)$$

if

$$\begin{aligned} & \left[ \alpha_1^2 R_{sp}'^2 - (1 - \alpha_1)^2 R_1^2 \right] S_{x_1}^2 + \left[ \alpha_2^2 R_{sp}'^2 - (1 - \alpha_2)^2 R_2^2 \right] S_{x_2}^2 \leq \\ & 2 \left\{ S_{yx_1} \left[ \alpha_1 R_{sp}' - (1 - \alpha_1) R_1 \right] + S_{yx_2} \left[ \alpha_2 R_{sp}' - (1 - \alpha_2) R_2 \right] \right. \\ & \quad \left. - S_{x_1 x_2} \left[ \alpha_1 \alpha_2 R_{sp}'^2 - R_1 R_2 (1 - \alpha_1)(1 - \alpha_2) \right] \right\} \end{aligned} \quad (26)$$

From expressions (20) and (13), the proposed estimators  $\hat{Y}_{SPT}$  are more efficient than the existing ratio estimator  $\hat{Y}_6$ . The derived conditions are:

$$MSE(\hat{Y}_{SPT}) \leq MSE(\hat{Y}_6)$$

if

$$\begin{aligned} & \left[ \alpha_1^2 R_{sp}'^2 - (1 - \alpha_1)^2 R_1^2 \right] S_{x_1}^2 + \left[ \alpha_2^2 R_{sp}'^2 - (1 - \alpha_2)^2 R_2^2 \right] S_{x_2}^2 \leq \\ & 2 \left\{ S_{yx_1} \left[ \alpha_1 R_{sp}' + (1 - \alpha_1) R_1 \right] + S_{yx_2} \left[ \alpha_2 R_{sp}' - (1 - \alpha_2) R_2 \right] \right. \\ & \quad \left. - S_{x_1 x_2} \left[ \alpha_1 \alpha_2 R_{sp}'^2 + R_1 R_2 (1 - \alpha_1)(1 - \alpha_2) \right] \right\} \end{aligned} \quad (27)$$

where  $R'_{sp} = \frac{\bar{Y}}{\alpha_1(\bar{X}_1 + T_1) + \alpha_2(\bar{X}_2 + T_2)}$

## Numerical Study

The performance of the proposed two parameter modified ratio estimators have been compared with that of the SRSWOR sample mean and some existing modified ratio estimators algebraically. However, the proposed estimators perform well compared to the existing estimators only under certain conditions and - for numerical comparisons - they are assessed for certain natural populations. In this connection, two natural populations were considered to assess the performance of the proposed estimators with that of existing estimators. Population 1 is from Singh and Chaudhary (1986, p. 177) and population 2 is from Kadilar and Cingi (2009, p. 117). The description of the study and auxiliary variables for the two populations are shown in Table 1.

**Table 1.** Description of the study variable and auxiliary variable

Population	Study Variable Y	Auxiliary Variable X <sub>1</sub>	Auxiliary Variable X <sub>2</sub>
1	Area under wheat in 1974	Area under wheat in 1971	Area under wheat in 1973
2	Length of the fish	Length of the head	Length of the fin

The population parameters and constants computed for the two populations are given in Tables 2-4.

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

**Table 2.** Parameters and Constants of the Populations

Parameter	$N$	$n$	$\bar{Y}$	$\bar{X}_1$	$\bar{X}_2$	$\rho_{yx_1}$	$\rho_{yx_2}$	$\rho_{x_1x_2}$	$\beta_{11}$
Pop. 1	34.00	20.00	856.41	208.88	199.44	0.45	0.45	0.98	0.87
Pop. 2	25.00	10.00	75.28	14.30	6.82	0.99	0.89	0.92	1.24

Parameter	$\beta_{12}$	$\beta_{21}$	$\beta_{22}$	$S_y$	$C_y$	$S_{x_1}$	$S_{x_2}$	$C_{x_1}$	$C_{x_2}$
Pop. 1	1.28	2.91	3.73	733.14	0.86	150.51	150.22	0.72	0.75
Pop. 2	0.86	4.26	4.35	17.27	0.23	3.17	1.53	0.22	0.22

**Table 3.** Variance/Mean squared error of the existing and proposed estimators for Population 1

Existing Estimators															
		$\hat{Y}_r$	$\hat{Y}_1$	$\hat{Y}_2$											
		37940.84	90847.02	40145.19	Proposed Estimators										
$\alpha_1$	$\alpha_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$					
0.0	1.0	67310.24	64818.97	64818.97	64818.97	37057.66	37047.45	37541.57	37466.93	37396.04					
0.1	0.9	62385.73	60005.70	60005.90	60005.94	36843.39	36834.06	37275.19	37210.98	37138.09					
0.2	0.8	58048.59	56317.41	56317.77	56317.84	36654.14	36645.64	37036.69	36982.32	36907.47					
0.3	0.7	54298.80	53754.11	53754.56	53754.66	36489.42	36481.71	36825.28	36780.23	36703.44					
0.4	0.6	51136.38	52315.78	52316.28	52316.42	36348.74	36341.77	36640.21	36604.01	36525.27					
0.5	0.5	48561.32	52002.43	52002.92	52003.10	36231.61	36225.37	36480.75	36452.97	36372.28					
0.6	0.4	46573.62	52814.07	52814.49	52814.70	36137.58	36132.02	36346.17	36326.46	36243.79					
0.7	0.3	45173.28	54750.68	54750.99	54751.23	36066.20	36061.30	36235.78	36223.84	36139.12					
0.8	0.2	44360.30	57812.27	57812.42	57812.69	36017.00	36012.74	36148.92	36144.46	36057.65					
0.9	0.1	44134.68	61998.84	61998.77	61999.08	35989.55	35985.91	36084.91	36087.73	35998.74					
1.0	0.0	44496.43	67310.39	67310.05	67310.39	35983.42	35980.39	36043.13	36053.04	35961.79					

**Table 4.** Variance/Mean squared error of the existing and proposed estimators for Population 2

Existing Estimators										
		$\hat{Y}_r$	$\hat{Y}_1$	$\hat{Y}_2$						
		17.90	17.58	17.58		Proposed Estimators				
$\alpha_1$	$\alpha_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$
0.0	1.0	35.07	34.61	34.61	34.61	5.32	5.54	3.89	3.90	5.72
0.1	0.9	32.15	31.58	31.62	31.64	4.50	4.72	2.84	2.84	4.72
0.2	0.8	29.57	29.24	29.31	29.34	3.85	4.07	2.12	2.12	3.92
0.3	0.7	27.33	27.58	27.67	27.71	3.32	3.53	1.62	1.62	3.28
0.4	0.6	25.42	26.60	26.71	26.75	2.89	3.10	1.26	1.26	2.77
0.5	0.5	23.85	26.31	26.41	26.47	2.54	2.74	1.01	1.01	2.36
0.6	0.4	22.62	26.71	26.79	26.86	2.26	2.44	0.83	0.83	2.03
0.7	0.3	21.72	27.78	27.83	27.92	2.02	2.19	0.70	0.70	1.76
0.8	0.2	21.16	29.55	29.55	29.65	1.83	1.99	0.61	0.61	1.55
0.9	0.1	20.94	31.99	31.94	32.05	1.67	1.81	0.55	0.55	1.38
1.0	0.0	21.05	35.12	35.00	35.12	1.53	1.67	0.51	0.51	1.25

From the values in Tables 3 and 4, the mean squared error of the proposed modified ratio estimators  $\hat{Y}_{SPj}, j=1,2,3,4,5$  are less than the variance of SRSWOR sample mean, the mean squared error of the existing modified ratio estimators  $\hat{Y}_j; j=1,2,3,\dots,6$ . Further, to show the efficiency of the proposed estimators, the percentage relative efficiencies (PRE's) of the proposed estimators with respect to the existing estimators is computed by:

$$PRE\left(\hat{Y}_{SPj}\right)=\frac{MSE(.)}{MSE\left(\hat{Y}_{SPj}\right)}*100.$$

# TWO PARAMETER MODIFIED RATIO ESTIMATORS

**Table 5.** PRE of the proposed estimator  $\hat{Y}_{SPj}$  for Population 1

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators					
			SRSWOR	Modified Ratio Estimators				
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_6$
0.0	1.0	$\hat{Y}_{SP1}$	102.38	245.15	108.33	181.64	174.91	174.91
		$\hat{Y}_{SP2}$	102.41	245.22	108.36	181.69	174.96	174.96
		$\hat{Y}_{SP3}$	101.06	241.99	106.94	179.30	172.66	172.66
		$\hat{Y}_{SP4}$	101.26	242.47	107.15	179.65	173.00	173.00
		$\hat{Y}_{SP5}$	101.46	242.93	107.35	179.99	173.33	173.33
0.1	0.9	$\hat{Y}_{SP1}$	102.98	246.58	108.96	169.33	162.87	162.87
		$\hat{Y}_{SP2}$	103.00	246.64	108.99	169.37	162.91	162.91
		$\hat{Y}_{SP3}$	101.79	243.72	107.70	167.37	160.98	160.98
		$\hat{Y}_{SP4}$	101.96	244.14	107.89	167.65	161.26	161.26
		$\hat{Y}_{SP5}$	102.16	244.62	108.10	167.98	161.57	161.58
0.2	0.8	$\hat{Y}_{SP1}$	103.51	247.85	109.52	158.37	153.65	153.65
		$\hat{Y}_{SP2}$	103.53	247.91	109.55	158.41	153.68	153.68
		$\hat{Y}_{SP3}$	102.44	245.29	108.39	156.73	152.06	152.06
		$\hat{Y}_{SP4}$	102.59	245.65	108.55	156.96	152.28	152.28
		$\hat{Y}_{SP5}$	102.80	246.15	108.77	157.28	152.59	152.59
0.3	0.7	$\hat{Y}_{SP1}$	103.98	248.97	110.02	148.81	147.31	147.32
		$\hat{Y}_{SP2}$	104.00	249.02	110.04	148.84	147.35	147.35
		$\hat{Y}_{SP3}$	103.03	246.70	109.02	147.45	145.97	145.97
		$\hat{Y}_{SP4}$	103.16	247.00	109.15	147.63	146.15	146.15
		$\hat{Y}_{SP5}$	103.37	247.52	109.38	147.94	146.46	146.46

Table 5, continued

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators						
			SRSWOR	Modified Ratio Estimators					
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$
0.4	0.6	$\hat{Y}_{SP1}$	104.38	249.93	110.44	140.68	143.93	143.93	143.93
		$\hat{Y}_{SP2}$	104.40	249.98	110.47	140.71	143.95	143.96	143.96
		$\hat{Y}_{SP3}$	103.55	247.94	109.57	139.56	142.78	142.78	142.78
		$\hat{Y}_{SP4}$	103.65	248.19	109.67	139.70	142.92	142.92	142.93
		$\hat{Y}_{SP5}$	103.88	248.72	109.91	140.00	143.23	143.23	143.23
0.5	0.5	$\hat{Y}_{SP1}$	104.72	250.74	110.80	134.03	143.53	143.53	143.53
		$\hat{Y}_{SP2}$	104.74	250.78	110.82	134.05	143.55	143.55	143.55
		$\hat{Y}_{SP3}$	104.00	249.03	110.04	133.11	142.55	142.55	142.55
		$\hat{Y}_{SP4}$	104.08	249.22	110.13	133.22	142.66	142.66	142.66
		$\hat{Y}_{SP5}$	104.31	249.77	110.37	133.51	142.97	142.97	142.97
0.6	0.4	$\hat{Y}_{SP1}$	104.99	251.39	111.09	128.88	146.15	146.15	146.15
		$\hat{Y}_{SP2}$	105.01	251.43	111.11	128.90	146.17	146.17	146.17
		$\hat{Y}_{SP3}$	104.39	249.95	110.45	128.14	145.31	145.31	145.31
		$\hat{Y}_{SP4}$	104.44	250.08	110.51	128.21	145.39	145.39	145.39
		$\hat{Y}_{SP5}$	104.68	250.66	110.76	128.50	145.72	145.72	145.72
0.7	0.3	$\hat{Y}_{SP1}$	104.99	251.38	111.09	125.00	151.50	151.50	151.50
		$\hat{Y}_{SP2}$	104.74	250.79	110.83	124.71	151.15	151.15	151.15
		$\hat{Y}_{SP3}$	104.71	250.71	110.79	124.66	151.10	151.10	151.10
		$\hat{Y}_{SP4}$	105.21	251.92	111.32	125.27	151.83	151.83	151.83
		$\hat{Y}_{SP5}$	105.20	251.89	111.31	125.25	151.81	151.81	151.81

# TWO PARAMETER MODIFIED RATIO ESTIMATORS

**Table 5**, continued

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators					
			SRSWOR	Modified Ratio Estimators				
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_6$
0.8	0.2	$\hat{Y}_{SP1}$	105.34	252.23	111.46	123.16	160.51	160.52
		$\hat{Y}_{SP2}$	105.35	252.26	111.47	123.18	160.53	160.53
		$\hat{Y}_{SP3}$	104.96	251.31	111.06	122.72	159.93	159.93
		$\hat{Y}_{SP4}$	104.97	251.34	111.07	122.73	159.95	159.95
		$\hat{Y}_{SP5}$	105.22	251.95	111.34	123.03	160.33	160.33
0.9	0.1	$\hat{Y}_{SP1}$	105.42	252.43	111.55	122.63	172.27	172.27
		$\hat{Y}_{SP2}$	105.43	252.45	111.56	122.64	172.29	172.29
		$\hat{Y}_{SP3}$	105.14	251.76	111.25	122.31	171.81	171.81
		$\hat{Y}_{SP4}$	105.14	251.74	111.24	122.30	171.80	171.80
		$\hat{Y}_{SP5}$	105.39	252.36	111.52	122.60	172.23	172.23
1.0	0.0	$\hat{Y}_{SP1}$	105.44	252.47	111.57	123.66	187.06	187.06
		$\hat{Y}_{SP2}$	105.45	252.49	111.58	123.67	187.08	187.08
		$\hat{Y}_{SP3}$	105.27	252.05	111.38	123.45	186.75	186.75
		$\hat{Y}_{SP4}$	105.24	251.98	111.35	123.42	186.70	186.70
		$\hat{Y}_{SP5}$	105.50	252.62	111.63	123.73	187.17	187.17



Table 5 shows the following ranges for the PRE of the proposed estimators:

- from 101.06 to 105.50 in comparison with the SRSWOR sample mean;
- from 241.99 to 252.62 in comparison with the existing estimator  $\hat{Y}_1$  defined in (4);
- from 106.94 to 111.63 in comparison with the existing estimator  $\hat{Y}_2$  defined in (6);
- from 122.30 to 181.69 in comparison with the existing estimator  $\hat{Y}_3$  defined in (8);
- from 142.55 to 187.17 in comparison with the existing estimator  $\hat{Y}_4, \hat{Y}_5, \hat{Y}_6$  defined in (10).

Based on these comparisons, it is concluded that the proposed estimators perform better than the SRSWOR sample mean and other existing ratio estimators for the natural population 1 considered in this study.

**Table 6.** PRE of the proposed estimator  $\hat{Y}_{SPj}$  for Population 2

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators					
			SRSWOR	Modified Ratio Estimators				
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$
0.0	1.0	$\hat{Y}_{SP1}$	336.47	330.45	330.45	659.21	650.56	650.56
		$\hat{Y}_{SP2}$	323.10	317.33	317.33	633.03	624.73	624.73
		$\hat{Y}_{SP3}$	460.15	451.93	451.93	901.54	889.72	889.72
		$\hat{Y}_{SP4}$	458.97	450.77	450.77	899.23	887.44	887.44
		$\hat{Y}_{SP5}$	312.94	307.34	307.34	613.11	605.07	605.07

# TWO PARAMETER MODIFIED RATIO ESTIMATORS

**Table 6, continued**

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators						
			SRSWOR	Modified Ratio Estimators					
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$
0.1	0.9	$\hat{Y}_{SP1}$	397.78	390.67	390.67	714.44	701.78	702.67	703.11
		$\hat{Y}_{SP2}$	379.24	372.46	372.46	681.14	669.07	669.92	670.34
		$\hat{Y}_{SP3}$	630.28	619.01	619.01	1132.04	1111.97	1113.38	1114.08
		$\hat{Y}_{SP4}$	630.28	619.01	619.01	1132.04	1111.97	1113.38	1114.08
		$\hat{Y}_{SP5}$	379.24	372.46	372.46	681.14	669.07	669.92	670.34
0.2	0.8	$\hat{Y}_{SP1}$	464.94	456.62	456.62	768.05	759.48	761.30	762.08
		$\hat{Y}_{SP2}$	439.80	431.94	431.94	726.54	718.43	720.15	720.88
		$\hat{Y}_{SP3}$	844.34	829.25	829.25	1394.81	1379.25	1382.55	1383.96
		$\hat{Y}_{SP4}$	844.34	829.25	829.25	1394.81	1379.25	1382.55	1383.96
		$\hat{Y}_{SP5}$	456.63	448.47	448.47	754.34	745.92	747.70	748.47
0.3	0.7	$\hat{Y}_{SP1}$	539.16	529.52	529.52	823.19	830.72	833.43	834.64
		$\hat{Y}_{SP2}$	507.08	498.02	498.02	774.22	781.30	783.85	784.99
		$\hat{Y}_{SP3}$	1104.94	1085.19	1085.19	1687.04	1702.47	1708.02	1710.49
		$\hat{Y}_{SP4}$	1104.94	1085.19	1085.19	1687.04	1702.47	1708.02	1710.49
		$\hat{Y}_{SP5}$	545.73	535.98	535.98	833.23	840.85	843.60	844.82
0.4	0.6	$\hat{Y}_{SP1}$	619.38	608.30	608.30	879.58	920.42	924.22	925.61
		$\hat{Y}_{SP2}$	577.42	567.10	567.10	820.00	858.06	861.61	862.90
		$\hat{Y}_{SP3}$	1420.63	1395.24	1395.24	2017.46	2111.11	2119.84	2123.02
		$\hat{Y}_{SP4}$	1420.63	1395.24	1395.24	2017.46	2111.11	2119.84	2123.02
		$\hat{Y}_{SP5}$	646.21	634.66	634.66	917.69	960.29	964.26	965.70

Table 6, continued

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators						
			SRSWOR	Modified Ratio Estimators					
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$
0.5	0.5	$\hat{Y}_{SP1}$	704.72	692.13	692.13	938.98	1035.83	1039.76	1042.13
		$\hat{Y}_{SP2}$	653.28	641.61	641.61	870.44	960.22	963.87	966.06
		$\hat{Y}_{SP3}$	1772.28	1740.59	1740.59	2361.39	2604.95	2614.85	2620.79
		$\hat{Y}_{SP4}$	2361.39	2604.95	2614.85	1772.28	1740.59	1740.59	1772.28
		$\hat{Y}_{SP5}$	758.47	744.92	744.92	1010.59	1114.83	1119.07	1121.61
0.6	0.4	$\hat{Y}_{SP1}$	792.04	777.88	777.88	1000.88	1181.86	1185.40	1188.50
		$\hat{Y}_{SP2}$	733.61	720.49	720.49	927.05	1094.67	1097.95	1100.82
		$\hat{Y}_{SP3}$	2156.63	2118.07	2118.07	2725.30	3218.07	3227.71	3236.14
		$\hat{Y}_{SP4}$	2156.63	2118.07	2118.07	2725.30	3218.07	3227.71	3236.14
		$\hat{Y}_{SP5}$	881.77	866.01	866.01	1114.29	1315.76	1319.70	1323.15
0.7	0.3	$\hat{Y}_{SP1}$	886.14	870.30	870.30	1075.25	1375.25	1377.72	1382.18
		$\hat{Y}_{SP2}$	817.35	802.74	802.74	991.78	1268.49	1270.78	1274.89
		$\hat{Y}_{SP3}$	2557.14	2511.43	2511.43	3102.86	3968.57	3975.71	3988.57
		$\hat{Y}_{SP4}$	2557.14	2511.43	2511.43	3102.86	3968.57	3975.71	3988.57
		$\hat{Y}_{SP5}$	1017.05	998.86	998.86	1234.09	1578.41	1581.25	1586.36
0.8	0.2	$\hat{Y}_{SP1}$	978.14	960.66	960.66	1156.28	1614.75	1614.75	1620.22
		$\hat{Y}_{SP2}$	899.50	883.42	883.42	1063.32	1484.92	1484.92	1489.95
		$\hat{Y}_{SP3}$	2934.43	2881.97	2881.97	3468.85	4844.26	4844.26	4860.66
		$\hat{Y}_{SP4}$	2934.43	2881.97	2881.97	3468.85	4844.26	4844.26	4860.66
		$\hat{Y}_{SP5}$	1154.84	1134.19	1134.19	1365.16	1906.45	1906.45	1912.90

# TWO PARAMETER MODIFIED RATIO ESTIMATORS

**Table 6**, continued

$\alpha_1$	$\alpha_2$	Proposed Estimators	Existing Estimators						
			<i>SRSWOR</i>	<i>Modified Ratio Estimators</i>					
			$\bar{y}_r$	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$
0.9	0.1	$\hat{Y}_{SP1}$	1071.86	1052.69	1052.69	1253.89	1915.57	1912.57	1919.16
		$\hat{Y}_{SP2}$	988.95	971.27	971.27	1156.91	1767.40	1764.64	1770.72
		$\hat{Y}_{SP3}$	3254.55	3196.36	3196.36	3807.27	5816.36	5807.27	5827.27
		$\hat{Y}_{SP4}$	3254.55	3196.36	3196.36	3807.27	5816.36	5807.27	5827.27
		$\hat{Y}_{SP5}$	1297.10	1273.91	1273.91	1517.39	2318.12	2314.49	2322.46
1.0	0.0	$\hat{Y}_{SP1}$	1169.93	1149.02	1149.02	1375.82	2295.42	2287.58	2295.42
		$\hat{Y}_{SP2}$	1071.86	1052.69	1052.69	1260.48	2102.99	2095.81	2102.99
		$\hat{Y}_{SP3}$	3509.80	3447.06	3447.06	4127.45	6886.27	6862.75	6886.27
		$\hat{Y}_{SP4}$	3509.80	3447.06	3447.06	4127.45	6886.27	6862.75	6886.27
		$\hat{Y}_{SP5}$	1432.00	1406.40	1406.40	1684.00	2809.60	2800.00	2809.60

Table 6 shows the following ranges for the PRE of the proposed estimators:

- from 312.94 to 3509.80 in comparison with SRSWOR sample mean;
- from 307.34 to 3447.06 in comparison with the existing estimator  $\hat{Y}_1$  defined in (4) and  $\hat{Y}_2$  defined in (6);
- from 613.11 to 4127.45 in comparison with the existing estimator  $\hat{Y}_3$  defined in (8);
- from 605.07 to 6886.27 in comparison with the existing estimator  $\hat{Y}_4$  defined in (10);
- from 605.07 to 6862.75 in comparison with the existing estimator  $\hat{Y}_5$  defined in (10);
- from 605.07 to 6886.27 in comparison with the existing estimator  $\hat{Y}_6$  defined in (10).

Based on these comparisons, it may be concluded that the proposed estimators perform better than the SRSWOR sample mean and other existing ratio estimators for the natural population 2 considered in this study.

## Conclusion

This article proposed two parameter modified ratio estimators with known correlation coefficient, skewness and kurtosis of the auxiliary variables and their linear combinations. The mean squared errors of the proposed estimators were derived and compared with that of SRSWOR sample mean, the classical ratio estimator and the existing modified ratio estimators. The performance of the proposed estimators was also assessed with that of the existing estimators for certain natural populations. It was observed from the numerical comparisons that the mean squared errors of the proposed estimators are less than the mean squared error of the existing estimators. Further it was shown that the PREs of the proposed estimators, with respect to existing estimators, range from 101.06 to 6886.27. Hence, the proposed modified ratio estimators are strongly recommended and may be preferred over existing estimators for practical applications.

## Acknowledgements

The authors wish to extend their gratitude and thanks for the financial assistance received through UGC-Major Research Project, and to the Editor and the Referees for their effort to improve the presentation of the paper.

## References

- Abu-Dayyeh, W. A., Ahmed, M. S., Ahmed, R. A., & Muttalak, H. A. (2003). Some estimators of finite population mean using auxiliary information. *Applied Mathematics and Computation*, 139: 287-298.
- Bandyopadhyay, S. (1980). Improved ratio and product estimators. *Sankhyā C*, 42: 45-49.
- Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *Journal of Agriculture Science*, 37: 199-212.

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

- Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.
- Kadilar, C., & Cingi, H. (2004). Estimator of a population mean using two auxiliary variables in simple random sampling. *International Mathematical Journal*, 5: 357-360.
- Kadilar, C., & Cingi, H. (2005). A new estimator using two auxiliary variables. *Applied Mathematics and Computation*, 162, 901-908.
- Kadilar, C., & Cingi, H. (2009). *Advances in sampling theory-ratio method of estimation*. Oak Park, IL: Bentham Science Publishers.
- Khare, B. B., Srivastava, U., & Kumar, K. (2013). A generalized chain ratio in regression estimator for population mean using two auxiliary characters in sample survey. *Journal of Scientific Research*, 57: 147-153.
- Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.
- Naik, V. D., & Gupta, P. C. (1991). A general class of estimators for estimating population mean using auxiliary information. *Metrika*, 38: 11-17.
- Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45: 154-165.
- Perri, P. F. (2004). Alcune considerazioni sull'efficienza degli stimatori rapporto-cum-prodotto. *Statistica & Applicazioni*, 2(2): 59-75.
- Perri, P. F. (2007). Improved ratio-cum-product type estimators. *Statistics in Transition-NS*, 8(1): 51-69.
- Raj, D. (1965). On a method of using multi-auxiliary information in sample surveys. *Journal of the American Statistical Association*, 60: 154-165.
- Rao, P. S. R. S., & Mudholkar, G. S. (1967). Generalized multivariate estimator for the mean of finite populations. *Journal of the American Statistical Association*, 62: 1009-1012.
- Sahoo, L. N., & Swain, A. K. P. C. (1980). Unbiased ratio-cum-product estimator. *Sankhyā C*, 42: 56-62.
- Singh, D., & Chaudhary, F. S. (1986). *Theory and analysis of sample survey designs*. New Delhi, India: New Age International Publisher.
- Singh, G. N. (2003). On the improvement of product method of estimation in sample surveys. *Journal of the Indian Society of Agricultural Statistics*, 56(3): 267-265.
- Singh, H. P., & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population means. *Statistics in Transition*, 6(4): 555-560.

- Singh, H. P., & Tailor, R. (2005). Estimation of finite population mean using known correlation coefficient between auxiliary characters. *Statistica*, 65: 407-418.
- Singh, H. P., Tailor, R., & Kakran, M. S. (2004). Improved estimators of population mean using power transformation. *Journal of the Indian Society of Agricultural Statistics*, 58(2): 223-230.
- Singh, M. P. (1965). On the estimation of ratio and product of the population parameters. *Sankhyā B*, 27: 321-328.
- Singh, M. P. (1967a). Multivariate product method of estimation for finite populations. *Journal of the Indian Society of Agricultural Statistics*, 31, 375-378.
- Singh, M. P. (1967b). Ratio cum product method of estimation. *Metrika*, 12: 34-42.
- Sisodia, B. V. S., & Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, 33(1):13-18.
- Srivenkataramana, T. (1980). A dual to ratio estimator in sample surveys. *Biometrika*, 67: 199-204.
- Srivenkataramana, T., & Tracy, D. S. (1981). An alternative to ratio method in sample surveys. *Annals of the Institute of Statistical Mathematics*, 32: 111-120.
- Subramani, J. (2013). Generalized Modified Ratio Estimator for Estimation of Finite Population Mean. *Journal of Modern Applied Statistical Methods*, 12(2), 121-155.
- Subramani, J., & Kumarapandiyan, G. (2012a). Estimation of population mean using known median and co-efficient of skewness. *American Journal of Mathematics and Statistics*, 2(5): 101-107.
- Subramani, J., & Kumarapandiyan, G. (2012b). Estimation of population mean using co-efficient of variation and median of an auxiliary variable. *International Journal of Probability and Statistics*, 1(4): 111-118.
- Subramani, J., & Kumarapandiyan, G. (2012c). Modified ratio estimators using known median and co-efficient of kurtosis, *American Journal of Mathematics and Statistics*, 2(4): 95-100.
- Subramani, J., & Kumarapandiyan, G. (2013). A new modified ratio estimator for estimation of population mean when median of the auxiliary variable is known. *Pakistan Journal of Statistics and Operation Research*, 9(2), 137-145.

## TWO PARAMETER MODIFIED RATIO ESTIMATORS

Tailor, R., Parmar, R., Kim, J. M., & Tailor, R. (2011). Ratio-cum-Product estimators of population mean using known population parameters of auxiliary variable. *Communication of the Korean Statistical Society*, 18(2): 155-164.

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: 375-387.

Upadhyaya, L. N., & Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, 41(5): 627-636.

Yan, Z., & Tian, B. (2010). Ratio method to the mean estimation using coefficient of skewness of auxiliary variable. *ICICA 2010, Part II, CCIS 106*: 103-110.



# Separate Ratio-type Estimators of Population Mean in Stratified Random Sampling

**Rajesh Tailor**  
Vikram University  
Ujjain, M.P., India

**Hilal A. Lone**  
Vikram, University  
Ujjain, M.P., India

---

Separate ratio-type estimators for population mean with their properties are considered. Some separate ratio-type estimators for population mean using known parameters of auxiliary variate are proposed. The bias and mean squared error of the proposed estimators are obtained up to the first degree of approximation. It is shown that the proposed estimators are more efficient than unbiased estimators in stratified random sampling and usual separate ratio estimators under certain obtained conditions. To judge the merits of the proposed estimators, an empirical study was conducted.

*Keywords:* Finite population mean, separate ratio estimator, auxiliary variable, bias, mean squared error, stratified random sampling

---

## Introduction

The use of auxiliary information improves the efficiency of estimators. Cochran (1940) used auxiliary information at the estimation stage and envisaged the ratio estimation method. This method provides a ratio estimator which assumes that the population mean of the auxiliary variate is known. The ratio estimator performs well when a study and auxiliary variate are positively correlated. When these variates are negatively correlated, Robson's (1957) product method, which was independently given by Murthy (1964), is used. Searls (1964) utilized the coefficient of variation of an auxiliary variate to estimate the population mean of a study variate. Based on the work of Searls (1964), Sisodia and Dwivedi (1981) used a coefficient of variation of an auxiliary variate. Singh, et al. (2004) proposed ratio and product type estimators using the coefficient of kurtosis of an auxiliary variate, whereas Upadhyaya and Singh (1999) utilized both the

---

*Dr. Tailor is a Reader in the School of Studies in Statistics. Email him at [tailorraj@gmail.com](mailto:tailorraj@gmail.com). H. Lone is a Research Scholar in the School of Studies in Statistics. Email at [hilalstat@gmail.com](mailto:hilalstat@gmail.com).*

## SEPARATE RATIO TYPE ESTIMATORS OF POPULATION MEAN

coefficients of variation and kurtosis of an auxiliary variate. Kadilar and Cingi (2003), Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999) and Singh, et al. (2004) defined estimators in stratified random sampling. This article develops separate ratio-type estimators along the lines of Kadilar and Cingi (2003).

Consider a population  $U$  of size  $N$  consisting of units  $U_1, U_2, U_3, \dots, U_N$ . Let  $x$  and  $y$  be the auxiliary variate and study variate, respectively. If population  $U$  is divided into  $L$  homogenous strata of sizes  $n_h$  ( $h=1, 2, 3, \dots, L$ ), and a sample of size  $n_h$  is drawn from the  $h^{th}$  stratum, then the usual separate ratio estimator for population mean  $\bar{Y}$  is defined as

$$\hat{\bar{Y}}_{RS} = \sum_{h=1}^L W_h \bar{y}_h \left( \frac{\bar{X}_h}{\bar{x}_h} \right),$$

where  $\bar{x}_h$  is the sample mean of the auxiliary variate in stratum  $h$ , and  $\bar{y}_h$  is the sample mean of a study variate of interest in stratum  $h$

To the first degree of approximation, the bias and mean squared of the usual separate ratio estimator are

$$B(\hat{\bar{Y}}_{RS}) = \sum_{h=1}^L W_h \gamma_h \bar{Y}_h (C_{xh}^2 - \rho_{yxh} C_{xh} C_{yh}) \quad (1)$$

and

$$MSE(\hat{\bar{Y}}_{RS}) = \sum_{h=1}^L W_h^2 \gamma_h (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh}). \quad (2)$$

### Proposed separate ratio-type estimator using coefficient of variation

Sisodia and Dwivedi (1981) defined a ratio-type estimator using coefficient of variation  $C_x$  of auxiliary variate  $x$  as

$$\hat{\bar{Y}}_{SD} = \bar{y} \left( \frac{\bar{X} + C_x}{\bar{x} + C_x} \right) \quad (3)$$

Kadilar and Cingi (2003) further defined Sisodia and Dwivedi's (1981) estimator in stratified random sampling as

$$\hat{Y}_{SD}^{ST} = \bar{y}_{st} \left( \frac{\sum_{h=1}^L W_h (\bar{X}_h + C_{xh})}{\sum_{h=1}^L W_h (\bar{x}_h + C_{xh})} \right) \quad (4)$$

Motivated by Sisodia and Dwivedi (1981), Kadilar and Cingi (2003), suggested a separate ratio-type estimator using coefficient of variation  $C_{xh}$  of auxiliary variate in  $h^{th}$  stratum as

$$\hat{Y}_{RS}^{SD} = \sum_{h=1}^L W_h \bar{y}_h \left( \frac{\bar{X}_h + C_{xh}}{\bar{x}_h + C_{xh}} \right) \quad (5)$$

To obtain the bias and mean squared error of the proposed separate ratio-type estimator  $\hat{Y}_{RS}^{SD}$  :

$$\bar{y}_h = \bar{Y}_h (1 + e_{oh}) \text{ and } \bar{x}_h = \bar{X}_h (1 + e_{1h}) ,$$

such that  $E(e_{oh}) = E(e_{1h}) = 0$  and

$$E(e_{oh}^2) = \left( \frac{1}{n_h} - \frac{1}{N_h} \right) C_{yh}^2 = \gamma_h C_{yh}^2 ,$$

$$E(e_{1h}^2) = \left( \frac{1}{n_h} - \frac{1}{N_h} \right) C_{xh}^2 = \gamma_h C_{xh}^2 ,$$

$$E(e_{oh} e_{1h}) = \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \rho_{yxh} C_{yh} C_{xh} = \gamma_h \rho_{yxh} C_{yh} C_{xh} .$$

Expressing  $\hat{Y}_{RS}^{SD}$  in terms of  $e_i$ 's results in

$$\hat{Y}_{RS}^{SD} = \sum_{h=1}^L W_h \bar{Y}_h + \sum_{h=1}^L W_h \bar{Y}_h (e_{oh} - \lambda_{1h} e_{1h} + \lambda_{1h}^2 e_{1h}^2 - \lambda_{1h} e_{oh} e_{1h}) .$$

## SEPARATE RATIO TYPE ESTIMATORS OF POPULATION MEAN

Thus, the bias and mean squared error of the proposed ratio estimator  $\hat{Y}_{RS}^{SD}$  up to the first degree of approximation is obtained as

$$B\left(\hat{Y}_{RS}^{SD}\right) = \sum_{h=1}^L W_h \bar{Y}_h \gamma_h \left( \lambda_{1h}^2 C_{xh}^2 - \lambda_{1h} \rho_{yxh} C_{yh} C_{xh} \right) \quad (6)$$

$$MSE\left(\hat{Y}_{RS}^{SD}\right) = \sum_{h=1}^L W_h^2 \gamma_h \left( S_{yh}^2 + \lambda_{1h}^2 R_h^2 S_{xh}^2 - 2R_h \lambda_{1h} S_{yxh} \right) \quad (7)$$

where  $\lambda_{1h} = \frac{\bar{X}_h}{\bar{X}_h + C_{xh}}$ ,  $R_h = \frac{\bar{Y}_h}{\bar{X}_h}$  and  $\rho_{yxh} = \frac{S_{yxh}}{S_{xh} S_{yh}}$ .

### Suggested separate ratio-type estimator using coefficient of kurtosis

Singh, et al. (2004) defined a modified ratio estimator using the coefficient of kurtosis  $\beta_2(x)$  of an auxiliary variate  $x$  as

$$\hat{Y}_{SE} = \bar{y} \left( \frac{\bar{X} + \beta_2(x)}{\bar{x} + \beta_2(x)} \right) \quad (8)$$

Kadilar and Cingi (2003) defined Singh, et al's (2004) estimator in stratified random sampling as

$$\hat{Y}_{SE}^{ST} = \bar{y}_{st} \left( \frac{\sum_{h=1}^L W_h (\bar{X}_h + \beta_{2h}(x))}{\sum_{h=1}^L W_h (\bar{x}_h + \beta_{2h}(x))} \right) \quad (9)$$

Motivated by Kadilar and Cingi (2003) and Singh, et al. (2004), the proposed estimator using the coefficient of kurtosis  $\beta_2(x)$  of auxiliary variate  $x$  in  $h^{th}$  stratum is

$$\hat{Y}_{RS}^{SE} = \sum_{h=1}^L W_h \bar{y}_h \left\{ \frac{\bar{X}_h + \beta_{2h}(x)}{\bar{x}_h + \beta_{2h}(x)} \right\} \quad (10)$$

The bias and mean squared error of the proposed estimator  $\hat{Y}_{RS}^{SE}$  are obtained as

$$B\left(\hat{Y}_{RS}^{SE}\right)=\sum_{h=1}^L W_h \bar{Y}_h \gamma_h \left(\lambda_{2h}^2 C_{xh}^2 - \lambda_{2h} \rho_{yxh} C_{yh} C_{xh}\right) \quad (11)$$

$$MSE\left(\hat{Y}_{RS}^{SE}\right)=\sum_{h=1}^L W_h^2 \gamma_h \left(S_{yh}^2 + \lambda_{2h}^2 R_h^2 S_{xh}^2 - 2R_h \lambda_{2h} S_{yxh}\right) \quad (12)$$

where  $\lambda_{2h} = \frac{\bar{X}_h}{\bar{X}_h + \beta_{2h}(x)}$ .

### Proposed separate ratio-type estimator using coefficient of variation and coefficient of kurtosis

Upadhyaya and Singh (1999) suggested two different ratio-type estimators using the parameters coefficient of variation and coefficient of kurtosis as

$$\hat{Y}_{US1} = \bar{y} \left( \frac{\bar{X} \beta_2(x) + C_x}{\bar{x} \beta_2(x) + C_x} \right) \quad (13)$$

and

$$\hat{Y}_{US2} = \bar{y} \left( \frac{\bar{X} C_x + \beta_2(x)}{\bar{x} C_x + \beta_2(x)} \right) \quad (14)$$

Kadilar and Cingi (2003) defined Upadhyaya and Singh's (1999) estimators in stratified random sampling as

$$\hat{Y}_{US1}^{ST} = \bar{y}_{st} \left( \frac{\sum_{h=1}^L W_h \left( \bar{X}_h \beta_{2h}(x) + C_{xh} \right)}{\sum_{h=1}^L W_h \left( \bar{x}_h \beta_{2h}(x) + C_{xh} \right)} \right) \quad (15)$$

and

## SEPARATE RATIO TYPE ESTIMATORS OF POPULATION MEAN

$$\hat{Y}_{US2}^{ST} = \bar{y}_{st} \frac{\sum_{h=1}^L W_h (\bar{X}_h C_{xh} + \beta_{2h}(x))}{\sum_{h=1}^L W_h (\bar{x}_h C_{xh} + \beta_{2h}(x))} \quad (16)$$

Based on Upadhyaya and Singh (1999) and Kadilar and Cingi (2003), the proposed separate ratio-type estimator using coefficients of kurtosis and variation in  $h^{th}$  stratum are

$$\hat{Y}_{RS}^{US1} = \sum_{h=1}^L W_h \bar{y}_h \left\{ \frac{\bar{X}_h \beta_{2h}(x) + C_{xh}}{\bar{x}_h \beta_{2h}(x) + C_{xh}} \right\} \quad (17)$$

and

$$\hat{Y}_{RS}^{US2} = \sum_{h=1}^L W_h \bar{y}_h \left\{ \frac{\bar{X}_h C_{xh} + \beta_{2h}(x)}{\bar{x}_h C_{xh} + \beta_{2h}(x)} \right\} \quad (18)$$

Using the standard procedure for finding the bias and mean squared errors shown previously, the bias and mean squared error of the proposed separate ratio-type estimators up to the first degree of approximation are obtained as:

$$B(\hat{Y}_{RS}^{US1}) = \sum_{h=1}^L W_h \bar{Y}_h \gamma_h (\lambda_{3h}^2 C_{xh}^2 - \lambda_{3h} \rho_{yxh} C_{yh} C_{xh}) \quad (19)$$

$$MSE(\hat{Y}_{RS}^{US1}) = \sum_{h=1}^L W_h^2 \gamma_h (S_{yh}^2 + \lambda_{3h}^2 R_h^2 S_{xh}^2 - 2R_h \lambda_{3h} S_{yxh}) \quad (20)$$

$$B(\hat{Y}_{RS}^{US2}) = \sum_{h=1}^L W_h \bar{Y}_h \gamma_h (\lambda_{4h}^2 C_{xh}^2 - \lambda_{4h} \rho_{yxh} C_{yh} C_{xh}) \quad (21)$$

$$MSE(\hat{Y}_{RS}^{US2}) = \sum_{h=1}^L W_h^2 \gamma_h (S_{yh}^2 + \lambda_{4h}^2 R_h^2 S_{xh}^2 - 2R_h \lambda_{4h} S_{yxh}) \quad (22)$$

where  $\lambda_{3h} = \frac{\bar{X}_h \beta_{2h}(x)}{\bar{X}_h \beta_{2h}(x) + C_{xh}}$  and  $\lambda_{4h} = \frac{\bar{X}_h C_{xh}}{\bar{X}_h C_{xh} + \beta_{2h}(x)}$ .

## Efficiency comparisons

The variance of the usual unbiased estimators in stratified random sampling is

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \gamma_h S_{yh}^2 \quad (23)$$

From (7) and (23), it is observed that the proposed estimator  $\hat{\bar{Y}}_{RS}^{SD}$  would be more efficient than the usual unbiased estimator  $\bar{y}_{st}$  if

$$A < 2B. \quad (24)$$

A comparison of (7) and (2) shows that the proposed estimator  $\hat{\bar{Y}}_{RS}^{SD}$  would be more efficient than the usual ratio estimator  $\hat{\bar{Y}}_{RS}$  if

$$C < 2D. \quad (25)$$

Comparing (12) and (23), it is observed that the proposed estimator  $\hat{\bar{Y}}_{RS}^{SE}$  would be more efficient than the usual unbiased estimator  $\bar{y}_{st}$  if

$$E < 2F. \quad (26)$$

From (12) and (2) it is observed that the proposed estimator  $\hat{\bar{Y}}_{RS}^{SE}$  would be more efficient than the usual separate ratio estimator  $\hat{\bar{Y}}_{RS}$  if

$$G < 2H. \quad (27)$$

Comparison of (20) and (23) shows that the proposed estimator  $\hat{\bar{Y}}_{RS}^{US1}$  would be more efficient than the usual unbiased estimator  $\bar{y}_{st}$  if

$$L < 2M. \quad (28)$$

## SEPARATE RATIO TYPE ESTIMATORS OF POPULATION MEAN

From (20) and (2), it is observed that the proposed estimator  $\hat{\bar{Y}}_{RS}^{US1}$  would be more efficient than the usual separate ratio estimator  $\hat{\bar{Y}}_{RS}$  if

$$N < 2P. \quad (29)$$

Comparison of (22) and (23) shows that the proposed estimator  $\hat{\bar{Y}}_{RS}^{US2}$  would be more efficient than the usual unbiased estimator  $\bar{y}_{st}$  if

$$Q < 2R. \quad (30)$$

From (22) and (2), it is observed that the proposed estimator  $\hat{\bar{Y}}_{RS}^{US2}$  would be more efficient than the usual separate ratio estimator  $\hat{\bar{Y}}_{RS}$  if

$$S < 2T, \quad (31)$$

$$\begin{aligned} \text{where } A &= \sum_{h=1}^L W_h^2 \gamma_h \lambda_{1h}^2 S_{xh}^2 R_h^2, \quad B = \sum_{h=1}^L W_h^2 \gamma_h R_h \lambda_{1h} S_{yxh}, \\ C &= \sum_{h=1}^L W_h^2 \gamma_h S_{xh}^2 R_h^2 (\lambda_{1h}^2 - 1), \quad D = \sum_{h=1}^L W_h^2 \gamma_h R_h S_{yxh} (\lambda_{1h} - 1), \quad E = \sum_{h=1}^L W_h^2 \gamma_h \lambda_{2h}^2 S_{xh}^2 R_h^2, \\ F &= \sum_{h=1}^L W_h^2 \gamma_h R_h \lambda_{2h} S_{yxh}, \quad G = \sum_{h=1}^L W_h^2 \gamma_h S_{xh}^2 R_h^2 (\lambda_{2h}^2 - 1), \quad H = \sum_{h=1}^L W_h^2 \gamma_h R_h S_{yxh} (\lambda_{2h} - 1), \\ L &= \sum_{h=1}^L W_h^2 \gamma_h \lambda_{3h}^2 S_{xh}^2 R_h^2, \quad M = \sum_{h=1}^L W_h^2 \gamma_h R_h \lambda_{3h} S_{yxh}, \quad N = \sum_{h=1}^L W_h^2 \gamma_h S_{xh}^2 R_h^2 (\lambda_{3h}^2 - 1), \\ P &= \sum_{h=1}^L W_h^2 \gamma_h R_h S_{yxh} (\lambda_{3h} - 1), \quad Q = \sum_{h=1}^L W_h^2 \gamma_h \lambda_{4h}^2 S_{xh}^2 R_h^2, \quad R = \sum_{h=1}^L W_h^2 \gamma_h R_h \lambda_{4h} S_{yxh} \\ S &= \sum_{h=1}^L W_h^2 \gamma_h S_{xh}^2 R_h^2 (\lambda_{4h}^2 - 1) \text{ and } T = \sum_{h=1}^L W_h^2 \gamma_h R_h S_{yxh} (\lambda_{4h} - 1). \end{aligned}$$

## Empirical study

To examine the performance of the proposed estimator in comparison to other estimators considered in this study, three natural population data sets were



considered (see Populations 1-3). The estimators based on the population data are compared in Table 1.

**Population 1.** (Singh and Mangat, 1996, p. 208)

	$n_1=14$	$n_2=9$	$n_3=12$	$n_4=17$
	$N_1=400$	$N_2=216$	$N_3=364$	$N_4=364$
	$\bar{X}_1=76.21$	$\bar{X}_2=58.11$	$\bar{X}_3=69.08$	$\bar{X}_4=63.71$
	$\bar{Y}_1=79.35$	$\bar{Y}_2=59.44$	$\bar{Y}_3=76.66$	$\bar{Y}_4=64.57$
$N = 1344,$ $n = 52$	$\beta_{21}(x)=2.22$	$\beta_{22}(x)=2.29$	$\beta_{23}(x)=1.96$	$\beta_{24}(x)=2.47$
	$C_{x_1}=0.1906$	$C_{x_2}=0.2416$	$C_{x_3}=0.201$	$C_{x_4}=0.1908$
	$S_{x_1}^2=210.9938$	$S_{x_2}^2=197.1041$	$S_{x_3}^2=192.7954$	$S_{x_4}^2=147.7651$
	$S_{y_1}^2=166.70$	$S_{y_2}^2=174.28$	$S_{y_3}^2=226.60$	$S_{y_4}^2=170.61$
	$S_{yx_1}=148.76$	$S_{yx_2}=161.19$	$S_{yx_3}=192.21$	$S_{yx_4}=143.83$

**Population 2.** (Murthy, 1967, p. 228)

	$n_1=2$	$n_2=2$
	$N_1=5$	$N_2=5$
	$\bar{X}_1=214.4$	$\bar{X}_2=333.8$
	$\bar{Y}_1=1925.8$	$\bar{Y}_2=3115.6$
$N = 10,$ $n = 4$	$\beta_{21}(x)=1.88$	$\beta_{22}(x)=2.32$
	$\rho_{yx_1}=0.85$	$\rho_{yx_2}=0.98$
	$C_{x_1}=0.34$	$C_{x_2}=0.19$
	$S_{x_1}^2=5605.84$	$S_{x_2}^2=4401.76$
	$S_{y_1}^2=379360.16$	$S_{y_2}^2=115860.24$
	$S_{yx_1}=39360.69$	$S_{yx_2}=22356.52$

## SEPARATE RATIO TYPE ESTIMATORS OF POPULATION MEAN

**Population 3.** (Singh and Mangat, 1996, p. 219)

<b>N = 10, n = 4</b>	$n_1=2$	$n_2=2$
	$N_1=5$	$N_2=5$
	$\bar{X}_1=214.4$	$\bar{X}_2=333.8$
	$\bar{Y}_1=1925.8$	$\bar{Y}_2=3115.6$
	$\beta_{21}(x)=1.88$	$\beta_{22}(x)=2.32$
	$\rho_{yx_1}=0.85$	$\rho_{yx_2}=0.98$
	$C_{x_1}=0.34$	$C_{x_2}=0.19$
	$S_{x_1}^2=5605.84$	$S_{x_2}^2=4401.76$
	$S_{y_1}^2=379360.16$	$S_{y_2}^2=115860.24$
	$S_{yx_1}=39360.69$	$S_{yx_2}=22356.52$

**Table 1.** Percent Relative Efficiency of  $\bar{y}_{st}$ ,  $\hat{\bar{Y}}_{RS}$ ,  $\hat{\bar{Y}}_{RS}^{SD}$ ,  $\hat{\bar{Y}}_{RS}^{SE}$ ,  $\hat{\bar{Y}}_{RS}^{US1}$  and  $\hat{\bar{Y}}_{RS}^{US2}$  with respect to  $\bar{y}_{st}$

Estimators	Population I	Population II	Population III
$\bar{y}_{st}$	100.00	100.00	100.00
$\hat{\bar{Y}}_{RS}$	350.08	239.76	254.99
$\hat{\bar{Y}}_{RS}^{SD}$	351.53	240.35	255.22
$\hat{\bar{Y}}_{RS}^{SE}$	364.51	244.55	258.23
$\hat{\bar{Y}}_{RS}^{US1}$	350.76	240.05	255.10
$\hat{\bar{Y}}_{RS}^{US2}$	397.29	260.33	275.69

## Conclusion

The conditions under which the proposed estimators have less mean squared error in comparison to the usual unbiased estimator in stratified random sampling and usual separate ratio estimator were described. Table 1 shows that the proposed estimators have the highest percent relative efficiency compared to the usual

unbiased estimator and separate ratio estimator, in all three populations. Thus, the proposed estimators  $\hat{Y}_{RS}^{SD}$ ,  $\hat{Y}_{RS}^{SE}$ ,  $\hat{Y}_{RS}^{US1}$  and  $\hat{Y}_{RS}^{US2}$  are recommended for use in practice for estimating the population mean when the described proper conditions are satisfied.

## References

- Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio gain to produce. *Journal of Agricultural Science*, 30: 262-275.
- Kadilar, C., & Cingi, H. (2003). Ratio estimators in stratified sampling. *Biometrical Journal*, 45: 218-225.
- Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.
- Robson, D. S. (1957). Application of multivariate polykeys to the theory of unbiased ratio-type estimation. *Journal of the American Statistical Association*, 52: 511-522.
- Searls, D. T. (1964). The utilization of a known coefficient of variation in the estimating procedure. *Journal of the American Statistical Association*, 59: 1225-1226. doi: [10.1080/01621459.1964.10480765](https://doi.org/10.1080/01621459.1964.10480765)
- 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 of Agricultural Statistics*, 58(2): 223-230.
- Singh, R., & Mangat, N. S. (1996). *Elements of survey sampling*. Boston, MA: Kluwer Academic.
- Sisodia, B. V. S., & Dwivedi, V. K. (1981). A modified Ratio Estimator using coefficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, 33(1): 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.

# Median Based Modified Ratio Estimators with Known Quartiles of an Auxiliary Variable

**Jambulingam Subramani**

Pondicherry University  
Puducherry, India

**G. Prabavathy**

Pondicherry University  
Puducherry, India

---

New median based modified ratio estimators for estimating a finite population mean using quartiles and functions of an auxiliary variable are proposed. The bias and mean squared error of the proposed estimators are obtained and the mean squared error of the proposed estimators are compared with the usual simple random sampling without replacement (SRSWOR) sample mean, ratio estimator, a few existing modified ratio estimators, the linear regression estimator and median based ratio estimator for certain natural populations. A numerical study shows that the proposed estimators perform better than existing estimators; in addition, it is shown that the proposed median based modified ratio estimators outperform the ratio and modified ratio estimators as well as the linear regression estimator.

*Keywords:* Bias, inter-quartile range, linear regression estimator, mean squared error, natural population, simple random sampling

---

## Introduction

Consider a finite population  $U = \{U_1, U_2, \dots, U_N\}$  of  $N$  distinct and identifiable units. Let  $Y$  be a study variable with value  $Y_i$  measured on  $U_i, i = 1, 2, 3, \dots, N$  giving a vector  $Y = \{Y_1, Y_2, \dots, Y_N\}$ . The goal is to estimate the population mean,

$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ , with some desirable properties on the basis of a random sample of size  $n$  selected from the population  $U$ . The simplest estimator of population mean is the sample mean, obtained by using simple random sampling without

---

*Dr. Subramani is Associate Professor and Head, Department of Statistics. Email him at: drjsubramani@yahoo.co.in. G. Prabavathy is a Project Fellow in the UGC-MRP and a Doctoral Student in the Department of Statistics. Email her at: praba.gopal.23@gmail.com.*

replacement (SRSWOR), when there is no information on the auxiliary variable available. Let  $X$  be an auxiliary variable that is positively correlated with the study variable  $Y$ : Sometimes the information on auxiliary variable  $X$ , positively correlated with  $Y$ , may be utilized to obtain a more efficient estimator of the population mean (for further details on ratio estimators see Cochran, 1977 and Murthy, 1967.) When the population parameters of an auxiliary variable  $X$ , such as, population mean, coefficient of variation, coefficient of kurtosis, coefficient of skewness and median are known, ratio, product and linear regression estimators (and their modifications) have been proposed in the literature – many of which perform better than the SRSWOR sample mean for estimating the population mean of a study variable.

Subramani (2013a) proposed a median based ratio estimator by using the median of a study variable as auxiliary information, and it has been shown that this median based ratio estimator outperforms the usual SRSWOR sample mean, ratio estimator, modified ratio estimator and linear regression estimator. Based on Subramani's (2013a) median based ratio estimator, some new median based modified ratio estimators with known quartiles of the auxiliary variable are proposed.

The first quartile, also called lower quartile, is denoted by  $Q_1$ ; the third quartile, also called the upper quartile, is denoted by  $Q_3$ . The lower quartile is a point where 25% of the observations are less than  $Q_1$  and 75% are above  $Q_1$ . The upper quartile is a point where 75% observations are less than  $Q_3$  and 25% are above  $Q_3$ . Quartiles are unaffected by extreme values unlike the population mean, variance, correlation coefficient, etc.

The inter-quartile range used as a measure of spread in a data set. The inter-quartile range of a distribution is the difference between the upper and lower quartiles. The formula for computing the inter-quartile range is

$$Q_r = Q_3 - Q_1. \quad (1)$$

The semi-quartile range of a distribution is half the difference between the upper and lower quartiles, or half the inter-quartile range. The formula for computing the semi-quartile range is

$$Q_d = \frac{Q_3 - Q_1}{2} \quad (2)$$

## MEDIAN BASED MODIFIED RATIO ESTIMATORS

Another measure, the quartile average, noted by  $Q_a$ , was suggested by Subramani and Kumarapandiyan (2012a) and is defined as

$$Q_a = \frac{Q_3 + Q_1}{2} \quad (3)$$

The notations and formulae used in this article are:

$N$ :	Population size
$n$ :	Sample size
$Y$ :	Study variable
$M$ :	Median of the study variable
$X$ :	Auxiliary variable
$Q_i$ :	$i^{\text{th}}$ Quartile of auxiliary variable, $i=1,3$
$\rho$ :	Correlation coefficient between $X$ and $Y$
$\bar{X}, \bar{Y}$ :	Population means
$\bar{x}, \bar{y}$ :	Sample means
$\bar{M}$ :	Average of sample medians of $Y$
$m$ :	Sample median of $Y$
$\beta$ :	Regression coefficient of $Y$ on $X$
$B(\cdot)$ :	Bias of the estimator
$V(\cdot)$ :	Variance of the estimator
$MSE(\cdot)$ :	Mean squared error of the estimator
$PRE(e, p) = \frac{MSE(e)}{MSE(p)} * 100$	Percent relative efficiency of the proposed estimator $p$ with respect to the existing estimator $e$

The formulae for computing various measures including the variance and the covariance of the SRSWOR sample mean and sample median are:

$$V(\bar{y}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{y}_i - \bar{Y})^2 = \frac{1-f}{n} S_y^2, \quad V(\bar{x}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{x}_i - \bar{X})^2 = \frac{1-f}{n} S_x^2,$$

$$MSE(m) = V(m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - M)^2,$$

$$Cov(\bar{y}, \bar{x}) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (\bar{x}_i - \bar{X})(\bar{y}_i - \bar{Y}) = \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}),$$

$$Cov(\bar{y}, m) = \frac{1}{N_{C_n}} \sum_{i=1}^{N_{C_n}} (m_i - M)(\bar{y}_i - \bar{Y}),$$

$$C'_{xx} = \frac{V(\bar{x})}{\bar{X}^2}, \quad C'_{mm} = \frac{V(m)}{M^2}, \quad C'_{ym} = \frac{Cov(\bar{y}, m)}{M\bar{Y}}, \quad C'_{yx} = \frac{Cov(\bar{y}, \bar{x})}{\bar{X}\bar{Y}},$$

$$\text{where } f = \frac{n}{N}; \quad S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, \quad S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2.$$

In the case of SRSWOR, the sample mean,  $\bar{y}$ , is used to estimate the population mean,  $\bar{Y}$ . That is, the estimator of  $\bar{Y} = \hat{\bar{Y}}_r = \bar{y}$  with variance

$$V(\hat{\bar{Y}}_r) = \frac{1-f}{n} S_y^2. \quad (4)$$

The classical ratio estimator for estimating the population mean  $\bar{Y}$  of a study variable  $Y$  is defined as  $\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R}\bar{X}$ . The bias and mean squared error of  $\hat{\bar{Y}}_R$  are:

$$B(\hat{\bar{Y}}_R) = \bar{Y} \{C'_{xx} - C'_{yx}\} \quad (5)$$

and

$$MSE(\hat{\bar{Y}}_R) = V(\bar{y}) + R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x}). \quad (6)$$

## MEDIAN BASED MODIFIED RATIO ESTIMATORS

The other commonly used estimator using the auxiliary variable  $X$  is the linear regression estimator. The linear regression estimator and its variance with known regression coefficient are:

$$\hat{Y}_{lr} = \bar{y} + \beta(\bar{X} - \bar{x}) \quad (7)$$

$$V(\hat{Y}_{lr}) = V(\bar{y})(1 - \rho^2) \text{ where } \rho = \frac{\text{Cov}(\bar{y}, \bar{x})}{\sqrt{V(\bar{x}) * V(\bar{y})}} \quad (8)$$

Subramani & Kumarapandiyan (2012a) suggested some modified ratio estimators using known quartiles and their functions of an auxiliary variable, these are:

$$\hat{Y}_{RM1} = \bar{y} \left( \frac{\bar{X} + Q_1}{\bar{x} + Q_1} \right) \quad (9)$$

$$\hat{Y}_{RM2} = \bar{y} \left( \frac{\bar{X} + Q_3}{\bar{x} + Q_3} \right) \quad (10)$$

$$\hat{Y}_{RM3} = \bar{y} \left( \frac{\bar{X} + Q_r}{\bar{x} + Q_r} \right) \quad (11)$$

$$\hat{Y}_{RM4} = \bar{y} \left( \frac{\bar{X} + Q_d}{\bar{x} + Q_d} \right) \quad (12)$$

$$\hat{Y}_{RM5} = \bar{y} \left( \frac{\bar{X} + Q_a}{\bar{x} + Q_a} \right) \quad (13)$$

The bias and the mean squared error of the modified ratio estimators in (9) to (13) are:

$$B(\hat{Y}_{RMi}) = \bar{Y} \{ \theta_i^2 C'_{xx} - \theta_i C'_{yx} \} \quad (14)$$



$$MSE\left(\hat{\bar{Y}}_{RMi}\right)=V(\bar{y})+R^2\theta_i^2V(\bar{x})-2R\theta_iCov(\bar{y},\bar{x}) \quad (15)$$

where  $R = \frac{\bar{Y}}{\bar{X}}$  and

$$\theta_1 = \frac{\bar{X}}{\bar{X}+Q_1}, \theta_2 = \frac{\bar{X}}{\bar{X}+Q_3}, \theta_3 = \frac{\bar{X}}{\bar{X}+Q_r}, \theta_4 = \frac{\bar{X}}{\bar{X}+Q_d}, \theta_5 = \frac{\bar{X}}{\bar{X}+Q_a}; i=1,2,3,4,5$$

Recently Subramani (2013a) suggested a median based ratio estimator for estimating  $\bar{Y}$  when the median of the study variable  $Y$  is known. The estimator with its bias and mean squared error are:

$$\hat{\bar{Y}}_M = \frac{\bar{y}}{m} M \quad (16)$$

$$B\left(\hat{\bar{Y}}_M\right)=\bar{Y}\left\{C'_{mm}-C'_{ym}-\frac{Bias(m)}{M}\right\} \quad (17)$$

$$MSE\left(\hat{\bar{Y}}_M\right)=V(\bar{y})+R'^2V(m)-2R'Cov(\bar{y},m) \text{ where } R'=\frac{\bar{Y}}{M}. \quad (18)$$

For further details on modified ratio estimators with known population parameters of an auxiliary variable, such as coefficient of variation, skewness, kurtosis, correlation coefficient, quartiles and their linear combinations, readers are referred to Kadilar and Cingi (2004, 2006a, b, 2009) Koyuncu and Kadilar (2009), Singh and Kakran (1993), Singh and Tailor (2003, 2005), Singh (2003), Sisodia and Dwivedi (1981), Subramani (2013a, b), Subramani and Kumarapandiyam (2012a, b, c, 2013), Tailor and Sharma (2009), Tin (1965), and Yan and Tian (2010).

The median based ratio estimator proposed by Subramani (2013a) is extended and, as a result, some new median based modified ratio estimators  $\hat{\bar{Y}}_{SP1}$ ,  $\hat{\bar{Y}}_{SP2}$ ,  $\hat{\bar{Y}}_{SP3}$ ,  $\hat{\bar{Y}}_{SP4}$  and  $\hat{\bar{Y}}_{SP5}$  with known quartiles and their functions of auxiliary variables are proposed.

### Proposed Median Based Modified Ratio Estimators

The proposed median based modified ratio estimators for estimating a population mean  $\bar{Y}$  based on Subramani's (2013a) ratio estimator are:

$$\hat{\bar{Y}}_{SP1} = \bar{y} \left( \frac{M + Q_1}{m + Q_1} \right) \quad (19)$$

$$\hat{\bar{Y}}_{SP2} = \bar{y} \left( \frac{M + Q_3}{m + Q_3} \right) \quad (20)$$

$$\hat{\bar{Y}}_{SP3} = \bar{y} \left( \frac{M + Q_r}{m + Q_r} \right) \quad (21)$$

$$\hat{\bar{Y}}_{SP4} = \bar{y} \left( \frac{M + Q_d}{m + Q_d} \right) \quad (22)$$

and

$$\hat{\bar{Y}}_{SP5} = \bar{y} \left( \frac{M + Q_a}{m + Q_a} \right). \quad (23)$$

To the first degree of approximation, the bias and mean squared error of  $\hat{\bar{Y}}_{SPj}$  are derived as:

$$B\left(\hat{\bar{Y}}_{SPj}\right) = \bar{Y} \left\{ \theta_j'^2 C_{mm}' - \theta_j' C_{ym}' - \theta_j' \frac{Bias(m)}{M} \right\}, j = 1, 2, 3, 4, 5, \quad (24)$$

$$MSE\left(\hat{\bar{Y}}_{SPj}\right) = V(\bar{y}) + R'^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m), j = 1, 2, 3, 4, 5 \quad (25)$$

where

$$R' = \frac{\bar{Y}}{M}, \theta_1' = \frac{M}{M + Q_1}, \theta_2' = \frac{M}{M + Q_3}, \theta_3' = \frac{M}{M + Q_r}, \theta_4' = \frac{M}{M + Q_d}, \theta_5' = \frac{M}{M + Q_a}.$$

See [Appendix A](#) for detailed derivation of the bias and the mean squared error of  $\hat{Y}_{SPj}$ .

## Efficiency Comparisons

### Comparison with SRSWOR Sample Mean

The conditions (see [Appendix B](#)) for which the proposed estimators  $\hat{Y}_{SPj}$ ,  $j = 1, 2, 3, 4, 5$  are more efficient than the SRSWOR sample mean  $\hat{Y}_r$  were derived from expressions (25) and (4) and are:

$$MSE\left(\hat{Y}_{SPj}\right) \leq V\left(\hat{Y}_r\right) \text{ if } 2C'_{ym} \geq \theta'_j C'_{mm}; j = 1, 2, 3, 4, 5. \quad (26)$$

### Comparison with Ratio Estimators

The conditions (see [Appendix B](#)) for which the proposed estimators  $\hat{Y}_{SPj}$ ,  $j = 1, 2, 3, 4, 5$  are more efficient than the usual ratio estimator  $\hat{Y}_R$  were derived from expressions (25) and (6) and are:

$$MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_R\right) \text{ if } \theta_j'^2 C'_{mm} - \theta_i'^2 C'_{xx} \leq 2\left(\theta'_j C'_{ym} - \theta'_i C'_{yx}\right); i, j = 1, 2, 3, 4, 5. \quad (27)$$

### Comparison with Modified Ratio Estimators

From expressions (25) and (15), the conditions (see [Appendix B](#)) for which the proposed estimators  $\hat{Y}_{SPj}$ ,  $j = 1, 2, 3, 4, 5$  are more efficient than the existing modified ratio estimator  $\hat{Y}_{RMi}$ ,  $i = 1, 2, 3, 4, 5$  were derived and are:

$$MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_{RMi}\right)$$

if

$$MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_{RMi}\right) \text{ if } \theta_j'^2 C'_{mm} - \theta_i'^2 C'_{xx} \leq 2\left(\theta'_j C'_{ym} - \theta'_i C'_{yx}\right); i, j = 1, 2, 3, 4, 5 \quad (28)$$

### Comparison with Linear Regression Estimator

From expressions (25) and (8), the conditions (see [Appendix B](#)) for which the proposed estimators  $\hat{Y}_{SPj}, j=1,2,3,4,5$  are more efficient than the usual linear regression estimator  $\hat{Y}_{lr}$  were derived and are:

$$MSE\left(\hat{Y}_{SPj}\right) \leq V\left(\hat{Y}_{lr}\right) \text{ if } 2\theta'_j C'_{ym} - \theta'^2_j C'_{mm} \geq \frac{\left[C'_{yx}\right]^2}{C'_{xx}}; j=1, 2. \quad (29)$$

### Comparison with Median Based Ratio Estimator

From expressions (25) and (18), the conditions (see [Appendix B](#)) for which the proposed estimators  $\hat{Y}_{SPj}, j=1,2,3,4,5$  are more efficient than the existing modified ratio type estimator  $\hat{Y}_M$  were derived and are:

$$MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_M\right) \text{ if } 2C'_{ym} \leq (\theta'_j + 1)C'_{mm}; j=1,2,3,4,5. \quad (30)$$

### Numerical Comparison

The conditions for which the proposed median based modified ratio estimators performed better than the other usual estimators considered in this study have been obtained. In order to show that the proposed estimators perform better than the other estimators, numerical comparisons were made to determine the efficiencies of the proposed estimators. Two populations were used to assess the efficiencies of the proposed median based modified ratio estimators with that of the existing estimators. Populations 1 and 2 are from Singh and Chaudhary (1986, p. 177). The parameter values and constants computed for the populations are given in [Table 1](#), the bias for the proposed and existing estimators computed for the two populations are given in [Table 2](#) and the mean squared errors are given in [Table 3](#).

**Table 1.** Parameter values and constants for 2 different populations

Parameter	$n = 3$		$n = 5$	
	Pop 1	Pop 2	Pop 1	Pop 2
$N$	34.0000	34.0000	34.0000	34.0000
$n$	3.0000	3.0000	5.0000	5.0000
$N_{c_n}$	5,984.0000	5,984.0000	278,256.0000	278,256.0000
$\bar{Y}$	856.4118	856.4118	856.4118	856.4118
$\bar{M}$	747.7223	747.7223	736.9811	736.9811
$M$	767.5000	767.5000	767.5000	767.5000
$\bar{X}$	208.8824	199.4412	208.8824	199.4412
$Q_1$	94.2500	99.2500	94.2500	99.2500
$Q_3$	254.7500	278.0000	254.7500	278.0000
$Q_r$	160.5000	178.7500	160.5000	178.7500
$Q_d$	80.2500	89.3750	80.2500	89.3750
$Q_a$	174.5000	188.6250	174.5000	188.6250
$R$	4.0999	4.2941	4.0999	4.2941
$R'$	1.1158	1.1158	1.1158	1.1158
$\theta_1$	0.6891	0.6677	0.6891	0.6677
$\theta_2$	0.4505	0.4177	0.4505	0.4177
$\theta_3$	0.5655	0.5274	0.5655	0.5274
$\theta_4$	0.7224	0.6905	0.7224	0.6905
$\theta_5$	0.5448	0.5139	0.5448	0.5139
$\theta_1^{'}$	0.8906	0.8855	0.8906	0.8855
$\theta_2^{'}$	0.7508	0.7341	0.7508	0.7341
$\theta_3^{'}$	0.8270	0.8111	0.8270	0.8111
$\theta_4^{'}$	0.9053	0.8957	0.9053	0.8957
$\theta_5^{'}$	0.8148	0.8027	0.8148	0.8027
$\text{var}(\bar{y})$	163,356.4086	163,356.4086	91,690.3713	91,690.3713
$\text{var}(\bar{x})$	6,884.4455	6,857.8555	3,864.1726	3,849.2480
$\text{var}(m)$	101,518.7738	101,518.7738	59,396.2836	59,396.2836
$\text{cov}(\bar{y}, m)$	90,236.2939	90,236.2939	48,074.9542	48,074.9542
$\text{cov}(\bar{y}, \bar{x})$	15,061.4011	14,905.0488	8,453.8187	8,366.0597
$\rho$	0.4491	0.4453	0.4491	0.4453

# MEDIAN BASED MODIFIED RATIO ESTIMATORS

**Table 2.** Bias of existing and proposed estimators

Estimators		<i>n</i> = 3		<i>n</i> = 5	
		Pop 1	Pop 2	Pop 1	Pop 2
Existing	$\hat{\bar{Y}}_R$	63.0241	72.9186	35.3748	40.9285
	$\hat{\bar{Y}}_{RM1}$	14.4774	15.9291	8.1261	8.9409
	$\hat{\bar{Y}}_{RM2}$	-5.0570	-5.4535	-2.8385	-3.0610
	$\hat{\bar{Y}}_{RM3}$	2.4369	1.6513	1.3678	0.9269
	$\hat{\bar{Y}}_{RM4}$	18.4357	18.8016	10.3478	10.5531
	$\hat{\bar{Y}}_{RM5}$	0.8276	0.5910	0.4645	0.3317
	$\hat{\bar{Y}}_M$	52.0924	52.0924	57.7705	57.7705
Proposed	$\hat{\bar{Y}}_{SP1}$	32.0179	31.1618	43.0405	42.3993
	$\hat{\bar{Y}}_{SP2}$	11.4953	9.4306	27.2167	25.5531
	$\hat{\bar{Y}}_{SP3}$	21.9708	19.6375	35.4268	33.6263
	$\hat{\bar{Y}}_{SP4}$	34.5121	32.8700	44.9012	43.6773
	$\hat{\bar{Y}}_{SP5}$	20.1662	18.4422	34.0355	32.6983

**Table 3.** Variance/mean squared error of existing and proposed estimators

Estimators		<i>n</i> = 3		<i>n</i> = 5	
		Pop 1	Pop 2	Pop 1	Pop 2
Existing	$\hat{\bar{Y}}$	163356.4086	163356.4086	91690.3713	91690.3713
	$\hat{\bar{Y}}_R$	155577.8155	161802.8878	87324.3215	90818.3961
	$\hat{\bar{Y}}_{RM1}$	133203.7861	134261.9210	74765.9957	75359.9173
	$\hat{\bar{Y}}_{RM2}$	131205.2291	131950.5079	73644.2252	74062.5432
	$\hat{\bar{Y}}_{RM3}$	130523.6191	131018.5135	73261.6440	73539.4239
	$\hat{\bar{Y}}_{RM4}$	134530.4901	135259.2456	75510.6618	75919.7060
	$\hat{\bar{Y}}_{RM5}$	130420.4186	130968.9816	73203.7186	73511.6221
	$\hat{\bar{Y}}_M$	130408.9222	130964.1249	73197.2660	73508.8959
	$\hat{\bar{Y}}_M$	88379.0666	88379.0666	58356.9234	58356.9234
Proposed	$\hat{\bar{Y}}_{SP1}$	84266.7092	84147.8927	54798.7634	54675.1252
	$\hat{\bar{Y}}_{SP2}$	83413.6960	83642.1970	52826.6580	52784.4688
	$\hat{\bar{Y}}_{SP3}$	83266.0122	83175.3231	53543.5010	53322.4123
	$\hat{\bar{Y}}_{SP4}$	84643.7479	84390.4264	55174.2962	54924.5239
	$\hat{\bar{Y}}_{SP5}$	83190.4430	83153.4557	53369.8070	53221.3731

The percentage relative efficiencies of the proposed estimators with respect to the existing estimators were also obtained and are shown in Tables 4-5.

**Table 4.** Percentage Relative Efficiency of  $\hat{Y}_{SPj}$  for Population 1

Existing Estimators	For sample size $n=3$					For sample size $n=5$				
	Proposed Estimators					Proposed Estimators				
	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$
$\hat{Y}_r$	193.86	195.84	196.19	192.99	196.36	167.32	173.57	171.24	166.18	171.80
$\hat{Y}_R$	184.63	186.51	186.84	183.80	187.01	159.35	165.30	163.09	158.27	163.62
$\hat{Y}_{RM1}$	158.07	159.69	159.97	157.37	160.12	136.44	141.53	139.64	135.51	140.09
$\hat{Y}_{RM2}$	155.70	157.29	157.57	155.01	157.72	134.39	139.41	137.54	133.48	137.99
$\hat{Y}_{RM3}$	154.89	156.48	156.75	154.20	156.90	133.69	138.68	136.83	132.78	137.27
$\hat{Y}_{RM4}$	159.65	161.28	161.57	158.94	161.71	137.80	142.94	141.03	136.86	141.49
$\hat{Y}_{RM5}$	154.77	156.35	156.63	154.08	156.77	133.59	138.57	136.72	132.68	137.16
$\hat{Y}_{lr}$	154.76	156.34	156.62	154.07	156.76	133.57	138.56	136.71	132.67	137.15
$\hat{Y}_M$	104.88	105.95	106.14	104.41	106.24	106.49	110.47	108.99	105.77	109.34

**Table 5:** Percentage Relative Efficiency of  $\hat{Y}_{SPj}$  for Population 2

Existing Estimators	For sample size $n=3$					For sample size $n=5$				
	Proposed Estimators					Proposed Estimators				
	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$	$\hat{Y}_{SP1}$	$\hat{Y}_{SP2}$	$\hat{Y}_{SP3}$	$\hat{Y}_{SP4}$	$\hat{Y}_{SP5}$
$\hat{Y}_r$	194.13	195.30	196.40	193.57	196.45	167.70	173.71	171.95	166.94	172.28
$\hat{Y}_R$	192.28	193.45	194.53	191.73	194.58	166.11	172.06	170.32	165.35	170.64
$\hat{Y}_{RM1}$	159.55	160.52	161.42	159.10	161.46	137.83	142.77	141.33	137.21	141.60
$\hat{Y}_{RM2}$	156.81	157.76	158.64	156.36	158.68	135.46	140.31	138.90	134.84	139.16
$\hat{Y}_{RM3}$	155.70	156.64	157.52	155.25	157.56	134.50	139.32	137.91	133.89	138.18
$\hat{Y}_{RM4}$	160.74	161.71	162.62	160.28	162.66	138.86	143.83	142.38	138.23	142.65
$\hat{Y}_{RM5}$	155.64	156.58	157.46	155.19	157.50	134.45	139.27	137.86	133.84	138.12
$\hat{Y}_{lr}$	155.64	156.58	157.46	155.19	157.50	134.45	139.26	137.86	133.84	138.12
$\hat{Y}_M$	105.03	105.66	106.26	104.73	106.28	106.73	110.56	109.44	106.25	109.65

Tables 4 and 5 show that the percent relative efficiencies of the proposed estimators, with respect to existing estimators, range in general from 104.41 to 196.45. In particular, the PRE ranges from 166.18 to 196.45 for comparing with the SRSWOR sample mean; ranging from 158.27 to 194.58 for comparing with the ratio estimator; ranging from 132.68 to 162.66 for comparing with the modified ratio estimators; ranging from 132.67 to 157.50 for comparing with the linear regression estimator and ranging from 104.41 to 110.56 for comparing with the median based ratio estimator. This demonstrates that the proposed estimators perform better than the existing SRSWOR sample mean, ratio, modified ratio and linear regression estimators for the two populations considered. Further it is observed from the numerical comparisons that the following inequalities hold:

$$MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_M\right) \leq V\left(\hat{Y}_{lr}\right) \leq MSE\left(\hat{Y}_{RMi}\right) \leq MSE\left(\hat{Y}_R\right) \leq V\left(\hat{Y}_r\right)$$

## Conclusion

This article proposed some new median based modified ratio estimators using known quartiles and their functions of the auxiliary variable. The conditions for which the proposed estimators are more efficient than the existing estimators were derived. Further the percentage relative efficiencies of the proposed estimators with respect to existing estimators were shown to range in general from 104.41 to 196.45 for certain natural populations available in the literature. It is usually believed that the linear regression estimator is the optimum estimator for estimating the population mean whenever an auxiliary variable exists that is positively correlated with that of a study variable. However, it was shown that the proposed median based modified ratio estimators outperform not only the ratio and modified ratio estimators but also the linear regression estimator. Based on results of this study, the proposed median based modified ratio estimators are recommended for estimating finite population means.

## Acknowledgements

The authors wish to express their gratitude and sincere thanks to the University Grants Commission, New Delhi for having given the financial assistance to carry out this research through the UGC-Major Research Project, and to the Editor and the referees for their effort to improve the presentation of this paper.



## References

- Cochran, W. G. (1977). *Sampling techniques, Third Edition*. Wiley Eastern Limited.
- Kadilar, C., & Cingi, H. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, 151: 893-902.
- Kadilar, C., & Cingi, H. (2006a). An improvement in estimating the population mean by using the correlation coefficient. *Hacettepe Journal of Mathematics and Statistics*, 35(1): 103-109.
- Kadilar, C., & Cingi, H. (2006b). Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19: 75-79.
- Kadilar, C., & Cingi, H. (2009). *Advances in Sampling Theory - Ratio Method of Estimation*. Bentham Science Publishers.
- Koyuncu, N., & Kadilar, C. (2009). Efficient Estimators for the Population Mean. *Hacettepe Journal of Mathematics and Statistics*, 38(2): 217-225
- Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta, India: Statistical Publishing Society.
- Singh, D., & Chaudhary, F. S. (1986). *Theory and analysis of sample survey designs*. New Age International Publishers.
- Singh, H. P., & Kakran, M. S. (1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. (Unpublished Manuscript).
- Singh, H. P., & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population means. *Statistics in Transition*, 6(4): 555-560 .
- Singh, H. P., & Tailor, R. (2005). Estimation of finite population mean with known co-efficient of variation of an auxiliary variable. *Statistica*, anno LXV(3): 301-313.
- Singh, G. N. (2003). On the improvement of product method of estimation in sample surveys. *Journal of Indian Society of Agricultural Statistics*, 56(3): 267-265.
- Sisodia, B. V. S., & Dwivedi, V. K. (1981). A modified ratio estimator using co-efficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, 33(2): 13-18.
- Subramani, J. (2013a). A new median based ratio estimator for estimation of the finite population mean. (Submitted for publication).

## MEDIAN BASED MODIFIED RATIO ESTIMATORS

Subramani, J. (2013b). Generalized modified ratio estimator for estimation of finite population mean. *Journal of Modern Applied Statistical Methods*, 12(2): 121-155.

Subramani, J., & Kumarapandiyan, G. (2012a). Modified ratio estimators for population mean using function of quartiles of auxiliary variable. *Bonfring International Journal of Industrial Engineering and Management Science*, 2(2): 19-23.

Subramani, J., & Kumarapandiyan, G. (2012b). Modified ratio estimators using known median and co-efficient of kurtosis. *American Journal of Mathematics and Statistics*, 2(4): 95-100.

Subramani, J., & Kumarapandiyan, G. (2012c). Estimation of population mean using known median and co-efficient of skewness. *American Journal of Mathematics and Statistics*, 2(5): 101-107.

Subramani, J., & Kumarapandiyan, G. (2013). A new modified ratio estimator of population mean when median of the auxiliary variable is known. *Pakistan Journal of Statistics and Operation Research*, 9(2): 137-145.

Tailor, R., & Sharma, B. (2009). A modified ratio-cum-product estimator of finite population mean using known coefficient of variation and coefficient of kurtosis. *Statistics in Transition - New Series*, 10(1): 15-24.

Tin, M. (1965). Comparison of some ratio estimators. *Journal of the American Statistical Association*, 60: 294-307.

Yan, Z., & Tian, B. (2010). Ratio method to the mean estimation using co-efficient of skewness of auxiliary variable. *ICICA 2010*, Part II, CCIS 106: 103-11.

## Appendix A

The derivation of the bias and the mean squared error of  $\bar{Y}_{SP1}$  are given below:  
Consider

$$\hat{\bar{Y}}_{SP1} = \bar{y} \left( \frac{M + Q_1}{m + Q_1} \right) \quad (A1)$$

$$\text{Let } e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \text{ and } e_1 = \frac{m - M}{M}$$

$$\Rightarrow E(e_0) = 0; E(e_1) = \frac{\bar{M} - M}{M} = \frac{\text{Bias}(m)}{M} \quad (A2)$$

$$\Rightarrow E(e_0^2) = \frac{V(\bar{y})}{\bar{Y}^2}; E(e_1^2) = \frac{V(m)}{M^2}; E(e_0 e_1) = \frac{\text{Cov}(\bar{y}, m)}{\bar{Y}M} \quad (A3)$$

The estimator  $\bar{Y}_{SP1}$  can be written in terms of  $e_0$  and  $e_1$  as

$$\hat{\bar{Y}}_{SP1} = \bar{Y} (1 + e_0) \left( \frac{M + Q_1}{M(1 + e_1) + Q_1} \right)$$

$$\Rightarrow \hat{\bar{Y}}_{SP1} = \bar{Y} (1 + e_0) \left( \frac{M + Q_1}{(M + Q_1) + M e_1} \right)$$

$$\Rightarrow \hat{\bar{Y}}_{SP1} = \bar{Y} (1 + e_0) \left( \frac{1}{1 + \left( \frac{M}{M + Q_1} \right) e_1} \right)$$

$$\Rightarrow \hat{\bar{Y}}_{SP1} = \bar{Y} (1 + e_0) \left( \frac{1}{1 + \theta'_1 e_1} \right); \text{ where } \theta'_1 = \frac{M}{M + Q_1}$$

$$\Rightarrow \hat{Y}_{SP1} = \bar{Y} (1 + e_0) (1 + \theta_1' e_1)^{-1}$$

Neglecting the terms of higher order, we have

$$\begin{aligned} \hat{Y}_{SP1} &= \bar{Y} (1 + e_0) (1 - \theta_1' e_1 + \theta_1'^2 e_1^2) \\ \Rightarrow \hat{Y}_{SP1} &= \bar{Y} + \bar{Y} e_0 - \bar{Y} \theta_1' e_1 - \bar{Y} \theta_1' e_0 e_1 + \bar{Y} \theta_1'^2 e_1^2 \\ \Rightarrow \hat{Y}_{SP1} - \bar{Y} &= \bar{Y} e_0 - \bar{Y} \theta_1' e_1 - \bar{Y} \theta_1' e_0 e_1 + \bar{Y} \theta_1'^2 e_1^2 \end{aligned} \quad (A4)$$

Taking expectations on both sides of (A4) we have,

$$\begin{aligned} E(\hat{Y}_{SP1} - \bar{Y}) &= \bar{Y} E(e_0) - \bar{Y} \theta_1' E(e_1) - \bar{Y} \theta_1' E(e_0 e_1) + \bar{Y} \theta_1'^2 E(e_1^2) \\ \Rightarrow E(\hat{Y}_{SP1} - \bar{Y}) &= \bar{Y} \left\{ \theta_1'^2 C'_{mm} - \theta_1' C'_{ym} - \theta_1' \frac{Bias(m)}{M} \right\} \text{ from (A2) and (A3)} \\ \Rightarrow Bias(\hat{Y}_{SP1}) &= \bar{Y} \left\{ \theta_1'^2 C'_{mm} - \theta_1' C'_{ym} - \theta_1' \frac{Bias(m)}{M} \right\} \end{aligned} \quad (A5)$$

The derivation of mean squared error of  $\bar{Y}_{SP1}$  is given below:

$$\begin{aligned} MSE(\hat{Y}_{SP1}) &= E(\hat{Y}_{SP1} - \bar{Y})^2 = E(\bar{Y} e_0 - \bar{Y} \theta_1' e_1)^2 \\ \Rightarrow MSE(\hat{Y}_{SP1}) &= \bar{Y}^2 \{ E(e_0^2) + \theta_1'^2 E(e_1^2) - 2\theta_1' E(e_0 e_1) \} \\ \Rightarrow MSE(\hat{Y}_{SP1}) &= \bar{Y}^2 \left\{ \frac{V(\bar{y})}{\bar{Y}^2} + \theta_1'^2 \frac{V(m)}{M^2} - 2\theta_1' \frac{Cov(\bar{y}, m)}{\bar{Y}M} \right\} \\ \Rightarrow MSE(\hat{Y}_{SP1}) &= V(\bar{y}) + \frac{\bar{Y}^2}{M^2} \theta_1'^2 V(m) - 2 \frac{\bar{Y}}{M} \theta_1' Cov(\bar{y}, m) \end{aligned}$$

$$\Rightarrow MSE\left(\hat{Y}_{SP1}\right)=V\left(\bar{y}\right)+R^2\theta_1^2V\left(m\right)-2R'\theta_1'Cov\left(\bar{y},m\right); R'=\frac{\bar{Y}}{M} \quad (A6)$$

In the Similar manner, the bias and mean squared error of  $\hat{Y}_{SP2}, \hat{Y}_{SP3}, \hat{Y}_{SP4}$  and  $\hat{Y}_{SP5}$  can be obtained.

## Appendix B

The conditions for which the proposed estimators perform better than the existing estimators are derived here and are given below:

### Comparison with that of SRSWOR sample mean

$$\text{Consider } MSE\left(\hat{Y}_{SPj}\right) \leq V\left(\hat{Y}_r\right)$$

$$\Rightarrow V\left(\bar{y}\right)+R^2\theta_j^2V\left(m\right)-2R'\theta_j'Cov\left(\bar{y},m\right) \leq V\left(\bar{y}\right)$$

$$\Rightarrow R^2\theta_j^2V\left(m\right)-2R'\theta_j'Cov\left(\bar{y},m\right) \leq 0$$

$$\Rightarrow R^2\theta_j^2V\left(m\right) \leq 2R'\theta_j'Cov\left(\bar{y},m\right)$$

$$\Rightarrow Cov\left(\bar{y},m\right) \geq \frac{R'\theta_j^2V\left(m\right)}{2}, j=1,2,3,4,5$$

$$\Rightarrow Cov\left(\bar{y},m\right) \geq \frac{\bar{Y}M\theta_j^2C_{mm}}{2}$$

$$\Rightarrow 2C'_{ym} \geq \theta_j^2C'_{mm}, j=1,2,3,4,5$$

### Comparison with that of Ratio Estimator

Consider  $MSE\left(\hat{\bar{Y}}_{SPj}\right) \leq MSE\left(\hat{\bar{Y}}_R\right)$

$$\Rightarrow V(\bar{y}) + R^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq V(\bar{y}) + R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x})$$

$$\Rightarrow R^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq R^2 V(\bar{x}) - 2RCov(\bar{y}, \bar{x})$$

$$\Rightarrow R^2 \theta_j'^2 V(m) - R^2 V(\bar{x}) \leq 2R' \theta_j' Cov(\bar{y}, m) - 2RCov(\bar{y}, \bar{x})$$

$$\Rightarrow \frac{\bar{Y}^2}{M^2} \theta_j'^2 V(m) - \frac{\bar{Y}^2}{\bar{X}^2} V(\bar{x}) \leq 2 \frac{\bar{Y}}{M} \theta_j' Cov(\bar{y}, m) - 2 \frac{\bar{Y}}{\bar{X}} Cov(\bar{y}, \bar{x})$$

$$\Rightarrow \theta_j'^2 \frac{V(m)}{M^2} - \frac{V(\bar{x})}{\bar{X}^2} \leq 2 \left\{ \theta_j' \frac{Cov(\bar{y}, m)}{\bar{Y}M} - \frac{Cov(\bar{y}, \bar{x})}{\bar{Y}\bar{X}} \right\}$$

$$\Rightarrow \theta_j'^2 C'_{mm} - C'_{xx} \leq 2 \left\{ \theta_j' C'_{ym} - C'_{yx} \right\}; j = 1, 2, 3, 4, 5$$

### Comparison with that of Modified Ratio Estimators

Consider  $MSE\left(\hat{\bar{Y}}_{SPj}\right) \leq MSE\left(\hat{\bar{Y}}_{RMi}\right)$

$$\Rightarrow V(\bar{y}) + R^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq V(\bar{y}) + R^2 \theta_i'^2 V(\bar{x}) - 2R\theta_i' Cov(\bar{y}, \bar{x})$$

$$\Rightarrow R^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq R^2 \theta_i'^2 V(\bar{x}) - 2R\theta_i' Cov(\bar{y}, \bar{x})$$

$$\Rightarrow R^2 \theta_j'^2 V(m) - R^2 \theta_i'^2 V(\bar{x}) \leq 2R' \theta_j' Cov(\bar{y}, m) - 2R\theta_i' Cov(\bar{y}, \bar{x})$$

$$\Rightarrow \frac{\bar{Y}^2}{M^2} \theta_j'^2 V(m) - \frac{\bar{Y}^2}{\bar{X}^2} \theta_i'^2 V(\bar{x}) \leq 2 \frac{\bar{Y}}{M} \theta_j' Cov(\bar{y}, m) - 2 \frac{\bar{Y}}{\bar{X}} \theta_i' Cov(\bar{y}, \bar{x})$$

$$\Rightarrow \theta_j'^2 \frac{V(m)}{M^2} - \theta_i'^2 \frac{V(\bar{x})}{\bar{X}^2} \leq 2 \left\{ \theta_j' \frac{Cov(\bar{y}, m)}{\bar{Y}M} - \theta_i' \frac{Cov(\bar{y}, \bar{x})}{\bar{Y}\bar{X}} \right\}$$

$$\Rightarrow \theta_j'^2 C'_{mm} - \theta_i'^2 C'_{xx} \leq 2 \{ \theta_j' C'_{ym} - \theta_i' C'_{yx} \}; i, j = 1, 2, 3, 4, 5$$

### Comparison with that of Linear Regression Estimator

Consider  $MSE(\hat{\bar{Y}}_{SPj}) \leq V(\hat{\bar{Y}}_{lr})$

$$\Rightarrow V(\bar{y}) + R'^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq V(\bar{y})(1 - \rho^2)$$

$$\Rightarrow R'^2 \theta_j'^2 V(m) - 2R' \theta_j' Cov(\bar{y}, m) \leq -V(\bar{y}) \left( \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x}) * V(\bar{y})} \right)$$

$$\Rightarrow 2R' \theta_j' Cov(\bar{y}, m) - R'^2 \theta_j'^2 V(m) \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})}; j = 1, 2, 3, 4, 5$$

$$\Rightarrow 2 \frac{\bar{Y}}{M} \theta_j' Cov(\bar{y}, m) - \frac{\bar{Y}^2}{M^2} \theta_j'^2 V(m) \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})}$$

$$\Rightarrow 2\bar{Y}^2 \theta_j' C'_{ym} - \bar{Y}^2 \theta_j'^2 C'_{mm} \geq \frac{[Cov(\bar{y}, \bar{x})]^2}{V(\bar{x})}$$

$$\Rightarrow 2\theta_j' C'_{ym} - \theta_j'^2 C'_{mm} \geq \frac{[C'_{yx}]^2}{C'_{xx}}, j = 1, 2, 3, 4, 5$$

**Comparison with that of Median Based Ratio Estimator**

Consider  $MSE\left(\hat{Y}_{SPj}\right) \leq MSE\left(\hat{Y}_M\right)$

$$\Rightarrow V(\bar{y}) + R'^2 \theta_j'^2 V(m) - 2R'\theta_j' Cov(\bar{y}, m) \leq V(\bar{y}) + R'^2 V(m) - 2R' Cov(\bar{y}, m)$$

$$\Rightarrow R'^2 \theta_j'^2 V(m) - 2R'\theta_j' Cov(\bar{y}, m) \leq R'^2 V(m) - 2R' Cov(\bar{y}, m)$$

$$\Rightarrow R'^2 \theta_j'^2 V(m) - R'^2 V(m) \leq 2R'\theta_j' Cov(\bar{y}, m) - 2R' Cov(\bar{y}, m)$$

$$\Rightarrow R' V(m) (\theta_j'^2 - 1) \leq 2(\theta_j' - 1) Cov(\bar{y}, m)$$

$$\Rightarrow R' V(m) (\theta_j' - 1) (\theta_j' + 1) \leq 2(\theta_j' - 1) Cov(\bar{y}, m)$$

$$\Rightarrow Cov(\bar{y}, m) \leq \frac{R' (\theta_j' + 1) V(m)}{2} \text{ Since } \theta_j' < 1; j = 1, 2, 3, 4, 5$$

$$\Rightarrow Cov(\bar{y}, m) \leq \frac{\bar{Y}M (\theta_j' + 1) C'_{mm}}{2} \text{ Since } \theta_j' < 1$$

$$\Rightarrow 2C'_{ym} \leq (\theta_j' + 1) C'_{mm}, j = 1, 2, 3, 4, 5$$



# Ridge Regression in Calibration Models with Symmetric Padding Extension-Daubechies Wavelet Transform Preprocessing

**Nurwiani**

STKIP PGRI Jombang  
Jombang, Indonesia

**S. Sunaryo**

Institut Teknologi  
Sepuluh Nopember  
Surabaya, Indonesia

**Setiawan**

Institut Teknologi  
Sepuluh Nopember  
Surabaya, Indonesia

**B. W. Otok**

Institut Teknologi  
Sepuluh Nopember  
Surabaya, Indonesia

---

Wavelet transformation is commonly used in calibration models as a preprocessing step. This preprocessing does not involve all results of a spectrum discretization; consequently, a lot of information can be missing. To avoid missing information, a symmetric padding extension (SPE) can be used to place all data points into dyadic scales, however, high dimensional discretization points need to be reduced. Dimension reduction can be performed with Daubechies wavelet transformation (DWT). Scale function and Daubechies wavelet are continuous functions, thus they perform a faster approximation. SPE-DWT preprocessing combines SPE and DWT. Multicollinearity often occurs in calibration models; the ridge regression (RR) method can be used to solve multicollinearity problems. This article proposes the RR method with SPE-DWT preprocessing. The proposed method is applied to determine a model for predicting the content of curcumin in turmeric. Selection of the best model is carried out by comparing coefficient of determinations, p-values of the Kolmogorov-Smirnov (KS) error models, and Root Mean Square Error Prediction (RMSEP). Results show that the RR method with SPE-DWT preprocessing gives an accurate prediction.

*Keywords:* Calibration models, Daubechies wavelet transform, symmetric padding extension

---

## Introduction

In calibration models, the number of observations is usually much smaller than the number of points resulted from the spectrum discretization obtained from Fourier Transform Infrared (FTIR). In preprocessing calibration models, some researchers use Daubechies wavelet transformation (DWT) without involving all

---

*Nurwiani is a lecturer in the Mathematics Department. Email her at: nurw\_13iem64@yahoo.com. Sunaryo, Setiawan and Otok are promotor in the Statistics Department.*

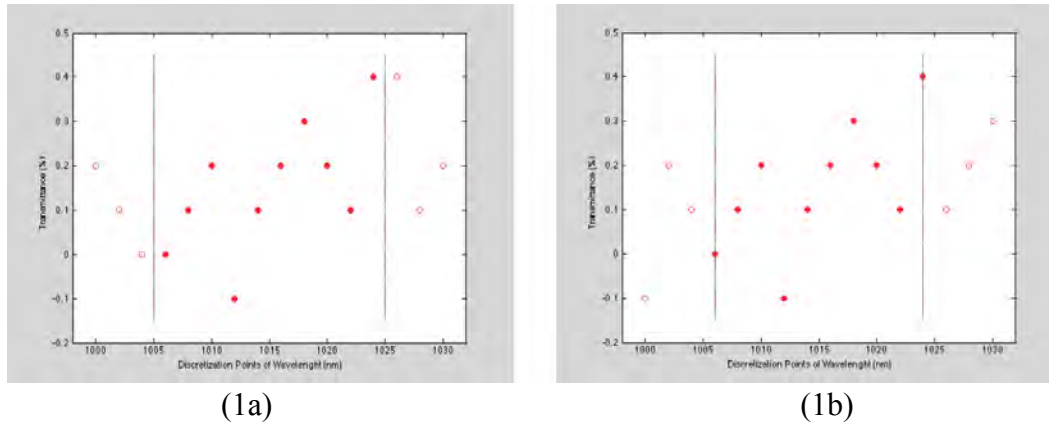
points resulted from discretization; this means that a lot of information will be missing in building models. Using DWT, Brown et al. (2001) involved  $2^8$  of 700 points in building models to estimate the content of fat, sugar, flour and water in bread dough. Using Haar wavelet transform, Sunaryo and Retnaningsih (2008) used  $2^{10}$  of 1,866 points to estimate the content of gingerol in ginger. In these two studies there are as many as 444 and 842 missing information points, respectively.

To avoid missing information, symmetric padding extension (SPE) can be used with all data points in dyadic scales. Spectrum discretization points are predictor variables with an original size of  $p$ , changed to  $q = 2^M$ , where  $M$  is a positive integer and  $q \geq p$ . Dimension reduction can be performed with DWT. The scale function and Daubechies wavelet are continuous functions; thus, they can perform a faster approximation.

To date, SPE-DWT preprocessing, which combines SPE and DWT in calibration models, has not been used. SPE-DWT preprocessing avoids information loss during preprocess and determines the orthogonal matrix in dimension reduction process. Multicollinearity often occurs in calibration models; the ridge regression (RR) method can be used to solve multicollinearity problem. This manuscript proposes a RR method with SPE-DWT preprocessing.

### Methodology

Several methods are available to categorize discretization points of wavelength into dyadic scales; one method is the SPE. According to Boggess & Narcowich (2001), SPE is defined as a spectrum that is evenly extended at the endpoints by reflection in two ways: (i) discretization points are reflected around mid-line between the end point and the next point, expressed by SPE1, and (ii) discretization points are reflected in the line through the two end points, expressed with SPE2. Figures 1(a) and 1(b) illustrate the SPE1 and SPE2 of 10 discretization points.



**Figure 1:** SPE1 (1a) and SPE2 (1b) of 10 Discretization Points

If the discretization points are matrix  $\mathbf{X}_{n_1 \times p}^*$ , then centering and SPE are performed on this matrix, the resulting matrix will be  $\mathbf{X}_{n_1 \times q}^*$ . The next step is determining the orthogonal matrix, size  $q \times q$ , for dimension reduction process by using the wavelet analysis.

There are two main functions in wavelet analysis: scale function  $\phi$  (father wavelet) and wavelet function  $\psi$  (mother wavelet). Both functions produce a family of functions that can be used to solve or reconstruct a spectrum (Boggess & Narcowich, 2001). Daubechies wavelet (Daubechies, 1992) is one of these wavelet functions. The scale function and Daubechies wavelet are continuous functions and, thus, can perform a faster approximation. Advantages of the Daubechies wavelet are compact support (closed and bounded), and that the width of support depends on the number of vanishing moments  $L$  (which limits the pedestal width) (Daubechies, 1992).

A smoother scaling function and the Daubechies wavelet function can be determined by choosing the power  $\ell = 2L - 1$  and filter length  $N = 2L$ . In the Daubechies wavelet, for each  $L$  number of vanishing moments, there will be  $2L$  coefficient scales with non-zero values. The scale and Daubechies wavelet functions are located on the interval  $0 \leq t \leq 2L - 1$ . Daubechies wavelet is commonly expressed by  $dbN$  for  $L = 2, \dots, 10$  or by  $db2L$  for  $L = 2, \dots, 10$  (Burrus et al., 1998; Boggess & Narcowich, 2001). To define the Daubechies wavelets, consider the two functions  $\phi(t)$  and  $\psi(t)$  which are solutions to the following equations:

equation of scale function,

$$\phi(t) = \sqrt{2} \sum_{k=0}^{N-1} h_k \phi(2t - k), \quad (1)$$

and equation of wavelet function,

$$\psi(t) = \sqrt{2} \sum_{k=0}^{N-1} g_k \phi(2t - k). \quad (2)$$

Filter coefficients of the scale function  $\phi(t)$  in (1) for the Daubechies wavelet must satisfy:

$$\sum_{k=0}^{N-1} h_k = \sqrt{2} \quad (3)$$

$$\sum_{k=0}^{N-1} (-1)^k k^m h_k = 0, \quad m = 0, 1, \dots, \frac{N}{2} - 1 \quad (4)$$

$$\sum_{k=0}^{N-1} h_k h_{k+2m} = 0, \quad m \neq 0 \quad (5)$$

$$\sum_{k=0}^{N-1} h_k^2 = 1 \quad (6)$$

A set sequence of scale filter numbers  $H = \{h_0, h_1, h_2, \dots, h_{N-1}\}$  called low-pass filters in a pyramid algorithm can be obtained from (3), (4), (5) and (6) (see Burrus et al. (1998) for a detailed discussion about the pyramid). The relationship between the scale filter coefficients  $h_k$  and wavelet filter coefficients  $g_k$ :

$$g_k = (-1)^k h_{N-1-k} \quad (7)$$

A set sequence of wavelet filter numbers  $G = \{g_0, g_1, \dots, g_{N-1}\}$  called high-pass filters in the pyramid algorithm is obtained from (7); based on the orthogonal wavelet matrix that satisfies (1) and (2) it can be determined using:

$$(\mathbf{W}^*)^t \mathbf{W} = \mathbf{W}^* (\mathbf{W}^*)^t = \mathbf{I}_q. \quad (8)$$

After centering and SPE are performed, the matrix  $\mathbf{W}^{*t}$  in (8) is used for dimension reduction. Dimension reduction can be done by determining the diagonal of matrix  $\Lambda_{q \times q}^*$  which are the eigenvalues of symmetric matrix  $(\mathbf{X}^t \mathbf{X})_{q \times q}$  (Anton & Rorres, 2005, p566, eq1) Hence, the diagonal of matrix  $\Lambda_{q \times q}^*$  is obtained from:

$$(\mathbf{W}^*)_{q \times q}^t (\mathbf{X}^t \mathbf{X})_{q \times q} (\mathbf{W}^*)_{q \times q} = \Lambda_{q \times q}^*.$$

Considering the proportion of

$$p(\lambda_r^*) = (\lambda_1^* + \lambda_2^* + \dots + \lambda_r^*) / (\lambda_1^* + \lambda_2^* + \dots + \lambda_q^*),$$

matrix  $\mathbf{W}_{q \times r}^t$  for dimension reduction is obtained, where  $r < q$ . Dimension reduction of predictor variables and parameters is determined with:

$$\mathbf{Z}_{n_1 \times r} = \mathbf{X}_{n_1 \times q} \mathbf{W}_{q \times r}^t \text{ and } \boldsymbol{\gamma}_{r \times 1} = \mathbf{W}_{r \times q} \boldsymbol{\beta}_{q \times 1}$$

Because the points of the spectrum discretization resulted in calibration models are generally highly correlated, it is necessary to carry out multicollinearity detection. According to Shi-ji & Zhi-bin (1993), by taking into account the type of condition number

$$\text{cond}(\mathbf{Z}^t \mathbf{Z}) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

multicollinearity can be detected through:

- i. If  $0 < \text{cond}(\mathbf{Z}^t \mathbf{Z}) < 100$ , there is no multicollinearity (Type I);
- ii. If  $100 \leq \text{cond}(\mathbf{Z}^t \mathbf{Z}) \leq 1000$ , there is some moderate or stronger degree of multicollinearity (Type II);
- iii. If  $\text{cond}(\mathbf{Z}^t \mathbf{Z}) > 1000$ , there is some serious degree of multicollinearity (Type III).

In this study, the active compound curcumin in turmeric is the response variable determined from extraction using High Performance Liquid Chromatography (HPLC). Because data for this response variable does not follow a normal distribution, Johnson Transformation (JT) is carried out. The original

## RIDGE REGRESSION IN CALIBRATION MODELS

response variable is  $\mathbf{y}_{n_1 \times 1}^*$ , whereas  $\mathbf{y}_{n_1 \times 1}$  is the response variable that has been transformed and centered. Given the normal distribution of  $y^*$ :

$$f(y^*) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y^* - \mu)^2 / 2\sigma^2},$$

for

$$-\infty < \mu < \infty, \sigma > 0, -\infty < y^* < \infty.$$

If  $y^*$  is not normally distributed, the JT can be determined using:

$$f(y^*) = \frac{1}{\sqrt{2\pi}} e^{-v^2/2}, -\infty < v < \infty, -\infty < y^* < \infty$$

where  $v = \alpha + \eta f_i(y^*, \delta, \xi)$ ,  $i = 1, 2, 3$ ,  $v$  is a standard normal random variable,  $\alpha$  and  $\eta$  are the shape parameters,  $\delta$  is a scale parameter and  $\xi$  is a location parameter. It is assumed that  $\eta > 0$  and  $\delta > 0$ . Based on its curve, the JT can be differentiated into three systems (George, 2007):

i.  $v_{S_b}$  with  $f_1$  Bounded system,  $f_1(y^*, \delta, \xi) = \ln\left(\frac{y^* - \xi}{\delta + \xi - y^*}\right)$ ,  $\xi \leq y^* \leq \delta + \xi$

ii.  $v_{S_l}$  with  $f_2$  Log-normal system and  $f_2(y^*, \delta, \xi) = \ln\left(\frac{y^* - \xi}{\delta}\right)$ ,  $y^* > \xi$

iii.  $v_{S_u}$  with  $f_3$  Unbounded system,  $f_3(y^*, \delta, \xi) = \sinh^{-1}\left(\frac{y^* - \xi}{\delta}\right)$ ,  $-\infty < y^* < \infty$

where:  $\sinh^{-1}\left(\frac{y^* - \xi}{\delta}\right) = \ln\left(\left(\frac{y^* - \xi}{\delta}\right) + \sqrt{1 + \left(\frac{y^* - \xi}{\delta}\right)^2}\right)$ . After the dimensional

reduction process, calibration models are obtained as (Naes et al., 2002):

$$\mathbf{X}_{n_1 \times q} = \mathbf{Z}_{n_1 \times r} \mathbf{W}_{r \times q} + \mathbf{F}_{n_1 \times q}, \quad (9)$$

$$\mathbf{y}_{n_1 \times 1} = \mathbf{Z}_{n_1 \times r} \boldsymbol{\gamma}_{r \times 1} + \boldsymbol{\varepsilon}_{n_1 \times 1}. \quad (10)$$

The eigenvalues

$$(\mathbf{Z}'\mathbf{Z})_{r \times r} = \boldsymbol{\Lambda}_{r \times r} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$$

can be obtained from (9), where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$  and  $\lambda_i > 0, i = 1, 2, \dots, r$ . The parameter estimation using ordinary least square can be defined as:

$$\hat{\boldsymbol{\gamma}}_{r \times 1} = \boldsymbol{\Lambda}_{r \times r}^{-1} \mathbf{Z}_{r \times n_1}' \mathbf{y}_{n_1 \times 1}$$

$$\text{where } \hat{\mathbf{Z}}_{n_1 \times r} = \mathbf{X}_{n_1 \times q} \hat{\mathbf{W}}_{q \times r}', \quad \hat{\boldsymbol{\beta}}_{q \times 1} = \mathbf{W}_{q \times r}' \hat{\boldsymbol{\gamma}}_{r \times 1},$$

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{Z}'\mathbf{Z})_{r \times r}^{-1} \text{ and}$$

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{i=1}^r \lambda_i^{-1}.$$

One of the methods to overcome the multicollinearity problem is the RR method (Hoerl & Kennard, 1970). From (10) ridge parameter estimation can be obtained as:

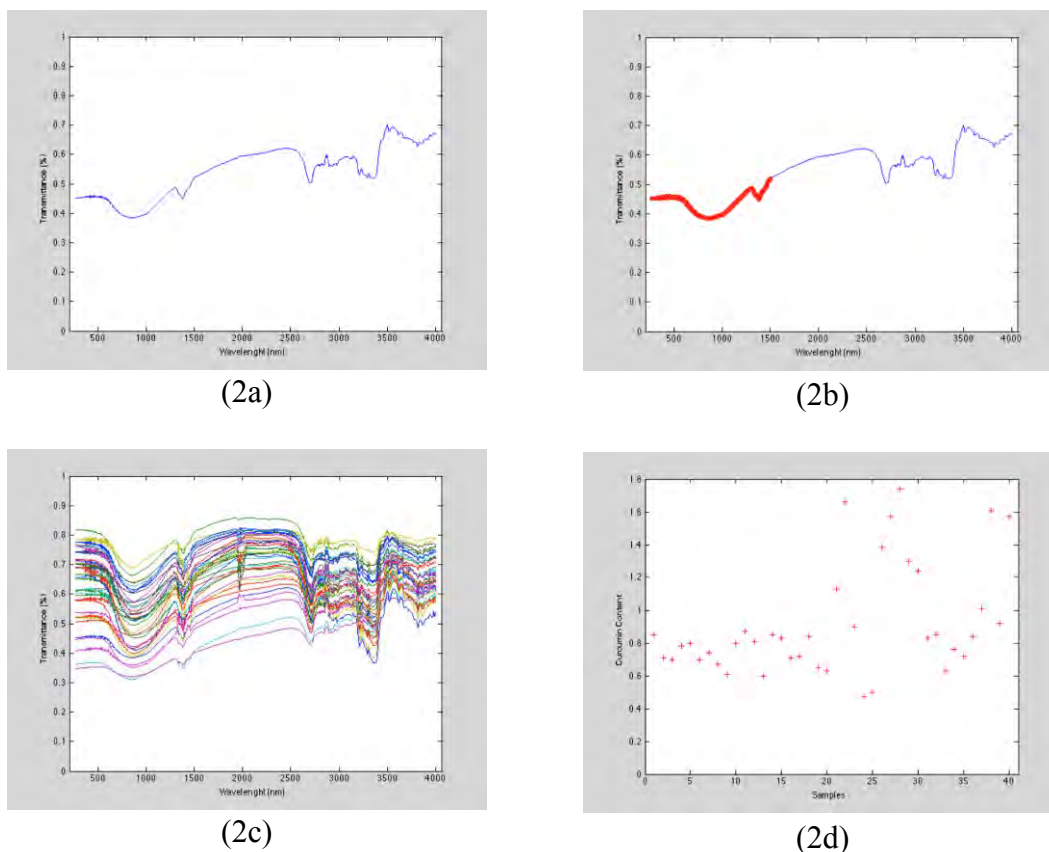
$$\hat{\boldsymbol{\gamma}}_{q \times 1}(\boldsymbol{\theta}) = (\mathbf{Z}'\mathbf{Z} + \boldsymbol{\theta}\mathbf{I})_{q \times q}^{-1} \mathbf{Z}_{q \times n_1}' \mathbf{y}_{q \times 1}$$

where  $\theta > 0$ , and the ridge parameter can be determined using  $k$  iterations (see Shi-ji & Zhi-bin (1993) for a detailed explanation of RR).

In the process of building models it is important to validate the selected regression models. According to Neter et al. (1989), the regression models can be validated by dividing the data into two parts ( $n = n_1 + n_2$ ). The first set of data  $n_1$ , called the model building set, is used to build the model. The second data set  $n_2$ , called the validation or prediction set, is applied to validate the model.

## Results

Calibration models using the RR method with SPE-DWT preprocessing are built to predict the content of curcumin in turmeric. The discretization points resulted from FTIR spectrum and the content of curcumin in turmeric determined by HPLC is shown in Figure 2. Figure 2(a) shows the FTIR spectrum of a sample with  $p=1866$  points, Figure 2(b) demonstrates the 616 first discretization points of the first sample data, Figure 2(c) illustrates the FTIR spectrum of 40 samples and Figure 2(d) shows the percentage of curcumin in turmeric as determined by the HPLC from 40 samples. These 40 samples are then divided into two parts, the first part of the data set consists of  $n_1 = 30$  samples as a model building set and the second part of data set comprises of,  $n_2 = 10$  samples as a prediction set.



**Figure 2.** Data of Curcumin in Turmeric



In general, the number of samples in the calibration models is limited, therefore, it is important to conduct normality test for the response variables. There are 30 observations of curcumin in turmeric that do not meet the normality assumption as shown with the Kolmogorov-Smirnov (KS)  $p$  value =  $3.7478e-13$  (see Figure 3(a)). The JT on the response variable yields:

$$v_{S_U} = -0.568871 + 0.784968 \sinh^{-1} \left( \frac{y^* - 0.708465}{0.0817548} \right)$$

with the KS  $p$  value = 0.6320. Further,  $y$ , a centered value of  $v_{S_U}$ , is defined and illustrated in Figure 3(b). In this study, the building set is carried out only for SPE1 on predictor variables, thus, the number of discretization points is  $q = 2^{11}$ .

Using DWT through the pyramid algorithm of  $h_k$  and  $g_k$  for  $N=10$ , an orthogonal matrix  $\mathbf{W}_{2048}^*$  is obtained. In data processing, the RR method requires  $r < n_1 - 1$  where  $n_1 = 30$ . Hence dimension reduction is done by determining the number of transformation matrices for  $r=1, 2, \dots, 28$ , and finally, the reduced matrix  $\hat{\mathbf{W}}_{2048 \times r}^t$  is obtained.

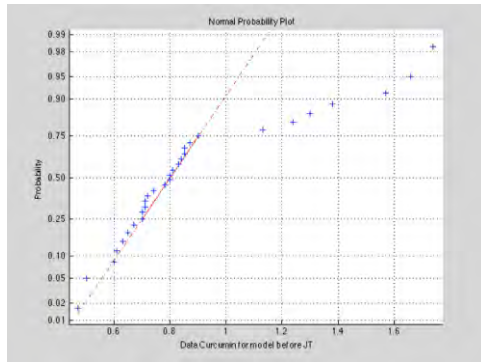
For  $r=1$  and  $r=2$ , the reduced matrix  $\hat{\mathbf{W}}_{2048 \times r}^t$  yields  $\text{cond}(\mathbf{Z}'\mathbf{Z})$  of Type I, while for  $r=3$ , it yields  $\text{cond}(\mathbf{Z}'\mathbf{Z})$  of Type II. For  $r=4, 5, \dots, 28$  the reduced matrix  $\hat{\mathbf{W}}_{2048 \times r}^t$  yields  $\text{cond}(\mathbf{Z}'\mathbf{Z})$  of Type III.

The RR method to overcome the multicollinearity problem among predictor variables is completed with multiple iterations. Table 1 presents the results of this study. The best model gives the coefficient of determination (Johnson & Whincern, 2002; Seber & Lee, 2003)  $R^2 > 0.85$ , has the smallest root mean square error (RMSE) (Naes et al., 2002) and has a KS  $p$  value error model more than 0.05 (Marsaglia et al., 2003). As Table 1 shows, the RR method with the SPE-DWT preprocessing can be used to build the best models for accurate prediction.

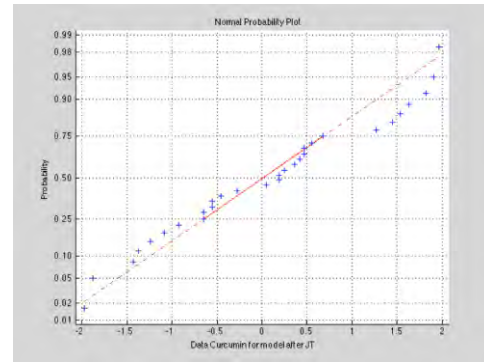
## RIDGE REGRESSION IN CALIBRATION MODELS

**Table 1:** SPE1-Wavelet Daubechies Ridge Regression Models

SPE1	Iteration	KS $p$ value error model	$R^2$	RMSEP
$r = 1$ $\text{cond}(\mathbf{Z}'\mathbf{Z}) = 1$ $p(\lambda_1^*) = 83.9266$	<b>12,000</b>	<b>0.3333</b>	<b>85.8942</b>	<b>0.4513</b>
	12,100	0.3123	91.4549	0.4551
	12,200	0.2922	97.5712	0.4592
	12,225	0.2873	99.1949	0.4603
	12,230	0.2864	99.5245	0.4605
	12,235	0.2854	99.8556	0.4607
$r = 2$ $p(\lambda_2^*) = 90.0857$ $\text{cond}(\mathbf{Z}'\mathbf{Z}) = 23.3715$	<b>11,650</b>	<b>0.3337</b>	<b>85.4124</b>	<b>0.4518</b>
	11,700	0.3228	88.1813	0.4537
	11,800	0.3017	94.1391	0.4578
	11,850	0.2915	97.3455	0.4599
	11,875	0.2865	99.0096	0.4610
	11,885	0.2846	99.6871	0.4615
	11,888	0.2840	99.8917	0.4616
$r = 3$ $p(\lambda_3^*) = 92.8664$ $\text{cond}(\mathbf{Z}'\mathbf{Z}) = 901.7588$	<b>8,625</b>	<b>0.3828</b>	<b>85.3135</b>	<b>0.3258</b>
	8,650	0.3727	87.2917	0.3261
	8,700	0.3517	91.4624	0.3268
	8,725	0.3416	93.6611	0.3272
	8,750	0.3316	95.9399	0.3276
	8,775	0.3218	98.3023	0.3280
	8,785	0.3179	99.2715	0.3282
	8,790	0.3160	99.7615	0.3283
	8,791	0.3156	99.8599	0.3283
	8,792	0.3152	99.9584	0.3283



(3a)



(3b)

**Figure 3:** Normal Probability Plot of 30 Curcumin Data

## Conclusion

In calibration models, response variables often do not meet the normal distribution assumption; therefore, the JT is necessary to fulfil model assumptions. The SPE-DWT with filter 10 is able to reduce the dimension, however, there is no guarantee that it can cope with multicollinearity problem. An effective method is needed to overcome the multicollinearity problem. This study shows that the combination of JT and SPE-DWT preprocessing in the RR method can be used to build models that will give accurate predictions. Further study is suggested by implementing the RR method with determination of optimum ridge parameter.

## References

- Anton, H., & Rorres, C. (2005). *Elementary linear algebra*. New York: John Wiley & Sons, Inc.
- Boggess, A., & Narcowich, F. J. (2001). *A first course in wavelets with fourier analysis*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Brown, P. J., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96: 398-408.
- Burrus, C. S., Gopinath, R. A., & Guo, H. (1998). *Introduction to wavelets and wavelet transforms*. Upper Saddle River, NJ: Prentice-Hall, Inc.
- Daubechies, I. (1992). *Ten lectures on wavelets* (5<sup>th</sup> ed.). Philadelphia, PA: Society for Industrial and Applied Mathematics.
- George, F. (2007). *Johnson's system of distributions and microarray data analysis*. (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12: 55-67.
- Johnson, R. A., & Winchern, D.W. (2002) *Applied multivariate statistical analysis* (5<sup>th</sup> ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.
- Marsaglia, G., Tsang, W., & Wang, J. (2003). Evaluating Kolmogorov's Distribution. *Journal of Statistical Software*, 8(18): 1-4.
- Naes, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to Multivariate Calibration and Classification*. Chichester, UK: IM Publications.

## RIDGE REGRESSION IN CALIBRATION MODELS

Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models* (2<sup>nd</sup> ed.). Homewood, IL: Irwin, Inc.

Seber, G. A., & Lee, A. J. (2003). *Linear regression analysis*. New York: John Wiley & Sons.

Shi-ji, C., & Zhi-bin, Z. (1993). Generalized multivariate ridge regression estimate and criteria  $Q(c)$  for choosing matrix  $K$ . *Applied Mathematics and Mechanics*, 14: 73-84.

Sunaryo, S., & Retnaningsih, S. M. (2008). The implementation of calibration models with discrete wavelet transformation-partial least squares (TWD-PLS) on gingerol data (in Indonesian language). *Ilmu Dasar*, 9: 51-56

# Estimation and Testing in Type-II Generalized Half Logistic Distribution

**R. R. L. Kantam**

Acharya Nagarjuna University  
Nagarjunanagar, India

**V. Ramakrishna**

K. L. University  
Vaddeswaram, India

**M. S. Ravikumar**

Acharya Nagarjuna University  
Nagarjunanagar, India

---

A generalization of the Half Logistic Distribution is developed through exponentiation of its survival function and named the Type II Generalized Half Logistic Distribution (GHLD). The distributional characteristics are presented and estimation of its parameters using maximum likelihood and modified maximum likelihood methods is studied with comparisons. Discrimination between Type II GHLD and exponential distribution in pairs is conducted via likelihood ratio criterion.

*Keywords:* Generalized Half Logistic Distribution (GHLD), maximum likelihood estimation (MLE), modified maximum likelihood estimation (MMLE), mean square error (MSE), likelihood ratio type criterion, percentiles, power of the test

---

## Introduction

In life testing and reliability studies a combination of monotone and constant failure rates over various segments of the range of lifetime of a random variable is also known as bath tub or non-monotone failure rate. In biological and engineering sciences, situations of non-monotone failure rates are common (see Rajarshi & Rajarshi (1988) for a comprehensive narration of these models). Mudholkar, et al. (1995) presented an extension of the Weibull family that contains unimodal distributions with bathtub failure rates and also allows for a broader class of monotone hazard rates. They named their extended version the Exponentiated Weibull Family.

Gupta and Kundu (1999) proposed a new model called the generalized exponential distribution. If  $\theta$  is a positive real number and  $F(x)$  is the cumulative

---

*Dr. R. R. L. Kantam is a Professor in the Department of Statistics. Email him at [kantam.rrl@gmail.com](mailto:kantam.rrl@gmail.com). V. Ramakrishna is an Associate Professor in the Department of Computer Science and Engineering. Email him at: [vramakrishna2006@gmail.com](mailto:vramakrishna2006@gmail.com). M. S. Ravikumar is a UGC Research Fellow in the Department of Statistics. Email him at: [msrk.raama@gmail.com](mailto:msrk.raama@gmail.com).*

distribution function (cdf) of a continuous positive random variable, then  $[F(x)]^\theta$  and the corresponding probability distribution may be termed an exponentiated or generalized version of  $F(x)$ .

A half logistic model obtained as the distribution of absolute standard logistic variate is a probability model of recent origin (Balakrishnan, 1985). Its standard probability density function, cumulative distribution function and hazard functions are given by:

$$f(x) = \frac{2e^{-x}}{(1+e^{-x})^2}, \quad x \geq 0 \quad (1)$$

$$F(x) = \left[ \frac{1-e^{-x}}{1+e^{-x}} \right], \quad x \geq 0 \quad (2)$$

$$F(x) = \left[ \frac{1-e^{-x}}{1+e^{-x}} \right], \quad x \geq 0. \quad (3)$$

Kantam et al. (2011) adopted this generalization to the well-known half logistic distribution, and named it the Type-I Generalized Half Logistic Distribution (GHLD).

Consider a series system of  $\theta$  components with individually and identically distributed (iid) individual lifetimes, for example,  $F(x)$ . The reliability function of such a system is given by  $[1 - F(x)]^\theta$ ; hence, the distribution function of the lifetime random variable of a series system is  $1 - [1 - F(x)]^\theta$ .

Taking  $F(x)$  as the half logistic model given by Equation (2), the corresponding distribution is termed the Type-II Generalized Half Logistic Distribution (GHLD-II). Its pdf, cdf and hazard function are given by:

$$f(x) = \frac{\theta(2e^{-x})^\theta}{(1+e^{-x})^{\theta+1}}, \quad x > 0, \quad \theta > 0 \quad (4)$$

$$F(x) = 1 - \left[ \frac{2e^{-x}}{1+e^{-x}} \right]^\theta, \quad x > 0, \quad \theta > 0 \quad (5)$$

$$h(x) = \frac{\theta}{1+e^{-x}}, \quad x > 0, \quad \theta > 0. \quad (6)$$

Balakrishnan and Sandhu (1995) suggested a new probability model with a standard pdf and cdf given by:

$$f(x) = \frac{2(1-kx)^{(1/k)-1}}{[1+(1-kx)^{1/k}]^2}, 0 \leq x \leq \frac{1}{k}, k \geq 0 \quad (7)$$

$$F(x) = \frac{1-(1-kx)^{1/k}}{1+(1-kx)^{1/k}}, 0 \leq x \leq \frac{1}{k}, k \geq 0. \quad (8)$$

The limits of (7) and (8) as  $k \rightarrow \infty$  are respectively (1) and (2) – the pdf and cdf of HLD. Balakrishnan and Sandhu (1995) called the distribution (7) and (8) Generalized HLD.

Olapade (2008) considered two distributions and discussed their distributional properties, order statistics in samples from these distributions: He named these distributions type-I and type-III GHLD, respectively. The types of generalized HLD of Olapade (2008) are through truncation of the type-I and type-III generalized logistic distributions from Balakrishnan and Leung (1988) at the origin. Thus, this type-II GHLD is conceptually different from the GHLDs of Balakrishnan and Sandhu (1995) and Olapade (2008). Hence, the proposed models motivated a separate research study.

## Estimation in Type-II Generalized Half Logistic Distribution (GHLD-II)

The probability density function and distribution function of GHLD-II with scale parameter  $\sigma$  and power parameter  $\theta$  are given by:

$$f(x) = \frac{\theta(2e^{-x/\sigma})^\theta}{\sigma(1+e^{-x/\sigma})^{\theta+1}}, 0 < x < \infty, \sigma > 0, \theta > 0 \quad (9)$$

$$F(x) = 1 - \left[ \frac{2e^{-x/\sigma}}{1+e^{-x/\sigma}} \right]^\theta, 0 < x < \infty, \sigma > 0, \theta > 0. \quad (10)$$

Let  $x_1 < x_2 < \dots < x_n$  be an ordered sample of size  $n$  from GHLD-II. The log likelihood function of the sample is

$$\log L = n(\log \theta - \log \sigma) + \sum_{i=1}^n \left[ \theta \log 2 - \theta \frac{x_i}{\sigma} - \theta \log(1 + e^{-x_i/\sigma}) - \log(1 + e^{-x_i/\sigma}) \right]$$

The log likelihood equations to estimate the parameters  $\sigma$  and  $\theta$  are given by

$$\frac{\partial \log L}{\partial \sigma} = 0, \quad \frac{\partial \log L}{\partial \theta} = 0,$$

$$\frac{\partial \log L}{\partial \sigma} = 0 \Rightarrow \sum_{i=1}^n \frac{x_i}{\sigma} \left[ \frac{\theta - e^{-x_i/\sigma}}{1 + e^{-x_i/\sigma}} \right] = n \quad (11)$$

$$\frac{\partial \log L}{\partial \theta} = 0 \Rightarrow \theta = \frac{n}{\sum_{i=1}^n \log(1 + e^{-x_i/\sigma}) + \frac{1}{\sigma} \sum_{i=1}^n x_i - \log 2^n} \quad (12)$$

It can be seen that these two equations must be solved iteratively for  $\theta$  and  $\sigma$  for a given sample. The asymptotic variances and covariances of MLEs of  $\sigma$  and  $\theta$  can be obtained by inverting the information matrix whose elements are the mathematical expectation of the following expressions:

$$-\left( \frac{\partial^2 \log L}{\partial \sigma^2} \right) = \sum_{i=1}^n \left[ \frac{x_i^2}{\sigma^4} \frac{(1 - \theta + 2e^{-x_i/\sigma})e^{-x_i/\sigma}}{(1 + e^{-x_i/\sigma})^2} \right] + \frac{n}{\sigma^2} \quad (13)$$

$$-\left( \frac{\partial^2 \log L}{\partial \theta^2} \right) = \frac{n}{\theta^2} \quad (14)$$

$$-\left( \frac{\partial^2 \log L}{\partial \theta \partial \sigma} \right) = -\frac{1}{\sigma^2} \sum_{i=1}^n \left[ \frac{x_i}{(1 + e^{-x_i/\sigma})} \right] \quad (15)$$



These equations, evaluated at estimates of  $\theta$  and  $\sigma$ , provide an estimated dispersion matrix. In order to obtain an analytical estimator for  $\sigma$ , its estimating equation is approximated by some admissible expression.

Equation (11) to get MLE of  $\sigma$ , after simplification would become

$$\sum_{i=1}^n \frac{z_i(\theta - e^{-z_i})}{1 + e^{-z_i}} - n = 0 \text{ where } z_i = \frac{x_i}{\sigma} \quad (16)$$

To obtain the analytical expression for  $\sigma$ , approximate the following expression in (16) by some linear function in the corresponding population quartile. Let,

$$G(z_i) = \frac{z_i(\theta - e^{-z_i})}{(1 + e^{-z_i})} \quad (17)$$

approximate

$$G(z_i) \approx \alpha_i + \beta_i z_i \quad (18)$$

where  $\alpha_i, \beta_i$  are to be suitably found. After using this approximation in (16) the solution for  $\sigma$  is

$$\hat{\sigma} = \frac{\sum_{i=1}^n \beta_i x_i}{n - \sum_{i=1}^n \alpha_i} \quad (19)$$

This estimator is named the MMLE of  $\sigma$ , which is a linear estimator in  $x_i$ 's

To obtain  $\alpha_i, \beta_i$ , let  $p_i = \frac{i}{n+1}; i=1,2,\dots,n$  and let  $t_i, t_i^*$  be the solutions of equations:

$$F(t_i) = p_j - \sqrt{\frac{p_i q_i}{n}} = p_i' \text{ (for example)} \quad (20)$$

$$F(t_i^*) = p_i + \sqrt{\frac{p_i q_i}{n}} = p_i'' \text{ (for example), where } q = 1 - p_i \quad (21)$$

where  $F(\cdot)$  is cdf of GHLD-II.

The intercept  $\alpha_i$  and slope  $\beta_i$  of linear approximation in the Equation (18) are respectively given by

$$\beta_i = \frac{G(t_i^*) - G(t_i)}{t_i^* - t_i} \quad (22)$$

$$\alpha_i = G(t_i^*) - \beta_i t_i^*. \quad (23)$$

Using distribution function  $F(\cdot)$  of GHLD-II, the expressions for  $t_i, t_i^*$  are given by

$$t_i = \log \left( \frac{2 - (1 - p_i')^{1/\theta}}{(1 - p_i')^{1/\theta}} \right), \quad t_i^* = \log \left( \frac{2 - (1 - p_i'')^{1/\theta}}{(1 - p_i'')^{1/\theta}} \right).$$

Table 1 shows the values of  $\alpha_i, \beta_i$  for various  $\theta$  and  $n$ . The MMLE of  $\sigma$  can be shown to be equivalent to the exact MLE with respect to the asymptotic variance. Their performance in small samples is also studied through simulation because the exact MLE is an iterative solution. The empirical sample characteristics are given in Table 2, which indicates the following:

1. The empirical sample characteristics bias, variance and MSE decrease as sample size increases.
2. MMLE is generally more biased than MLE; with reference to variance as well as MSE, MMLE is better than MLE for small samples.

**Table 1.** Intercept and Slope of the Approximation  $G(Z_i) = \alpha_i + \beta_i z_i$  (GHLD –II)

<i>n</i>	<i>i</i>	$\theta = 2$		$\theta = 3$		$\theta = 4$	
		$\alpha_i$	$\beta_i$	$\alpha_i$	$\beta_i$	$\alpha_i$	$\beta_i$
5	1	0.0000	0.7752	0.0000	1.2528	0.0000	1.7410
	2	-0.0612	1.0780	-0.0391	1.5442	-0.0286	2.0251
	3	-0.2071	1.4122	-0.1407	1.8906	-0.1060	2.3751
	4	-0.4780	1.7832	-0.3573	2.3293	-0.2823	2.8464
	5	-0.4013	1.5789	-0.3942	2.2353	-0.3891	2.8947
10	1	0.0000	0.6432	0.0000	1.1294	0.0000	1.6223
	2	-0.0150	0.7934	-0.0092	1.2679	-0.0066	1.7545
	3	-0.0477	0.9512	-0.0299	1.4170	-0.0216	1.8985
	4	-0.1010	1.1170	-0.0650	1.5785	-0.0476	2.0569
	5	-0.1789	1.2912	-0.1187	1.7552	-0.0883	2.2334
	6	-0.2866	1.4743	-0.1974	1.9509	-0.1496	2.4334
	7	-0.4310	1.6668	-0.3111	2.1713	-0.2413	2.6662
	8	-0.6219	1.8683	-0.4780	2.4268	-0.3826	2.9484
	9	-0.8720	2.0755	-0.7370	2.7393	-0.6201	3.3206
	10	-0.5324	1.7681	-0.5479	2.4572	-0.5547	3.1336
15	1	0.0000	0.5960	0.0000	1.0870	0.0000	1.5020
	2	-0.0066	0.6969	-0.0040	1.1781	-0.0028	1.6684
	3	-0.0207	0.8003	-0.0127	1.2736	-0.0091	1.7596
	4	-0.0430	0.9072	-0.0268	1.3739	-0.0193	1.8563
	5	-0.0744	1.0176	-0.0472	1.4797	-0.0343	1.9591
	6	-0.1161	1.1318	-0.0749	1.5917	0.0550	2.0691
	7	-0.1698	1.2498	-0.1116	1.7106	0.0827	2.1875
	8	-0.2357	1.3719	-0.1589	1.8377	-0.1191	2.3158
	9	-0.3171	1.4981	-0.2196	1.9744	-0.1668	2.4564
	10	-0.4159	1.6286	-0.2972	2.1226	-0.2292	2.6121
	11	-0.5350	1.7632	-0.3970	2.2852	-0.3120	2.7877
	12	-0.6781	1.9015	-0.5274	2.4665	-0.4245	2.9907
	13	-0.8492	2.0417	-0.7032	2.6737	-0.5843	3.2348
	14	-1.0496	2.1780	-0.9559	2.9223	-0.8344	3.5533
	15	-0.5925	1.8548	-0.6260	2.5684	-0.6430	3.2610

# ESTIMATION AND TESTING IN TYPE II GHLD

**Table 1, continued**

<i>n</i>	<i>i</i>	$\theta = 2$		$\theta = 3$		$\theta = 4$	
		$\alpha_i$	$\beta_i$	$\alpha_i$	$\beta_i$	$\alpha_i$	$\beta_i$
20	1	0.0000	0.5732	0.0000	1.0656	0.0000	1.5617
	2	0.0037	0.6482	-0.0022	1.1334	0.0016	1.6259
	3	0.0115	0.7251	-0.0070	1.2037	0.0050	1.6927
	4	-0.0237	0.8039	-0.0146	1.2766	-0.0104	1.7623
	5	-0.0406	0.8847	-0.0252	1.3522	-0.0182	1.8351
	6	-0.0627	0.9675	-0.0394	1.4309	-0.0285	1.9112
	7	-0.0903	1.0525	-0.0575	1.5129	-0.0419	1.9913
	8	-0.1241	1.1395	-0.0802	1.5986	0.0588	2.0755
	9	-0.1645	1.2288	-0.1080	1.6883	-0.0798	2.1647
	10	-0.2123	1.3204	-0.1417	1.7825	-0.1057	2.2593
	11	-0.2683	1.4143	-0.1825	1.8818	-0.1373	2.3603
	12	-0.3332	1.5107	-0.2314	1.9869	-0.1760	2.4686
	13	-0.4084	1.6094	-0.2904	2.0987	-0.2234	2.5859
	14	-0.4948	1.7105	-0.3615	2.2183	-0.2819	2.7138
	15	-0.5942	1.8139	-0.4479	2.3474	-0.3548	2.8551
	16	-0.7081	1.9192	-0.5543	2.4881	-0.4475	3.1039
	17	-0.8383	2.0254	-0.6878	2.6439	-0.5687	3.1967
	18	-0.9857	2.1303	-0.8609	2.8203	-0.7349	3.4153
	19	-1.1453	2.2265	-1.0994	3.0286	-0.9857	3.6982
	20	-0.6282	1.9063	-0.6757	2.6416	-0.7012	3.3450

**Table 2.** Empirical Sample Characteristics (Type-II GHLD)

$\theta$	<i>n</i>	Bias		Variance		MSE	
		MLE	MMLE	MLE	MMLE	MLE	MMLE
2	5	0.1077	0.0910	0.0651	0.0121	0.0766	0.0203
	10	0.0551	0.0659	0.0320	0.0079	0.0350	0.0122
	15	0.0364	0.0522	0.0206	0.0060	0.0219	0.0087
	20	0.0273	0.0427	0.0153	0.0048	0.0160	0.0066
3	5	0.1064	0.0977	0.0643	0.0125	0.0756	0.0220
	10	0.0549	0.0667	0.0320	0.0081	0.0350	0.0125
	15	0.0364	0.0530	0.0207	0.0061	0.0220	0.0089
	20	0.0274	0.0435	0.0154	0.0049	0.0161	0.0067
4	5	0.1055	0.0926	0.0636	0.0131	0.0747	0.0216
	10	0.0546	0.0676	0.0318	0.0085	0.0347	0.0130
	15	0.0362	0.0538	0.0206	0.0064	0.0219	0.0092
	20	0.0273	0.0443	0.0153	0.0051	0.0160	0.0070

## GHLD-II vs. Exponential Model

The discrimination between GHLD-II and the exponential model is made using the likelihood ratio (LR) criterion. Specify GHLD-II as null population ( $P_0$ ) and the exponential model as alternative population ( $P_1$ ). A null hypothesis is proposed as  $H_0$ : a given sample belongs to GHLD-II ( $P_0$ ) versus an alternative hypothesis  $H_1$ : the sample belongs to the population Exponential model ( $P_1$ ). Let  $L_1$ ,  $L_0$ , respectively, stand for the likelihood function of a sample with population  $P_1$  and  $P_0$ . The percentiles of the LR criterion  $L_1/L_0$  are obtained by simulation as:

10,000 random samples of sizes  $n = 5, 10, 15, 20$  are generated from the null population  $P_0$  and its parameters are estimated using each sample. The value of the likelihood function of the null population is computed at the generated sample observations and the corresponding parameter estimates; this value is denoted by  $L_0$ . Using the same sample, generated from  $P_0$ , the parameters and likelihood function value of the alternative population are calculated, for example,  $L_1$ . The values of  $L_1/L_0$  over 10,000 runs are sorted and selected percentiles are identified for a given  $n, \theta$  (see Table 3).

**Table 3.** Percentiles of  $L_1/L_0$  ( $P_0$ : GHLD-II,  $P_1$ : Exponential)

$\theta$	$n \setminus p$	0.00135	0.01	0.025	500	0.95	0.975	0.99865
2	5	0.7468	0.9743	1.3335	1.7250	2.5433	2.6067	4.4061
	10	0.4786	0.7651	1.2327	1.6918	4.7663	4.8496	6.6528
	15	0.3369	0.7567	1.1770	1.6473	6.0976	7.4550	8.6546
	20	0.2520	0.7344	1.0456	1.5327	8.9127	8.9845	10.7528
3	5	1.6877	1.9325	2.2646	2.6432	3.4623	4.6379	20.6042
	10	2.5396	3.0615	4.0111	5.0243	8.7750	9.0357	39.9667
	15	3.1753	5.4376	7.9391	9.8175	18.5364	18.7628	50.6341
	20	3.9089	9.7390	14.7436	19.4296	54.4206	69.0316	80.0497
4	5	3.6630	4.5150	5.1328	4.0879	18.2493	18.4501	73.5894
	10	10.1778	12.7498	16.3453	12.4361	30.6046	31.1481	81.5585

The entries under the column headings 0.95 in Table 3 may be taken as 5% level of significance critical values for discriminating between the GHLD-II and exponential models. The powers of the test statistic  $L_1/L_0$  are also evaluated through simulation by calculating  $L_1/L_0$  with samples generated from exponential

## ESTIMATION AND TESTING IN TYPE II GHLD

population ( $P_1$ ) and estimating, the parameters calculating the values of the likelihood functions  $L_1, L_0$  with sample from  $P_1$ . The proportion of  $L_1/L_0$  values falling above 95<sup>th</sup> percentile of  $L_1/L_0$  would become the power of the LR test criterion (see Table 4). It is observed that the discrimination between GHLD-II and exponential models falls with increased sample size, indicating less distinguishability between the exponential model and GHLD-II.

**Table 4.** Powers of LR Test Criterion at  $\alpha = 0.05$

$\theta$	$n \setminus$ Distributions	GHLD-II vs. Exponential
2	5	0.9123
	10	0.9239
	15	0.9373
	20	0.9441
3	5	0.9135
	10	0.9159
	15	0.9176
	20	0.9161
4	5	0.9072
	10	0.9057
	15	0.9053
	20	0.9025

## References

- Balakrishnan, N. (1985). Order statistics from the half logistic distribution. *Journal of Statistical Computation and Simulation*, 20: 287-309.
- Balakrishnan, N., & Leung, M. Y. (1988). Order statistics from type-I generalized logistic distribution. *Communication Statistics – Simulation and Computing*, 17(1): 25-50.
- Balakrishnan, N., & Sandhu, R. A. (1995). Recurrence relations for single and product moments of order Statistics from a generalized half logistic distribution with applications to inference. *Journal of Statistical Computation and Simulation*, 52(4): 385-398.
- Gupta, R. D., & Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, 41: 173-188.

- Kantam, R. R. L., & Srinivasa Rao, G. (1993). Reliability estimation in Rayleigh distribution with censoring some approximations to ML Method. *Proceedings of II Annual Conference of Society for Development of Statistics, Acharya Nagarjuna University*: 56-63.
- Kantam, R. R. L., & Srinivasa Rao, G. (2002). Log-logistic distribution: Modified Maximum likelihood estimation. *Gujarat Statistical Review*, 29(1-2): 25-36.
- Kantam, R. R. L., and Sriram, B. (2003). Maximum likelihood estimation from censored samples: Some modifications in length biased version of exponential model. *Statistical methods*, 5(1): 63-78.
- Mehrotra, K. G., & Nanda, P. (1974). Unbiased estimation of parameters by order statistics in the case of censored samples. *Biometrika*, 61: 601-606.
- Mudholkar, G. S., Srivastava, D., & Freimer, M. (1995). Exponentiated Weibull family: A reanalysis of the bus-motor failure data. *Technometrics*, 37(4): 436-445.
- Olapade, A. K. (2008). On Type III Generalized Half Logistic Distribution. [arXiv:0806.1580v1 \[math.ST\]](https://arxiv.org/abs/0806.1580v1) 10 Jun 2008.
- Pearson, T., & Rootzen, H. (1977). Simple highly efficient estimators for a type-I Censored Normal sample, *Biometrika*, 64: 123-128.
- Rajarshi, S., & Rajarshi, M. B. (1988). Bathtub distributions: A review. *Communication in Statistics –Theory & Methods*, 17: 2597-2621.
- Rosaiah, K., Kantam, R. R. L., & Narasimham, V. L. (1993a). ML and Modified ML Estimation in gamma distribution with known prior relation among the Parameters. *Pakistan Journal of Statistics*, 9(3)B: 37-48.
- Rosaiah, K., Kantam, R. R. L., & Narasimham, V. L. (1993b). On modified maximum likelihood estimation of gamma parameters. *Journal of Statistical Research*, 27(1-2): 15-28.
- Tiku, M. L. (1967). Estimating the mean and standard deviation from a censored Normal sample. *Biometrika*, 54: 155-165.
- Tiku, M. L., & Suresh, R. P. (1992). A new method of estimation for location and scale parameters. *Journal of Statistical Planning and Inference*, 30: 281-92.

# A Compound of Geeta Distribution with Generalized Beta Distribution

**Adil Rashid**

University of Kashmir  
Srinagar, India

**T. R. Jan**

University of Kashmir  
Srinagar, India

---

A compound of Geeta distribution with Generalized Beta distribution (GBD) is obtained and the compound is specialized for different values of  $\beta$ . The first order factorial moments of some special compound distributions are also obtained. A chronological overview of recent developments in the compounding of distributions is provided in the introduction.

*Keywords:* Compound distribution, Geeta distribution, Generalized Beta Distribution (GBD), factorial moments

---

## Introduction

Regarding the problem of the compounding of the probability distributions, work has been conducted in this area since 1920. It is well known that the parameter in a Poisson distribution is considered to be a gamma variate in the famous article by Greenwood and Yule (1920). Skellam (1948) derived a probability distribution from the binomial distribution by regarding the probability of success as a beta variable between sets of trials. The interrelationships among compound and generalized distributions were first explored by Gurland (1957) after which, Molenaar (1965) discussed some important remarks on mixtures of distributions.

Dubey (1970) derived compound gamma, beta and F distributions by compounding a gamma distribution with another gamma distribution and reducing it to the beta 1<sup>st</sup> and 2<sup>nd</sup> kind and to the F distribution via suitable transformations. The application of compounding of distributions to calculate moments was explored by Dyczka (1973). The problem of compounding of distributions was further addressed by Gerstenkorn (1993, 1996) who proposed several compound distributions; Gerstenkorn obtained a compound of gamma

---

*Adil Rashid is a PhD Scholar in the Department of Statistics. Email at: [adilstat@gmail.com](mailto:adilstat@gmail.com). T. R. Jan is a Faculty member in the Department of Statistics. Email at: [drtrjan@gmail.com](mailto:drtrjan@gmail.com).*



distribution with exponential distribution by treating the parameter of a gamma distribution as an exponential variate and obtained a compound of polya with beta. Gerstenkorn (2004) also found a compound of a generalized negative binomial distribution with generalized beta distribution by treating the parameter of generalized negative binomial distribution as a generalized beta distribution. Ali, Aslam and Kazmi (2011) improved the informative prior for the mixture of a Laplace distribution under different loss functions. Rashid and Jan (2013) recently obtained a compound of zero truncated generalized negative binomial distribution with that of generalized beta distribution. A broad range of relevant references can be found in studies by Johnson, Kotz and Kemp (1992).

## The compounding of probability distributions

The following definition and relations are needed for compounding probability distributions. A certain compound distribution arises when all (or some) parameters of a distribution vary according to some probability distribution, called the compounding distribution. Suppose  $X|y$  is a random variable with a distribution function  $F(x|y)$  that depends on parameter  $y$ . If parameter  $y$  is considered to be a random variable  $Y$  with distribution function  $G(y)$ , then the distribution that has the distribution function of  $X$  is defined by

$$H(x) = \int_{-\infty}^{\infty} F(x|cy) dG(y) \quad (1)$$

which is called compound, where  $c$  is an arbitrary constant or a constant bounded on some interval (Gurland, 1957).

The occurrence of the constant  $c$  in (1) has a practical justification inasmuch as the distribution of a random variable, in describing a phenomenon, often depends on a parameter that is itself a realization of another random variable multiplied by a certain constant. A variable that has distribution function (1) will be symbolized by  $X \wedge Y$  and will be called a compound of the variable  $X$  with respect to the compounding  $Y$ .

Relation (1) is symbolized as follows:

$$H(x) \equiv F(x|cy) \underset{Y}{\wedge} G(y) \quad (2)$$

Consider the case when one variable is discrete with probability function  $P(X = x_i | cy)$ , if parameter  $y$  is a random variable  $Y$  with density  $g(y)$ , then (1) is expressed by

$$h(x_i) = P(X = x_i) = \int_{-\infty}^{\infty} g(y)P(X = x_i | cy)dy \quad (3)$$

### Compounding the Geeta Distribution with the Generalized Beta distribution

Suppose  $X$  is a discrete random variable defined over positive integers. The random variable  $X$  is said to have a Geeta distribution with parameters  $\theta$  and  $\beta$  if

$$P_{\beta}(x; \theta) = \begin{cases} \frac{1}{\beta x - 1} \binom{\beta x - 1}{x} \theta^{x-1} (1 - \theta)^{\beta x - x} & ; x = 1, 2, \dots \\ 0 & ; \text{otherwise} \end{cases} \quad (4)$$

where  $0 < \theta < 1$  and  $1 < \beta < \frac{1}{\theta}$ . The upper limit on  $\beta$  has been imposed for the existence of the mean. When  $\beta \rightarrow 1$ , the Geeta distribution degenerates and its probability mass is concentrated at point  $x = 1$  (Consul, 1990).

The Generalized Beta Distribution (GBD) is a distribution given by the density function

$$GB(y; a, b, w, r) = \begin{cases} \frac{ay^{r-1}}{(bw)^{r/a} B(r/a, w)} \left(1 - y^a/bw\right)^{w-1} & ; 0 < y < (bw)^{1/a} \\ 0 & ; y \leq 0 \text{ or } y \geq (bw)^{1/a} \end{cases} \quad (5)$$

where  $a, b, w, r > 0$  and  $B(r/a, w)$  is a beta function. Distribution (5) is a special limit case of the Bessel distribution (Srodka, 1973; Seweryn, 1986) that has been applied in reliability theory (Oginski, 1979).

Consider a Geeta distribution (4) that depends on  $cy$  :

$$P_{\beta}(x; cy) = \frac{1}{\beta x - 1} \binom{\beta x - 1}{x} (cy)^{x-1} (1 - cy)^{\beta x - x}, \quad x = 1, 2, 3, \dots \quad (6)$$

where  $0 < cy < 1$ ,  $1 < \beta < \frac{1}{\theta}$  and  $Y$  is a random variable with GBD (5).

### Theorem 1

The probability function of the compound of Geeta distribution with GBD is

$$P_{\beta GB}(x) = D_1 \sum_{K=0}^{\infty} \binom{\beta x - x}{k} (-c)^k (bw)^{k/a} B\left(\frac{x + r + k - 1}{a}, w\right) \quad (7)$$

$$\text{where } D_1 = \frac{\frac{1}{\beta x - 1} \binom{\beta x - 1}{x} c^{x-1} (bw)^{\frac{x-1}{a}}}{B(r/a, w)}$$

and  $x = 1, 2, 3, \dots, a, b, w, r > 0$ ,  $0 < cy < 1$  and  $\beta cy < 1$

**Proof:** From (3), (5) and (6)

$$\begin{aligned} P_{\beta GB}(x) &= a D_1^* \int_0^{(bw)^{1/a}} y^{x+r-2} \left(1 - \frac{y^a}{bw}\right)^{w-1} (1 - cy)^{\beta x - x} dy \\ &= a D_1^* \sum_{K=0}^{\infty} (-c)^K \binom{\beta x - x}{k} \int_0^{(bw)^{1/a}} y^{x+r+k-2} \left(1 - \frac{y^a}{bw}\right)^{w-1} dy \end{aligned}$$

$$\text{where } D_1^* = \frac{\frac{1}{\beta x - 1} \binom{\beta x - 1}{x} c^{x-1}}{(bw)^{r/a} B(r/a, w)}.$$

Substituting,  $\frac{y^a}{bw} = t$ , results in

$$P_{\beta}GB(x) = D_1 \sum_{k=0}^{\infty} (-c)^k \binom{\beta x - x}{k} (bw)^{k/a} \int_0^1 t^{\frac{x+r+k-1}{a}-1} (1-t)^{w-1} dt \quad (8)$$

where ,  $x=1,2,3,\dots,a,b$ ,  $w,r>0$ ,  $\beta \geq 1$  and  $0 < c \leq \frac{1}{\beta(bw)^{\frac{1}{a}}}$ .

Using the definition of beta function, (7) is obtained.

### Special Cases

**Case I:** When  $\beta = 2$  in (4), Haight's distribution results and a compound of the Haight distribution with generalized beta follows from (8):

$$P_2GB(x) = D_2^* \sum_{k=0}^{\infty} (-c)^k \binom{x}{k} (bw)^{k/a} B\left(\frac{x+r+k-1}{a}, w\right) \quad (9)$$

where  $D_2^* = \frac{1}{2x-1} \binom{2x-1}{x} c^{x-1} (bw)^{\frac{x-1}{a}}$ ;  $x=1,2,3,\dots,a,b$ ,  $w,r>0$ ,  $0 < cy < 1$ .

**Case II:** If  $b = 1/w$  and  $a = 1$ , in (5), the a beta distribution and a compound of Geeta distribution with beta distribution follow from (8):

$$P_{\beta}B(x) = D_3^* \sum_{k=0}^{\infty} (-c)^k \binom{\beta x - x}{k} B(x+r+k-1, w) \quad (10)$$

where  $D_3^* = \frac{1}{\beta x - 1} \binom{\beta x - 1}{x} c^{x-1}$ .

**Case III:** When  $\beta = 2$  and  $b = 1/w$ ,  $a = 1$  in (4) and (5), respectively obtained are the Haight and beta distributions and a compound of the Haight distribution with beta distribution follows from (8):

$$P_2 B(x) = D_4^* \sum_{k=0}^x (-c)^k \binom{x}{k} B(x+r+k-1, w) \quad (11)$$

$$\text{where } D_4^* = \frac{\frac{1}{2x-1} \binom{2x-1}{x} c^{x-1}}{B(r, w)}$$

### Factorial moments of the Compound of Geeta distribution with Generalized Beta distribution and some special cases

Let  $X_y$  and  $X$  be a random variable with distribution function  $F(x|y)$  and  $H(X)$ , respectively (see (1)), and let parameter  $y$  have distribution  $G(y)$ . Keeping in mind the formula for the so-called factorial polynomial

$$x^{[l]} = x(x-1)(x-2)\dots(x-(l-1))$$

$$m_{[l]} = E(X^{[l]}) = \int_{-\infty}^{+\infty} E(X_y^{[l]}) dG(y) \quad (12)$$

is called a factorial moment of order  $l$  of the variable  $X$  with compound distribution (1).

Relation (12) is symbolized as

$$E(X_y^{[l]})_y^{\wedge} G(y). \quad (13)$$

### Theorem 2

The first order factorial moments of the compound of Geeta distribution with GBD is given by

$$m_{[1]}(\beta; cy)_y^{\wedge} GB(y; a, b, r, w) = \frac{\sum_{k=0}^{\infty} (\beta c)^k (bw)^{k/a}}{B(r/a, w)} \left( B\left(\frac{r+k}{a}, w\right) - c(bw)^{\frac{1}{a}} B\left(\frac{r+k+1}{a}, w\right) \right) \quad (14)$$

**Proof:** The first order factorial moments of the Geeta distribution is given by

$$m_{[1]}(\beta, \theta) = \frac{1-\theta}{1-\beta\theta}.$$

Thus, from (13), the 1<sup>st</sup> order factorial moment of the compound of the Geeta distribution with a Generalized beta distribution if  $\theta = cy$  is

$$\begin{aligned} m_{[1]}(\beta; cy) \wedge_y GB(y, a, b, r, w) &= \frac{a}{(bw)^{r/a} B(r/a, w)} \int_0^{(bw)^{1/a}} \left( \frac{1-cy}{1-\beta cy} \right) y^{r-1} \left( 1 - \frac{y^a}{bw} \right)^{w-1} dy \\ &= \frac{a \sum_{k=0}^{\infty} (\beta c)^k}{(bw)^{r/a} B(r/a, w)} \left( \int_0^{(bw)^{1/a}} y^{r+k-1} \left( 1 - \frac{y^a}{bw} \right)^{w-1} dy - c \int_0^{(bw)^{1/a}} y^{r+k} \left( 1 - \frac{y^a}{bw} \right)^{w-1} dy \right). \end{aligned}$$

Substituting,  $\frac{y^a}{bw} = t$  results in

$$\begin{aligned} &= \frac{\sum_{k=0}^{\infty} (\beta c)^k}{B(r/a, w)} \left( (bw)^{k/a} \int_0^1 t^{\frac{r+k}{a}-1} (1-t)^{w-1} dt - c (bw)^{\frac{k+1}{a}} \int_0^1 t^{\frac{r+k+1}{a}-1} (1-t)^{w-1} dt \right) \\ &= \frac{\sum_{k=0}^{\infty} (\beta c)^k}{B(r/a, w)} (bw)^{k/a} \left( \int_0^1 t^{\frac{r+k}{a}-1} (1-t)^{w-1} dt - c (bw)^{\frac{1}{a}} \int_0^1 t^{\frac{r+k+1}{a}-1} (1-t)^{w-1} dt \right) \end{aligned} \quad (15)$$

Using the definition of beta function, (14) is obtained.

### Special Case

When  $b = 1/w$  and  $a = 1$  in (15), the 1<sup>st</sup> order factorial moment of the compound of Geeta distribution with beta distribution is obtained as

$$m_{[1]}(\beta; cy) \wedge_y B(y; 1, 1/w, r, w) = \frac{\sum_{k=0}^{\infty} (\beta c)^k}{B(r, w)} (B(r+k, w) - cB(r+k+1, w)).$$

### Theorem 3

First order factorial moments of the compound of Haight with generalized beta distribution.

$$m_{[1]}(2; cy) \wedge_y GB(y; a, b, r, w) = \frac{\sum_{k=0}^{\infty} (2c)^k (bw)^{k/a}}{B(r/a, w)} \left( B\left(\frac{r+k}{a}, w\right) - c(bw)^{1/a} B\left(\frac{r+k+1}{a}, w\right) \right). \quad (16)$$

**Proof:** The result follows directly from (15) for  $\beta = 2$ ,

$$m_{[1]}(2; cy) \wedge_y GB(y; a, b, r, w) = \frac{\sum_{k=0}^{\infty} (2c)^k (bw)^{k/a}}{B(r/a, w)} \left( \int_0^1 t^{\frac{r+k}{a}-1} (1-t)^{w-1} dt - c(bw)^{1/a} \int_0^1 t^{\frac{r+k+1}{a}-1} (1-t)^{w-1} dt \right) \quad (17)$$

which yields (16).

### Special case

When  $b=1/w$ ,  $a=1$  in (17) the following result is obtained:

$$m_{[1]}(2; cy) \wedge_y B(y; 1, 1/w, r, w) = \frac{\sum_{k=0}^{\infty} (2c)^k}{B(r, w)} (B(r+k, w) - cB(r+k+1, w))$$

which gives the 1<sup>st</sup> order factorial moment of the compound of the Haight distribution with beta distribution.

### References

- Ali, S., Aslam, M., & Kazmi, S. M. (2011). Improved informative prior for the mixture of LaPlace distribution under different loss functions. *Journal of Reliability and Statistical Studies*, 4(2): 57-82.
- Consul, P. C. (1990). Geeta distribution and its properties. *Communications in Statistics—Theory and Methods*, 19: 3051-3068.

## COMPOUND GEETA AND GENERALIZED BETA DISTRIBUTION

- Dubey, D. S. (1970). Compound gamma, beta and F distributions. *Metrika*, 16(1): 27-31.
- Dyczka, W. (1973). Application of compounding of distribution to determination of moments in Polish. *Matematyka*, 3: 205-230.
- Gerstenkorn, T. (1993). A compound of the generalized gamma distribution with the exponential one. *Recherches sur les deformations*, 16(1): 5-10.
- Gerstenkorn, T. (1996). A compound of the Polya distribution with the beta one. *Random Operators and Stochastic Equations*, 4(2): 103-110.
- Gerstenkorn, T. (2004). A compound of the generalized negative binomial distribution with the generalized beta one. *Central European Journal of Mathematics*, 2(4): 527-537.
- Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 83: 255-279.
- Gurland, J. (1957). Some interrelations among compound and generalized distributions. *Biometrika*, 44: 265-268.
- Johnson, N. L., and Kotz, S. (1969). *Discrete distributions* (First Edition). Boston, MA: Houghton Mifflin.
- Molenaar, W. (1965). Some remarks on mixtures of distributions. In *Proceedings of the 35<sup>th</sup> Session of the International Statistical Institute, Belgrade* (*Bulletin de l'Institut international de statistique*, 41), pp. 764-765.
- Oginski, L. (1979). Application of a distribution of the Bessel type in reliability theory. *Matematyka*, 12: 31-42, (in Polish).
- Rashid, A., & Jan, T. R. (2013). A compound of zero truncated generalized negative binomial distribution with the generalized beta distribution. *Journal of Reliability and Statistical Studies*, 6(1): 11-19.
- Seweryn J. G. (1986). Some probabilistic properties of Bessel distribution. *Matematyka*, 19: 69-87.
- Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, 10: 257-261.
- Srodka, T. (1973). On some generalized Bessel- type probability distribution. *Matematyka*, 4: 5-31.



# Inference for the Rayleigh Distribution Based on Progressive Type-II Fuzzy Censored Data

**Abbas Pak**

Shahid Chamran University  
Ahvaz, Iran

**Gholam Ali Parham**

Shahid Chamran University  
Ahvaz, Iran

**Mansour Saraj**

Shahid Chamran University  
Ahvaz, Iran

---

Classical statistical analysis of the Rayleigh distribution deals with precise information. However, in real world situations, experimental performance results cannot always be recorded or measured precisely, but each observable event may only be identified with a fuzzy subset of the sample space. Therefore, the conventional procedures used for estimating the Rayleigh distribution parameter will need to be adapted to the new situation. This article discusses different estimation methods for the parameters of the Rayleigh distribution on the basis of a progressively type-II censoring scheme when the available observations are described by means of fuzzy information. They include the maximum likelihood estimation, highest posterior density estimation and method of moments. The estimation procedures are discussed in detail and compared via Monte Carlo simulations in terms of their average biases and mean squared errors. Finally, one real data set is analyzed for illustrative purposes.

*Keywords:* Progressive type-II censoring, fuzzy information, maximum likelihood principle, highest posterior density estimation

---

## Introduction

The Rayleigh distribution was originally introduced by Lord Rayleigh (1880) in the field of acoustics; since its introduction, many researchers have used the distribution in different fields of science and technology. The Rayleigh distribution is frequently used to model wave heights in oceanography, in communication engineering and it also has a wide application in lifetime data analysis, especially in reliability theory and survival analysis. An important characteristic of the Rayleigh distribution is that its hazard rate is a linearly increasing function of time at constant rate, which makes it a suitable model for

---

*Abbas Pak is in the Department of Statistics. Email at: [a-pak@scu.ac.ir](mailto:a-pak@scu.ac.ir). Gholam Ali Parham is in the Department of Mathematics. Email at: [parham\\_g@scu.ac.ir](mailto:parham_g@scu.ac.ir). Mansour Saraj is in the Department of Mathematics. Email at: [seraj.a@scu.ac.ir](mailto:seraj.a@scu.ac.ir).*

## INFERENCE FOR THE RAYLEIGH DISTRIBUTION

the lifetime of components/items that age rapidly with time. Thus, as time increases, the reliability function of the Rayleigh distribution decreases at a much higher rate than the exponential reliability function does. The probability density function (pdf) and the cumulative distribution function (cdf) of a Rayleigh random variable  $X$  can be written as:

$$f(x) = 2\lambda x e^{-\lambda x^2}; \quad x > 0, \lambda > 0, \quad (1)$$

and

$$F(x) = 1 - e^{-\lambda x^2}; \quad x > 0, \quad (2)$$

respectively. Inferences for the Rayleigh distribution have been discussed by several authors. Dyer and Whisenand (1973) demonstrated the importance of this distribution in communication engineering. Bhattacharya and Tyagi (1990) mentioned that in some clinical studies dealing with cancer patients, the survival pattern follows the Rayleigh distribution. Chung (1995) obtained the best invariant estimator and the Bayes estimator of the parameter of Rayleigh distribution under entropy loss. Fernandez (2010) addressed the problems of estimating the parameter, hazard rate and reliability function of the Rayleigh distribution on the basis of sample quantiles. Dey and Maiti (2012) derived Bayes estimator of the Rayleigh parameter and its associated risk based on extended Jeffrey's prior.

In many life testing and reliability experiments, a sample of  $n$  items is tested, and the experiment is terminated when all of them fail. This procedure may take a long time when the lifetime distribution of items has a thick tail. Moreover, if the items are expensive, such as medical equipment, it is costly to gather information from the whole sample. There are many situations where experimental units are lost or removed from the test before complete failure. For example, individuals in a clinical trial may drop out of the study, the study may have to be terminated early for lack of funds or the test units may accidentally break. In other scenarios, the experiment may have to be terminated in order to free up testing facilities for other purposes.

In view of above, censoring is used in life testing to save time and cost of testing units. The removal of units in a test may be unintentional or pre-planned. Data obtained from such experiments are called censored sample. There are many types of censoring schemes used in lifetime analysis. The two most common

censoring schemes are termed type-I and type-II censoring schemes. In the conventional type-I censoring scheme, the experiment continues up to a pre-specified time  $T$ ; the conventional type-II censoring scheme requires the experiment to continue until a pre-specified number of failures occur. These schemes, however, do not allow removal of units before the termination of the experiment; thus, a more general kind of censoring scheme called progressive type-II censoring is considered, which is as follows: Suppose that  $n$  units are placed on a life test and the experimenter decides beforehand a quantity  $m$ , the number of units to be failed. Now at the time of the first failure,  $R_1$  of the remaining  $n - 1$  surviving units are randomly removed from the experiment. Continuing on, at the time of the second failure,  $R_2$  of the remaining  $n - R_1 - 2$  units are randomly removed from the experiment. Finally, at the time of the  $m^{\text{th}}$  failure, all the remaining  $n - m - R_1 - \dots - R_{m-1}(=R_m)$  surviving units are removed from the experiment. The work on progressive censoring has become popular in life-testing and reliability studies. Kim and Han (2009) studied the problem of estimating the scale parameter of the Rayleigh distribution under general progressive censoring. Krishna and Kumar (2011) discussed reliability estimation for the Lindley distribution with progressive type-II censored data. Lee et al. (2011) obtained a Bayes estimator under the Rayleigh distribution with a progressive type-II right censored sample. Raqab and Madi (2011) addressed inference for the generalized Rayleigh distribution based on progressively censored data. Azimi et al. (2012) considered the Bayesian estimation of the parameter and reliability function of Rayleigh distribution based on a progressively type-II censored sample. Rastogi and Tripathi (2012) studied parameter estimation of the Burr type XII distribution on the basis of a progressively type-II censored sample. Pradhan and Kundu (2009) considered the statistical inference of the unknown parameters of the generalized exponential distribution in presence of progressive censoring. A recent account on progressive censoring schemes can be obtained in the monograph by Balakrishnan and Aggarwala (2000) or in the excellent review article by Balakrishnan (2007).

The above referenced studies for estimating parameters of different lifetime distributions under progressive type-II censoring are limited to precise data. However, in real world situations, experiments do not provide exact information. For example, the reaction time of a person to a certain stimulus in a psychological experience cannot be exactly determined, but the psychologist is able to determine it by means of the following imprecise information, such as: The time of reaction is approximately 25 to 35 seconds. To deal with the lack of precision of the data, it is necessary to incorporate the fuzzy concept to statistical techniques. Recently,

Pak et al. (2013) proposed a new method to determine the maximum likelihood estimate of the scale parameter of a Rayleigh distribution under doubly type-II censored sample from fuzzy data. Further, in a life testing experiment, some test units may need to be removed at different stages in the study for various reasons. This would lead to progressive censoring. The purpose of this article is to develop the inferential procedures for the Rayleigh distribution under a progressive type-II censoring scheme when the available observations are reported by means of fuzzy information. The maximum likelihood estimate (MLE) of the parameter  $\lambda$  is obtained by using EM algorithm and the highest posterior density (HPD) estimate of the unknown parameter is computed. The estimation via method of moments is discussed, a Monte Carlo simulation study is presented, and a comparison of all estimation procedures developed and one real data set is analyzed for illustrative purposes.

First, the fundamental notation and basic definitions of fuzzy set theory used herein is reviewed. Consider an experiment characterized by a probability space  $S = (\Omega, F, P_\theta)$  where  $(\Omega, F)$  is a measurable space and  $P_\theta$  belongs to a specified family of probability measures  $\{P_\theta, \theta \in \Theta\}$  on  $(\Omega, F)$ . Assume that the observer cannot distinguish or transmit with exactness the outcome in the performance of  $S$ , but that rather the available observation may be described in terms of fuzzy information, which is defined as:

**Definition 1**

A fuzzy event  $\tilde{\omega}$  on  $\Omega$ , characterized by a Borel measurable membership function  $\mu_{\tilde{\omega}}(\omega)$  from  $\Omega$  to  $[0,1]$ , where  $\mu_{\tilde{\omega}}(\omega)$  represents the grade of membership of  $\omega$  to  $\tilde{\omega}$ , is called *fuzzy information* associated with the experiment  $\mathcal{S}$ .

The set consisting of all observable events from the experiment  $\mathcal{S}$  determines a fuzzy information system associated with it, which is defined as:

**Definition 2**

A *fuzzy information system* (f.i.s.)  $\tilde{\mathcal{S}}$  associated with the experiment  $\mathcal{S}$  is a fuzzy partition  $\{\tilde{\omega}_1, \dots, \tilde{\omega}_K\}$  of  $\mathcal{S}$ , i.e., a set of  $K$  fuzzy events on  $\mathcal{S}$  satisfying the orthogonality condition (see Tanaka et al., 1979):

$$\sum_{k=1}^K \mu_{\tilde{\omega}_k}(\omega) = 1,$$

where  $\mu_{\tilde{\omega}_k}$  denotes the membership function of  $\tilde{\omega}_k$ .

According to Zadeh (1968), given the experiment  $\mathcal{S} = (\Omega, F, P_\theta)$ ,  $\theta \in \Theta$ , and a f.i.s.  $\tilde{\mathcal{S}}$  associated with it, each probability measure  $P_\theta$  on  $(\Omega, F)$  induces a probability measure on  $\tilde{\mathcal{S}}$  defined as:

### Definition 3

The probability distribution on  $\tilde{\mathcal{S}}$  induced by  $P_\theta$  is the mapping  $P$  from  $\tilde{\mathcal{S}}$  to  $[0, 1]$  such that

$$P(\tilde{\omega}) = \int_{\Omega} \mu_{\tilde{\omega}}(\omega) dP_\theta(\omega), \quad \tilde{\omega} \in \tilde{\mathcal{S}}. \quad (3)$$

In particular, the conditional density of a continuous random variable  $Y$  with p.d.f.  $g(y)$  given the fuzzy event  $\tilde{\omega}$  can be defined as

$$g(y|\tilde{\omega}) = \frac{\mu_{\tilde{\omega}}(y)g(y)}{\int \mu_{\tilde{\omega}}(u)g(u)du}. \quad (4)$$

In order to model imprecise lifetimes, a generalization of real numbers is necessary. These lifetimes can be represented by fuzzy numbers. A fuzzy number is a subset, denoted by  $\tilde{x}$ , of the set of real numbers (denoted by  $\mathbb{R}$ ) and is characterized by the so called membership function  $\mu_{\tilde{x}}(\cdot)$ . Fuzzy numbers satisfy the constraints (Dubois and Prade, 1980):

1.  $\mu_{\tilde{x}} : \mathbb{R} \rightarrow [0, 1]$  is Borel-measurable;
2.  $\exists x_0 \in \mathbb{R} : \mu_{\tilde{x}}(x_0) = 1$ ; and

3. The so-called  $\lambda$ -cuts ( $0 < \lambda \leq 1$ ), defined as  $B_\lambda(\tilde{x}) = \{x \in \mathbb{R} : \mu_{\tilde{x}}(x) \geq \lambda\}$ , are all closed interval, i.e.,  $B_\lambda(\tilde{x}) = [a_\lambda, b_\lambda]$ ,  $\forall \lambda \in (0, 1]$ .

Among the various types of fuzzy numbers, the triangular and trapezoidal fuzzy numbers are most convenient and useful in describing fuzzy lifetime data. The triangular fuzzy number can be defined as  $\tilde{x} = (a, b, c)$  and its membership function is defined by the following expression:

$$\mu_{\tilde{x}}(x) = \begin{cases} \frac{x-a}{b-a} & a \leq x \leq b, \\ \frac{c-x}{c-b} & b \leq x \leq c, \\ 0 & \text{otherwise.} \end{cases}$$

The trapezoidal fuzzy number can be defined as  $\tilde{x} = (a, b, c, d)$  with the membership function:

$$\mu_{\tilde{x}}(x) = \begin{cases} \frac{x-a}{b-a} & a \leq x \leq b, \\ 1 & b \leq x \leq c, \\ \frac{d-x}{d-c} & c \leq x \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

## Maximum likelihood estimation

Suppose that  $n$  identical units are put on a life testing experiment and that the lifetime distribution of each unit is given by (1). Prior to the experiment, a number  $m < n$  is determined and the censoring scheme  $(R_1, \dots, R_m)$  with  $R_i \geq 0$  and  $\sum_{i=1}^m R_i + m = n$  is specified. Let  $\mathbf{x} = (x_1, \dots, x_m)$  denote the corresponding progressively type-II censored sample. The likelihood function for the parameter  $\lambda$  becomes proportional to

$$L(\mathbf{x}; \lambda) = \lambda^m e^{-\lambda \sum_{i=1}^m (1+R_i) x_i^2} \quad (5)$$

Now consider the problem where  $\mathbf{x}$  is not observed precisely and only partial information about  $\mathbf{x}$  is available in the form of fuzzy numbers  $\tilde{x}_i = (a_i, c_i, b_i)$ ,  $i = 1, \dots, m$ , with the corresponding membership functions  $\mu_{\tilde{x}_1}(x_1), \dots, \mu_{\tilde{x}_m}(x_m)$ . Let  $c_{(1)} \leq c_{(2)} \leq \dots \leq c_{(m)}$  denote the ordered values of the means of these fuzzy numbers. The lifetime of  $R_i$  surviving units, which are removed from the test after the  $i^{\text{th}}$  failure, can be encoded as fuzzy numbers  $\tilde{z}_{i1}, \dots, \tilde{z}_{iR_i}$  with the membership functions

$$\mu_{\tilde{z}_{ij}}(z) = \begin{cases} 0 & z \leq c_{(i)} \\ 1 & z > c_{(i)} \end{cases}, \quad j = 1, \dots, R_i.$$

The fuzzy data  $\tilde{\mathbf{w}} = (\tilde{x}_1, \dots, \tilde{x}_m, \tilde{z}_1, \dots, \tilde{z}_m)$  where  $\tilde{z}_i$  is a  $1 \times R_i$  vector with  $\tilde{z}_i = (\tilde{z}_{i1}, \dots, \tilde{z}_{iR_i})$  for  $i = 1, \dots, m$ , is thus the set of observed lifetimes. The corresponding observed data log-likelihood function can be obtained by using the expression (3) as follows:

$$\begin{aligned} L_o(\tilde{\mathbf{w}}; \lambda) &= \sum_{i=1}^m \log \int 2\lambda x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx + \sum_{i=1}^m \sum_{j=1}^{R_i} \log \int 2\lambda z e^{-\lambda z^2} \mu_{\tilde{z}_{ij}}(z) dz \\ &= m \log \lambda + \sum_{i=1}^m \log \int 2x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx - \lambda \sum_{i=1}^m R_i c_{(i)}^2. \end{aligned} \quad (6)$$

The maximum likelihood estimate of the parameter  $\lambda$  can be obtained by maximizing the log-likelihood  $L_o(\tilde{\mathbf{w}}; \lambda)$ . Equating the partial derivative of the log-likelihood (6) with respect to  $\lambda$  to zero, the resulting equation is:

$$\frac{\partial L_o(\tilde{\mathbf{w}}; \lambda)}{\partial \lambda} = \frac{m}{\lambda} - \sum_{i=1}^m \frac{\int x^3 e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx} - \sum_{i=1}^m R_i c_{(i)}^2 = 0. \quad (7)$$

Because there is no closed form of the solution to equation (7), an iterative numerical search can be used to obtain the MLE. The Expectation Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates and useful in a variety of incomplete-data

problems. Because the observed fuzzy data  $\tilde{\mathbf{w}}$  can be seen as an incomplete specification of a complete data, the EM algorithm is applicable to obtain the maximum likelihood estimate of the parameter. In the following, the fuzzy EM algorithm (see Denoeux, 2011) is used to determine the MLE of  $\lambda$ .

First, denote the lifetimes of the failed and censored units by  $\mathbf{X} = (X_1, \dots, X_m)$  and  $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_m)$ , respectively, where  $\mathbf{Z}_i$  is a  $1 \times R_i$  vector with  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iR_i})$ , for  $i = 1, \dots, m$ . The combination of  $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$  forms the complete lifetimes and the corresponding log-likelihood function is denoted by  $L_c(\mathbf{W}, \lambda)$ , then, ignoring the additive constant,

$$L_c(\mathbf{W}, \lambda) = n \log \lambda - \lambda \left[ \sum_{i=1}^m x_i^2 + \sum_{i=1}^m \sum_{j=1}^{R_i} z_{ij}^2 \right]. \quad (8)$$

For the E-step, it is necessary to compute the pseudo log-likelihood function. It can be obtained from (8) as follows:

$$n \log \lambda - \lambda \left[ \sum_{i=1}^m E_\lambda(X_i^2 | \tilde{x}_i) + \sum_{i=1}^m \sum_{j=1}^{R_i} E_\lambda(Z_{ij}^2 | \tilde{z}_{ij}) \right] \quad (9)$$

By using (4), the conditional expectations  $E_\lambda(X_i^2 | \tilde{x}_i)$  and  $E_\lambda(Z_{ij}^2 | \tilde{z}_{ij})$  can be computed as:

$$E_\lambda(X_i^2 | \tilde{x}_i) = \frac{\int x^3 e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}, \quad i = 1, \dots, m,$$

$$E_\lambda(Z_{ij}^2 | \tilde{z}_{ij}) = c_{(i)}^2 + \frac{1}{\lambda}, \quad i = 1, \dots, m, \quad j = 1, \dots, R_i.$$

Next, the M-step involves the maximization of the pseudo function (9). Therefore, if at the  $h^{\text{th}}$  stage, the estimate of  $\lambda$  is  $\lambda^{(h)}$ , then  $\lambda^{(h+1)}$  can be obtained by maximizing

$$L_c^*(\mathbf{W}, \lambda) = n \log \lambda - \lambda \left[ \sum_{i=1}^m E_{\lambda^{(h)}}(X_i^2 | \tilde{x}_i) + \sum_{i=1}^m \sum_{j=1}^{R_i} E_{\lambda^{(h)}}(Z_{ij}^2 | \tilde{z}_{ij}) \right] \quad (10)$$



with respect to  $\lambda$ . From

$$\frac{\partial}{\partial \lambda} L_c^*(\mathbf{W}, \lambda) = 0, \quad (11)$$

$$\hat{\lambda}^{(h+1)} = \frac{n}{\sum_{i=1}^m [E_{\lambda^{(h)}}(X^2 | \tilde{x}_i) + R_i(c_{(i)}^2 + 1 / \lambda^{(h)})]}$$

The iteration process continues until convergence, i.e., until  $|L_O(\tilde{\mathbf{w}}; \lambda^{(h+1)}) - L_O(\tilde{\mathbf{w}}; \lambda^{(h)})| < \varepsilon$  for some pre-fixed  $\varepsilon > 0$ .

### HPD estimation

Consider the highest posterior density (HPD) estimation of the Rayleigh parameter based on observed fuzzy sample  $\tilde{\mathbf{w}}$ . As a conjugate prior for  $\lambda$ , take the *Gamma*( $a, b$ ) density with pdf given by

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-\lambda b}, \quad \lambda > 0, \quad (12)$$

where  $a > 0$  and  $b > 0$ . Based on this prior, the posterior density function of  $\lambda$  given the data can be written as follows:

$$\pi(\lambda | \tilde{\mathbf{w}}) \propto \lambda^{m+a-1} e^{-\lambda(b + \sum_{i=1}^m R_i c_{(i)}^2)} \prod_{i=1}^m \int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx \quad (13)$$

The method of HPD estimation then estimates  $\lambda$  as the mode of the posterior density  $\pi(\lambda | \tilde{\mathbf{w}})$ ; therefore, the HPD estimate of  $\lambda$  can be obtained by solving the equation

$$\frac{\partial \log \pi(\lambda | \tilde{\mathbf{w}})}{\partial \lambda} = 0 \quad (14)$$

where

$$\frac{\partial \log \pi(\lambda | \tilde{\mathbf{w}})}{\partial \lambda} = \frac{m+a-1}{\lambda} - b - \sum_{i=1}^m R_i c_{(i)}^2 - \sum_{i=1}^m \frac{\int x^3 e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}. \quad (15)$$

However, the solution cannot be obtained explicitly. [Theorem 1](#) discusses the existence and uniqueness of the HPD estimate of  $\lambda$ .

### Theorem 1

Let  $g(\lambda)$  denote the function on the right-hand side of the expression in (15). Then the root of the equation  $g(\lambda) = 0$  exists and is unique.

**Proof.** From (15), it is seen that  $\lim_{\lambda \rightarrow 0} g(\lambda) = \infty$ . Also, note that  $g(\lambda) < \frac{m+a-1}{\lambda}$ ,  $\forall \lambda \in (0, \infty)$ , and consequently

$$\lim_{\lambda \rightarrow 0} g(\lambda) < \lim_{\lambda \rightarrow 0} \frac{m+a-1}{\lambda} = 0, \quad \forall \lambda \in (0, \infty).$$

Thus, the equation  $g(\lambda) = 0$  has at least one root in  $(0, \infty)$ . To prove that the root is unique, consider the first derivative of  $g$ ,  $g'(\lambda)$  given by

$$g'(\lambda) = -\frac{m+a-1}{\lambda^2} + \sum_{i=1}^m \frac{\partial^2}{\partial \lambda^2} \log \int 2x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx \quad (16)$$

Because the integrand of the second term in (16) is a log-concave function of  $\lambda$ , and  $g'(\lambda) < 0$ . It follows that  $g$  is a strictly decreasing function w.r.t.  $\lambda$  and hence the equation  $g(\lambda) = 0$  has exactly one solution. The HPD estimate of  $\lambda$  must be derived numerically. In the following, the Newton-Raphson algorithm to determine the HPD estimate is described.

The Newton-Raphson algorithm is a direct approach for estimating the relevant parameters in a likelihood function. In this algorithm, the solution of the likelihood equation is obtained through an iterative procedure. Let  $\hat{\lambda}^{(h)}$  be the parameter value from the  $h^{\text{th}}$  step. Then, at the  $(h+1)^{\text{th}}$  step of iteration process, the updated parameter is obtained as

$$\hat{\lambda}^{(h+1)} = \hat{\lambda}^{(h)} - \frac{\frac{\partial \log \pi(\lambda|\tilde{\mathbf{w}})}{\partial \lambda} \Big|_h}{\frac{\partial^2 \log \pi(\lambda|\tilde{\mathbf{w}})}{\partial \lambda^2} \Big|_h} \quad (17)$$

where the notation  $A|_h$ , for any partial derivative  $A$ , means the partial derivative evaluated at  $\hat{\lambda}^{(h)}$ . The second-order derivative of  $\log \pi(\lambda|\tilde{\mathbf{w}})$  with respect to the parameter, required for proceeding with the Newton-Raphson method, is obtained as:

$$\frac{\partial^2 \log \pi(\lambda|\tilde{\mathbf{w}})}{\partial \lambda^2} = -\frac{m+a-1}{\lambda^2} + \sum_{i=1}^n \left\{ \frac{\int x^5 e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx} - \left[ \frac{\int x^3 e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda x^2} \mu_{\tilde{x}_i}(x) dx} \right]^2 \right\}. \quad (18)$$

The iteration process then continues until convergence, i.e., until  $|\hat{\lambda}^{(h+1)} - \hat{\lambda}^{(h)}| < \varepsilon$ , for some pre-fixed  $\varepsilon > 0$ .

## Method of moments

Let  $X$  be a random variable which has the Rayleigh distribution with pdf given by (1). It is known that the  $k^{\text{th}}$  moment of the Rayleigh model with parameter  $\lambda$  is

$$E(X^k) = \Gamma(1 + \frac{k}{2}) \lambda^{\frac{k}{2}}. \quad (19)$$

Equating the first sample moment to the corresponding population moment, the following equation can be used to find the estimate of moment method:

$$\lambda = \frac{\pi n^2}{4} \left\{ \sum_{i=1}^m E_{\lambda}(X|\tilde{x}_i) + \sum_{i=1}^m \sum_{j=1}^{R_i} E_{\lambda}(Z|\tilde{z}_{ij}) \right\}^{-2}. \quad (20)$$

Because the closed form of the solution to (20) could not be obtained, an iterative numerical process to obtain the parameter estimate is described as:

## INFERENCE FOR THE RAYLEIGH DISTRIBUTION

1. Let the initial estimate of  $\lambda$  be  $\lambda^{(0)}$ , with  $h = 0$ .
2. In the  $(h+1)^{\text{th}}$  iteration, first compute

$$E_{1i} = E_{\lambda^{(h)}}(X|\tilde{x}_i) = \frac{\int x^2 e^{-\lambda^{(h)} x^2} \mu_{\tilde{x}_i}(x) dx}{\int x e^{-\lambda^{(h)} x^2} \mu_{\tilde{x}_i}(x) dx}, \quad i = 1, \dots, m,$$

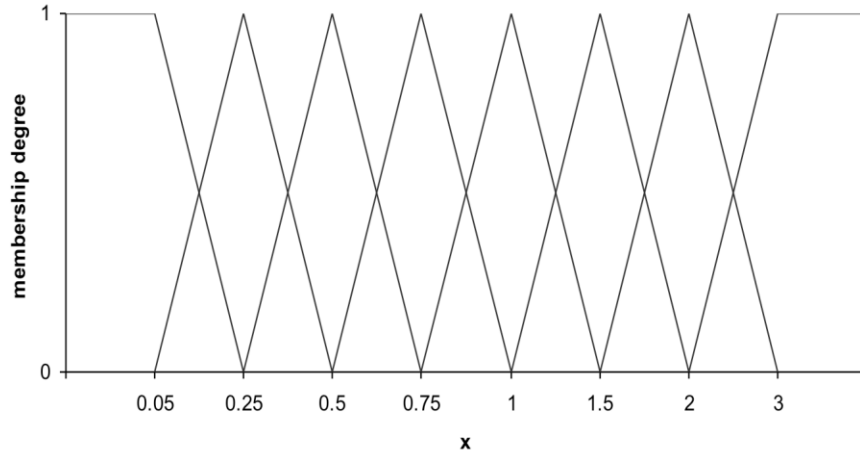
and

$$E_{2i} = E_{\lambda^{(h)}}(Z|\tilde{z}_{ij}) = \frac{\int z^2 e^{-\lambda^{(h)} z^2} \mu_{\tilde{z}_{ij}}(z) dz}{\int z e^{-\lambda^{(h)} z^2} \mu_{\tilde{z}_{ij}}(z) dz}, \quad i = 1, \dots, m, \quad j = 1, \dots, R_i.$$

The new estimate of  $\lambda$ , for example  $\lambda^{(h+1)}$ , can be obtained from:

$$\lambda^{(h+1)} = \frac{\pi n^2}{4} \left\{ \sum_{i=1}^m (E_{1i} + R_i E_{2i}) \right\}^{-2}.$$

3. Checking convergence, if the convergence occurs then the current  $\lambda^{(h+1)}$  is the estimate of  $\lambda$  by the method of moments; otherwise, set  $h = h+1$  and go to Step 2.



**Figure 1.** Fuzzy information system used to encode the simulated data

## Numerical Study

Experimental results illustrate how the different methods behave for varying sample sizes. First, for fixed  $\lambda = 1$  and different choices of  $n$ ,  $m$  and censoring scheme  $(R_1, \dots, R_m)$ , progressively censored samples from the Rayleigh distribution were generated, using the method proposed by Balakrishnan and Sandhu (1995), as follows:

1. Generate  $Z_i$  from  $U(0,1)$  for  $i = 1, \dots, m$ .
2. For given values of the progressive censoring scheme  $(R_1, \dots, R_m)$ , set  $V_i = Z_i^{1/a_i}$ ,  $a_i = i + \sum_{j=m-i+1}^m R_j$ ,  $i = 1, \dots, m$ .
3. Set  $U_i = 1 - V_{m-i+1}V_{m-i+2} \dots V_m$ ,  $i = 1, \dots, m$ .
4. Thus,  $X_i = F^{-1}(U_i)$ ,  $i = 1, \dots, m$ , is the desired progressive type-II censored sample from the Rayleigh distribution.

Each realization of  $\mathbf{x}$  was then fuzzified using the f.i.s. shown in Figure 1, and the ML, HPD and moment estimates (MME) of  $\lambda$  for the fuzzy sample were computed using the methods provided in the preceding sections. For computing the HPD estimate of the unknown parameter, assume that  $\lambda$  has  $Gamma(a, b)$  prior. To make the comparison meaningful, it is assumed that the priors are non-informative, and they are  $a = b = 0$ . Note that in this case the priors are non-proper also. Press (2001) suggested using very small non-negative values of the hyperparameters in this case, and it will make the priors proper. This study uses  $a = b = 0.0001$ . The results are not significantly different than the corresponding results obtained using non-proper priors, and are not reported due to space. The average values and mean squared errors of the estimates, computed over 1,000 replication, are presented in Tables 1-3.

# INFERENCE FOR THE RAYLEIGH DISTRIBUTION

**Table 1.** Average value (AV) and mean squared error (MSE) of the estimates of  $\lambda$  for different censoring schemes. ( $n = 20$ )

$m$	Censoring scheme	MLE		HPD		MME	
		AV	MSE	AV	MSE	AV	MSE
8	(0,...,0,12)	1.181	0.219	1.163	0.185	1.184	0.221
	(12,0,...,0)	1.178	0.211	1.159	0.173	1.175	0.207
	(0,12,0,...,0)	1.163	0.203	1.151	0.169	1.160	0.202
10	(0,...,0,10)	1.140	0.172	1.127	0.148	1.141	0.172
	(10,0,...,0)	1.153	0.182	1.136	0.166	1.155	0.184
	(0,10,0,...,0)	1.145	0.177	1.130	0.154	1.148	0.179
15	(0,...,0,5)	1.127	0.155	1.114	0.132	1.122	0.151
	(5,0,...,0)	1.132	0.161	1.119	0.137	1.130	0.159
	(0,5,0,...,0)	1.138	0.164	1.125	0.140	1.133	0.162

**Table 2.** Average value (AV) and mean squared error (MSE) of the estimates of  $\lambda$  for different censoring schemes. ( $n = 30$ )

$m$	Censoring scheme	MLE		HPD		MME	
		AV	MSE	AV	MSE	AV	MSE
8	(0,...,0,12)	1.163	0.185	1.149	0.162	1.166	0.187
	(12,0,...,0)	1.167	0.188	1.152	0.169	1.165	0.184
	(0,12,0,...,0)	1.155	0.179	1.143	0.157	1.152	0.170
10	(0,...,0,10)	1.138	0.160	1.127	0.145	1.137	0.163
	(10,0,...,0)	1.125	0.154	1.119	0.130	1.123	0.147
	(0,10,0,...,0)	1.132	0.159	1.122	0.138	1.130	0.156
15	(0,...,0,5)	1.112	0.136	1.103	0.117	1.114	0.139
	(5,0,...,0)	1.116	0.138	1.105	0.125	1.119	0.131
	(0,5,0,...,0)	1.121	0.133	1.109	0.115	1.120	0.133

**Table 3.** Average value (AV) and mean squared error (MSE) of the estimates of  $\lambda$  for different censoring schemes. ( $n = 50$ )

$m$	Censoring scheme	MLE		HPD		MME	
		AV	MSE	AV	MSE	AV	MSE
8	(0,...,0,12)	1.134	0.115	1.124	0.098	1.137	0.118
	(12,0,...,0)	1.141	0.122	1.130	0.113	1.145	0.125
	(0,12,0,...,0)	1.139	0.119	1.127	0.105	1.136	0.114
10	(0,...,0,10)	1.085	0.097	1.062	0.071	1.082	0.095
	(10,0,...,0)	1.079	0.093	1.056	0.065	1.077	0.090
	(0,10,0,...,0)	1.073	0.088	1.051	0.059	1.070	0.084
15	(0,...,0,5)	1.038	0.069	1.021	0.037	1.042	0.072
	(5,0,...,0)	1.025	0.045	1.011	0.026	1.025	0.053
	(0,5,0,...,0)	1.031	0.062	1.018	0.031	1.036	0.066

Several points are clear from the experiment: Even for small sample sizes, the performances of the estimates are satisfactory in terms of AVs and MSEs. For all the methods, it is observed that for fixed  $n$  as  $m$  increases, the MSEs of the estimates decrease. Among the three estimation procedures developed in the paper, the HPD procedure gives the most precise parameter estimates as shown by MSEs in Tables 1-3.

### Application example

To demonstrate the application of proposed methods to real data, consider the data collected during the experiment reported by Pak et al. (2013). In this experiment, a sample of 25 ball bearings is placed on a life test. A ball bearing may work perfectly over a certain period but be breaking for some time and finally be unusable at a certain time. So, the observed failure times of the ball bearings are reported by fuzzy numbers  $\tilde{x}_i = (a_i, x_i, b_i)$ , where  $a_i = 0.05x_i$  and  $b_i = 0.03x_i$  with the membership functions

$$\mu_{\tilde{x}_i}(x) = \begin{cases} \frac{x - (x_i - a_i)}{a_i} & x_i - a_i \leq x \leq x_i, \\ \frac{x_i + b_i - x}{b_i} & x_i \leq x \leq x_i + b_i, \end{cases} \quad i = 1, \dots, 25.$$

Progressively censored samples of size  $m = 10$  were considered from these fuzzy data using three different sampling schemes, namely:

Scheme 1:  $R_1 = \dots = R_{m-1} = 0$  and  $R_m = 15$ .

Scheme 2:  $R_1 = 15$  and  $R_2 = \dots = R_m = 0$ .

Scheme 3:  $R_1 = \dots = R_{m-1} = 1$  and  $R_m = 6$ .

the estimate of the parameter  $\lambda$  was then computed using the ML, HPD and moment methods. For computing the HPD estimates, it was assumed that  $\lambda$  has  $Gamma(a, b)$  prior, including the non-informative gamma prior, i.e.  $a = b = 0$  and informative gamma prior, i.e.  $a = b = 2$ . All the results are summarized in Table 4.

## INFERENCE FOR THE RAYLEIGH DISTRIBUTION

**Table 4.** ML, HPD and moment estimates of the parameter for Example 2

Scheme	MLE	MME	HPD ( $a = b = 0$ )	HPD ( $a = b = 2$ )
1	0.00016	0.00015	0.00027	0.00033
2	0.00043	0.00048	0.00058	0.00061
3	0.00019	0.00021	0.00032	0.00039

## Conclusions

Some work has been done in the past on the estimation of the parameter of Rayleigh distribution based on complete and censored samples, but traditionally it is assumed that the data available are performed in exact numbers. In real world situations, however, some collected lifetime data might be imprecise and are represented in the form of fuzzy numbers. Therefore, suitable statistical methodology is needed to handle these data. This article proposed different procedures for estimating the parameter of Rayleigh distribution under progressive type-II censoring when the available observations are described by means of fuzzy information. They are maximum likelihood estimation (MLE), highest posterior density (HPD) estimation and method of moments (MME). A simulation study was conducted to assess the performance of these procedures. Based on the results of the simulation study, it may be observed that, the performance of the HPD estimates is generally best followed by the MLE and MME. Thus, it would seem reasonable to recommend the use of the HPD procedure for estimating the unknown parameter  $\lambda$  from the Rayleigh distribution.

## References

- Azimi, R., Yaghmaei, F., & Azimi, D. (2012). Comparison of Bayesian estimation methods for Rayleigh progressive censored data under the different asymmetric loss function. *International Journal of Applied Mathematical Research*, 1(4): 452-461.
- Balakrishnan, N. (2007). Progressive censoring methodology: an appraisal. *Test*, 16(2): 211-296.
- Balakrishnan, N. & Aggarwala, R. (2000). *Progressive censoring, theory, methods and applications*. Boston, MA: Birkhauser.



- Balakrishnan, N. & Sandhu, R. A. (1995). A simple algorithm for generating progressively Type-II censored samples. *The American Statistician*, 49(2): 229-230.
- Bhattacharya, S. K., & Tyagi, R. K. (1990). Bayesian survival analysis based on the Rayleigh model. *Trabajos de Estadística*, 5: 81-92.
- Chung, Y. (1995). Estimation of scale parameter from Rayleigh distribution under entropy loss. *Journal of Applied Mathematics and Computing*, 2(1): 33-40.
- Denoeux, T. (2011). Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy Sets and Systems*, 183(1): 72-91.
- Dey, S., & Maiti, S. S. (2012). Bayesian estimation of the parameter of Rayleigh distribution under the extended Jeffrey's prior. *Electronic Journal of Applied Statistical Analysis*, 5(1), 44-59.
- Dubois, D., & Prade, H. (1980). *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, New York.
- Dyer, D. D. & Whisenand, C. W. (1973). Best linear unbiased estimator of the parameter of the Rayleigh distribution. *IEEE Transactions on Reliability R-22*: 27-34, 455-466.
- Fernandez, A. J. (2010). Bayesian estimation and prediction based on Rayleigh sample quantiles. *Quality & Quantity*, 44: 1239-1248.
- Kim, C., and Han, K. (2009). Estimation of the scale parameter of the Rayleigh distribution under general progressive censoring. *Journal of the Korean Statistical Society*, 38: 239-246.
- Lee, W., Wu, J., Hong, M., Lin, L., & Chan, R. (2011). Assessing the lifetime performance index of Rayleigh products based on the Bayesian estimation under progressive type II right censored samples. *Journal of Computational and Applied Mathematics*, 235: 1676-1688.
- Krishna, H. & Kumar, K. (2011). Reliability estimation in Lindley distribution with progressively type-II right censored sample. *Mathematics and Computers in Simulation*, 82: 281-294.
- 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: 303-320.
- Pradhan, B., & Kundu, D. (2009). On progressively censored generalized exponential distribution. *Test*, 18: 497-515.

## INFERENCE FOR THE RAYLEIGH DISTRIBUTION

Raqab, M. Z., & Madi, T. M. (2011). Inference for the generalized Rayleigh distribution based on progressively censored data. *Journal of Statistical Planning and Inference*, 141: 3313-3322.

Rastogi, M. K., & Tripathi, Y. M. (2012). Estimating the parameters of Burr distribution under progressive type II censoring. *Statistical Methodology*, 9: 381-391.

Rayleigh, J. W. S. (1880). On the resultant of a large number of vibration of the same pitch and of arbitrary phase. *Philosophical Magazine*, 5th series, 10: 73-78.

Press, S. J. (2001). *The subjectivity of scientists and the Bayesian approach*. New York: Wiley.

Tanaka, H. Okuda, T. & Asai, K. (1979). Fuzzy information and decision in statistical model. In M. M. Gupta et al., Eds. *Advances in Fuzzy Sets Theory and Applications*, pp. 303-310. Amsterdam: North-Holland Publishing Co.

Zadeh, L. A. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications* 10: 421-427.

# Evaluation of Area under the Constant Shape Bi-Weibull ROC Curve

**Sudesh Pundir**

Pondicherry University  
Puducherry, India

**R. Amala**

Pondicherry University  
Puducherry, India

The Receiver Operating Characteristic (ROC) curve generated based on assuming a constant shape Bi-Weibull distribution is studied. In the context of ROC curve analysis, it is assumed that biomarker values from controls and cases follow some specific distribution and the accuracy is evaluated by using the ROC model developed from that specified distribution. This article assumes that the biomarker values from the two groups follow Weibull distributions with equal shape parameter and different scale parameters. The ROC model, area under the ROC curve (AUC), asymptotic and bootstrap confidence intervals for the AUC are derived. Theoretical results are validated by simulation studies.

**Keywords:** Constant shape Bi-Weibull ROC model, area under the ROC curve, asymptotic variance of accuracy, confidence interval, parametric bootstrap variance

## Notations and Terminologies

$X$	Random variable representing controls	$t$	Cut-off point of classification, $t \in x \cup y$
$Y$	Random variable representing cases	$I(\theta)$	Fisher Information matrix
$m$	Number of controls	$y(x)$	ROC model
$n$	Number of cases	$MLE$	Maximum Likelihood Estimate
$f(x)$	Probability Density Function (PDF) of X	$x(t)$	False Positive Rate (FPR) at cut-off $t$
$g(y)$	PDF of Y	$y(t)$	True Positive Rate (TPR) at cut-off $t$
$F(x)$	Distribution function of X	$TPR$	Probability that cases are correctly identified (Sensitivity)
$G(y)$	Distribution function of Y	$FPR$	Probability that controls are wrongly identified as cases (1-Specificity)
$AUC$	Population Area under the ROC curve	$\alpha_0, \alpha_1$	Shape parameters of X and Y, respectively
$\hat{AUC}$	Observed Area Under the ROC curve	$\beta_0, \beta_1$	Scale parameters of X and Y, respectively

*Dr. Sudesh Pundir is an Assistant Professor in the Department of Statistics. Email her at [sudeshpundir19@gmail.com](mailto:sudeshpundir19@gmail.com). R. Amala is a Doctoral Student in Applied Statistics. Email her at [amalar.statistics@gmail.com](mailto:amalar.statistics@gmail.com).*

## Introduction

A Receiver Operating Characteristic (ROC) curve provides quick access to the quality of classification in many medical diagnoses. In ROC curve analysis, the accuracy has been analyzed in terms of a model relating the parameters of cases and controls called as the ROC model. ROC model can be defined as the TPR obtained as a function of FPR which takes the form

$$y(x) = 1 - G\left(F^{-1}(1 - x(t))\right); 0 \leq x(t) \leq 1 \quad (1)$$

where  $x(t)$  and  $y(t)$  are defined as follows:

$$\left. \begin{aligned} x(t) &= P(X > t) = \int_t^{\infty} f(x)dx = 1 - \int_0^t f(x)dx = 1 - F(t) \\ y(t) &= P(Y > t) = \int_t^{\infty} g(y)dy = 1 - \int_0^t g(y)dy = 1 - G(t) \end{aligned} \right\} \quad (2)$$

Graphically, a ROC curve is a graph of TPR versus FPR for all possible threshold values. The ROC curve can be plotted by three approaches viz. parametric, non-parametric and semi-parametric. This article considers the parametric way of plotting the ROC curve. After the ROC curve is generated the intrinsic accuracy provided by the biomarker must be interpreted. To summarize the information contained in a ROC curve, many indices have been used. Among them, area under the ROC curve is most commonly adopted index. In this article, the inference about the area under the ROC curve is of primary interest.

The problem of assessing the accuracy of diagnosis/Biomarker has been studied by several authors by assuming various distributions to the biomarker values. They are Bi-Normal ROC model (Zhou, Obuchowski & McClish, 2002), Bi-Logistic ROC model (Oglive & Creelman, 1968), Bi-Lomax ROC model (Campbell & Ratnaparkhi, 1993), Bi-Gamma ROC model (Dorfman et al., 1996), Bi-Exponential ROC model (Betinec, 2008), Generalized Bi-Exponential ROC model (Hussain, 2011), Bi-Rayleigh ROC model and its comparison with Bi-Normal model (Pundir & Amala, 2012), comparison of Bi-Rayleigh ROC model with Bi-Normal and Bi-Gamma ROC models (Pundir & Amala, 2012) and a review of all parametric ROC models in case of continuous data (Pundir & Amala, 2014), Normal-Exponential (Pundir & Amala, 2014).

A constant shape Bi-Weibull ROC model is proposed for non-normal data. Two parameter Weibull distribution is a most widely used life distribution in various fields viz. Survival analysis, Reliability engineering and recently in ROC curve analysis. Let  $X \sim W(\alpha_0, \beta_0)$  and  $Y \sim W(\alpha_1, \beta_1)$ , then the ROC model developed from two parameter Bi-Weibull distribution is given by

$$y(x) = \text{Exp} \left\{ - \frac{(-\beta_0 \ln(x(t)))^{\frac{\alpha_1}{\alpha_0}}}{\beta_1} \right\}, \alpha_1, \alpha_0 > 0, \beta_1, \beta_0 > 0 \quad (3)$$

One major disadvantage of assuming two parameter Weibull distribution to the biomarker is that the accuracy cannot be expressed in closed form. By substituting the MLE's  $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0$  and  $\hat{\beta}_1$ , the accuracy can be evaluated numerically using Monte Carlo integration or any other numerical procedure. In the absence of closed form expression, the statistical inference on the accuracy measure will not be possible. To overcome this problem and to obtain a closed form expression, equal shape parameter and different scale parameters are assumed. Moreover, the original accuracy of the diagnosis is not affected by taking equal shape parameter. The ROC model developed from this assumption is called the constant shape Bi-Weibull ROC model.

Research interest may lie in comparing the effectiveness of two separate diagnostic tests or the efficiency of biomarkers in predicting the disease. The comparison can be accomplished either by AUC or sensitivity of the test. In order to compare the AUC and to construct the confidence interval, the Standard Error (SE) of AUC are needed. Here, the standard error of accuracy is studied by different methods viz. Monte Carlo, asymptotic MLE, parametric bootstrap and non-parametric methods. For parametric, the delta method will yield variance and SE with the help of asymptotic expressions for the variance and co-variances of the parameters.

### Constant Shape Bi-Weibull ROC Model

The constant shape Bi-Weibull ROC model assumes that the biomarker values from controls and cases follow two parameter Weibull distribution with same shape parameter and different scale parameters.

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

The PDF of controls and cases take the form

$$f(x) = \frac{\alpha}{\beta_0} x^{\alpha-1} \exp\left\{-\frac{x^\alpha}{\beta_0}\right\}, x > 0, \alpha, \beta_0 > 0 \quad (4)$$

and

$$g(y) = \frac{\alpha}{\beta_1} y^{\alpha-1} \exp\left\{-\frac{y^\alpha}{\beta_1}\right\}, y > 0, \alpha, \beta_1 > 0 \quad (5)$$

respectively.

The probabilities, Sensitivity and 1-Specificity for constant shape Bi-Weibull distribution can be given as follows:

$$Sensitivity = \int_t^\infty \frac{\alpha}{\beta_1} y^{\alpha-1} \exp\left\{-\frac{y^\alpha}{\beta_1}\right\} dy = \exp\left\{\frac{-t^\alpha}{\beta_1}\right\} \quad (6)$$

$$1 - Specificity = \int_t^\infty \frac{\alpha}{\beta_0} x^{\alpha-1} \exp\left\{-\frac{x^\alpha}{\beta_0}\right\} dx = \exp\left\{\frac{-t^\alpha}{\beta_0}\right\} \quad (7)$$

Hence, the ROC model is given by

$$y(x) = x(t)^{\left(\frac{\beta_0}{\beta_1}\right)}; 0 \leq x(t) \leq 1. \quad (8)$$

The ROC curve can be estimated by substituting the MLE of parameters in equation (8) and plotted by taking  $x(t)$  in equation (7) on  $x$ -axis and  $y(x)$  in equation (8) on  $y$ -axis. Also, one can plot the ROC curve by taking 1-Specificity on  $X$  axis and Sensitivity on  $Y$  axis. The area under the ROC curve is obtained by integrating the joint density function of  $X$  and  $Y$  and it has the following form.

$$\begin{aligned}
A = P(X < Y) &= \int_0^\infty \int_0^y \frac{\alpha^2}{\beta_0 \beta_1} (xy)^{\alpha-1} \text{Exp} \left[ - \left( \frac{y^\alpha}{\beta_1} + \frac{x^\alpha}{\beta_0} \right) \right] dx dy \\
&= \frac{\beta_1}{\beta_1 + \beta_0}
\end{aligned} \tag{9}$$

The MLEs of  $\beta_0$  and  $\beta_1$  can be used again to estimate the AUC. And the performance of the estimator  $\widehat{AUC}$  can be assessed through variance estimate.

### Maximum Likelihood Estimate of Parameters

The MLE of two parameter Weibull distribution has been discussed by (Kundu & Gupta, 2006) in the context of Reliability estimation. Let  $X_1, X_2, \dots, X_m$  be a random sample of size  $m$  from  $W(\alpha, \beta_0)$  and  $Y_1, Y_2, \dots, Y_n$  be a random sample of size  $n$  from  $W(\alpha, \beta_1)$ . The likelihood function of the selected sample is given by

$$L(x_i, y_j / \theta) = \prod_{i=1}^m f_X(x_i / \alpha, \beta_0) \prod_{j=1}^n f_Y(y_j / \alpha, \beta_1) \tag{10}$$

where  $\theta = (\alpha, \beta_0, \beta_1)$ ,

The log-likelihood function is

$$\begin{aligned}
LnL &= (m+n) \ln \alpha + (\alpha-1) \left[ \sum_{j=1}^n \ln y_j + \sum_{i=1}^m \ln x_i \right] - n \ln \beta_1 \\
&\quad - m \ln \beta_0 - \frac{1}{\beta_1} \sum_{j=1}^n y_j^\alpha - \frac{1}{\beta_0} \sum_{i=1}^m x_i^\alpha
\end{aligned} \tag{11}$$

Differentiating (11) with respect to  $\alpha$  results in

$$\begin{aligned}
\frac{\partial LnL}{\partial \alpha} &= \frac{(m+n)}{\alpha} + \left[ \sum_{j=1}^n \ln y_j + \sum_{i=1}^m \ln x_i \right] \\
&\quad - \frac{1}{\beta_1} \sum_{j=1}^n y_j^\alpha \ln y_j - \frac{1}{\beta_0} \sum_{i=1}^m x_i^\alpha \ln x_i
\end{aligned} \tag{12}$$

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

By differentiating the log-likelihood function with respect to  $\beta_0$ ,  $\beta_1$  and equating to zero, we get the estimates. The MLE's of  $\beta_1$  and  $\beta_0$  are determined as,

$$\hat{\beta}_1(\alpha) = \frac{\sum_{j=1}^n y_j^\alpha}{n} \quad \text{and} \quad \hat{\beta}_0(\alpha) = \frac{\sum_{i=1}^m x_i^\alpha}{m} \quad (13)$$

Substituting  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in equation (12) and equating it to 0, results in a non-linear equation:

$$h(\hat{\alpha}) = \frac{m + n + \sum_{i=1}^n y_j^\alpha + \sum_{i=1}^m x_i^\alpha}{\frac{m \sum_{i=1}^m x_i^\alpha \ln x_i}{\sum_{i=1}^m x_i^\alpha} + \frac{n \sum_{i=1}^n y_i^\alpha \ln y_j}{\sum_{i=1}^n y_j^\alpha}} \quad (14)$$

Hence,  $\hat{\alpha}$  can be determined as a solution of non-linear equation (14). By substituting equation (13) and (14) in equation (9), an estimate of  $AUC$  ( $\hat{AUC}$ ) will result.

### Asymptotic Distribution of area under constant shape Bi-Weibull ROC Model

To evaluate the significance of the statistic  $AUC$ , its variance and standard error must be computed. The following theorem evaluates the variance of the estimate,  $\hat{AUC}$ .



**Theorem 1**

The area under the constant shape Bi-Weibull ROC curve will converge in distribution to a Normal random variable with mean zero and variance

$$\tau = \frac{\beta_0^2 \beta_l^2}{(\beta_0 + \beta_l)^4} \left[ \frac{(m+n)}{mn} + \frac{\left[ \ln \left( \frac{\beta_0}{\beta_l} \right) \right]^2}{(m+n)(1 + \Gamma_2'' - \Gamma_2'^2)} \right] \text{ for large } N, \text{ where } N = m+n.$$

**Proof:** Let  $L(\theta / x, y)$ ;  $\theta = (\alpha, \beta_0, \beta_l)'$  be the likelihood function of the sample observations from  $X$  and  $Y$  which is given by

$$\begin{aligned} \ln L(\theta / x, y) &= (m+n) \ln \alpha - m \ln \beta_0 \\ &\quad - n \ln \beta_l + (\alpha - 1) \left[ \sum_{i=1}^m \ln x_i + \sum_{j=1}^n \ln y_j \right] \\ &\quad - \frac{1}{\beta_0} \sum_{i=1}^m x_i^\alpha - \frac{1}{\beta_l} \sum_{j=1}^n y_j^\alpha \end{aligned} \quad (15)$$

Asymptotic normality of MLE states that a consistent solution of the likelihood equation is asymptotically normally distributed about the true value  $\theta$  i.e.  $\hat{\theta} \sim N(\theta, I^{-1}(\theta))$ .

$$\Rightarrow \sqrt{N}(\hat{\theta} - \theta) \rightarrow N(0, I^{-1}(\theta)). \quad (16)$$

where  $I(\theta)$  is the Fisher Information matrix which is given by

$$I(\theta) = - \begin{bmatrix} E\left(\frac{\partial^2 \ln L}{\partial \alpha^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial \alpha \partial \beta_0}\right) & E\left(\frac{\partial^2 \ln L}{\partial \alpha \partial \beta_1}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial \beta_0 \partial \alpha}\right) & E\left(\frac{\partial^2 \ln L}{\partial \beta_0^2}\right) & E\left(\frac{\partial^2 \ln L}{\partial \beta_0 \partial \beta_1}\right) \\ E\left(\frac{\partial^2 \ln L}{\partial \beta_1 \partial \alpha}\right) & E\left(\frac{\partial^2 \ln L}{\partial \beta_1 \partial \beta_0}\right) & E\left(\frac{\partial^2 \ln L}{\partial \beta_1^2}\right) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. \quad (17)$$

where

$$\begin{aligned} a_{11} &= \frac{1}{\alpha^2} \left[ (m+n) \left[ 1 + \Gamma_2'' \right] + 2(n \ln \beta_1 + m \ln \beta_0) \Gamma_2' + n(\ln \beta_1)^2 + m(\ln \beta_0)^2 \right], \\ a_{22} &= \frac{m}{\beta_0^2}, \quad a_{33} = \frac{n}{\beta_1^2}, \\ a_{23} &= a_{32} = 0, \quad a_{12} = a_{21} = \frac{-m}{\alpha \beta_0} (\Gamma_2' + \ln \beta_0), \\ a_{13} &= a_{31} = \frac{-n}{\alpha \beta_1} (\Gamma_2' + \ln \beta_1), \\ V(\hat{\beta}_0) &= \frac{\beta_0^2 \left[ n(m+n)(1 + \Gamma_2'') + 2mn \log(\beta_0) \Gamma_2' + mn [\log(\beta_0)]^2 - n^2 (\Gamma_2')^2 \right]}{mn(m+n) \left( 1 - \Gamma_2'' - (\Gamma_2')^2 \right)}, \\ V(\hat{\beta}_1) &= \frac{\beta_1^2 \left[ n(m+n)(1 + \Gamma_2'') + 2mn \log(\beta_1) \Gamma_2' + mn [\log(\beta_1)]^2 - m^2 (\Gamma_2')^2 \right]}{mn(m+n) \left( 1 - \Gamma_2'' - (\Gamma_2')^2 \right)}. \end{aligned}$$

The  $I^{-1}(\theta)$  is calculated as:

$$I^{-1}(\theta) = \frac{1}{a_{11}a_{22}a_{33} - a_{12}^2a_{33} - a_{22}a_{13}^2} \begin{bmatrix} a_{22}a_{33} & -a_{21}a_{33} & -a_{22}a_{31} \\ -a_{12}a_{33} & a_{11}a_{33} - a_{13}^2 & a_{12}a_{31} \\ -a_{22}a_{13} & a_{21}a_{13} & a_{11}a_{22} - a_{12}^2 \end{bmatrix}. \quad (18)$$

$$= \begin{bmatrix} V(\hat{\alpha}) & Cov(\hat{\alpha}, \hat{\beta}_0) & Cov(\hat{\alpha}, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\alpha}) & V(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_1, \hat{\alpha}) & Cov(\hat{\beta}_1, \hat{\beta}_0) & V(\hat{\beta}_1) \end{bmatrix} \quad (19)$$

where

$$V(\hat{\alpha}) = \frac{\alpha^2}{(m+n)(1-\Gamma_2''-(\Gamma_2')^2)}, \quad Cov(\hat{\alpha}, \hat{\beta}_0) = \frac{\alpha\beta_0(\Gamma_2' + \ln \beta_0)}{(m+n)[1-\Gamma_2''-(\Gamma_2')^2]},$$

$$Cov(\hat{\alpha}, \hat{\beta}_1) = \frac{\alpha\beta_1(\Gamma_2' + \ln \beta_1)}{(m+n)[1-\Gamma_2''-(\Gamma_2')^2]},$$

$$\text{and } C(\hat{\beta}_0, \hat{\beta}_1) = \frac{\beta_0\beta_1(\Gamma_2' + \ln \beta_0)(\Gamma_2' + \ln \beta_1)}{(m+n)[1-\Gamma_2''-(\Gamma_2')^2]}$$

Because the area under the ROC curve is a function of parameters  $\theta = (\alpha, \beta_0, \beta_1)'$ , the Delta method will be adopted for finding the approximate variance.  $V(\hat{AUC})$  can be defined as:

$$V(\hat{AUC}) = \left( \frac{\partial AUC}{\partial \beta_1} \right)^2 V(\hat{\beta}_1) + \left( \frac{\partial AUC}{\partial \beta_0} \right)^2 V(\hat{\beta}_0) + 2 \left( \frac{\partial AUC}{\partial \beta_0} \right) \left( \frac{\partial AUC}{\partial \beta_1} \right) Cov(\hat{\beta}_0, \hat{\beta}_1). \quad (20)$$

$$\tau = V(\hat{AUC}) = \frac{\beta_0^2 \beta_1^2}{(\beta_0 + \beta_1)^4} \left[ \frac{m+n}{mn} + \frac{\left[ \ln \left( \frac{\beta_0}{\beta_1} \right) \right]^2}{(m+n)(1+\Gamma_2''-\Gamma_2'^2)} \right]. \quad (21)$$

where  $V(\hat{\beta}_1)$ ,  $V(\hat{\beta}_0)$  and  $Cov(\hat{\beta}_0, \hat{\beta}_1)$  are taken from the matrix  $I^{-1}(\theta)$ . The estimate of variance is obtained by substituting the estimates of the parameters  $\beta_0, \beta_1$ . Hence, the estimate of accuracy follows that

$$\frac{\sqrt{N}(\hat{AUC} - AUC)}{\sqrt{\tau}} \rightarrow N(0,1). \quad (22)$$

where  $\tau$  is obtained in equation (20) and it is proven that  $\hat{AUC} \sim N(0, \tau)$ ,  $\Gamma'_n = -(n-1)! \left[ \frac{1}{n} + \gamma - \sum_{k=1}^n \frac{1}{k} \right]$  where  $\gamma$  is Euler-Mascheroni constant approximately equal to 0.5772. Note:  $\hat{AUC}$  is an Unbiased Estimator of AUC (See [Appendix D](#) for the proof).

## Confidence Interval for $\hat{AUC}$

### Asymptotic Confidence Interval

The asymptotic  $100(1-\alpha)\%$  confidence interval for accuracy is given by

$$\left[ \hat{AUC} - Z_{\alpha/2} SE(\hat{AUC}), \hat{AUC} + Z_{\alpha/2} SE(\hat{AUC}) \right]. \quad (23)$$

where  $SE(\hat{AUC})$  can be obtained from equation (21),  $\alpha$  is the level of significance and  $Z_{\alpha/2}$  is the critical value. For example,  $Z_{\alpha/2}$  for a 5% level of significance is 1.96.

### Bootstrap Confidence Interval

The parametric bootstrap is a resampling technique which can be used to find the variance of any estimator. The idea of bootstrap is to create or resample an artificial dataset from an empirical distribution with same sample size and structure as the original for large number of times. Once the dataset is created, the parameters of interest are to be estimated for each data set. The bootstrap variance of parameter is nothing but the variance of all estimated parameters.

Parametric bootstrap is very similar to the non-parametric bootstrap method. In non-parametric bootstrap the sample is simulated from empirical distribution but in parametric bootstrap it is simulated from specified parametric distribution. The following are the steps involved in finding the parametric bootstrap estimate:

**Step 1:** Let  $X_1, X_2, \dots, X_m$  be a random sample of size  $m$  from  $W(\alpha_0, \beta_0)$  and  $Y_1, Y_2, \dots, Y_n$  be a random sample of size  $n$  from  $W(\alpha_1, \beta_1)$ . By using equation (13) and (14), the ML estimates of the parameters  $\alpha, \beta_0, \beta_1$  are estimated.

**Step 2:** By using the estimated parameters  $\hat{\alpha}, \hat{\beta}_0$  and  $\hat{\beta}_1$ , the random observations  $X_b$  of size  $m$  and  $Y_b$  of size  $n$  (Bootstrap samples) are generated. From  $X_b$  and  $Y_b$ , the bootstrap estimates viz.  $\hat{\alpha}_b, \hat{\beta}_{b0}$  and  $\hat{\beta}_{b1}$  are obtained. Using these bootstrap estimates the accuracy ( $A\hat{U}C_b$ ) is obtained.

**Step 3:** Step 2 is repeated 10,000 times. The mean of all 10,000 estimates of  $\hat{\alpha}_b$ 's,  $\hat{\beta}_{b0}$ 's and  $\hat{\beta}_{b1}$ 's are called the bootstrap estimates of parameters  $\alpha, \beta_0$  and  $\beta_1$  respectively and mean of all  $A\hat{U}C_b$ 's is called the estimated bootstrap accuracy. The standard deviation of all estimates  $A\hat{U}C_b$  is called the standard error of  $A\hat{U}C_b$ .

**Step 4:** The  $100(1-\alpha)\%$  confidence interval for  $A\hat{U}C_b$  is obtained as follows:

$$\left[ A\hat{U}C_b - Z_{\alpha/2} SE(A\hat{U}C_b), A\hat{U}C_b + Z_{\alpha/2} SE(A\hat{U}C_b) \right].$$

where  $\alpha$  is the level of significance and  $Z_{\frac{\alpha}{2}}$  is the critical value.

## Simulation Studies

Thus, the accuracy, standard error of  $A\hat{U}C$  and 95 % confidence interval for  $A\hat{U}C$  have been computed through four different techniques via Monte Carlo method, asymptotic MLE method, parametric bootstrap and non-parametric method.

### Monte Carlo Method

The model in equation (3) does not possess a closed form, so Monte Carlo integration of equation (3) is necessary. A Monte Carlo simulation was performed to inspect the accuracy obtained by Monte Carlo integration. The Monte Carlo estimate of AUC, SE ( $A\hat{U}C$ ) and 95% confidence interval for  $A\hat{U}C$  is presented in Table 1. The R codes for the Monte Carlo simulation is provided in Appendix A.

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

**Table 1.** Accuracy, standard error and Confidence interval of  $\hat{AUC}$  based on Constant Shape Bi-Weibull ROC model through Monte Carlo Simulation

SL. No.	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{AUC}$	$V(\hat{AUC})$	95% Confidence Interval	Band Width
1	3.0	2.0	9	45	0.9188	0.051969	[0.816923, 1]	0.1831
2	3.0	2.0	9	30	0.8835	0.072986	[0.740444, 1]	0.2596
3	2.5	1.5	9	12	0.7590	0.124569	[0.514798, 1]	0.4852
4	3.5	2.5	9	10	0.6727	0.180434	[0.319085, 1]	0.6809

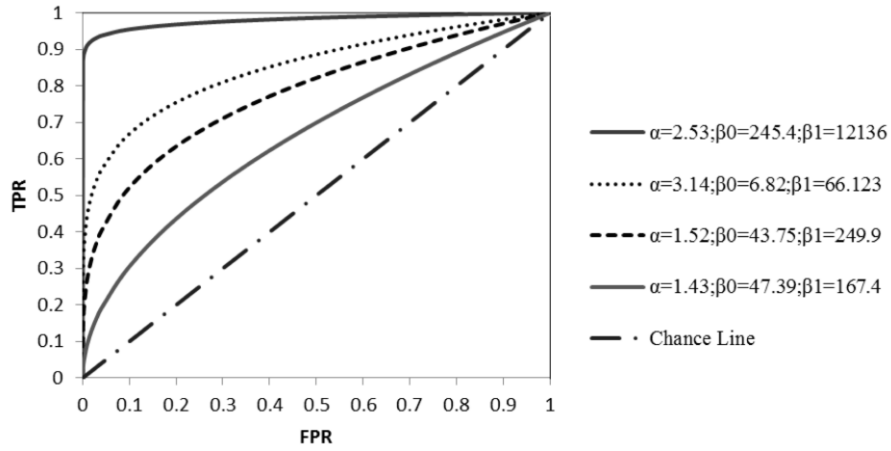
### Asymptotic MLE Method

Numerical experiments were carried out to inspect how the MLE's of AUC and their asymptotic results work for simulated data sets. Four different samples of size  $(m, n) = (30, 30)$  with different parametric values were considered as mentioned in column 2, 3 and 4 of Table 2. The corresponding accuracy, SE, 95% confidence interval and the band width are shown in 5, 6, 7, 8 columns of Table 1. As the accuracy increases, the SE tend to decrease, simultaneously, the coverage area of the confidence band are tends to decrease as accuracy increases. Because the asymptotic distribution is independent of  $\alpha$ ,  $\alpha$  may be kept constant or it may vary. From the sample  $\alpha$  is estimated using iterative procedure from equation (14) and using  $\alpha$ , the other two parameters using were found using equation (13). Hence, the ML estimate of AUC is obtained. The 95% asymptotic confidence interval and the confidence width are also calculated.

**Table 2.** Accuracy, standard error and Confidence interval of  $\hat{A}$  based on Constant Shape Bi-Weibull ROC model through Asymptotic MLE method

Sl. No.	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{A}$	$V(\hat{A})$	95% Confidence Interval	Band Width
1	2.530	12136	245.4000	0.980	0.00913	[0.9623, 0.99133]	0.02903
2	3.140	66.123	6.8201	0.907	0.02924	[0.8491, 0.96380]	0.11460
3	1.520	249.980	43.7500	0.850	0.03960	[0.7735, 0.92860]	0.15510
4	1.430	167.430	47.3900	0.778	0.04950	[0.6824, 0.87638]	0.19398
5	1.085	36.290	18.7200	0.660	0.05990	[0.5425, 0.77700]	0.23450

Table 3 shows simulated independent samples of  $m$  controls and  $n$  cases ( $m = n = 5, 10, 40, 50, 80, 100$ ) to assess the behavior of asymptotic MLE's and confidence interval over different sample sizes by fixing  $\hat{\beta}_0 = 5$  and for different values of  $\hat{\beta}_1$  viz. 8, 12, 20, 100. In Tables 3 and 4, first row represents the AUC, second row gives the SE, third row gives the lower confidence limit and the fourth row represents the upper confidence limit. It is observed that, as the sample size increases the variance decreases and the coverage area of confidence interval is narrow.



**Figure 1.** Constant Shape Bi-Weibull ROC model plotted for different AUC

Table 4 shows simulated independent samples of  $m$  controls and  $n$  cases ( $m = n = 40, 50, 80, 100$ ) to inspect the behavior of asymptotic MLE and confidence interval over different sample sizes by fixing  $\hat{\beta}_1 = 45$  and for different values of  $\hat{\beta}_0$  viz. 3, 8, 10, 20. It is observed that, as the sample size increases the variance decreases and the coverage area of confidence interval is narrow.

# EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

**Table 3.** Accuracy, Variance and 95% confidence Interval for AUC when  $\hat{\beta}_0 = 5$  for different sample size

Sample Size	$m = n = 5$	$m = n = 10$	$m = n = 40$	$m = n = 50$	$m = n = 80$	$m = n = 100$
$\hat{\beta}_1 = 8$	0.6154	0.6137	0.6137	0.6137	0.6137	0.6137
	0.0232	0.0116	0.0029	0.0023	0.0012	0.0012
	0.3171	0.4045	0.5099	0.5210	0.5408	0.5487
	0.9137	0.8263	0.7192	0.7080	0.6883	0.6804
$\hat{\beta}_1 = 12$	0.7059	0.7057	0.7057	0.7057	0.7057	0.70570
	0.0193	0.0096	0.0024	0.0019	0.0012	0.00096
	0.4339	0.5136	0.6097	0.6199	0.6379	0.64510
	0.9778	0.8982	0.8020	0.7918	0.7738	0.76610
$\hat{\beta}_1 = 20$	0.8000	0.8000	0.8000	0.8000	0.80000	0.8000
	0.0132	0.0066	0.0017	0.0013	0.00083	0.0007
	0.5745	0.6406	0.7203	0.7287	0.74360	0.7496
	1.0000	0.9594	0.8797	0.8730	0.85640	0.8504
$\hat{\beta}_1 = 100$	0.9524	0.9500	0.95000	0.9500	0.95000	0.95000
	0.0019	0.0028	0.00024	0.0001	0.00012	0.00009
	0.8659	0.7961	0.92180	0.9250	0.93080	0.93310
	1.0000	1.0000	0.98060	0.9773	0.97160	0.96930

**Table 4:** Accuracy, SE and 95% confidence Interval for AUC when  $\hat{\beta}_1 = 45$  for different sample size

Sample Size	$m = n = 5$	$m = n = 10$	$m = n = 40$	$m = n = 50$	$m = n = 80$	$m = n = 100$
$\hat{\beta}_0 = 3$	0.9375	0.9375	0.93750	0.93750	0.93750	0.93750
	0.0029	0.0015	0.00036	0.00029	0.00018	0.00015
	0.8318	0.8628	0.90010	0.90410	0.91110	0.91390
	1.0000	1.0000	0.97490	0.97090	0.96390	0.96110
$\hat{\beta}_0 = 8$	0.8490	0.8491	0.849100	0.84910	0.8491	0.84910
	0.0096	0.0048	0.001194	0.00096	0.0006	0.00047
	0.6575	0.7136	0.781300	0.78850	0.8012	0.80620
	1.0000	0.9845	0.916782	0.90960	0.8969	0.89190
$\hat{\beta}_0 = 10$	0.8182	0.8182	0.8182	0.8182	0.81820	0.818200
	0.0119	0.0059	0.0015	0.0012	0.00074	0.000595
	0.6044	0.6670	0.7426	0.7506	0.76470	0.770400
	1.0000	0.9694	0.8938	0.8858	0.87160	0.865900
$\hat{\beta}_0 = 20$	0.6923	0.6923	0.9500	0.6923	0.6923	0.692300
	0.0200	0.0100	0.0025	0.0019	0.0012	0.000998
	0.4154	0.8881	0.5944	0.6048	0.6231	0.630400
	0.9693	0.4965	0.7902	0.7799	0.7615	0.754200



### Estimation of Bootstrap Variance

For parametric bootstrapping, the data was generated from a uniform distribution using  $(m, n)$  as specified in Table 5. Then by inverse transformation method, it is converted into Weibull variate with the values of  $\alpha_0, \alpha_1, \beta_0$  and  $\beta_1$ . Using Step 1 results in the estimates as  $\hat{\alpha}, \hat{\beta}_0, \hat{\beta}_1$ . By using Step 2, the estimate of bootstrap sample is obtained as  $\hat{\alpha}_b, \hat{\beta}_{b0}$  and  $\hat{\beta}_{b1}$ . From these 10,000 estimates of parameters, one can find an estimate of AUC by using the equation (10). By averaging these 10,000 numbers of estimates of AUC, one can estimate the bootstrap estimate AUC. Standard error of  $\hat{AUC}_b$  is nothing but the standard deviation of the  $b$  number  $\hat{AUC}_b$ 's. By Step 4, the 95% confidence interval for bootstrap AUC is obtained as usual. Table 5 shows the bootstrap area under the curve, SE and confidence interval for  $\hat{AUC}_b$ .

**Table 5.** Accuracy, standard error and Confidence interval of  $\hat{AUC}$  based on Constant Shape Bi-Weibull ROC model through Bootstrap Simulation

$(m, n)$	$\hat{\alpha}_b$	$\hat{\beta}_{b0}$	$\hat{\beta}_{b1}$	$\hat{AUC}_b$	$SE(\hat{AUC}_b)$	95% Confidence Interval	Band Width
(10, 10)	2.6709	7.5414	201.8100	0.9249	0.04700	[0.8328, 1.0000]	0.1672
(20, 20)	2.5274	6.3739	103.9060	0.9240	0.03366	[0.8581, 0.9899]	0.1318
(30, 30)	2.4662	5.8770	84.2817	0.9220	0.02680	[0.8695, 0.9745]	0.1050
(50, 50)	2.4340	5.6493	74.3820	0.9222	0.02110	[0.8581, 0.9636]	0.1055
(100, 100)	2.3948	5.4283	66.5260	0.9211	0.01450	[0.8926, 0.9496]	0.0570

Comparing asymptotic and bootstrap variance, both perform at the same level. The asymptotic variance does not perform well for small samples such as (5, 5) and (10, 10) where the bound for accuracy has reached below 0.5 which is not regarded as a good estimate. Hence, the asymptotic variance holds for large samples only.

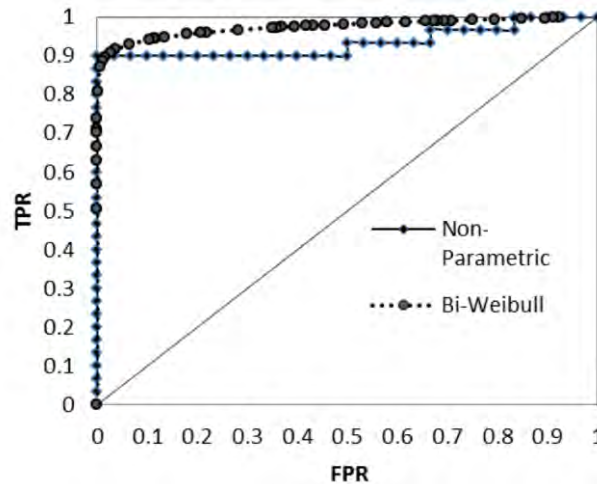
### Sensitivity and Specificity

To generate a Weibull random variate with parametric values  $\alpha_0 = 3; \alpha_1 = 2; \beta_0 = 9$  and  $\beta_1 = 45$ . The data is

$X = \{0.76261, 0.803019, 0.863084, 0.905439, 1.146029, 1.338408, 1.366008, 1.39672, 1.415312, 1.432053, 1.592267, 1.608494, 1.673259, 1.710255, 1.81614, 1.899346, 1.903763, 1.991144, 2.011153, 2.024541, 2.05607, 2.31567, 2.36017, 2.376429, 2.516461, 2.660371, 2.663695, 2.669402, 2.73371, 3.092265\}$

$Y = \{1.183838, 1.472276, 1.849655, 3.121439, 3.298009, 3.478297, 3.512602, 3.853157, 4.751021, 5.094757, 5.143248, 5.263026, 5.682114, 5.824499, 6.555983, 6.71353, 6.747835, 7.373468, 7.736402, 7.743548, 8.111, 8.393854, 9.171785, 9.313726, 9.789551, 10.28716, 10.63431, 11.08168, 12.01407, 12.10905\}$

Using equations (12) and (13), ML estimates are found to be  $\alpha = 2.705, \beta_0 = 6.539$  and  $\beta_1 = 245.0269$ . Using equations (6) and (7) the sensitivity and specificity of the test were also calculated: the sensitivity of the test is 94% and specificity is 89%. To the data generated above all the four methods were applied and compared (see Table 6). The non-parametric estimates are obtained by the method of Hanley and McNeil (1982), and the R codes are given in Appendix F.



**Figure 2.** Constant Shape Bi-Weibull ROC curve plotted for simulated data

## Conclusion

This article considered a ROC model developed from two parameter Weibull distributions for evaluating the accuracy of biomarkers in predicting disease status. It did not yield a closed form expression for area under the ROC curve. For this reason, equal shape parameter and different scale parameter were assumed. It should be noted that, the accuracy remains unchanged by this assumption. Hence, estimation of area under the constant shape Bi-Weibull ROC curve is a main objective for this study.

The Maximum Likelihood technique is adopted for estimating the parameters. The technique yielded an asymptotically unbiased estimate of the accuracy. The asymptotic distribution of  $A\hat{U}C$ ,  $SE(A\hat{U}C)$  and 95% confidence interval were found. The behavior of asymptotic SE and confidence interval is studied through simulation. The parametric AUC is higher than the AUC obtained by other methods including Monte Carlo, non-parametric and parametric bootstrap.

## References

- Amala, R., & Pundir, S. (2012). Statistical inference on AUC from a bi-lognormal ROC model for continuous data. *International Journal of Engineering Science and Innovative Technology*, 1(2): 283-295.
- Betinec, M. (2008). Testing the difference of the ROC curves in biexponential model. *Tatra Mountains Mathematical Publications*, 39: 215-223.
- Campbell, G., & Ratnaparkhi, M. V. (1993). An application of Lomax distributions in receiver operating characteristic (ROC) curve analysis. *Communication in Statistics*, 22(6): 1681-1687.
- Dorfman, D. D., et al. (1996). Proper receiver operating characteristics analysis: The bi-Gamma model. *Academic Radiology*, 4: 138-149.
- Hussain, E. (2011). The ROC curve model from generalized-exponential distribution. *Pakistan Journal of Statistics and Operations Research*, 7(2): 323-330.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29-36.
- Kundu, D., & Gupta, R. D. (2006). Estimation of  $P[Y < X]$  for Weibull Distribution. *IEEE Transactions on Reliability*, 55(2): 270-280.

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

Oglove, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology*, 5: 377-391.

Pundir, S., & Amala, R. (2012a). A study on the Bi-Rayleigh ROC model. *Bonfring International Journal of Data Mining*, 2(2): 42-47.

Pundir, S., & Amala, R. (2012b). A study on the comparison of bi-Rayleigh ROC model with bi-Gamma ROC model. In *Application of Reliability Theory and Survival Analysis*, 196-209. Coimbatore, India: Bonfring Publications.

Pundir, S., & Amala, R. (2013). Parametric receiver operating characteristic modeling for continuous data: A glance. *Model Assisted Statistics and Application*, 9(2): 121-135. doi: [10.3233/MASA-130284](https://doi.org/10.3233/MASA-130284).

Pundir, S. & Amala, R. (2014). Statistical Inference on AUC in Normal-Exponential ROC model. *Aligarh Journal of Statistics*, 34: 67-84.

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. New York, NY: John Wiley & Sons, Inc.

## Appendix A. R Code for Evaluation of AUC and Estimation of Standard Error Using Monte Carlo Simulation

```
m<-100; a0<-2.9753; a1=2.30387;
b0<-10295.0304;b1<-20646.898;x<-runif(m)
auc<-mean(exp(-(1/b1)* ( (-b0*log(x))^(a1/a0) ) ))
print(auc)
v.auc<-var(exp(-(1/b1)* ( (-b0*log(x))^(a1/a0) ) ))
print(v.auc); print(sqrt(v.auc))
```

## Appendix B. Evaluation of AUC

$$\int_0^y \int_0^x \frac{\alpha}{\beta_1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta_1}\right] * \frac{\alpha}{\beta_0} * x^{\alpha-1} * \text{Exp}\left[\frac{-x^\alpha}{\beta_0}\right] dx dy$$

$$\text{Conditional Expression} \left[ \frac{\beta_1}{\beta_0 + \beta_1}, \text{Re}\left[\frac{1}{\beta_0} + \frac{1}{\beta_1}\right] > 0 \&\& \text{Re}[\beta_1] > 0 \&\& \text{Re}[\alpha] > 0 \right]$$

## Appendix C. Evaluation of Asymptotic Distribution of AUC

$$\text{i) } E[X^\alpha] = \int_0^\infty x^\alpha * \frac{\alpha}{\beta_0} * x^{\alpha-1} * \text{Exp}\left[\frac{-x^\alpha}{\beta_0}\right] dx$$

$$\text{Conditional Expression} [\beta_0, \text{Re}[\alpha] > 0 \&\& \text{Re}[\beta_0] > 0]$$

$$\text{ii) } E[X^\alpha (\text{Log} X)^2] = \int_0^\infty x^\alpha (\text{Log}[x])^2 * \frac{\alpha}{\beta_0} * x^{\alpha-1} * \text{Exp}\left[\frac{-x^\alpha}{\beta_0}\right] dx$$

$$\text{Conditional Expression} \left[ \frac{1}{6\alpha^2} \beta_0 \begin{pmatrix} 6(-2 + \text{EulerGamma}) \\ \text{EulerGamma} + \pi^2 + 6\text{Log}[\beta_0] \\ (2 - 2\text{EulerGamma} + \text{Log}[\beta_0]) \end{pmatrix}, \right.$$

$$\left. \text{Re}[\alpha] > 0 \&\& \text{Re}[\beta_0] > 0 \right]$$

$$\text{iii) } E[Y^\alpha] = \int_0^\infty y^\alpha * \frac{\alpha}{\beta_1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta_1}\right] dy$$

$$\text{Conditional Expression} [\beta_1, \text{Re}[\alpha] > 0 \&\& \text{Re}[\beta_1] > 0]$$

# EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

- iv)  $E\left[Y^\alpha (\text{Log } Y)^2\right] = \int_0^\infty y^\alpha * (\text{Log}[y])^2 * \frac{\alpha}{\beta 1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta 1}\right] dy$   
 Conditional Expression  $\left[\frac{1}{6\alpha^2} \beta 1 \left( \frac{6(-2 + \text{EulerGamma})}{\text{EulerGamma} + \pi^2 + 6\text{Log}[\beta 1]} \right), \right.$   
 $\left. (2 - 2\text{EulerGamma} + \text{Log}[\beta 1]) \right)$   
 $\text{Re}[\alpha] > 0 \ \& \ \& \text{Re}[\beta 1] > 0]$
- v)  $E\left[X^\alpha \text{Log } X\right] = \int_0^\infty x^\alpha * \text{Log}[x] * \frac{\alpha}{\beta 0} * x^{\alpha-1} * \text{Exp}\left[\frac{-x^\alpha}{\beta 0}\right] dx$   
 $\beta 0 \left( -1 + \text{EulerGamma} + \text{Log}\left[\frac{1}{\beta 0}\right] \right)$   
 Conditional Expression  $\left[ -\frac{\beta 0 \left( -1 + \text{EulerGamma} + \text{Log}\left[\frac{1}{\beta 0}\right] \right)}{\alpha}, \right.$   
 $\left. \text{Re}[\alpha] > 0 \ \& \ \& \text{Re}[\beta 0] > 0] \right)$
- vi)  $E\left[Y^\alpha \text{Log } Y\right] = \int_0^\infty y^\alpha * \text{Log}[y] * \frac{\alpha}{\beta 1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta 1}\right] dy$   
 $\beta 1 \left( -1 + \text{EulerGamma} + \text{Log}\left[\frac{1}{\beta 1}\right] \right)$   
 Conditional Expression  $\left[ -\frac{\beta 1 \left( -1 + \text{EulerGamma} + \text{Log}\left[\frac{1}{\beta 1}\right] \right)}{\alpha}, \right.$   
 $\left. \text{Re}[\alpha] > 0 \ \& \ \& \text{Re}[\beta 1] > 0] \right)$
- vii)  $E\left[Y \log Y\right] = \int_0^\infty y * \text{Log}[y] * \frac{\alpha}{\beta 1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta 1}\right] dy$   
 Conditional Expression  $\left[ -\frac{1}{\alpha^2} \beta 1^{\frac{1}{\alpha}} \text{Gamma}\left[\frac{1}{\alpha}\right] \right.$   
 $\left. \left( \text{EulerGamma} - \text{HarmonicNumber}\left[\frac{1}{\alpha}\right] + \text{Log}\left[\frac{1}{\beta 1}\right] \right) \right),$   
 $\text{Re}[\alpha] > 0 \ \& \ \& \text{Re}[\beta 1] > 0]$

$$\begin{aligned} \text{viii)} \quad & E[Y \log Y] \int_0^\infty y * \text{Log}[y] * \frac{\alpha}{\beta 1} * y^{\alpha-1} * \text{Exp}\left[\frac{-y^\alpha}{\beta 1}\right] dy \\ & \text{ConditionalExpression}\left[-\frac{1}{\alpha^2} \beta 1^{\frac{1}{\alpha}} \text{Gamma}\left[\frac{1}{\alpha}\right] \right. \\ & \left. \left( \text{EulerGamma} - \text{HarmonicNumber}\left[\frac{1}{\alpha}\right] + \text{Log}\left[\frac{1}{\beta 1}\right] \right) \right], \\ & \text{Re}[\alpha] > 0 \ \& \ \text{Re}[\beta 1] > 0 \end{aligned}$$

- ix) The first order differentiation of  $\Gamma_n$  is given by  $\Gamma_n \psi(n)$  where  $\psi(n)$  is called the digamma function. The value of  $\Gamma'_n$  at  $n$  is equal to  $1-\gamma$ ; where  $\gamma$  is the Euler-Mascheroni constant has the approximate value 0.5772. The second order differentiation of  $\Gamma_n$  can be represented

$$\text{as } \Gamma''_n = \int_0^\infty x^{n-1} e^{-x} (\log x)^2 dx \text{ has the value } -1 + (1-\gamma)^2 + \frac{\pi^2}{6}.$$

In general the  $m^{\text{th}}$  derivative of  $\Gamma_n$  is obtained by

$$\Gamma''^m_n = \int_0^\infty x^{n-1} e^{-x} (\log x)^m dx.$$

$$\text{x)} \quad \psi(n)$$

$$\text{xi)} \quad -1 + (1-\gamma)^2 + \frac{\pi^2}{6}.$$

## Appendix D. Unbiasedness of Estimated AUC

An estimator  $T$  is said to be an unbiased estimator if it satisfies the condition  $E(T) = \mu$ . The estimated accuracy is

$$A\hat{U}C = \frac{\hat{\beta}_1}{\hat{\beta}_1 + \hat{\beta}_0} = \frac{\frac{\sum_{j=1}^n y_j^\alpha}{n}}{\frac{\sum_{j=1}^n y_j^\alpha}{n} + \frac{\sum_{i=1}^m x_i^\alpha}{m}}$$

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

Taking the expectation results in

$$E(\hat{AUC}) = \frac{\sum_{j=1}^n E(y_j^\alpha)}{\sum_{j=1}^n E(y_j^\alpha) + \sum_{i=1}^m E(x_i^\alpha)} = \frac{\beta_1}{\beta_1 + \beta_0} = AUC$$

Hence  $\hat{AUC}$  is an unbiased estimator of  $AUC$ .

### Appendix E. R Code for Evaluation of Bootstrap AUC and Confidence Interval

```
k<-10000 ;a10<-3.9; a11<-2.84;be1<-38.56;be0<-11.31;m<-30;n<-30;
df1 <-data.frame(array(dim=c(n,k))); df0 <-data.frame(array(dim=c(m,k)))
dfw0<-data.frame(array(dim=c(m,k))); dfw1<-data.frame(array(dim=c(n,k)))
a<-array(dim=k); ave<-array(dim=k); b1<-array(dim=k); b0<-array(dim=k)
auc<-array(dim=k); SE<-array(dim=k); for(i in 1:k)
{
df1[i]<-runif(30); df0[i]<-runif(30);
dfw0[i]<-(-be0*log(1-df0[i]))^(1/a10);
dfw1[i]<-(-be1*log(1-df1[i]))^(1/a11);
loglik<-function(param)
{
a[i]<-param[1]; b0[i]<-param[2]; b1[i]<-param[3]
ll<-(m+n)*log(a[i])+(a[i]-1)*(sum(log(dfw1[i]))+sum(log(dfw0[i]))) -n*log(b1[i])-
m*log(b0[i])-(sum(dfw1[i]^a[i])/b1[i])-(sum(dfw0[i]^a[i])/b0[i])
ll
}
}
M0<-maxNR(loglik,start=c(1,2,3))
a[i]<-M0$estimate[1]; b0[i]<-M0$estimate[2]; b1[i]<-M0$estimate[3]
auc[i]<-(b1[i]/(b1[i]+b0[i])); dt<-data.frame(a[i],b0[i],b1[i],auc[i])
}
print(dt); b.auc<-mean(auc); b.se.auc<-sd(auc);
cat("Bootstrap Accuracy=", "\n", b.auc, "\n")
cat("Bootstrap Standard Error=", "\n", b.se.auc)
lcl<-(b.se.auc-(1.96*b.se.auc)); ucl<-(b.se.auc+(1.96*b.se.auc))
```

### Appendix F. R code for Sensitivity and Specificity Analysis

```
s<-sort(c(h,d)); n<-length(d); m<-length(h); X<-array(dim=m+n-1)
k<-m+n-1;
for(i in 1:k)
{
X[i]<-(s[i]+s[i+1])/2;
```



```

    }
t<-c(s[1]-1,X,s[m+n]+1); print(t); a<-2.705492 ; b0<-6.538623 ;
b1<-245.026947; # Estimated by MLE from data
Sen<-exp((-t^a)/b1); Sp<-1-exp((-t^a)/b0); dt<-data.frame(t, Sen, Sp);

```

## Appendix G. R code for Non-Parametric Method

```

NP.ROC<-function(h,d) # Creating a function named NP.ROC()
{
  s<-sort(c(h,d)); n<-length(d); m0<-mean(h);m1<-mean(d);m<-length(h)
  X<-array(dim=m+n-1);k<-m+n-1;
  for(i in 1:k)
  {
    X[i]<-(s[i]+s[i+1])/2;
  }
  t<-c(s[1]-1,X,s[m+n]+1); print(t);
  TPR<-array(dim=length(t)) # Defining empty array to save calculations
  FPR<-array(dim=length(t)); TP<-array(dim=length(t)); TN<-array(dim=length(t));
  FN<-array(dim=length(t)); FP<-array(dim=length(t)); AUC<-array(dim=length(t));
  SP<-array(dim=length(t)); TNR<-array(dim=length(t)); SplusS<-
  array(dim=length(t));
  se<-array(dim=length(t));q1<-array(dim=length(t));q2<-array(dim=length(t));v<-
  array(dim=length(t));
  for(i in 1:length(t))
  {
    A<-d[d>=t[i]]# observations greater than or equal to t among diseased i.e. True
    Positives
    B<-d[d<t[i]] # observations less than t among diseased i.e. False Negatives
    C<-h[h>=t[i]]# observations greater than or equal to t among healthy i.e. False
    Positives
    D<-h[h<t[i]] # observations less than t among healthy i.e. True Negatives
    TP[i]<-length(A) # No. of TPs
    FP[i]<-length(C) # No. of FPs
    FN[i]<-length(B) # No. of FNs
    TN[i]<-length(D) # No. of TNs
    TPR[i]<-(TP[i]/n)
    FPR[i]<-(FP[i]/m)
    TNR[i]<-1-FPR[i] # or TN[i]/m
    AUC[i]<-(TP[i]+TN[i])/(TP[i]+TN[i]+FN[i]+FP[i])
    SplusS[i]<-TPR[i]+TNR[i] # TNR+TPR
    q1[i]<-AUC[i]/(2-AUC[i]); q2[i]<-(2*AUC[i]^2)/(1+AUC[i])
    v[i]<-(AUC[i]*(1-AUC[i])+(n-1)*(q1[i]-AUC[i]^2)+(m-1)*(q2[i]-AUC[i]^2))/(m*n)
    se[i]<-sqrt(v[i]);
  }
  library(utils)
  write.csv(dt,"msanalysis.csv") # writing the data frame in CSV format for usage
  m<-length(h); n<-length(d); l<-m*n;
  sum=0;
  for(i in 1 : m)
  {
    s<-c(0);

```

## EVALUATION OF AREA UNDER BI-WEIBULL ROC CURVE

```

        for(j in 1 : n)
        {
            if(d[j]>h[i])
            {
                s[j]=1;
            }
            else if (d[j]==h[i])
            {
                s[j]=0.5;
            }
            else
            s[j]=0
        }
        output=data.frame(s)
        sum=sum+sum(output)
    }
    value= sum/(m*n)
    print(value)
    dt<-data.frame (t, FPR, TPR, TP, TN, FP, FN, AUC, se)
    print(dt); Q1<-value/(2-value); Q2<-(2*value^2)/(1+value)
    V<-(value*(1-value)+(n-1)*(Q1-value^2)+(m-1)*(Q2-value^2))/(m*n)
    SE<-sqrt(V); # Standard Error of AUC
    lc<-value-(SE*1.96) # Lower Confidence Limit of AUC
    uc<-value+(SE*1.96) # Upper Confidence Limit of AUC
    if(uc>1){ # Sometimes if the standard error is high, the upper CI may go greater
    that one in which case approximating it to one.
    uc<-1.0
    }
    cat("-----",
    "\n", "Healthy Mean","\t",":","\t",m0,
    "\n", "Diseased Mean","\t",":","\t",m1,
    "\n", "AUC","\t","\t",":","\t", value,
    "\n", "SE","\t","\t",":","\t", SE,
    "\n", "CI","\t","\t",":","\t", "[", lc, ",", "\t",uc,"]",
    "\n", "-----","\n")
    plot(TPR~FPR,type="b",main="",xlab="FPR",ylab="TPR",xlim=c(0,1),ylim=c(0,1))
    abline(lm(c(0:1)~c(0,1)))
    }
    NP.ROC(h,d);

```

# Hierarchical Clustering with Simple Matching and Joint Entropy Dissimilarity Measure

**A. Mete Çilingirtürk**  
Marmara University  
Istanbul, Turkey

**Özlem Ergüt**  
Marmara University  
Istanbul, Turkey

---

Conventional clustering algorithms are restricted for use with data containing ratio or interval scale variables; hence, distances are used. As social studies require merely categorical data, the literature is enriched with more complicated clustering techniques and algorithms of categorical data. These techniques are based on similarity or dissimilarity matrices. The algorithms are using density based or pattern based approaches. A probabilistic nature to similarity structure is proposed. The entropy dissimilarity measure has comparable results with simple matching dissimilarity at hierarchical clustering. It overcomes dimension increase through binarization of the categorical data. This approach is also functional with the clustering methods, where a-priori cluster number information is available.

*Keywords:* Categorical data, clustering, dissimilarity, entropy

---

## Introduction

Clustering analysis is a process used for classifying objects so that homogeneous subsets are built in heterogeneous groups. A variety of distance/similarity criteria are used when classifying objects in groups according to their similarity. One important criterion for choosing the distance or similarity measure, when classifying objects into groups, is the type of the data. In the literature it can be seen that most studies examine the clustering of continuous data. If the data set consists of continuous data, Euclid and Manhattan are the distance measures most widely used in applications. However, in a data set with categorical data it is not possible to use this type of distance measures. These variables are first transformed into binary data and then the analysis is applied, which increases the

---

*A. Mete Çilingirtürk is Professor of Statistics in the Department of Econometrics. Email him at [acilingi@marmara.edu.tr](mailto:acilingi@marmara.edu.tr). Özlem Ergüt is an Assistant in the Department of Economics. Email at [ozlem.ergut@marmara.edu.tr](mailto:ozlem.ergut@marmara.edu.tr).*

## HIERARCHICAL CLUSTERING WITH SIMPLE MATCHING

number of dimensions when there are multinomial variables in the data. This procedure increases memory allocation.

There are different techniques and approaches for finding clusters with categorical data. One includes transformation of categorical variable into dummy variable, independently from calculation of distances. Applications of such hierarchical method algorithms are single linkage, complete linkage, average linkage, etc. (Chaturvedi et al., 2001).

Another approach uses the  $k$ -means algorithm for clustering of categorical data developed by Ralambondrainy (1995). In this approach multiple category attributes are turned into binary variables, which are assumed to be numeric variables and thus, the  $k$ -means algorithm is applied. The drawback of this approach is the increase in the number of binary variables when there are too many categories in variables. Further, cluster centres, given as 0 and 1, do not reflect the real characteristics of clusters (Huang, 1998). The basics of  $K$ -medoids algorithm is founded on finding  $k$  number of objects representative of several structural features of data. A Medoid is the most central point of the cluster with minimum average distance to other objects that are located in the same cluster (Kaufman and Rousseeuw, 2005; Xu and Wunsch, 2009). Due to the distance measure used in  $K$ -means algorithm, this method is not used in clustering categorical data. As the data set consists of categorical data,  $k$ -modes method, which is an extension of  $k$ -means model, is used for clustering categorical data, which was developed by Huang (1998). In this algorithm,

1. simple matching dissimilarity measure for categorical objects,
2. mod is used for clusters instead of mean,
3. frequency-based method is used for updating modes (Huang, 1998).

An extended-modes algorithm was proposed by Aranganayagi and Thangwell (2010), which uses a probability weighted single matching dissimilarity function.

Initially, expectation maximization algorithm assigns randomly different possibilities to each class or category. These probabilities are determined with consecutive iterations so as to maximize the similarity value of the data, which will also fit a pre-set number of clusters. The EM algorithm assumes that the model is suitable for a non-observable latent variable and that the stochastic model performs maximum likelihood estimations of the parameters (Agarwal et al., 2010). The optimization algorithm determines the convergence of the parameters.

ROCK (RObust Clustering using linKs) is an adaptation of the hierarchical clustering algorithm developed for clustering of categorical data. In this algorithm similarity value between two objects is calculated using Jaccard coefficient, then the threshold value ( $\theta$ ), defined between 0 and 1 by the researcher, is compared to decide adjacent points. In order that a given point  $q_i$  is adjacent to a point  $q_j$  for an  $i^{\text{th}}$  object in an  $m$ -dimensional space, similarity value has to exceed threshold value ( $\theta$ ) (Guha et al., 1999).

$$\text{sim}(q_i, q_j) \geq \theta$$

If this condition is met, it can be said that the points are neighbours. This algorithm classifies the objects into clusters according to their link ability. The link ability between two clusters gives the number of common adjacent points between  $q_i$  and  $q_j$ . The higher the linkability of  $q_i$  and  $q_j$ , the higher is the possibility of  $q_i$  and  $q_j$  being in the same cluster.

COOLCAT is proposed for categorical clustering analysis as an entropy-based algorithm (Barbara et al., 2002). The entropy-based algorithm consists of two steps, namely initialization and incremental steps. In the initialization step  $K$  most dissimilar records are selected from the sample. In the next step remaining records in the data set are assigned to appropriate clusters. The algorithm groups objects in the data set trying to minimize the expected entropy of the clusters. Similarly, He et al. (2005) maximized Ensemble algorithms with the average normalized mutual information [0,1] function based on entropy in separating of units with the purpose of categorical clustering.

## Definitions and Notations

$X$  and  $Y$  are two categorical objects defined by  $n$  and  $m$  attributes, the dissimilarity measure between  $X$  and  $Y$  is the sum of mismatches in relevant variable attributes of the two objects. The smaller the number of mismatches, the more similar are two objects. This measure is also a kind of generalized Hamming distance (Ng et al., 2007).

$$d(X, Y) = \sum_{k=1}^m \delta(x_k, y_k) \quad (1)$$

$$\delta(x_k, y_k) = \begin{cases} 0 & (x_k = y_k) \\ 1 & (x_k \neq y_k) \end{cases} \quad (2)$$

As statistics is applied to physics, the development of statistical physics earned entropy new meanings with entropy, which is an indicator of the irregularity and uncertainty in a physical system. The increase in irregularity in the system is proportionate to the increase in entropy. The uncertainty of occurrence of  $x_i$  situation in system  $X$ , which is the entropy of  $x_i$  situation, is shown as  $c(p_i) = -\log p(x_i)$ , while the entropy of the system is expressed as (Roy, 2002; Müller, 2003)

$$H(X) = -\sum_{i=1}^n P(x) \log P(x). \quad (3)$$

As the logarithmic operations are performed, the entropy becomes an additive quantity for independent systems (Georgii, 2003).

For a given  $n$ , when  $p(x_1) = p(x_2) = \dots = p(x_n) = \frac{1}{n}$ ,

$$\begin{aligned} H_{max} &= -\sum \frac{1}{n} \log \frac{1}{n} = -n \frac{1}{n} \log \frac{1}{n} \\ H_{max} &= \log(n) \end{aligned} \quad (4)$$

is obtained. This means that  $H$  reaches its maximum value when it is equal to  $\log(n)$ . When a two-dimensional  $(X, Y)$  random variable is in question,  $P$  joint probability matrix becomes  $P = \{p_{ij}\} = P(X = x_i, Y = y_j)$  and thus the entropy becomes

$$H(i, j) = -\sum_{i=1}^n \sum_{j=1}^m p_{ij} \log p_{ij}. \quad (5)$$

The uncertainty coefficient calculated asymmetrically and symmetrically based on entropy in cross-tables is more appropriate for use. The uncertainty coefficient for symmetric structures is calculated as

$$U(X, Y) = 2 \left[ \frac{H(x_i) + H(y_j) - H(x_i, y_j)}{H(x_i) + H(y_j)} \right]. \quad (6)$$

## Proposed Method

Taking observation units as variables, the proposed method ensures that calculation of combined entropy values remains on the same constant ( $\log m$ ) for  $m$  number of categorical attributes.

The  $S$  matrix, which shows that  $n$  number of objects take identical values, provides the basis of entropy dissimilarity measure approach, unlike the simple matching dissimilarity measures matrix. Each row/column in this matrix shows the number of similar objects for each  $m$  variables. Therefore each row/column of the matrix is the frequency distribution of its similarity with another observation. The uncertainty coefficient given in equation (6) aims that a single value is generated for a cross-table; thus, the formula has been organized with the help of the following equations with the purpose of measuring uncertainty based on entropy.

$$\begin{aligned} H(i.) &= p_{i.} \log(p_{i.}), \\ H(.j) &= p_{.j} \log(p_{.j}) \text{ and} \\ H(i, j) &= p_{ij} \log(p_{ij}). \end{aligned} \quad (7)$$

If  $X$  and  $Y$  are independent random variables, combined entropy is equal to the sum of the entropies of these two random variables

$$\begin{aligned} H(i, j) &= H(i.) + H(.j), \\ U(i, j) &= \begin{cases} 2 \left[ \frac{H(i) + H(j) - H(i, j)}{H(i) + H(j)} \right] & p_{ij} \neq 0. \\ 2 & p_{ij} = 0 \end{cases} \end{aligned} \quad (8)$$

Equation (8) displays a symmetric dissimilarity matrix which does not consist of constant values: the reason for this is that the entropy of an object with itself depends on the frequency of encountering the characteristics in the total

## HIERARCHICAL CLUSTERING WITH SIMPLE MATCHING

distribution. The uncertainty of an object with frequently observable characteristics will be proportionately low. If two objects have no common features,  $p_{ij} = 0$  and as logarithm is non-defined, maximum entropy dissimilarity value of  $-2$  is used. However, as algorithm software used for clustering will accept a symmetric dissimilarity matrix with constant diagonal (0),  $U(i,j)$  values are proportioned and corrected in 0-1 interval.

$$U^*(i,j) = \frac{U(i,j) - \text{diag}U(i,j)}{2 - \text{diag}U(i,j)} \quad (9)$$

The numerator of fraction brings the diagonal values, which are the smallest values of each row and column, to zero, whereas denominator proportions the dissimilarity of other values according to the maximum value and earns the value 1 for maximum dissimilarity.

### Empirical Results

Dissimilarity matrices were formed based on simple matching dissimilarity measure and entropy in this article. The results obtained by using hierarchical methods in both dissimilarity matrices were compared with each other. The data used in the study was Teaching Assistant data obtained from UCI database (Loh, W. -Y. & Lim, T. -S., 1997). It was collected for evaluation of the performances of 151 research assistants at statistics department of Wisconsin-Madison University during three semesters and two summer schools. The scores were divided into 3 roughly equal-sized categories (low, medium, high) to form the class variable. The four variables chosen for determining the performance of 151 research assistants is:

1. Whether or not the TA is a native English speaker? (2 categories)
2. Course instructor (25 categories)
3. Course (26 categories)
4. Summer or regular semester (2 categories)

Within the scope of the study, Stata 11.0 program was used for application of hierarchical methods for entropy and simple matching dissimilarity measures. The results obtained from simple matching dissimilarity measure and hierarchical methods using single linkage, complete linkage and average linkage methods



were interpreted. In single linkage method, the two closest objects or clusters (minimum distance or biggest similarity) using distance/similarity values are combined. In complete linkage, the maximum of the distance between the new cluster formed after combining two clusters (objects) and the other cluster is taken. In the average linkage method, which is suggested as an alternative as it provides results between these two extreme techniques, the distance between two clusters is equal to the average values of the distances between observed couples located in two clusters.

One of the measures used in evaluating the success and quality of clustering results is F measure. This measure consists of a combination of precision and recall measures. F measure is basically the harmonic mean of precision and recall (Işık and Çamurcu, 2007). F measure, which is one of the measures that ensures (i) comparison of the classification which is known in advance and the clusters obtained as a result of clustering analysis (Loh and Shin, 1997) and (ii) evaluation of clustering, is calculated as follows for j.cluster and i.class.

$$F(i, j) = \frac{2 * r(i, j) * p(i, j)}{r(i, j) + p(i, j)}$$

where  $r$  means recall and  $p$  means precision.

$$r(i, j) = \frac{n_{ij}}{n_i} \quad p(i, j) = \frac{n_{ij}}{n_j}$$

In  $n_{ij}$ , the number of observations in j.cluster and i.class, namely  $n_j$  and  $n_i$ , are respectively the magnitudes of j.cluster and i.class. Total  $F$  measure for a data set consisting of  $n$  number of observations is calculated as follows (Dalli, 2003):

$$F = \sum_i \frac{n_i}{n} \max [F(i, j)]$$

If single linkage is used with simple matching dissimilarity measure, as there are considerable number of connections, observations are not classified into clusters and combined in a single cluster. In complete linkage method while observations are assigned to maximum three clusters; however, if average linkage method is preferred, observations can form maximum 34 clusters but the  $F$

## HIERARCHICAL CLUSTERING WITH SIMPLE MATCHING

measure obtained in the case that there are three clusters is as,  $F = 0,360$  with simple matching dissimilarity and  $F = 0,387$  with entropy dissimilarity.

In single linkage, which is one of the three hierarchical methods, observations are classified into four clusters in the case that entropy dissimilarity is used. In the case that there are three clusters, 146 of the observations are assigned to the first cluster, four are assigned to the second cluster and one is assigned to the third cluster. In simple matching, dissimilarity observations cannot be classified into clusters, whereas clusters consisting of small number of observations occur in entropy dissimilarity. If the same measure is used in complete linkage method, as there are considerable number of connections, observations are not classified into clusters and combined in a single cluster. In average linkage method, however, observations are concentrated in the first cluster if there are three clusters.

Performance in the data set was evaluated in three categories namely good, mediocre and poor. The results obtained according to both dissimilarity measures and two were compared with these three categories, the level of concordance were determined. Accordingly,

In the average linkage method, 49 observations were correctly assigned (33 percent) if entropy dissimilarity measure was used.

In the average linkage method, 47 observations were correctly assigned (31 percent) if simple matching dissimilarity measure was used.

In the average linkage method, the F measure value obtained using simple matching dissimilarity, entropy measure were 0.36 and 0.38, respectively.

## Conclusion

In categorical data, with the exception of data mining algorithms, clustering algorithms are applied with two-step clustering method and simple matching measure is used. Two-step clustering first digitalizes the categorical variables and then performs distance calculations. Parameter estimations require optimized solutions with iterations. The simple matching method however does not take the frequency of observing a certain characteristic in categorical variables and the possibility of a unit for having this unique characteristic into the consideration.

The selection of distance and/or similarity measure lies in the foundation of all clustering methods. The findings are based on the selection of both clustering

methods and distance measure. Therefore, this study offers an estimation of entropy matrix based on dissimilarity of categorical variables. The method also provides a solution to the problem of increase in the number of variables by using dummy variable in the case of existence of categorical variables. The study can also be used for developing a different clustering algorithm with a non-constant diagonal, which therefore will take into consideration the low level of uncertainty that is caused by having frequently encountered characteristics.

## References

- Agarwal, P., Alam, M. A., & Biswas, R. (2010). Analyzing the Agglomerative Hierarchical Clustering Algorithm for Categorical Attributes. *International Journal of Innovation, Management and Technology*, 1: 186-190.
- Aranganayagi, S., & Thangavel, K. (2010). Extended K-modes with Probability Measure. *International Journal of Computer Theory and Engineering*, 2: 431-435.
- Barbara, D., Couto, J., & Li, Y. (2002). COOLCAT: An Entropy-based Algorithm for Categorical Data. In C. Nicholas (Chair). *Proceedings of the 11th International ACM Conference on Information and Knowledge Management*, McLean, VA, pp. 582-589.
- Chaturvedi, A., Green, P. E. & Carroll, J. D. (2001). K-modes Clustering. *Journal of Classification*, 18: 35-55.
- Dalli, A. (2003). Adaptation of the F-Measure to Cluster Based Lexion Quality Evaluation. *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Available at <http://aclweb.org/anthology/W/W03/W03-2807.pdf>
- Georgii, H. (2003). Probabilistic Aspects of Entropy. In A. Greven, G. Keller & G. Warnecke (Eds.). *Entropy*. New Jersey: Princeton University Press, pp. 37-52.
- Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A Robust Clustering Algorithm for Categorical Attributes. In M. Kitsuregawa, L. Maciaszek, M. Papazoglou & C. Pu (Eds.). *Proceedings of the 15<sup>th</sup> IEEE International Conference on Data Engineering*, Sydney, Australia. Available at [theory.stanford.edu/~sudipto/mypapers/categorical.pdf](http://theory.stanford.edu/~sudipto/mypapers/categorical.pdf).
- He, Z., Xu, X., & Deng, S., (2005). A Cluster Ensemble Method for Clustering Categorical Data. *Information Fusion*, 6: 143-151.

## HIERARCHICAL CLUSTERING WITH SIMPLE MATCHING

Huang, Z. (1998). Extensions to the  $k$ -Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 304: 283-304.

Işık, M., & Çamurcu, A. Y. (2007). K-Means, K-Medoids ve Bulanık C-Means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 6: 31-45.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons.

Loh, W. -Y, & Shin, Y. -S (1997). Split Selection Methods for Classification Trees, *Statistica Sinica*, 7: 815-840. Available at [citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.7375\[1\].pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.7375[1].pdf)

Loh, W. -Y. & Lim, T. -S. (1997). Teaching Assistant Evaluation Data Set. *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. Available at <http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation>

Müller, I. (2003). Entropy: A Subtle Concept in Thermodynamics. In A. Greven, G. Keller & G. Warnecke (Eds.). *Entropy*. New Jersey: Princeton University Press, pp. 19-35.

Ng, M., Li, M. J., Huang, J. Z., & He, Z. (2007). On the Impact of Dissimilarity Measure in  $k$ -Modes Clustering Algorithm. *Pattern Analysis and Machine Intelligence*, 29: 503-507. Available at <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2007.53>.

Ralambondrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, 16: 1147-1157.

Roy, B. N. (2002). *Fundamentals of Classical and Statistical Thermodynamics*. New York: John Wiley and Sons.

Xu, R., & Wunsch, D. (2009). *Clustering*. New York: John Wiley and Sons.

# Change Point Estimation for Pareto Type-II Model

**Gyan Prakash**  
S. N. Medical College  
Agra, U.P., India

---

Some Bayes estimators of the change point for the Pareto Type-II model under right item failure-censoring scheme are proposed. The Bayes estimators are obtained here in two cases, the first is when one parameter is known and second when both parameters are considered as the random variable. The performances of the procedures are illustrated by simulation technique.

*Keywords:* Change point, Pareto Type-II model, Bayes estimation

---

## Introduction

The Pareto distribution and its close relatives provide a flexible family of fat-tailed distributions, which may be used as a model for income distribution of higher income group and in socio-economic studies. This distribution has played important role in variety of other problems such as size of cities and firms, business mortality, service time in queuing system. It is often used as a model for analyzing areas including city population distribution, stock price fluctuation, oil field locations and military areas.

It has been found to be suitable for approximating the right tails of distribution with positive skewness. Pareto distribution has a decreasing failure rate, so it has often been used for model survival after some medical procedures (the ability to survive for a longer time appears to increase, the longer one survives after certain medical procedures).

Harries (1968) used this distribution in determining times of maintenance service while Dyer (1981) found that two-parameter Pareto distribution transformation is equivalent to the two-parameter exponential distribution. Madi & Raqab (2004) discussed about the forecasting of the temperatures records by

---

*Gyan Prakash is an Assistant Professor in the Department of Community Medicine.  
Email him at: [ggyanji@yahoo.com](mailto:ggyanji@yahoo.com).*

Pareto distribution. Singh et al (2007) discussed about different types of test-estimation for the Pareto Model. The length of Bayes prediction limits have been obtained recently by Prakash & Singh (2013) for the Pareto model. Panahi & Asadi (2011) presented stress-Strength model for a Lomax distribution. Some inferences regarding the Lomax distribution under the generalized order statistics has discussed by Moghadam et al. (2012). Nasiri & Hosseini (2012) presented Bayesian and classical statistical inferences for Lomax model based on record values. Recently, Al-Zahrani & Al-Sobhi (2013) presents some parameter estimation for Lomax distribution under general progressive censoring criterion.

The probability density function of the considered Pareto Type-II model is given as

$$f(x; \sigma, \theta) = \theta \sigma^\theta (x + \sigma)^{-(\theta+1)} ; x \geq 0, \theta > 0, \sigma > 0 \quad (1)$$

Here,  $\theta$  is the shape parameter and  $\sigma$  is the scale parameter. The proposed Pareto Type-II model is the result of mixture of the Exponential distribution with the parameter  $\alpha$ , and the exponential scale parameter  $\alpha$  is distributed as a Gamma with parameters  $\theta$  and  $\sigma$ .

This article discusses the Bayes estimation of change point for Pareto Type-II model. The Bayes estimator has been obtained under the right item failure censoring criteria in two cases: the first is when the scale parameter is known and second when both parameters are considered as the random variable. A numerical study was carried out for illustration of the procedures in next section by MCMC technique.

## The Change Point

In order to obtain information on their endurance, manufactured items such as mechanical or electronic components are often put to life tests and life times are observed periodically. Physical systems manufacturing the items are often subject to random fluctuations. It may happen that at some point of time, there is a change in the parameter. The objective of study is to find out when and where this change has started occurring, which is called the change point inference problem.

Bayesian model may play an important role in the study of such change point estimation problem and have been studied by Broemeling & Tsurumi (1987), Jani & Pandya (1999), Ebrahimi & Ghose (2001), Goldenshluger, et al. (2006). Pandya & Jadav (2010) presents Bayesian estimation of change point in mixture of left truncated exponential and degenerate distribution. Some Bayes estimation

of shift point in Poisson model was presented by Srivastava (2012). Recently, Pandya (2013) presented Bayes estimation of auto regressive model with change point.

Consider a sequence of independent random sample of size  $n(\geq 3)$  such as  $x_1, x_2, \dots, x_{m-1}, x_m, x_{m+1}, \dots, x_n$  from the considered model with survival function  $\psi_1(t)$  at time  $t (> 0)$  but later it is found that there is a change in the system at some point of time  $m$  and it is reflected in the sequence after the observation  $x_m$  by the change in the survival function. The probability density function and survival function of the first  $m$  observations  $x_1, x_2, \dots, x_m$  are given from model (1) as:

$$f(x_i; \sigma, \theta_1) = \theta_1 \sigma^{\theta_1} (x_i + \sigma)^{-(\theta_1+1)}; x_i \geq 0, \theta_1 > 0, \sigma > 0, i = 1, 2, \dots, m \quad (2)$$

and

$$\psi_1(t) = \sigma^{\theta_1} (t + \sigma)^{-\theta_1}; t > 0, \theta_1 > 0, \sigma > 0.$$

Similarly, the probability density function and survival function of remaining  $(n-m)$  components  $x_{m+1}, x_{m+2}, \dots, x_n$  are

$$\begin{aligned} f(x_i; \sigma, \theta_2) &= \theta_2 \sigma^{\theta_2} (x_i + \sigma)^{-(\theta_2+1)}; \\ x_i \geq 0, \theta_2 > 0, \sigma > 0, i &= m+1, m+2, \dots, n \end{aligned} \quad (3)$$

and

$$\psi_2(t) = \sigma^{\theta_2} (t + \sigma)^{-\theta_2}; t > 0, \theta_2 > 0, \sigma > 0.$$

In life testing, the observations usually occur in ordered manner such that the weakest items fail first and then second one and so on. Suppose that  $n$  items are put to test under the considered model without replacement and only  $k (\leq n)$  items are fully measured, while the remaining  $(n-k)$  items are censored. These  $(n-k)$  censored items will be ordered separately. This censoring scheme is known as the right item failure-censoring criteria.

The change point criteria was introduced inside the right item-censoring scheme; assume a sequence of ordered independent random sample of size  $n$  such as  $x_{(1)}, x_{(2)}, \dots, x_{(k-1)}, x_{(k)}, x_{(k+1)}, \dots, x_{(n)}$  from the model (1), with the

## CHANGE POINT ESTIMATION FOR PARETO TYPE-II MODEL

parameters  $\theta_1$  and  $\sigma$ . All  $n$  items are tested without replacement and first  $k$  ordered items are fully measured while remaining  $(n-k)$  items are censored. From the first fully measured  $k$  ( $=x_{(1)}, x_{(2)}, \dots, x_{(k-1)}, x_{(k)}$ ) items, it is found that there is a change in the system at some point of time  $m$  and it is reflected in the sequence after  $x_{(m)}$  ( $m \leq k$ ) by the change in the survival function.

The probability density function of first  $m$  ( $m \leq k \leq n$ ) random samples  $x_{(1)}, x_{(2)}, \dots, x_{(m)}$  with parameters  $\theta_1$  and  $\sigma$ , are

$$f(x_{(i)}; \sigma, \theta_1) = \theta_1 \sigma^{\theta_1} (x_{(i)} + \sigma)^{-(\theta_1+1)}; \quad (4)$$

$$x_{(i)}, \theta_1, \sigma > 0, i = 1, 2, \dots, m (m \leq k, k \leq n).$$

The first remaining group of random samples  $x_{(m+1)}, x_{(m+2)}, \dots, x_{(k)}$  with size  $(k-m)$  using a considered Pareto model has a probability density function with parameters  $\theta_2$  and  $\sigma$

$$f(x_{(i)}; \sigma, \theta_2) = \theta_2 \sigma^{\theta_2} (x_{(i)} + \sigma)^{-(\theta_2+1)}; \quad (5)$$

$$x_{(i)}, \theta_2, \sigma > 0, i = m+1, m+2, \dots, k (k \geq m, k \leq n).$$

The last remaining group of random samples  $x_{(k+1)}, x_{(k+2)}, \dots, x_{(n)}$  of size  $(n-k)$  distributed again a Pareto model with parameters  $\theta_1$  and  $\sigma$  – has the probability density function

$$f(x_{(i)}; \sigma, \theta_1) = \theta_1 \sigma^{\theta_1} (x_{(i)} + \sigma)^{-(\theta_1+1)}; \quad (6)$$

$$x_{(i)}, \theta_1, \sigma > 0, i = k+1, k+2, \dots, n (\geq k).$$

Under the above scenario the likelihood function for the random sample  $\underline{x} (=x_{(1)}, x_{(2)}, \dots, x_{(n)})$  is defined as

$$L(\underline{x} | \theta_1, \theta_2, \sigma, m) = \left( \prod_{i=1}^m f(x_{(i)}; \sigma, \theta_1) \right) \cdot \left( \prod_{i=m+1}^k f(x_{(i)}; \sigma, \theta_2) \right) \\ \cdot \prod_{i=k+1}^n \left( 1 - \int_0^{x_{(i)}} f(x_{(i)}; \sigma, \theta_1) dx_{(i)} \right)$$



$$\Rightarrow L(\underline{x} | \theta_1, \theta_2, \sigma, m) = T_0 \theta_1^m \theta_2^{k-m} \sigma^{\theta_1(n-k+m) + \theta_2(k-m)} e^{-\theta_1 T_1} e^{-\theta_2 T_2}; \quad (7)$$

where  $T_0 = \prod_{i=1}^k (x_{(i)} + \sigma)^{-1}$ ,  $T_1 = \sum_{i=1}^m \log(x_{(i)} + \sigma) + \sum_{i=k+1}^n \log(x_{(i)} + \sigma)$  and  $T_2 = \sum_{i=m+1}^k \log(x_{(i)} + \sigma)$ .

### Remark

1. Substitute  $\theta_1 = \theta = \theta_2$  in (7)

$$L(\underline{x} | \theta, \sigma) = T_0 \theta^k \sigma^{n\theta} e^{-\theta T_3}; T_3 = \sum_{i=1}^n \log(x_{(i)} + \sigma).$$

Here,  $L(\underline{x} | \theta, \sigma)$  shows the likelihood function under the right item-failure censoring criterion without consideration of change point.

2. Substitute  $\theta_1 = \theta = \theta_2$  and  $k = n$  in (7) to obtain the likelihood function for complete sample case without consideration of change point.

$$L(\underline{x} | \theta, \sigma) = T_0^* \theta^n \sigma^{n\theta} e^{-\theta T_3}; T_0^* = \prod_{i=1}^n (x_{(i)} + \sigma)^{-1}.$$

### Change Point Estimation (Scale Parameter Is Known)

From a Bayesian viewpoint; there is clearly no way in which one can say that one prior is better than other. It is more frequently the case that, that a prior is selected to restrict attention to a given natural family of priors, and one is chosen from that family, which seems to match best with one's personal beliefs. A natural family of conjugate prior for shape parameter  $\theta$  is considered here as a Gamma distribution (when scale parameter is known) with probability density function

$$g(\theta | \sigma) \propto \theta^{a-1} e^{-\beta\theta}; \theta > 0, a > 0, \beta > 0. \quad (8)$$

Based on change point criterion the prior density (8) is re-parameterized as

## CHANGE POINT ESTIMATION FOR PARETO TYPE-II MODEL

$$g_j(\theta_j | \sigma) \propto \theta_j^{a_j-1} e^{-\beta_j \theta_j} ; \theta_j > 0, a_j > 0, \beta_j > 0, j=1, 2. \quad (9)$$

A discrete uniform over the set  $(1, 2, \dots, k-1)$ , is considered as the prior distribution of change point  $m$  and defined as

$$g_3(m) = \frac{1}{k-1}. \quad (10)$$

The joint prior distribution when scale parameter is considered to be known, is defined as

$$h_1(\theta_1, \theta_2, m) = g_1(\theta_1 | \sigma) \cdot g_2(\theta_2 | \sigma) \cdot g_3(m).$$

The joint posterior density function is now obtained as

$$\begin{aligned} \pi_1(\theta_1, \theta_2, m | \underline{x}) &= \frac{L(\underline{x} | \theta_1, \theta_2, \sigma, m) \cdot h_1(\theta_1, \theta_2, m)}{\sum_m \int_{\theta_1} \int_{\theta_2} L(\underline{x} | \theta_1, \theta_2, \sigma, m) \cdot h_1(\theta_1, \theta_2, m) d\theta_2 d\theta_1} \\ \Rightarrow \pi_1(\theta_1, \theta_2, m | \underline{x}) &= \bar{\sigma} \theta_1^{m+\theta_1-1} \theta_2^{k-m+\theta_2-1} e^{-T_1^* \theta_1} e^{-T_2^* \theta_2}; \end{aligned} \quad (11)$$

$$\text{where } \bar{\sigma} = \left( \sum_{m=1}^{k-1} \Delta \right)^{-1}, \Delta = \left( \frac{\Gamma(m+a_1)}{(T_1^*)^{m+a_1}} \frac{\Gamma(k-m+a_2)}{(T_2^*)^{k-m+a_2}} \right), T_1^* = T_1 + \beta_1 - (n-k+m)$$

$$\log \sigma \text{ and } T_2^* = T_2 + \beta_2 - (k-m) \log \sigma.$$

Hence, the marginal posterior density for change point  $m$  is

$$\pi^*(m | \underline{x}) = \int \int_{\theta_1 \theta_2} \pi_1(\theta_1, \theta_2, m | \underline{x}) d\theta_2 d\theta_1 = \bar{\sigma} \Delta. \quad (12)$$

The choice of the loss function may be crucial in Bayesian analysis. It has always been recognized that the most commonly used loss function, squared error loss function (SELF), is inappropriate in many situations. The Bayes estimator of a parameter under SELF is the posterior mean. If SELF is taken as a measure of inaccuracy then the resulting risk is often too sensitive to the assumptions about the behavior of the tail of the probability distribution. To overcome this difficulty,

a useful asymmetric loss function based on the squared error loss function (ISELF) is defined for any estimate  $\hat{\theta}$  corresponding to the parameter  $\theta$  as

$$L(\hat{\theta}, \theta) = (\theta^{-1} \partial)^2; \partial = \hat{\theta} - \theta.$$

The Bayes estimator of the change point  $m$  under ISELF is obtained as

$$\begin{aligned} \hat{m}_I &= E_P(m^{-1}) / E_P(m^{-2}) \\ \Rightarrow \hat{m}_I &= \sum_{m=1}^{k-1} (m^{-1} \Delta) / \sum_{m=1}^{k-1} (m^{-2} \Delta). \end{aligned} \quad (13)$$

Here, the suffix  $P$  indicates the expectation taken under the posterior density.

When positive and negative errors have different consequences, the use of squared error loss function (SELF) in Bayesian estimation may not be appropriate. In addition, in some estimation problems overestimation is more serious than the underestimation, or vice-versa. To deal with such cases, a useful and flexible class of asymmetric loss function (LINEX loss function (LLF)) is given as

$$L(\partial) = e^{a\partial} - a\partial - 1.$$

The shape parameter of LLF is denoted by ' $a$ '. Negative (positive) value of shape parameter ' $a$ ', gives more weight to overestimation (underestimation) and its magnitude reflect the degree of asymmetry. It is also observed that, for  $a=1$ , the function is very asymmetric with overestimation being more costly than underestimation. For small values of  $|a|$ , the LLF is almost symmetric and is not far from the SELF.

Bayes estimator of  $m$  under the LLF is obtained as

$$\begin{aligned} \hat{m}_L &= -\frac{1}{a} \log E_P \{ e^{-am} \} \\ &= -\frac{1}{a} \log \left\{ \bar{\sigma} \sum_{m=1}^{k-1} (\Delta e^{-am}) \right\}. \end{aligned} \quad (14)$$

A close form of both the estimators does not exist. A numerical method is applied for obtaining the values of their estimates.

### Change Point Estimation (Both Parameter Unknown)

In the case of when both the parameters  $\theta$  and  $\sigma$  are unknown for the considered Pareto model, there does not exist any joint conjugate prior. Assume that the prior beliefs about the parameters  $\theta$  and  $\sigma$  are independent. The natural family of conjugate priors for parameters  $\theta$  and non-informative prior for parameter  $\sigma$  are considered independently here. The non-informative prior of the parameter  $\sigma$  is the limiting form of the appropriate natural conjugate prior. The joint prior distribution when both parameters are unknown is defined as

$$g(\theta, \sigma) = g(\theta | \sigma) \cdot h(\sigma); h(\sigma) = \frac{1}{\sigma}, \sigma > 0. \quad (15)$$

The prior distribution  $g(\theta | \sigma)$  is given in equation (8). The likelihood function in present case is redefined as

$$\begin{aligned} L(\underline{x} | \theta_1, \theta_2, \sigma, m) &= \left( \prod_{i=1}^m f(x_{(i)}; \sigma, \theta_1) \right) \\ &\quad \cdot \left( \prod_{i=m+1}^k f(x_{(i)}; \sigma, \theta_2) \right) \cdot \prod_{i=k+1}^n \left( 1 - \int_0^{x_{(i)}} f(x_{(i)}; \sigma, \theta_1) dx_{(i)} \right) \\ \Rightarrow L(\underline{x} | \theta_1, \theta_2, \sigma, m) &= T_0 \theta_1^m \theta_2^{k-m} e^{\theta_1(n-k+m) \log \sigma} e^{\theta_2(k-m) \log \sigma} e^{-\theta_1 T_1} e^{-\theta_2 T_2}. \quad (16) \end{aligned}$$

The joint prior density for the parameters  $\theta_1, \theta_2, \sigma$  and  $m$  is written as

$$h_2(\theta_1, \theta_2, \sigma, m) = g_1(\theta_1 | \sigma) \cdot g_2(\theta_2 | \sigma) \cdot h(\sigma) \cdot g_3(m).$$

Hence, the joint posterior density is obtained as

$$\begin{aligned} \pi_2(\theta_1, \theta_2, \sigma, m | \underline{x}) &= \frac{L(\underline{x} | \theta_1, \theta_2, \sigma, m) \cdot h_2(\theta_1, \theta_2, \sigma, m)}{\sum_m \int_{\sigma} \int_{\theta_1} \int_{\theta_2} L(\underline{x} | \theta_1, \theta_2, \sigma, m) \cdot h_2(\theta_1, \theta_2, \sigma, m) d\theta_2 d\theta_1 d\sigma} \\ \Rightarrow \pi_2(\theta_1, \theta_2, \sigma, m | \underline{x}) &= \bar{\sigma} \frac{T_0}{\sigma} \theta_1^{m+a_1-1} \theta_2^{k-m+a_2-1} e^{-T_1^* \theta_1} e^{-T_2^* \theta_2}; \quad (17) \end{aligned}$$

where  $\bar{\sigma} = \left( \sum_{m=1}^{k-1} \bar{\Delta} \right)^{-1}$  and  $\bar{\Delta} = \int_{\sigma} \frac{\Delta T_0}{\sigma} d\sigma$ .

Hence, the marginal posterior density for the change point  $m$  is

$$\pi^{**}(m|\underline{x}) = \int_{\sigma} \int_{\theta_1} \int_{\theta_2} \pi_2(\theta_1, \theta_2, \sigma, m|\underline{x}) d\theta_2 d\theta_1 d\sigma = \bar{\sigma} \bar{\Delta}, \quad (18)$$

and, the Bayes estimator under ISELF and LLF for the change point  $m$  are obtained as

$$\hat{m}_I = \sum_{m=1}^{k-1} (\bar{\Delta}/m) / \sum_{m=1}^{k-1} (\bar{\Delta}/m^2) \quad (19)$$

and

$$\hat{m}_L = -\frac{1}{a} \log \left\{ \bar{\sigma} \sum_{m=1}^{k-1} \left( e^{-am} \bar{\Delta} \right) \right\}. \quad (20)$$

## Numerical Analysis

### One Parameter Known Case

To assess and study the properties of the Bayes estimator for the change point  $m$ , a simulation study was performed. The random samples were generated as:

Generate  $\theta_i; i=1, 2$ ; through prior density  $g_i(\theta_i); i=1, 2$ ; for the given values of prior parameters  $\alpha_i, \beta_i; i=1, 2$ ; as  $(\alpha_i, \beta_i) = (0.25, 0.50), (4, 2), (9, 3); i=1, 2$ . The selections of prior parametric values meet the criterion that the prior variance should be unity.

Using generated values of  $\theta_i; i=1, 2$ ; and  $\sigma = 0.50, 1.00, 1.50, 3.00$ ; generate 10,000 random samples of size  $n = 15$  by using the model (2) and (3).

The values of the Bayes estimate  $\hat{m}_I$  under the ISELF have been obtained and presented them in the Table 1, for selected set of censored sample size  $k = 04, 06, 08, 10$ .

# CHANGE POINT ESTIMATION FOR PARETO TYPE-II MODEL

**Table 1.** Bayes Estimate of  $m$  under ISELF (Scale Parameter Known)

$n=15$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	3.7066	3.7177	3.7251	3.7326	3.7400
0.50	04, 02	3.6908	3.7019	3.7092	3.7166	3.7239
0.50	09, 03	3.3381	3.3481	3.3549	3.3615	3.3681
1.00	0.25, 0.50	3.7336	3.7447	3.7522	3.7597	3.7672
1.00	04, 02	3.7052	3.7163	3.7236	3.7311	3.7385
1.00	09, 03	3.6519	3.6627	3.6700	3.6773	3.6846
1.50	0.25, 0.50	3.7371	3.7483	3.7558	3.7633	3.7707
1.50	04, 02	3.7177	3.7289	3.7363	3.7436	3.7510
1.50	09, 03	3.6906	3.7016	3.7091	3.7164	3.7238
3.00	0.25, 0.50	3.4849	3.4953	3.5024	3.5094	3.5165
3.00	04, 02	3.4701	3.4806	3.4875	3.4944	3.5012
3.00	09, 03	3.4088	3.4189	3.4257	3.4325	3.4393

Table 1 shows that when censored sample size  $k$  increases, the magnitude of the estimate increases, but increment in magnitude is nominal (robust). A similar trend also noted when scale parameter  $\sigma$  increases, however for large value of  $\sigma$  ( $>1.5$ ) the magnitude of the estimate decreases. The opposite trend has been seen when set of prior parameter increases.

Using above considered set of parametric values with  $a = 0.25, 0.50, 1.00, 2.00$ ; (shape parameter of LLF) the magnitude of the Bayes estimate under LLF have been obtained and present in the Table 2, only for  $a = 0.25, 1.00$ .

**Table 2.** Bayes Estimate of  $m$  under LLF (Scale Parameter Known)

$n = 15, a = 0.25$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	3.3060	3.3159	3.3226	3.3292	3.3358
0.50	04, 02	3.2778	3.2877	3.2942	3.3007	3.3073
0.50	09, 03	2.9531	2.9620	2.9678	2.9737	2.9797
1.00	0.25, 0.50	3.2839	3.2937	3.3002	3.3069	3.3135
1.00	04, 02	3.2699	3.2798	3.2863	3.2927	3.2992
1.00	09, 03	3.2698	3.2795	3.2860	3.2926	3.2991
1.50	0.25, 0.50	3.2331	3.2426	3.2491	3.2556	3.2621
1.50	04, 02	3.2192	3.2289	3.2353	3.2417	3.2480
1.50	09, 03	3.1623	3.1717	3.1780	3.1842	3.1906
3.00	0.25, 0.50	2.9560	2.9648	2.9708	2.9768	2.9827
3.00	04, 02	2.9433	2.9522	2.9581	2.9639	2.9699
3.00	09, 03	2.8913	2.8999	2.9057	2.9115	2.9173
$n = 15, a = 1.00$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	4.0109	4.0228	4.0309	4.0390	4.0470
0.50	04, 02	3.9937	4.0058	4.0137	4.0216	4.0295
0.50	09, 03	3.9231	3.9348	3.9425	3.9504	3.9583
1.00	0.25, 0.50	3.7329	3.7440	3.7515	3.7590	3.7665
1.00	04, 02	3.7010	3.7121	3.7194	3.7269	3.7343
1.00	09, 03	3.3344	3.3444	3.3510	3.3577	3.3644
1.50	0.25, 0.50	3.7030	3.7140	3.7214	3.7289	3.7363
1.50	04, 02	3.6871	3.6982	3.7056	3.7129	3.7202
1.50	09, 03	3.6869	3.6980	3.7054	3.7127	3.7201
3.00	0.25, 0.50	2.7579	2.7661	2.7716	2.7772	2.7828
3.00	04, 02	2.7460	2.7544	2.7598	2.7652	2.7707
3.00	09, 03	2.6975	2.7055	2.7109	2.7163	2.7217

## CHANGE POINT ESTIMATION FOR PARETO TYPE-II MODEL

Table 2 shows that when shape parameter  $\sigma$  increases, the magnitude of the estimator decreases (except for large prior parametric value). An increasing trend in the magnitude of the estimate also is also noted when ' $a$ ' increases but the increment in magnitude is robust. Others properties are similar to ISELF.

### When Both Parameters Unknown

When both parameters are considered as a random variable, a simulation study was carried out to study the properties of Bayes estimators of change point.

Similarly, a 10,000 random sample of size  $n = 15$  was generated. The Bayes estimate of  $m$  under the ISELF and LLF were obtained and are presented in Tables 3-4 respectively for different selected set of values.

**Table 3.** Bayes Estimate of  $m$  under ISELF (Both Parameter Unknown)

$n = 15$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	3.3640	3.3741	3.3808	3.3877	3.3944
0.50	04, 02	3.3497	3.3599	3.3664	3.3731	3.3798
0.50	09, 03	3.2223	3.2320	3.2384	3.2448	3.2513
1.00	0.25, 0.50	3.6074	3.6183	3.6254	3.6327	3.6399
1.00	04, 02	3.5766	3.5873	3.5945	3.6017	3.6088
1.00	09, 03	3.2905	3.3003	3.3068	3.3134	3.3200
1.50	0.25, 0.50	3.6150	3.6258	3.6331	3.6403	3.6477
1.50	04, 02	3.5996	3.6105	3.6176	3.6248	3.6319
1.50	09, 03	3.5982	3.6090	3.6161	3.6233	3.6304
3.00	0.25, 0.50	3.6787	3.6896	3.6971	3.7045	3.7118
3.00	04, 02	3.6630	3.6740	3.6813	3.6886	3.6959
3.00	09, 03	3.5995	3.6103	3.6174	3.6246	3.6318



**Table 4.** Bayes Estimate of  $m$  under LLF (Both Parameter Known)

$n = 15, a = 0.25$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	3.4309	3.4411	3.4480	3.4550	3.4617
0.50	04, 02	3.4162	3.4265	3.4333	3.4401	3.4469
0.50	09, 03	3.3558	3.3658	3.3725	3.3792	3.3859
1.00	0.25, 0.50	3.2678	3.2775	3.2841	3.2907	3.2972
1.00	04, 02	3.2398	3.2496	3.2560	3.2624	3.2689
1.00	09, 03	3.1907	3.2002	3.2066	3.2130	3.2194
1.50	0.25, 0.50	3.2046	3.2141	3.2205	3.2269	3.2334
1.50	04, 02	3.1907	3.2002	3.2066	3.2130	3.2194
1.50	09, 03	2.9189	2.9277	2.9335	2.9392	2.9452
3.00	0.25, 0.50	2.4142	2.4214	2.4262	2.4312	2.4361
3.00	04, 02	2.4039	2.4112	2.4159	2.4207	2.4255
3.00	09, 03	2.3614	2.3684	2.3732	2.3779	2.3826
$n = 15, a = 1.00$		$k \downarrow$				
$\sigma$	$(\alpha, \beta) \downarrow$	4	6	8	10	15
0.50	0.25, 0.50	3.7918	3.8031	3.8108	3.8183	3.8260
0.50	04, 02	3.7756	3.7871	3.7946	3.8020	3.8095
0.50	09, 03	3.7088	3.7199	3.7272	3.7347	3.7421
1.00	0.25, 0.50	3.5352	3.5458	3.5528	3.5599	3.5671
1.00	04, 02	3.5201	3.5308	3.5378	3.5448	3.5517
1.00	09, 03	3.5199	3.5305	3.5375	3.5446	3.5516
1.50	0.25, 0.50	3.2398	3.2496	3.2560	3.2624	3.2689
1.50	04, 02	3.1908	3.2005	3.2068	3.2131	3.2195
1.50	09, 03	3.0107	3.0194	3.0207	3.0328	3.0475
3.00	0.25, 0.50	2.6585	2.6665	2.6717	2.6772	2.6826
3.00	04, 02	2.6472	2.6552	2.6603	2.6657	2.6710
3.00	09, 03	2.6003	2.6080	2.6133	2.6185	2.6237

## CHANGE POINT ESTIMATION FOR PARETO TYPE-II MODEL

The behavior of  $\hat{m}_I$  was shown to be similar as compare to  $\hat{m}_I$  under ISELF. It is also noted that the magnitude of the estimator  $\hat{m}_I$  increases as  $\sigma$  increases for all selected parametric set of values. Further, the magnitude of the estimate of  $\hat{m}_I$  is closer than the estimate of  $\hat{m}_I$  except for large value of  $\sigma$ .

All properties of estimator  $\hat{m}_L$  were similar as compared to  $\hat{m}_L$  under LLF. For small values of 'a', the magnitude of estimate of  $\hat{m}_L$  is wider than  $\hat{m}_L$  for all considered values of  $\sigma$  (except for  $\sigma=0.50$ ). For large values of 'a', the magnitude of estimate of  $\hat{m}_L$  becomes narrower than  $\hat{m}_L$  for all considered values of  $\sigma$  (except for  $\sigma=1.00$ ). Other properties are the same, as in the case of a known shape parameter.

### Remark

In the case when the censored sample size  $r=15$ , the censoring criterion reduces to the complete sample size criterion and, hence, all results are valid for the complete sample case.

### References

- Al-Zahrani, B. & Al-Sobhi, M. (2013). On parameters estimation of Lomax distribution under general progressive censoring. *Journal of Quality and Reliability Engineering*, 2013: 1-7.
- Broemeling, L. D. & Tsurumi, H. (1987). *Econometrics and structural change*. New York: Marcel Dekker.
- Dyer, D. (1981). Structural probability bounds for the strong Pareto law. *Canadian Journal Statistics*, 9: 71-77.
- Ebrahimi, N. & Ghosh, S. K. (2001). Bayesian and frequentist methods in change-point problems. In N. Balkrishna & C. R. Rao (Eds). *Handbook of statistics, Vol. 20: Advances in Reliability*, (Elsevier), 777-787.
- Goldenshluger, A., Tsybakov, A. & Zeevi, A. (2006). Optimal change point estimation from indirect observations. *The Annals of Statistics*, 34(1): 350-372.
- Harris, C. M. (1968). The Pareto distribution as a queue discipline. *Operations Research*, 16(2): 307-313.

- Jani, P. N. & Pandya, M. (1999). Bayes estimation of shift point in left truncated exponential sequence. *Communications in Statistics-Theory and Methods*, 28(11): 2623-2639.
- Madi, M. T. & Raqab, M. Z. (2004). Bayesian prediction of temperature records using the Pareto model. *Environmetrics*, 15(7): 701–710.
- Moghadam, M. S., Farhad, Y. & Manoochehr, B. (2012). Inference for Lomax distribution under generalized order statistics. *Applied Mathematical Sciences*, 6(105): 5241-5251
- Nasiri, P. & Hosseini, S. (2012). Statistical inferences for Lomax distribution based on record values (Bayesian and Classical). *Journal of Modern Applied Statistical Methods*, 11(1): 179-189.
- Panahi, H. & Asadi, S. (2011). Inference of stress-strength model for a Lomax distribution. *World Academy of Science, Engineering and Technology*, 55: 275-278.
- Pandya, M. (2013). Bayesian estimation of AR (1) with change point under asymmetric loss functions. *Statistical Research Letters*, 2(2): 53-62.
- Pandya, M. & Jadav, P. (2010). Bayesian estimation of change point in mixture of left truncated exponential and degenerate distribution. *Communication in statistics-Theory and Methods*, 39(15): 2725-2742.
- Prakash, G. & Singh, D. C. (2013). Bayes prediction intervals for the Pareto model. *Journal of Probability and Statistical Science*, 11(1): 109-122.
- Singh, D. C., Prakash, G. & Singh, P. (2007). Shrinkage testimators for the shape parameter of Pareto distribution using LINEX loss function. *Communications in Statistics-Theory and Methods*, 36(4): 741-753.
- Srivastava, U. (2012). Bayesian estimation of shift point in Poisson model under asymmetric loss functions. *Pakistan Journal of Statistics and Operation Research*, 8(1): 31-42.

# Distance Correlation Coefficient: An Application with Bayesian Approach in Clinical Data Analysis

**Atanu Bhattacharjee**

Malabar Cancer Centre  
Kerala, India

---

The distance correlation coefficient – based on the product-moment approach – is one method by which to explore the relationship between variables. The Bayesian approach is a powerful tool to determine statistical inferences with credible intervals. Prior information about the relationship between BP and Serum cholesterol was applied to formulate the distance correlation between the two variables. The conjugate prior is considered to formulate the posterior estimates of the distance correlations. The illustrated method is simple and is suitable for other experimental studies.

*Keywords:* Conjugate prior, credible interval, distance covariance, canonical correlation

---

## Introduction

The correlation coefficient is a widely used tool to observe the association between two random variables in experimental research. The assessment of relation between two variables ( $X$ ,  $Y$ ) is a common problem. The Pearson and Spearman correlation coefficients are wonderful tools to explore the relationship between two variables. Canonical, rank and Renyi correlations are the most widely used tools to investigate the strength of relation between random vectors (Bickel & Xu, 2009). The Renyi (1959) correlation becomes zero if  $X$  and  $Y$  are independent, thus, the Renyi (1959) is examined on maximal correlation. The Pearson correlation coefficient computation is simpler than the Renyi correlation coefficient. It is well-known that Pearson's product correlation coefficient  $\rho$  becomes zero for bivariate normal independence. In the multivariate case, the diagonal matrix  $\Sigma$  becomes independent, but it is unable to specify dependence

---

*Dr. Atanu Bhattacharjee is in the Division of Clinical Research and Biostatistics. Email at: [atanustat@gmail.com](mailto:atanustat@gmail.com).*

for general case. It may be concluded the  $\rho$  and  $\Sigma$  are not able to characterize independence in general.

The joint independence of random variables can be explored through distance correlation and is measured with product moment correlation  $\rho$ . It is the measures of correlation with multivariate dependence coefficients through arbitrary random vectors. Basically, the distance correlation is a product-moment correlation and a generalized form of bivariate measures of dependency. It is a very useful and unexplored area for statistical inference.

A new type of coefficient applicable to measure the dependence between random vectors of equal or unequal distance is useful for complicated dependence structures in multivariate data. The introduction of distance correlation is well detailed (Szekely, et al., 2007) and it can be computed with a simple formula of sample size  $n > 2$ . It is free with matrix inversion and estimation of parameters. The distance correlation has the advantage over there. The literature on testing measures of dependence is rich (Anderson, 2003; Blomqvist, 1950; Hollander & Wolfe, 1999; Blum, et al., 1961). The Likelihood Ratio Test (LRT) and Wilks Lambda are applicable for multivariate data but fail if dimension exceeds the sample size. The proposed method distance correlation with Bayesian approach is completely new.

In another aspect, it is general practice to ignore prior information about the relation between variables and establish the new correlation. As an alternative, the Bayesian approach takes the opportunity to incorporate the prior information of the variables to establish the inference about correlation. The posterior estimate of the correlation coefficient is applied to explore the relation between maternal weight and infant birth-weight (Bashir, 1997). The Bayesian approach is an attractive method for estimating tools because it incorporates previous studies' observations into its calculation. The aim of this article is to elaborate the application of a Bayesian approach in distance correlation. The work is illustrated with the estimation of distance correlation between serum cholesterol and BP. The data are captured from two different studies detailed below.

## Distance Covariance and Distance Correlation

Distance covariance between the random variables  $X$  and  $Y$  can be defined with the marginal characteristic function  $f_X(t)$  and  $f_Y(s)$  by:

$$V^2(X, Y) = \left[ f_{(X, Y)}(t, s) - f_X(t)f_Y(s) \right]^2. \quad (1)$$

## DISTANCE CORRELATION COEFFICIENT: BAYESIAN APPROACH

The function  $f_{(X,Y)}$  is a joint characteristic function of  $X$  and  $Y$ . The terms  $s$  and  $t$  are vectors and the product of  $t$  and  $s$  is  $\langle t, s \rangle$ . The distance covariance measures the distance  $\|f_{(X,Y)}(t, s) - f_X(t)f_Y(s)\|$  between the joint characteristic and marginal characteristics functions. The random vectors  $X$  and  $Y$  are in  $R^p$  and  $R^q$  respectively. The hypotheses are  $H_0 : f_{X,Y} = f_X f_Y$  and  $H_1 : f_{X,Y} \neq f_X f_Y$ . The distance variance is:

$$V(X) = \left[ f_{(X,X)}(t, s) - f_X(t)f_X(s) \right]. \quad (2)$$

The distance correlation between  $X$  and  $Y$  is defined with finite first moments  $R(X, Y)$  by

$$R^2(X, Y) = \frac{V^2(X, Y)}{\sqrt{V^2(X)V^2(Y)}} > 0, \text{ otherwise } = 0.$$

The distance covariance  $V_n(X, Y)$  is defined with

$$V_n^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl}. \quad (3)$$

Similarly it can be defined with

$$V_n^2(X, X) = \frac{1}{n^2}. \quad (4)$$

The parameter  $a_{kl} = |X_k - Y_l|$ ,  $\bar{a}_{k.} = \frac{1}{n} \sum_{l=1}^n a_{kl}$ ,  $\bar{a}_{.l} = \frac{1}{n} \sum_{k=1}^n a_{kl}$  and  $\bar{a}_{..} = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n a_{kl}$

$$A = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}. \quad (5)$$

$B_{kl}$  is defined similarly.

### Properties of Distance Correlation Coefficient

The distance correlation provides the scope to generalize the correlation between variables ( $X$  and  $Y$ ) by  $R$  is defined on arbitrary dimensions  $R=0$  independent of  $X$  and  $Y$ . The range of the distance correlation is  $0 \leq R \leq 1$ .  $R$  can be defined as the function of Pearson correlation coefficient  $\rho$  with  $R(X, Y) < |\rho(X, Y)|$  with equality when  $\rho = \pm 1$ .

### Importance of Distance Covariance

The random variables  $X$  and  $Y$  are expressed as  $A_i = X_i + \varepsilon_i$  and  $B_j = Y_j + \varepsilon_j$  respectively. The error terms  $\varepsilon_i$  and  $\varepsilon_j$  are independent with the variables  $X_i$  and  $Y_j$ . The relation between random functions  $A_i$  and  $B_j$  is irrelevant, but the relation between  $X_i$  and  $Y_j$  is important and a matter of concern. The strength of relation between  $X$  and  $Y$  can be measured through distance correlation in this scenario.

### Distance Correlation in One-sided Test

The frequency approach tests the problem through  $p(X)$  value of the null hypothesis  $H_0$ . By contrast, Bayesian measures through posterior probability  $p(H_0 / X)$ . Let the data follow a normal distribution  $(\theta, \sigma^2)$  with null hypothesis  $H_0 : \theta \leq 0$  and alternate  $H_1 : \theta > 0$ . The frequency and robust Bayesian often coincide (Casella & Berger, 2002). Let the marginal distance correlation  $\rho$  be applied between  $p(X) = 1 - \Phi(X / \sigma)$  and  $p(H_0 / X)$ . The distance correlation should be greater than or equal to zero. Because  $p(X)$  and  $p(H_0 / X)$  both decrease with respect to  $X$ .

### Distance Correlation between Parameter and Unbiased Estimator

Suppose,  $(\theta, X)$  are random variables with joint characteristics function  $f_{(X, Y)}(t, s)$  and the marginal distribution of  $\theta$  is  $\pi$ . The estimator of  $\theta$  is  $\delta(X)$  and square error loss is  $r(\pi, \delta) = E[\delta(X) - \theta]^2$  and risk is  $\delta_\pi(X) = E(\theta / X)$ . The distance correlation between  $\theta$  and  $\delta(X)$  is

$$\rho(\theta, \delta(X)) = \frac{\text{var}(\theta) + \text{cov}\{\theta, b(\theta)\}}{\sqrt{\text{var}(\theta)} \sqrt{\text{var}\{\theta + (\theta)\} + \tau(\pi, \delta) - E\{b^2(\theta)\}}} . \quad (6)$$

## Statistical Methods

The Bayes' Theorem provides prior information about the relevant parameter for a specific statistical analysis. It is helpful to test the hypothesis in presence of posterior probability of the parameter of interest. The parameter of interest  $R(X, Y)$  can be computed with posterior probability through Bayes' theorem:

$$P(R(X, Y) / \text{Information}) = \frac{(P(\text{Information} / R(X, Y))) P(R(X, Y))}{(P(\text{Information}))} . \quad (7)$$

The term  $P(R(X, Y))$  is the prior probability of  $R(X, Y)$  observed from the previous study. The term  $P(\text{information}/R(X, Y))$  is the likelihood of  $R(X, Y)$  that occurred in a previous study or is in data collected by an investigator. The sum of the function  $1/(P(\text{Information}))$  should be equal to 1 as the theory of total Bayes theorem. The relation between posterior and prior is:

$$\text{PosteriorProbability} \propto \text{Likelihood} \times \text{PriorProbability} . \quad (8)$$

The posterior density of  $R(X, Y)$  is generated with

$$P(R(X, Y) / x, y) \propto P(R(X, Y)) \frac{(1 - R(X, Y)^2)^{(n-1)/2}}{(1 - R(X, Y) * r)^{n-(\frac{3}{2})}} . \quad (9)$$

The mean and variance of  $X$  and  $Y$  are  $\mu_1, \mu_2, \sigma_1^2$  and  $\sigma_2^2$  respectively. The mean ( $z$ ) is derived from

$$e^z = \frac{\mu_1 \sigma_2}{\mu_2 \sigma_1} . \quad (10)$$

The term  $R(X, Y)$  is defined by  $\tanh \varepsilon$  and is assumed  $\varepsilon \sim N(z, \frac{1}{n})$ . The mathematical formulations are detailed in Fisher (1915). The hyperbolic



transformation plays a role in considering the conjugate prior with normal distributions. The posterior mean can be represented with

$$\mu_{Posterior} = \varepsilon_{Posterior}^2 \left[ \eta_{Prior} \tanh^{-1} R(X, Y)_{Prior} + \eta_{Likelihood} \tanh^{-1} R(X, Y)_{Likelihood} \right] \quad (11)$$

$$\sigma_{Posterior}^2 = \frac{1}{\eta_{Prior} + \eta_{Likelihood}} \quad (12)$$

and the prior with the form

$$P(R(X, Y)) \propto (1 - R(X, Y)^2)^c. \quad (13)$$

The prior is dependent on the choice of  $c$ ;  $c = 0$  gives  $P(R(X, Y)) \propto 1$ .

### Illustrated Example

Among different types of risk factors hypertension and abnormalities of lipid profiles are established reasons for coronary artery disease as observed through epidemiological and genetics studies (for details see Williams, et al., 1988). Serum cholesterol is related with blood pressure (BP) values (Ferrannini, et al., 1987; Hunt, et al., 1986, Floras, et al., 1987; Simone, et al., 1992; Sung, et al., 1997). The present work is undertaken to check whether serum cholesterol is an influencing factor of BP. The BP measurement was taken in 24 hours close observation. Study 1 was conducted to observe two drug treatment effects among liver cirrhosis patients in St. Stephen hospital during 2009 to 2011. The data on serum cholesterol and BP were observed during the study of 179 patients with follow up observations. In this article, data is considered to illustrate the application of a distance correlation coefficient between serum cholesterol and BP. In Study 2, a total of 100 patients of type 2 diabetes were observed with two types of drug treatments in Madurai Menakshi Mission Hospital in 2009. The different biochemical parameters through follow up periods were observed as an effect of drug treatment. The measurements of serum cholesterol and BP were observed through the follow up periods. The work is explored on the data to illustrate the distance correlation coefficient among the patients.

### Frequentist Test for Distance Correlation

To check the distance correlation between BP and serum cholesterol, the variables were defined. For every  $i^{\text{th}}$  person the BP is denoted with  $x_i$  and serum cholesterol by  $y_i$ . In Study 1, the null hypothesis to test the distance correlation coefficient is assumed as zero, i.e.  $R(X, Y) = 0$ . The `dcor.ttest(x,y)` function in the *energy* package of R i386 3.0.1 is applied to test the null hypothesis. Results show that the distance correlation rejects the null hypothesis that  $p = 0.01$ . Consequently, researchers may feel that it is possible to reject the null hypothesis of no correlation between BP and serum cholesterol.

### Bayesian test for Distance Correlation

The measure of evidence of

$$p(H_0 / x)$$

is the probability of

$H_0$  is true with  $X = x$  is

$$P(H_0 / x) = P(\theta \leq 0 / x) = \frac{\int_0^{\infty} f(x - \theta) \pi(\theta) d\theta}{\int_{-\infty}^{\infty} f(x - \theta) \pi(\theta) d\theta}. \quad (14)$$

In both studies, it was assumed that the BP and serum cholesterol are correlated to each of the others. The relation with distance covariance was examined using a Bayesian approach. The relation between BP and serum cholesterol during surgery in patients was observed from anesthesia data. The estimated distance correlation between serum cholesterol and BP may be measured with error. The error arises due to the presence of small sample size. The fluctuation of observed correlation in different studies may be due to different sample sizes. Using several studies, a meta-analysis can be conducted to estimate the real correlation between BP and cholesterol. However, if lacking several studies, the Bayesian posterior estimate is applied to estimate the robust distance correlation between BP and serum cholesterol.

$$\sigma_{posterior}^2 = \frac{1}{\eta_{Prior} + \eta_{Likelihood}} = \frac{1}{179+100} = 0.035 \quad (15)$$

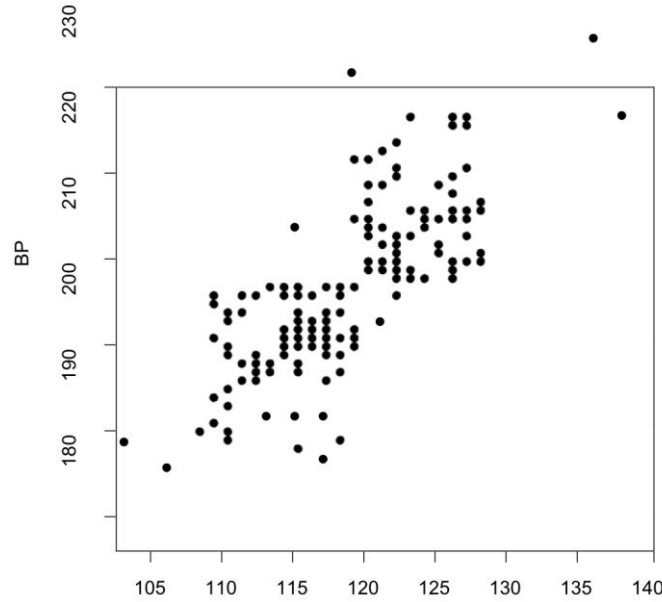
$$\mu_{Posterior} = 0.035(179 \tanh^{-1} 0.58 + 100 \tanh^{-1} 0.43) \quad (16)$$

$$\mu_{posterior} = 0.0035(179*0.67 + 100*0.47) \quad (17)$$

The confidence interval is

$$\mu_{Posterior} \pm 1.96\sqrt{(\sigma_{post}^2)} = 0.58 \pm 1.96(0.0035)^{\frac{1}{2}} \quad (18)$$

i.e. (0.69, 0.47). This shows the posterior estimate of distance correlation  $R(X, Y)$  is 0.58 with credible interval (0.69, 0.47). The estimate observed at the 95 confidence interval is 0.23 (0.28, 0.18). The values can be compared and a conclusion can be drawn. The simple approach for distance correlation can be extended to other experimental research.



**Figure 1.** Relationship between BP and serum cholesterol. A positive correlation suggests that serum cholesterol is the influencing factor for high BP.

## Discussion

The risk of coronary artery disease is depends on different risk factors (Levy et al., 1988; Assmann, et al., 1988). Different risk factors can be classified into global and individual risk factors; however, global risk factors are more important than individual risk factors for cardiovascular disease (Ferrara, 2002). Distance correlation is applied to explore the relation between BP and cholesterol. Distance correlation is applicable to all types of metric spaces. The idea is to apply the distance correlation by replace the Euclidean distance into metric distance. Distance covariance is also applicable to test the linear model  $Y = X\beta + \varepsilon$ ; where  $(x, \varepsilon)$  are i.i.d. Distance covariance is defined on arbitrary dimension and it can be extended for multivariate responses. It is simple like the Pearson product moment covariance. If  $X$  and  $Y$  are sample from different metric space, then distance covariance can be measured. The Pearson correlation is the best choice to explore the relation between variables. However, it is not feasible to apply it to non-normal data. Distance covariance can also be applied to non-normal data.

The correlation coefficient for the sample average was examined with uniform prior by Daniels (1999). Extensions of the work were carried with shrinkage priors by Daniels Kass (2001). The measurements of correlation tested through logarithmic transformation of the eigenvalue (Leonard et al., 1992). Barnard et al. (2000) proposed a normal prior for a transformation of correlation coefficients. Wong, et al., (2003) give a prior probability model for graphical models and partial correlations through the sparseness of the precision matrix. Gabor et al., (2007) discussed the advantage of distance correlation over Pearson correlation: It is the generalized form of the Pearson correlation in two ways (1) its ability to measure the linear relation with consideration of all types of dependency, and (2) exposure to measure the dependency through random vectors in the arbitrary dimension. Tracz, et al., (1992) showed that the distance correlation is more suitable as a dependent index than the product moment correlation coefficient.

The Affine invariance property is important for the transformation of data in statistical inference. The Affine invariance with a group is detailed by Eaton (1989) and Giri (1996). Gabor (2007) proved distance correlation is free from Affine invariance. Correlation analysis is strong a filler to draw statistical inference in any medical research. Distance correlation is another useful tool to explore the relation between variables. Distance correlation with confidence interval is a statistical tool to sketch the inference about the relation between variables. In this study, the confidence interval between serum cholesterol and BP

was observed. The Bayesian approach was applied with credible intervals and observed with less interval estimates. Small sample sizes tend to be a problem in clinical trials due to cost and time. The Bayesian approach gives a practical concession. It is also useful choice to deal with random measurement error in weight gain relation. It is simple and accurate. The approach is helpful to explore the relation between variables more intuitively.

## **Conclusion**

Distance correlation with a Bayesian approach is not the only choice of correlation analysis, but can be considered in many cases as an alternative of Pearson's covariance. An example with clinical trials illustrated where distance correlation can give more information not captured by traditional correlation analysis. In exploratory analysis with small sample size data, the Bayesian distance correlation is an alternative choice for the low dimensional marginal distribution of two variables. The Bayesian distance correlation can be useful to test the linear relation between variables and it can be a first choice to explore the relation between variables to made decisions about specific tools for further data analysis. Distance correlation having high value of one (or near to one) shows a strong relation between variables. The Bayesian approach is suitable tool for calculating distance correlation coefficient among variables. The work can be extended to explore the relation between bivariate observations in different experimental research. Like the correlation coefficient, distance correlation can be applied to understand the relation between variables by clinician. It can serve clinicians to know the real strength of variables and, as a result, interpretation of the results in the real life practice. In any experimental research relations between variables is unavoidable. Distance correlation can be considered as easily interpretable tool to discover the relations.

## **Acknowledgement**

We would like to thank Dr. Shuarav in St. Stephen Hospital and Mr. Rakesh in Menakshi Mission Hospital for permission to apply data with initial review to initiate the work of this paper.

## References

- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd Ed.). New York, Wiley.
- Assmann G. & Schulte H. (1988). The Prospective Cardiovascular Münster (PROCAM) study: prevalence of hyperlipid aemia in persons with hypertension and/or diabetes 343 mellitus and the relationship to coronary heart disease. *American Heart Journal*, 116(6 Pt. 2): 1713-1724.
- Barnard, J., McCulloch, R. & Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, 10: 1281-311.
- Bashir S. A. & Duffy S. W. (1997). The Correction of Risk Estimates for Measurement Error. *Annals of Epidemiology*, 7: 154-164.
- Bickel, P. J & Ying, X. (2009). Discussion of Brownian Distance Covariance. *The Annals of Applied Statistics*, 3(4): 1266-1269.
- Blomqvist, N. (1950). On a measure of dependence between two random variables. *The Annals of Mathematical Statistics*, 21: 593-600.
- Blum, J. R., Kiefer, J. & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32(2): 485-498.
- Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.
- Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27: 567-78.
- Daniels, M. J. & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics*, 57(4): 1174-84.
- Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Hayward, CA, IMS.
- Ferrannini, E., Buzzigoli, G., Bonadonna, R., Giorico, M. A., Oleggini, M., Gradizdei, L., Pedrinelli, R., Brandi, L. & Bevilacqua, S. (1987). Insulin resistance in essential hypertension. *The New England Journal of Medicine*, 317: 350-357.
- Ferrara L. A., Guida, L., Iannuzzi, R., Celentano, A. & Lionello, F. (2002). Serum cholesterol affects blood pressure Regulation. *Journal of Human Hypertension*, 16: 337-343.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 19(4): 507-521.

Floras, J. S., Hassan, M. O., Jones, J. V., & Sleight, P. (1987). Pressor responses to laboratory stresses and daytime blood pressure variability. *American Journal of Hypertension*, 5(6): 715-719.

Giri, N. C. (1996). *Group Invariance in Statistical Inference*. Edge, NJ, World River Scientific.

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric Statistical Methods* (2nd Ed.). New York, Wiley.

Hunt, S. C., Williams, R. R., Smith, J. B., & Ash, K. O. (1986). Associations of three erythrocytes cation transport system with plasma lipids in Utah subjects. *American Journal of Hypertension*, 8: 30-36.

Leonard, T. & Hsu, J. S. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics*, 20: 1669-96.

Levy, D., Anderson, K. M., Savage, D. D., Kannel, W. B., Christiansen, J. C. & Castelli, W. P. (1988). Echocardiographically detected left ventricular hypertrophy: prevalence and risk factors. The Framingham Heart Study. *Annals of Internal Medicine*, 108(1): 7-13.

Renyi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10: 441-451.

de Simone, G., Daniels, S. R., Devereux, R. B., Meyer, R. A., Roman, M. J., de Divitiis, O., & Alderman, M. H. (1992). Left ventricular mass and body size in normotensive children and adults: assessment of allometric relations and impact of overweight. *Journal of the American College of Cardiology*, 20: 1251-1260. doi: [10.1016/0735-1097\(92\)90385-Z](https://doi.org/10.1016/0735-1097(92)90385-Z)

Sung, B. H., Izzo, J. L., & Wilson, M. F. (1997). Effects of cholesterol reduction on BP response to mental stress in patients with high cholesterol. *American Journal of Hypertension*, 10: 592-599.

Szekely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by Correlation of distances. *The Annals of Statistics*, 35(6): 2769-2794.

Tracz, S. M., Elomore, P. B., & Pohlmann, J. T. (1992). Correlation meta-analysis: Independent and no independent cases. *Educational and Psychological Measurement*, 52: 879-888.

## DISTANCE CORRELATION COEFFICIENT: BAYESIAN APPROACH

Williams, R. R., Hunt, S. C., Hopkins, P. N., Stults, B. M., Wu, L. L., Hasstedt, S. J., Barlow, G. K., Stephenson, S. H., Lalouel, J. M., & Kuida, H. (1988). Familial dyslipidemic hypertension. Evidence from 58 Utah families for a syndrome present in approximately 12% of patients with essential hypertension. *The Journal of the American Medical Association*, 259(24): 3579-3586. doi: [10.1001/jama.259.24.3579](https://doi.org/10.1001/jama.259.24.3579)

Wong, F., Carter, C. K. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90: 809-830.



# Estimation of Reliability in Multicomponent Stress-Strength Based on Generalized Rayleigh Distribution

**Gadde Srinivasa Rao**

University of Dodoma  
Dodoma, Tanzania

---

A multicomponent system of  $k$  components having strengths following  $k$ - independently and identically distributed random variables  $x_1, x_2, \dots, x_k$  and each component experiencing a random stress  $Y$  is considered. The system is regarded as alive only if at least  $s$  out of  $k$  ( $s < k$ ) strengths exceed the stress. The reliability of such a system is obtained when strength and stress variates are given by a generalized Rayleigh distribution with different shape parameters. Reliability is estimated using the maximum likelihood (ML) method of estimation in samples drawn from strength and stress distributions; the reliability estimators are compared asymptotically. Monte-Carlo simulation is used to compare reliability estimates for the small samples and real data sets illustrate the procedure.

*Keywords:* Generalized Rayleigh distribution, reliability estimation, stress-strength, ML estimation, confidence intervals

---

## Introduction

Surles and Padgett (1998, 2001) introduced the two-parameter Burr Type X distribution and named it the generalized Rayleigh distribution. Note that the two-parameter generalized Rayleigh distribution is a particular member of the generalized Weibull distribution, originally proposed by Mudholkar and Srivastava (1993). The two-parameter Burr Type X distribution is referred to as the generalized Rayleigh distribution (GRD). For  $\alpha > 0$  and  $\lambda > 0$ , the two-parameter GRD has the density function;

$$f(x; \alpha, \lambda) = 2\alpha\lambda^2 x e^{-(\lambda x)^2} \left(1 - e^{-(\lambda x)^2}\right)^{\alpha-1} \quad \text{for } x > 0 \quad (1)$$

---

*Dr. G. Srinivasa Rao is a Professor of Statistics. Email at: gaddesrao@yahoo.com.*

and the distribution function is given by

$$F(x; \alpha, \lambda) = \left(1 - e^{-(\lambda x)^2}\right)^\alpha \quad \text{for } x > 0. \quad (2)$$

Here  $\alpha$  and  $\lambda$  are the shape and scale parameters respectively. The GRD has been studied extensively by Kundu and Raqab (2005) and Raqab and Kundu (2005). The two-parameter GRD is denoted by  $GR(\alpha, \lambda)$ . Surles and Padgett (2001) showed that the two-parameter GR distribution can be used effectively in modeling strength as well as general lifetime data.

This article studies reliability in a multicomponent stress-strength based on  $X, Y$ , two independent random variables, where  $X$  and  $Y$  follow generalized Rayleigh distributions with shape parameters  $\alpha$  and  $\beta$  respectively and with common scale parameter  $\lambda$ .

Let the random samples  $y, x_1, x_2, \dots, x_k$  be independent,  $G(y)$  be the continuous distribution function of  $Y$  and  $F(x)$  be the common continuous distribution function of  $x_1, x_2, \dots, x_k$ . The reliability in a multicomponent stress-strength model developed by Bhattacharyya and Johnson (1974) is

$$\begin{aligned} R_{s,k} &= P[\text{at least } s \text{ of the } (x_1, x_2, \dots, x_k) \text{ exceed } Y] \\ &= \sum_{i=s}^k \binom{k}{i} \int_{-\infty}^{\infty} [1 - G(y)]^i [G(y)]^{k-i} dF(y), \end{aligned} \quad (3)$$

where  $x_1, x_2, \dots, x_k$  are independently and identically distributed (*iid*) with common distribution function  $F(x)$ , this system is subjected to common random stress  $Y$ . The probability in (3) is called reliability in a multicomponent stress-strength model (Bhattacharyya & Johnson, 1974). The survival probability of single component stress-strength versions have been considered by several authors assuming various lifetime distributions for the stress-strength random variates (Enis & Geisser, 1971; Downton, 1973; Awad & Gharraf, 1986; McCool, 1991; Nandi & Aich, 1994; Surles & Padgett, 1998; Raqab & Kundu, 2005; Kundu & Gupta, 2005, 2006; Raqab, et al., 2008; Kundu & Raqab, 2009). Reliability in a multicomponent stress-strength was developed by Bhattacharyya and Johnson (1974) and Pandey and Borhan Uddin (1985) and the references therein cover the study of estimating  $P(Y < X)$  in many standard distributions assigned to one or both of stress, strength variates. Recently Srinivasa Rao and Kantam (2010)

studied estimation of reliability in multicomponent stress-strength for log-logistic distribution.

Suppose a system, with  $k$  identical components, functions if  $s(1 \leq s \leq k)$  or more of the components simultaneously operate. In its operating environment, the system is subjected to a stress  $Y$  which is a random variable with distribution function  $G(\cdot)$ . The strengths of the components, that is the minimum stresses to cause failure, are independent and identically distributed random variables with distribution function  $F(\cdot)$ . Then the system reliability, which is the probability that the system does not fail, is the function  $R_{s,k}$  given in (3). The estimation of survival probability in a multicomponent stress-strength system when the stress, strength variates are following Rayleigh distribution is not paid much attention. Therefore, this article studies the estimation of reliability in multicomponent stress-strength model with reference to Rayleigh distribution.

### Maximum Likelihood Estimator of $R_{s,k}$

Let  $X \sim GR(\alpha, \lambda)$  and  $Y \sim GR(\beta, \lambda)$  with unknown shape parameters  $\alpha, \beta$  and common scale parameter  $\lambda$ , where  $X$  and  $Y$  are independently distributed. The reliability in multicomponent stress-strength for generalized Rayleigh distribution using (3) results in:

$$\begin{aligned} R_{s,k} &= \sum_{i=s}^k \binom{k}{i} \int_0^{\infty} \left[ 1 - \left( 1 - e^{-(\lambda y)^2} \right)^{\beta} \right]^i \left[ \left( 1 - e^{-(\lambda y)^2} \right)^{\beta} \right]^{k-i} 2\alpha\lambda^2 y e^{-(\lambda y)^2} \left( 1 - e^{-(\lambda y)^2} \right)^{\alpha-1} dy \\ &= \sum_{i=s}^k \binom{k}{i} \int_0^1 [1-t^{\beta}]^i [t^{\beta}]^{k-i} \alpha t^{\alpha-1} dt \quad \text{where } t = 1 - e^{-(\lambda y)^2} \\ &= \sum_{i=s}^k \binom{k}{i} \nu \int_0^1 z^{k-i+\nu-1} [1-z]^i dz \quad \text{if } z = t^{\beta}, \nu = \frac{\alpha}{\beta} \\ &= \nu \sum_{i=s}^k \binom{k}{i} \beta(k+\nu-i, i+1). \end{aligned}$$

After simplification this reduces to

$$R_{s,k} = \nu \sum_{i=s}^k \frac{k!}{(k-i)!} \left[ \prod_{j=0}^i (k+\nu-j) \right]^{-1} \quad (4)$$

because  $k$  and  $i$  are integers. The probability in (4) is called reliability in a multicomponent stress-strength model. If  $\alpha$  and  $\beta$  are not known, it is necessary to estimate  $\alpha$  and  $\beta$  to estimate  $R_{s,k}$ . In this article  $\alpha$  and  $\beta$  are estimated using the ML method. The estimates are substituted in  $\nu$  to obtain an estimate of  $R_{s,k}$  using equation (4).

It is known that the method of Maximum Likelihood Estimation (MLE) has invariance property. In this direction, this article proposes the ML estimator for the reliability of a multicomponent stress-strength model by considering the estimators of the parameters of stress, strength distributions by ML method of estimation in a generalized Rayleigh distribution.

Let  $x_1 < x_2 < \dots < x_n$ ;  $y_1 < y_2 < \dots < y_m$  be two ordered random samples of size  $n, m$  respectively on strength, stress variates each following GRD with shape parameters  $\alpha$  and  $\beta$ , common scale parameter  $\lambda$ . The log-likelihood function of the observed sample is

$$\begin{aligned} L(\alpha, \beta, \lambda) = & (n+m) \ln 2 + n \ln \alpha + m \ln \beta \\ & + 2(n+m) \ln \lambda - \lambda^2 \left[ \sum_{i=1}^n x_i^2 + \sum_{j=1}^m y_j^2 \right] + \sum_{i=1}^n \ln x_i + \sum_{j=1}^m \ln y_j \\ & + (\alpha-1) \sum_{i=1}^n \ln \left( 1 - e^{-(\lambda x_i)^2} \right) + (\beta-1) \sum_{j=1}^m \ln \left( 1 - e^{-(\lambda y_j)^2} \right) \end{aligned} \quad (5)$$

The MLEs of  $\alpha, \beta$  and  $\lambda$ , for example,  $\hat{\alpha}, \hat{\beta}$  and  $\hat{\lambda}$ , respectively can be obtained as the iterative solution of

$$\frac{\partial L}{\partial \alpha} = 0 \Rightarrow \frac{n}{\alpha} + \sum_{i=1}^n \ln \left( 1 - e^{-(\lambda x_i)^2} \right) = 0 \quad (6)$$

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \frac{m}{\beta} + \sum_{j=1}^m \ln \left( 1 - e^{-(\lambda y_j)^2} \right) = 0 \quad (7)$$

$$\begin{aligned} \frac{\partial L}{\partial \lambda} = 0 \Rightarrow & \frac{2(m+n)}{\lambda} - 2\lambda \left[ \sum_{i=1}^n x_i^2 + \sum_{j=1}^m y_j^2 \right] + (\alpha-1) \sum_{i=1}^n \frac{2\lambda x_i^2 e^{-(\lambda x_i)^2}}{1 - e^{-(\lambda x_i)^2}} \\ & + (\beta-1) \sum_{j=1}^m \frac{2\lambda y_j^2 e^{-(\lambda y_j)^2}}{1 - e^{-(\lambda y_j)^2}} = 0 \end{aligned} \quad (8)$$

and then from (6), (7) and (8)

$$\hat{\alpha} = \frac{-n}{\sum_{i=1}^n \ln(1 - e^{-(\hat{\lambda} x_i)^2})} \quad (9)$$

$$\hat{\beta} = \frac{-m}{\sum_{j=1}^m \ln(1 - e^{-(\hat{\lambda} y_j)^2})} \quad (10)$$

where  $\hat{\lambda}$  can be obtained as the solution of non-linear equation

$$\begin{aligned} g(\lambda) = 0 \Rightarrow & \frac{m+n}{\lambda} - \frac{n\lambda \sum_{i=1}^n \frac{x_i^2 e^{-(\lambda x_i)^2}}{1 - e^{-(\lambda x_i)^2}}}{\sum_{k=1}^n \ln(1 - e^{-(\lambda x_k)^2})} - \frac{m \sum_{j=1}^m \frac{y_j^2 e^{-(\lambda y_j)^2}}{1 - e^{-(\lambda y_j)^2}}}{\sum_{k=1}^m \ln(1 - e^{-(\lambda y_k)^2})} \\ & - \sum_{i=1}^n \frac{\lambda x_i^2}{1 - e^{-(\lambda x_i)^2}} - \sum_{j=1}^m \frac{\lambda y_j^2}{1 - e^{-(\lambda y_j)^2}} = 0 \end{aligned} \quad (11)$$

Therefore,  $\hat{\lambda}$  is simple iterative solution of non-linear equation  $g(\lambda) = 0$ . Once  $\hat{\lambda}$  is known,  $\hat{\alpha}$  and  $\hat{\beta}$  can be obtained from (9) and (10) respectively. Therefore, the MLE of  $R_{s,k}$  becomes

$$R_{s,k} = \hat{\nu} \sum_{i=s}^k \frac{k!}{(k-i)!} \left[ \prod_{j=0}^i (k + \hat{\nu} - j) \right]^{-1} \quad \text{where } \hat{\nu} = \frac{\hat{\alpha}}{\hat{\beta}}. \quad (12)$$

## ESTIMATION OF RELIABILITY IN STRESS-STRENGTH

The asymptotic confidence interval for  $R_{s,k}$ , is calculated as: First, the asymptotic variance of the MLE is given by

$$V(\hat{\alpha}) = \left[ E \left( -\frac{\partial^2 L}{\partial \alpha^2} \right) \right]^{-1} = \frac{\alpha^2}{n} \text{ and } V(\hat{\beta}) = \left[ E \left( -\frac{\partial^2 L}{\partial \beta^2} \right) \right]^{-1} = \frac{\beta^2}{m} \quad (13)$$

The asymptotic variance (AV) of an estimate of  $R_{s,k}$  which a function of two independent statistics, for example,  $\alpha, \beta$  is given by Rao (1973).

$$AV(\hat{R}_{s,k}) = V(\hat{\alpha}) \left( \frac{\partial R_{s,k}}{\partial \alpha} \right)^2 + V(\hat{\beta}) \left( \frac{\partial R_{s,k}}{\partial \beta} \right)^2 \quad (14)$$

From the asymptotic optimum properties of MLEs (Kendall & Stuart, 1979) and of linear unbiased estimators (David, 1981), it is known that MLEs are asymptotically equally efficient having the Cramer-Rao lower bound as their asymptotic variance as given in (13). Thus, from Equation (14), the asymptotic variance of  $\hat{R}_{s,k}$  can be obtained.

To avoid the difficulty of derivation of  $R_{s,k}$ , the derivatives of  $R_{s,k}$  are obtained for  $(s,k)=(1,3)$  and  $(2,4)$  separately, they are given by

$$\begin{aligned} \frac{\partial R_{1,3}}{\partial \alpha} &= \frac{-3}{\beta(3+\hat{\nu})^2} \text{ and } \frac{\partial R_{1,3}}{\partial \beta} = \frac{3\hat{\nu}}{\beta(3+\hat{\nu})^2} \cdot \\ \frac{\partial R_{2,4}}{\partial \alpha} &= \frac{-12(7+2\hat{\nu})}{\beta[(3+\hat{\nu})(4+\hat{\nu})]^2} \text{ and } \frac{\partial R_{2,4}}{\partial \beta} = \frac{12\hat{\nu}(7+2\hat{\nu})}{\beta[(3+\hat{\nu})(4+\hat{\nu})]^2} \cdot \end{aligned}$$

$$\text{Thus, } AV(\hat{R}_{1,3}) = \frac{9\hat{\nu}^2}{(3+\hat{\nu})^4} \left( \frac{1}{n} + \frac{1}{m} \right) \text{ and}$$

$$AV(\hat{R}_{2,4}) = \frac{144\hat{\nu}^2(7+2\hat{\nu})^2}{[(3+\hat{\nu})(4+\hat{\nu})]^4} \left( \frac{1}{n} + \frac{1}{m} \right).$$

$$\text{As } n \rightarrow \infty, m \rightarrow \infty, \frac{\hat{R}_{s,k} - R_{s,k}}{AV(\hat{R}_{s,k})} \xrightarrow{d} N(0,1),$$

and the asymptotic  $100(1-\alpha)\%$  confidence interval for  $R_{s,k}$  is given by

$$\hat{R}_{s,k} \mp Z_{(1-\alpha/2)} \sqrt{AV(\hat{R}_{s,k})}.$$

The asymptotic  $100(1-\alpha)\%$  confidence interval for  $R_{1,3}$  is given by

$$\hat{R}_{1,3} \mp Z_{(1-\alpha/2)} \frac{3\hat{\nu}}{(3+\hat{\nu})^2} \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}, \text{ where } \hat{\nu} = \hat{\alpha} / \hat{\beta}.$$

The asymptotic  $100(1-\alpha)\%$  confidence interval for  $R_{2,4}$  is given by

$$\hat{R}_{2,4} \mp Z_{(1-\alpha/2)} \frac{12\hat{\nu}(7+2\hat{\nu})}{[(3+\hat{\nu})(4+\hat{\nu})]^2} \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}, \text{ where } \hat{\nu} = \hat{\alpha} / \hat{\beta}.$$

where  $Z_{(1-\alpha/2)}$  is the  $(1-\alpha/2)^{\text{th}}$  percentile of the standard normal distribution.

## Simulation Study and Data Analysis

### Simulation Study

Results based on Monte Carlo simulations to compare the performance of the  $R_{s,k}$  using different sample sizes are presented. 3,000 random sample of size 10(5)35 each from stress population, strength population were generated for  $(\alpha, \beta) = (3.0, 1.0), (2.5, 1.0), (2.0, 1.0), (1.5, 1.0), (1.0, 1.0), (1.5, 2.0), (1.5, 2.5)$  and  $(1.5, 3.0)$  on lines of Bhattacharyya and Johnson (1974). The ML estimators of  $\alpha$  and  $\beta$  were then substituted in  $\nu$  to obtain the reliability in a multicomponent stress-strength for  $(s, k) = (1, 3), (2, 4)$ . The average bias and average mean square error (MSE) of the reliability estimates over the 3,000 replications are given in Tables 1 and 2. Average confidence length and coverage probability of the simulated 95% confidence intervals of  $R_{s,k}$  are given in Tables 3 and 4. The true value of reliability in multicomponent stress-strength with the given combinations of  $(\alpha, \beta)$  for  $(s, k) = (1, 3)$  are 0.563, 0.600, 0.643, 0.692, 0.750, 0.800, 0.833, 0.857, 0.875 and for  $(s, k) = (2, 4)$  are 0.355, 0.400, 0.454, 0.519, 0.600, 0.674, 0.725,

# ESTIMATION OF RELIABILITY IN STRESS-STRENGTH

0.762, 0.790. Thus the true value of reliability in multicomponent stress- strength increases as  $\beta$  increases for a fixed  $\alpha$  whereas reliability in multicomponent stress- strength decreases as  $\alpha$  increases for a fixed  $\beta$  in both the cases of  $(s, k)$ . Therefore, the true value of reliability is increases as  $\nu$  decreases and vice-versa. The average bias and average MSE are decreases as sample size increases for both  $(s, k)$ . It verifies the consistency property of the MLE of  $R_{s,k}$ . Also the bias is negative in both situations of  $(s, k)$ . Whereas, among the parameters the absolute bias and MSE are increases as  $\alpha$  increases for a fixed  $\beta$  in both the cases of  $(s, k)$  and the absolute bias and MSE are decreases as  $\beta$  increases for a fixed  $\alpha$  in both the cases of  $(s, k)$ . The average length of the confidence interval also decreases as the sample size increases. The coverage probability is close to the nominal value in all cases but slightly less than 0.95 in most of the combinations. Overall, the performance of the confidence interval is good for all combinations of parameters. Whereas, among the parameters observed, the same phenomenon for average length and average coverage probability were observed in the case of average bias and MSE.

**Table 1.** Average bias of the simulated estimates of  $R_{s,k}$

		$(\alpha, \beta)$								
$(s,k)$	$(n,m)$	(3.5,1.5)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)	(1.5,3.5)
(1,3)	(10,10)	-0.02036	-0.01872	-0.01650	-0.01357	-0.00984	-0.00649	-0.00429	-0.00360	-0.00249
	(15,15)	-0.01517	-0.01412	-0.01268	-0.01075	-0.00825	-0.00592	-0.00428	-0.00311	-0.00227
	(20,20)	-0.00860	-0.00773	-0.00669	-0.00548	-0.00367	-0.00277	-0.00179	-0.00114	-0.00101
	(25,25)	-0.00851	-0.00766	-0.00657	-0.00521	-0.00357	-0.00215	-0.00122	-0.00060	-0.00017
	(30,30)	-0.00679	-0.00613	-0.00528	-0.00421	-0.00290	-0.00175	-0.00098	-0.00046	-0.00010
	(35,35)	-0.00655	-0.00610	-0.00517	-0.00413	-0.00255	-0.00147	-0.00073	-0.00021	-0.00008
(2,4)	(10,10)	-0.01113	-0.01143	-0.01124	-0.01027	-0.00819	-0.00657	-0.00539	-0.00472	-0.00348
	(15,15)	-0.00908	-0.00945	-0.00950	-0.00903	-0.00776	-0.00601	-0.00447	-0.00324	-0.00229
	(20,20)	-0.00601	-0.00599	-0.00571	-0.00508	-0.00400	-0.00176	-0.00120	-0.00089	-0.00138
	(25,25)	-0.00512	-0.00505	-0.00473	-0.00405	-0.00293	-0.00168	-0.00073	-0.00054	-0.00045
	(30,30)	-0.00420	-0.00416	-0.00391	-0.00338	-0.00247	-0.00144	-0.00063	-0.00043	-0.00040
	(35,35)	-0.00386	-0.00390	-0.00377	-0.00237	-0.00160	-0.00125	-0.00058	-0.00019	-0.00035



**Table 2.** Average MSE of the simulated estimates of  $R_{s,k}$

		$(\alpha, \beta)$								
$(s,k)$	$(n,m)$	(3.5,1.5)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)	(1.5,3.5)
(1,3)	(10,10)	0.01529	0.01437	0.01300	0.01109	0.00854	0.00627	0.00477	0.00375	0.00303
	(15,15)	0.01052	0.00988	0.00894	0.00764	0.00590	0.00436	0.00334	0.00264	0.00214
	(20,20)	0.00713	0.00666	0.00599	0.00509	0.00392	0.00289	0.00221	0.00175	0.00143
	(25,25)	0.00592	0.00551	0.00494	0.00418	0.00322	0.00236	0.00181	0.00144	0.00117
	(30,30)	0.00460	0.00428	0.00383	0.00323	0.00248	0.00182	0.00139	0.00110	0.00090
	(35,35)	0.00402	0.00374	0.00337	0.00286	0.00220	0.00162	0.00125	0.00099	0.00081
(2,4)	(10,10)	0.01801	0.01877	0.01900	0.01831	0.01611	0.01320	0.01077	0.00889	0.00744
	(15,15)	0.01285	0.01337	0.01351	0.01298	0.01140	0.00932	0.00762	0.00630	0.00529
	(20,20)	0.00906	0.00936	0.00938	0.00895	0.00781	0.00635	0.00518	0.00429	0.00361
	(25,25)	0.00754	0.00778	0.00778	0.00740	0.00643	0.00522	0.00426	0.00352	0.00296
	(30,30)	0.00594	0.00612	0.00610	0.00578	0.00500	0.00404	0.00328	0.00271	0.00228
	(35,35)	0.00522	0.00538	0.00538	0.00512	0.00445	0.00361	0.00295	0.00244	0.00205

**Table 3.** Average confidence length of the simulated 95% confidence intervals of  $R_{s,k}$

		$(\alpha, \beta)$								
$(s,k)$	$(n,m)$	(3.5,1.5)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)	(1.5,3.5)
(1,3)	(10,10)	0.4091	0.4021	0.3880	0.3632	0.3224	0.2763	0.2400	0.2113	0.1884
	(15,15)	0.3399	0.3334	0.3210	0.2999	0.2658	0.2279	0.1981	0.1747	0.1559
	(20,20)	0.2975	0.2911	0.2794	0.2601	0.2296	0.1961	0.1701	0.1497	0.1335
	(25,25)	0.2675	0.2617	0.2512	0.2338	0.2063	0.1762	0.1529	0.1346	0.1201
	(30,30)	0.2453	0.2398	0.2300	0.2140	0.1887	0.1612	0.1398	0.1232	0.1099
	(35,35)	0.2277	0.2226	0.2135	0.1986	0.1753	0.1498	0.1301	0.1146	0.1023
(2,4)	(10,10)	0.4569	0.4719	0.4802	0.4760	0.4498	0.4055	0.3637	0.3274	0.2966
	(15,15)	0.3834	0.3953	0.4012	0.3968	0.3739	0.3366	0.3019	0.2719	0.2466
	(20,20)	0.3396	0.3492	0.3533	0.3479	0.3260	0.2920	0.2610	0.2346	0.2123
	(25,25)	0.3056	0.3143	0.3180	0.3131	0.2934	0.2628	0.2350	0.2112	0.1912
	(30,30)	0.2813	0.2891	0.2923	0.2875	0.2691	0.2409	0.2153	0.1935	0.1752
	(35,35)	0.2611	0.2684	0.2714	0.2670	0.2500	0.2240	0.2003	0.1801	0.1631

## ESTIMATION OF RELIABILITY IN STRESS-STRENGTH

**Table 4.** Average coverage probability of the simulated 95% confidence intervals of  $R_{s,k}$

		$(\alpha, \beta)$								
$(s,k)$	$(n,m)$	(3.5,1.5)	(3.0,1.5)	(2.5,1.5)	(2.0,1.5)	(1.5,1.5)	(1.5,2.0)	(1.5,2.5)	(1.5,3.0)	(1.5,3.5)
(1,3)	(10,10)	0.9090	0.9140	0.9193	0.9280	0.9317	0.9303	0.9267	0.9260	0.9267
	(15,15)	0.9120	0.9150	0.9187	0.9213	0.9250	0.9273	0.9290	0.9287	0.9253
	(20,20)	0.9303	0.9347	0.9370	0.9390	0.9390	0.9377	0.9380	0.9347	0.9303
	(25,25)	0.9227	0.9267	0.9317	0.9360	0.9383	0.9373	0.9357	0.9333	0.9277
	(30,30)	0.9353	0.9403	0.9423	0.9463	0.9490	0.9463	0.9433	0.9400	0.9390
	(35,35)	0.9317	0.9330	0.9353	0.9387	0.9387	0.9400	0.9363	0.9337	0.9297
(2,4)	(10,10)	0.9113	0.9153	0.9203	0.9243	0.9273	0.9277	0.9257	0.9250	0.9243
	(15,15)	0.9103	0.9160	0.9190	0.9193	0.9233	0.9263	0.9297	0.9260	0.9223
	(20,20)	0.9310	0.9340	0.9370	0.9380	0.9367	0.9367	0.9360	0.9323	0.9287
	(25,25)	0.9237	0.9287	0.9313	0.9353	0.9370	0.9350	0.9323	0.9307	0.9257
	(30,30)	0.9357	0.9397	0.9437	0.9437	0.9477	0.9457	0.9420	0.9403	0.9397
	(35,35)	0.9297	0.9317	0.9367	0.9360	0.9407	0.9393	0.9367	0.9327	0.9300

### Data Analysis

Strength data, which was originally reported by Badar and Priest (1982), represents the strength measured in GPA for single carbon fibers and impregnated 1,000-carbon fiber tows. Single fibers were tested under tension at gauge lengths of 20 mm (Data Set I) and 10 mm (Data Set II), with sample sizes  $n = 69$  and  $m = 63$  respectively (see Data sets I and II).

Data Set I (gauge lengths of 20 mm).

1.312, 1.314, 1.479, 1.552, 1.700, 1.803, 1.861, 1.865, 1.944, 1.958, 1.966, 1.997, 2.006, 2.021, 2.027, 2.055, 2.063, 2.098, 2.140, 2.179, 2.224, 2.240, 2.253, 2.270, 2.272, 2.274, 2.301, 2.301, 2.359, 2.382, 2.382, 2.426, 2.434, 2.435, 2.478, 2.490, 2.511, 2.514, 2.535, 2.554, 2.566, 2.570, 2.586, 2.629, 2.633, 2.642, 2.648, 2.684, 2.697, 2.726, 2.770, 2.773, 2.800, 2.809, 2.818, 2.821, 2.848, 2.880, 2.809, 2.818, 2.821, 2.848, 2.880, 2.954, 3.012, 3.067, 3.084, 3.090, 3.096, 3.128, 3.233, 3.433, 3.585, 3.585.

Data Set II (gauge lengths of 10 mm).

1.901, 2.132, 2.203, 2.228, 2.257, 2.350, 2.361, 2.396, 2.397, 2.445, 2.454, 2.474, 2.518, 2.522, 2.525, 2.532, 2.575, 2.614, 2.616, 2.618, 2.624, 2.659, 2.675, 2.738, 2.740, 2.856, 2.917, 2.928, 2.937, 2.937, 2.977, 2.996, 3.030, 3.125, 3.139, 3.145, 3.220, 3.223, 3.235, 3.243, 3.264, 3.272, 3.294, 3.332, 3.346, 3.377, 3.408, 3.435, 3.493, 3.501, 3.537, 3.554, 3.562, 3.628, 3.852, 3.871, 3.886, 3.971, 4.024, 4.027, 4.225, 4.395, 5.020.

Surles and Padgett (1998, 2001) observed that generalized Rayleigh works well for strength data. Raqab and Kundu (2005) analyzed the data by subtracting 1.0 and 1.8 from the first and second data set respectively. The transformed data sets correspond to 20 mm and 10 mm gauge lengths are assumed to follow  $GR(\alpha, \lambda)$  and  $GR(\beta, \lambda)$  respectively. The obtained final estimates for these two data sets are  $\hat{\alpha} = 2.4421$ ,  $\hat{\beta} = 1.4216$ , and  $\hat{\lambda} = 0.8598$ . Also they checked the validity of the models using the Kolmogorov-Smirnov (K-S) tests for each data set. It was observed that for Data Sets I and II, the K-S distances are 0.09 and 0.12 with the corresponding  $p$  values of 0.6069 and 0.2845 respectively. It indicates that the GR model provides reasonable fit to the transformed data sets.

Based on estimates of  $\alpha$  and  $\beta$  the MLE of  $R_{s,k}$  become  $\hat{R}_{1,3} = 0.63588$  and  $\hat{R}_{2,4} = 0.44484$ . The 95% confidence intervals for  $R_{1,3}$  become (0.55680, 0.71496) and for  $R_{2,4}$  become (0.34387, 0.54581).

## Conclusions

This article used real data sets to investigate multicomponent stress-strength reliability for a generalized Rayleigh distribution when both stress, strength variates follow the same population. Asymptotic confidence intervals for multicomponent stress-strength reliability were estimated using the ML method. Simulation results indicate that the average bias and average MSE decreases as sample size increases in both cases of  $(s, k)$ . Among the parameters the absolute bias and MSE are increases (decreases) as  $\alpha$  increases ( $\beta$  increases) in both the cases of  $(s, k)$ . The length of the confidence interval also decreases as the sample size increases and coverage probability is close to the nominal value in all sets of parameters considered.

## References

- Awad, M. & Gharraf, K. (1986). Estimation of  $p(Y < X)$  in Burr case: A comparative study. *Communications in Statistics - Simulations & Comp.*, 15: 389-403.
- Badar, M. G. & Priest, A. M. (1982). Statistical aspects of fiber and bundle strength in hybrid composites. In T. Hayashi, K. Kawata, and S. Umekawa (eds.), *Progress in Science and Engineering Composites*, (pp. 1129-1136). Tokyo: ICCM-IV.

## ESTIMATION OF RELIABILITY IN STRESS-STRENGTH

- Bhattacharyya, G. K. & Johnson, R. A. (1974). Estimation of reliability in multicomponent stress – strength model. *JASA*, 69: 966-970.
- David, H. A. (1981). *Order Statistics*. New York: John Wiley and Sons.
- Downtown, F. (1973). The estimation of  $p(X>Y)$  in the normal case. *Technometrics*, 15: 551-558.
- Enis, P. & Geisser, S. (1971). Estimation of the probability that  $Y<X$ . *JASA*, 66: 162- 168.
- Kendall, M. G. & Stuart, A. (1979). *The Advanced Theory of Statistics*, (Vol. 2). London: Charles Griffin and Company Limited.
- Kundu, D. & Gupta, R. D. (2005). Estimation of  $p(Y<X)$  for the generalized exponential distribution. *Metrika*, 61(3): 291-308.
- Kundu, D. & Gupta, R. D. (2006). Estimation of  $p(Y<X)$  for Weibull distribution, *IEEE Transactions on Reliability*, 55(2): 270-280.
- Kundu, D. & Raqab, M. Z. (2005). Generalized Rayleigh distribution: different methods of estimation. *Computational Statistics and Data Analysis*, 49: 187 - 200.
- Kundu, D. and Raqab, M. Z. (2009). Estimation of  $R=p(Y<X)$  for three-parameter Weibull distribution, *Statistics and Probability Letters*, 79: 1839-1846.
- McCool, J. I. (1991). Inference on  $p(Y<X)$  in the Weibull case, *Communications in Statistics - Simulations & Comp.*, 20: 129-148.
- Mudholkar, G. S. & Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure data, *IEEE Transactions on Reliability*, 42: 299-302.
- Nandi, S. B. & Aich, A. B. (1994). A note on estimation of  $p(X>Y)$  for some distributions useful in life- testing, *IAPQR Transactions*, 19(1): 35 - 44.
- Pandey, M. & Uddin, B. (1985). Estimation of reliability in multicomponent stress – strength model following Burr distribution. *Proceedings of the First Asian congress on Quality and Reliability*, (pp. 307 – 312). New Delhi, India.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. India: Wiley Eastern Limited.
- Raqab, M. Z. & Kundu, D. (2005). Comparison of different estimators of  $p(Y<X)$  for a scaled Burr type X distribution, *Communications in Statistics – Simulation and Computation*, 34(2): 465 - 483.
- Raqab, M. Z., Madi, M. T. & Kundu, D. (2008). Estimation of  $p(Y<X)$  for the 3-parameter generalized exponential distribution, *Communications in Statistics - Theory and Methods*, 37(18): 2854 - 2864.

Srinivasa Rao, G. & Kantam, R. R. L. (2010). Estimation of reliability in multicomponent stress–strength model: log-logistic distribution, *Electronic Journal of Applied Statistical Analysis*, 3(2): 75-84.

Surles, J. G. & Padgett, W. J. (1998). Inference for  $p(Y < X)$  in the Burr Type X model, *Journal of Applied Statistical Sciences*, 7: 225 - 238.

Surles, J. G. & Padgett, W. J. (2001). Inference for reliability and stress-strength for a scaled Burr Type X distribution, *Lifetime Data Analysis*, 7: 187-200.

# Stochastic Randomized Response Model for a Quantitative Sensitive Random Variable

**Sarjinder Singh**

Texas A&M University-Kingsville  
Kingsville, TX

**Stephen A. Sedory**

Texas A&M University-Kingsville  
Kingsville, TX

---

A new stochastic randomized response model is introduced that is useful for estimating the population mean of a sensitive quantitative variable. The proposed stochastic randomized response model is an extension of the stochastic randomized response model from a qualitative sensitive variable to a quantitative variable found in Singh (2002). The stochastic nature of a randomized response device helps increase a respondent's cooperation while collecting information on sensitive variables in a society. The Bar-Lev, Bobovitch, and Boukai (2004) model is shown to be a special case of the proposed model.

*Keywords:* Sensitive variable; estimation of population mean, stochastic randomized response device

---

## Introduction

The collection of data through personal interview surveys on sensitive issues, such as induced abortion, drug abuse and family income, is a serious issue. For example, some questions are sensitive:

- By how much did you underreport your income on your 2009 tax return?
- Are you a Baath Party Member?
- How many abortions have you had?
- How many children have you molested?
- Do you use illegal drugs?

Randomized response techniques are one way to encourage people to answer truthfully. Warner (1965) considered the case where the respondents in a

---

*Dr. Singh is an Associate Professor in the Department of Mathematics. Email him at: [sarjinder@yahoo.com](mailto:sarjinder@yahoo.com). Dr. Sedory is a Professor in the Department of Mathematics. Email him at [stephen.sedory@tamuk.edu](mailto:stephen.sedory@tamuk.edu).*

population can be divided into two mutually exclusive groups: one group with stigmatizing or otherwise sensitive characteristic  $A$ , and the other group without it. For estimating  $\pi$ , the proportion of respondents in the population belonging to the sensitive group  $A$ , a simple random sample of  $n$  respondents is selected with replacement from the population. For collecting information on the sensitive characteristic, Warner (1965) made use of a randomization device. One such device could be a deck of cards. On each card is written one of the following two statements: "I belong to group  $A$ ", or "I do not belong to group  $A$ ." The statements occur with relative frequencies  $p_0$  and  $(1-p_0)$  respectively in the deck of cards. Each respondent in the sample is asked to select a card at random from the well-shuffled deck. Without showing the card to the interviewer, the interviewee answers the question, "Is the statement true for you?" The number of people,  $n_1$ , who answer yes is binomially distributed with parameters  $p_0\pi + (1-p_0)(1-\pi)$  and  $n$ . The maximum likelihood estimator of  $\pi$  exists for  $p_0 \neq 0.5$  is given by

$$\hat{\pi}_w = \frac{(n_1/n) - (1-p_0)}{2p_0 - 1}. \quad (1)$$

The estimator is unbiased with variance:

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p_0(1-p_0)}{n(2p_0-1)^2}. \quad (2)$$

In a randomized response procedure, the cooperation of respondents depends on the confidentiality of their responses – the greater the confidentiality, the greater the cooperation from the respondents. Conversely, if the magnitude of response confidentiality is increased, the efficiency of the estimator of population proportion  $\pi$  is adversely affected. It is necessary, therefore, to strike a balance between response confidentiality and estimator efficiency. Several researchers have tried to modify data collection procedures to increase the confidentiality of responses. Horvitz, et al. (1967) felt that by providing the respondent with the opportunity of replying to one of two questions in which one question is completely innocuous and unrelated to the sensitive attribute, the sense of confidentiality among the respondents could possibly be increased. The theoretical framework for their approach was developed by Greenberg, et al. (1969).

Singh (2002) considered another procedure that may result in a greater sense of response confidentiality among the sampled individuals. The procedure can be used in surveys where respondents selected in the sample assemble at common place for the conduction of the survey. This could be a situation of collecting data from a small town, community or organization. The procedure invokes  $K$  decks of cards (called a stochastic randomization device) with different proportions of cards carrying the statement, "I belong to group A." After explaining to the respondents how the randomization device provides confidentiality to their responses, the investigator asks one of the assembled respondents to randomly select a deck of cards from the box containing  $K$  decks of cards. The deck is then used to collect information on the sensitive attribute from the respondents. Every sampled respondent draws one card from the selected deck of cards and reads the statement on it. In the proposed procedure every respondent is provided with two identical slips of paper with yes or no printed on them. According to his status in relation to the statement printed on the card drawn, each respondent is requested to put one of the two slips of paper into an empty box. After the survey is completed, the number of yes answers is counted from the box and the proportion,  $p^*$ , for the deck used in the survey is noted. Random selection of one randomization device from several such devices may help in increasing the sense of confidentiality among the respondents. The choice of values of  $p$  for preparing  $K$  decks of cards for the survey is important in this procedure. These  $K$  values of  $p$  could either be purposively selected by an investigator, or they could be taken as a random sample from a known discrete or continuous density function. Let this density function be denoted by  $f(p)$ . The value of  $p$  corresponding to the deck used in the survey will be selected from this random sample of  $p$ -values with equal probabilities. Thus, the value of  $p^*$  used in the survey is a random variable with  $f(p)$  as its probability density function. When  $f(p)$  is a one-point distribution, the proposed procedure reduces to Warner (1965). Singh (2002) assumes let  $n_1$  persons in the sample answered yes and  $(n-n_1)$  answer no. Because the probability of a yes answer for a particular choice of  $p^*$  is given by

$$\theta = p^* \pi + (1 - \pi)(1 - p^*). \quad (3)$$

Singh (2002) considers the unbiased estimator of  $\pi$  as

$$\hat{\pi}_R = \frac{\hat{\theta} - (1 - p^*)}{2p^* - 1}, \quad (4)$$



where  $\hat{\theta} = n_1/n$  is the proportion of yes answers in the sample, with variance

$$V(\hat{\pi}_R) = \frac{\pi(1-\pi)}{n} + \int_a^b \frac{p(1-p)}{n(2p-1)^2} f(p) dp \dots a \leq p \leq b, p \neq 0.5 \quad (5)$$

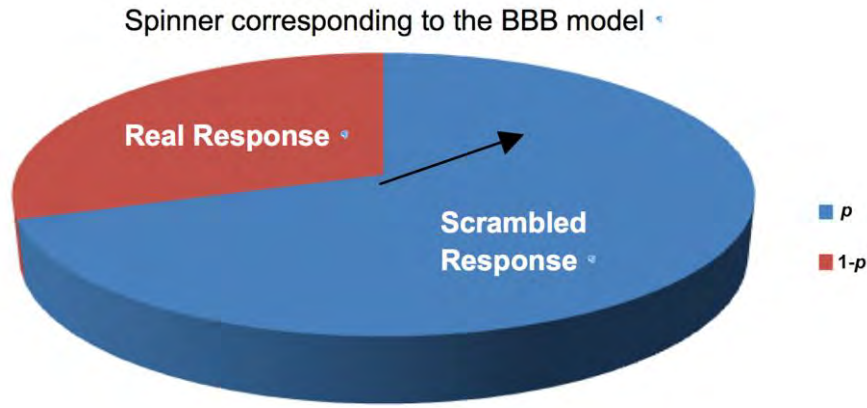
where  $f(p)$  denotes the probability density function (p.d.f.) of  $p$ . Singh (2002) showed that the stochastic version of the Warner (1965) model remains more efficient than the pioneer Warner (1965) model. In the same article, Singh (2002) also considers stochastic version of the Kuk (1990) model and showed its benefits over the original Kuk (1990) model. Recent work on randomized response techniques is found in Abdelfatah, et al. (2013).

### Quantitative Randomized Response Model

In the randomized response model due to Bar-Lev, Bobovitch, and Boukai (2004), hereafter the BBB model, the distribution of responses is given by:

$$Y_i = \begin{cases} X_i S & \text{with probability } (1-p) \\ X_i & \text{with probability } p \end{cases} \quad (6)$$

In other words, each respondent is requested to rotate a spinner unobserved by the interviewer and if the spinner stops in the shaded area then he/she is requested to report the real response on the sensitive variable, for example  $X_i$ ; and if the spinner stops in the non-shaded area then he/she is requested to report the scrambled response, for example  $X_i S$ , where  $S$  is any scrambling variable and its distribution is assumed to be known. In other words,  $E(S) = \theta$  and  $V(S) = \gamma^2$  are assumed to be known. Let  $p$  be the proportion of the shaded area of the spinner and  $(1-p)$  be the non-shaded area of the spinner as shown in the Figure 1.



**Figure 1.** BBB Randomized Response Device.

An unbiased estimator of population mean  $\mu_x$  is given by

$$\hat{\mu}_{X(BBB)} = \frac{1}{n\{(1-p)\theta + p\}} \sum_{i=1}^n Y_i \quad (7)$$

with the variance of the estimator  $\hat{\mu}_{X(BBB)}$  given by

$$V[\hat{\mu}_{X(BBB)}] = \frac{1}{n} \left[ \sigma_x^2 + \mu_x^2 (1 + C_\gamma^2) C_s^2(p) \right] \quad (8)$$

where  $C_s^2(p) = \frac{(1-p)\theta^2(1+C_\gamma^2)+p}{[(1-p)\theta+p]^2} - 1$ , and  $C_\gamma^2 = \gamma^2/\theta^2$ .

### Proposed Stochastic Quantitative Randomized Response Model

Let  $p^*$  be the stochastic proportion of cards in a deck bearing the statement, “Please report the real response  $X_i$ ” and  $(1-p^*)$  be the stochastic proportion of cards in the same deck bearing the statement, “Please report the scrambled

response  $X_i S^*$ , where  $S^*$  is also a stochastic scrambling variable. Let  $\theta^*$  be the mean value, between pre-decided limits  $a$  and  $b$ , of a scrambling variable  $S^*$ . Again this procedure can be used in surveys where the respondents selected in the sample assemble at a common place to take a survey on a sensitive quantitative variable. This could be a situation of collecting data from a small town, community or organization, or a homogeneous stratum.

In practice, it is suggested that, at the gathering place, there is a collection of  $K_1$  decks of pink cards in a box. Every pink deck of cards consists of two types of cards bearing the two statements (a) and (b) with stochastic proportions  $p^*$  and  $(1-p^*)$  respectively. In another box, there are  $K_2$  green decks of cards and each green deck can produce stochastic scrambling variable with different mean values of  $\theta^*$  in the range  $a < \theta^* < b$ . In the presence of all the respondents and the interviewer, a lottery method is used. A huge number of pink decks of cards are left in box I, and a huge number of green decks of cards are left in box II. One green deck is selected and another pink deck is selected by the lottery method. The values of  $p^*$  and  $\theta^*$  remain unknown during and after the survey. Both decks are either returned back to the boxes or are destroyed without looking at the particular values of  $p^*$  and  $\theta^*$  used in the survey. This ensures respondents cooperation and privacy. The decks selected by the lottery method are used in the entire survey. Also note that the values of  $a$  and  $b$  are assumed to be known to the interviewer and interviewees. Thus, in the proposed stochastic randomized response model the distribution of the responses is given by

$$Y_i^* = \begin{cases} X_i S^* & \text{with probability } (1-p^*) \\ X_i & \text{with probability } p^* \end{cases} \quad (9)$$

The following theorems result.

### Theorem 1

An unbiased estimator of the population mean  $\mu_x$  is given by

$$\hat{\mu}_{S(BBB)} = \frac{1}{n\{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n Y_i^* \quad (10)$$

where the values of  $p^*$  and  $\theta^*$  remains unknown to both the interviewer and interviewees, unlike Warner (1965) model. These values  $p^*$  and  $\theta^*$  are derived from the known joint density of  $p^*$  and  $\theta^*$  to get an estimate from the observed responses.

**Proof.** Let  $E_1$  denote the expected value over all possible samples and  $E_2$  denote the expected value over the randomization device for given values of  $\theta^*$  and  $p^*$ . Taking expected value on both sides of (10), results in

$$\begin{aligned} E[\hat{\mu}_{S(BBB)}] &= E\left[\frac{1}{n\{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n Y_i^* \right] = E_1 E_2 \left[ \frac{1}{n\{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n Y_i^* \mid p^*, \theta^* \right] \\ &= E_1 \left[ \frac{1}{n\{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n E_2 \{Y_i^* \mid p^*, \theta^*\} \right] = E_1 \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] = \mu_x, \end{aligned}$$

thus proving the theorem.

## Theorem 2

The variance of the unbiased estimator  $\hat{\mu}_{S(BBB)}$  of the population mean  $\mu_x$  is given by

$$V(\hat{\mu}_{S(BBB)}) = \frac{1}{n} \left[ \sigma_x^2 + \mu_x^2 (1 + C_x^2) \int_a^b \int_0^1 \left\{ \frac{p(1-p)(1-\theta)^2 + (1-p)\gamma^2}{\{p + (1-p)\theta\}^2} \right\} f(p, \theta) dp d\theta \right] \quad (11)$$

where  $\gamma^2 = V(S^*)$  is constant and known.

**Proof.** Let  $V_1$  be the variance over all possible samples and  $V_2$  denote the variance for the given values of the randomization device  $p^*$  and  $\theta^*$ . By the definition of the variance,

$$\begin{aligned}
 V(\hat{\mu}_{S(BBB)}) &= E_1 V_2(\hat{\mu}_{S(BBB)} | p^*, \theta^*) + V_1 E_2(\hat{\mu}_{S(BBB)} | p^*, \theta^*) \\
 &= E_1 V_2 \left[ \frac{1}{n \{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n Y_i^* | p^*, \theta^* \right] + V_1 E_2 \left[ \frac{1}{n \{(1-p^*)\theta^* + p^*\}} \sum_{i=1}^n Y_i^* | p^*, \theta^* \right] \\
 &= E_1 \left[ \frac{1}{n^2 \{(1-p^*)\theta^* + p^*\}^2} \sum_{i=1}^n V_2(Y_i^*) \right] + V_1 \left[ \frac{\sum_{i=1}^n E_2(Y_i^*)}{n \{(1-p^*)\theta^* + p^*\}} \right]
 \end{aligned} \tag{12}$$

Now

$$\begin{aligned}
 V_2(Y_i^* | p^*, \theta^*) &= E_2(Y_i^{*2}) - (E_2(Y_i^*))^2 \\
 &= p^* X_i^2 + (1-p^*) X_i^2 E_2(S^2) - [p^* X_i + (1-p^*) X_i \theta]^2 \\
 &= p^* X_i^2 + (1-p^*) X_i^2 (\gamma^2 + \theta^2) \\
 &\quad - [p^{*2} X_i^2 + (1-p^*)^2 X_i^2 \theta^{*2} + 2p^*(1-p^*) X_i^2 \theta^*] \\
 &= p^*(1-p^*) X_i^2 + (1-p^*) X_i^2 \gamma^2 + X_i^2 \theta^{*2} p^*(1-p^*) \\
 &\quad - 2p^*(1-p^*) X_i^2 \theta^* \\
 &= p^*(1-p^*) X_i^2 [1 + \theta^{*2} - 2\theta^*] + (1-p^*) X_i^2 \gamma^2 \\
 &= p^*(1-p^*) X_i^2 (1-\theta^*)^2 + (1-p^*) X_i^2 \gamma^2 \\
 &= [p^*(1-p^*)(1-\theta^*)^2 + (1-p^*)\gamma^2] X_i^2
 \end{aligned} \tag{13}$$

Thus, plugging (13) into (12):

$$\begin{aligned}
 V(\hat{\mu}_{S(BBB)}) &= E_1 \left[ \frac{p^*(1-p^*)(1-\theta^*)^2 + (1-p^*)\gamma^2 \sum_{i=1}^n X_i^2}{n^2 \{p^* + (1-p^*)\theta^*\}^2} \right] + V_1 \left[ \frac{\sum_{i=1}^n \{p^* + (1-p^*)\theta^*\} X_i}{n(p^* + (1-p^*)\theta^*)} \right] \\
 &= E_1 \left[ \frac{[p^*(1-p^*)(1-\theta^*)^2 + (1-p^*)\gamma^2]}{n \{p^* + (1-p^*)\theta^*\}^2} \right] \left[ \frac{1}{N} \sum_{i=1}^N X_i^2 \right] + V_1 \left[ \frac{1}{n} \sum_{i=1}^n X_i \right]
 \end{aligned}$$

$$\begin{aligned}
 &= E_1 \left[ \frac{[p^*(1-p^*)(1-\theta^*)^2 + (1-p^*)\gamma^2]}{n\{p^* + (1-p^*)\theta^*\}^2} \right] \left( \sigma_x^2 + \mu_x^2 \right) + \frac{\sigma_x^2}{n} \\
 &= \frac{1}{n} \left[ \sigma_x^2 + \mu_x^2 (1 + C_x^2) E_1 \left\{ \frac{p^*(1-p^*)(1-\theta^*)^2 + (1-p^*)\gamma^2}{[p^* + (1-p^*)\theta^*]^2} \right\} \right] \\
 &= \frac{1}{n} \left[ \sigma_x^2 + \mu_x^2 (1 + C_x^2) \int_a^b \int_0^1 \left\{ \frac{p(1-p)(1-\theta)^2 + (1-p)\gamma^2}{[p + (1-p)\theta]^2} \right\} f(p, \theta) dp d\theta \right]
 \end{aligned}$$

which proves the theorem.

### Theorem 3

A joint probability density function of  $p$  and  $\theta$  is given by

$$f(p, \theta) = \frac{[p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1}}{(b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \quad (14)$$

$$0 \leq p \leq 1, \quad a \leq \theta \leq b$$

**Proof.** Consider

$$f(p, \theta) = k [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1}, \quad 0 < p < 1, a < \theta < b \quad (15)$$

Therefore,

$$\begin{aligned}
 \int_a^b \int_0^1 f(p, \theta) dp d\theta &= k \int_a^b \left[ \int_0^1 [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1} dp \right] d\theta, \quad a \leq \theta \leq b \\
 &= k \int_a^b \left[ \int_0^1 \{p^2 + (1-p)^2 \theta^2 + 2\theta p(1-p)\} p^{\alpha-1} (1-p)^{\beta-1} dp \right] d\theta \\
 &= k \int_a^b \left[ \int_0^1 p^{(\alpha+2)-1} (1-p)^{\beta-1} dp + \theta^2 \int_0^1 p^{\alpha-1} (1-p)^{(\beta+2)-1} dp + 2\theta \int_0^1 p^{(\alpha+1)-1} (1-p)^{(\beta+1)-1} dp \right] d\theta
 \end{aligned}$$

$$\begin{aligned}
 &= k \int_a^b \left[ B(\alpha+2, \beta) + \theta^2 B(\alpha, \beta+2) + 2\theta B(\alpha+1, \beta+1) \right] d\theta \\
 &= k \left[ B(\alpha+2, \beta) \int_a^b d\theta + B(\alpha, \beta+2) \int_a^b \theta^2 d\theta + 2B(\alpha+1, \beta+1) \int_a^b \theta d\theta \right]
 \end{aligned}$$

which implies

$$k \left[ (b-a)B(\alpha+2, \beta) + \frac{1}{3}(b^3 - a^3)B(\alpha, \beta+2) + B(\alpha+1, \beta+1)(b^2 - a^2) \right] = 1$$

or

$$k(b-a) \left[ B(\alpha+2, \beta) + \frac{(a^2 + ab + b^2)}{3} B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right] = 1$$

or

$$k = \frac{1}{(b-a) \left[ B(\alpha+2, \beta) + \frac{(a^2 + ab + b^2)}{3} B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \quad (16)$$

Substituting (16) into (15), proves the theorem.

#### Theorem 4

Under the joint probability density function  $f(p, \theta)$ , the variance of the estimator  $\hat{\mu}_{S(BBB)}$  is given by

$$V(\hat{\mu}_{S(BBB)}) = \frac{1}{n} \left[ \sigma_x^2 + \frac{\mu_x^2(1 + C_x^2) \left\{ B(\alpha+1, \beta+1) \left( 1 + \frac{1}{3}(a^2 + ab + b^2) - (b+a) \right) + \gamma^2 B(\alpha, \beta+1) \right\}}{B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1)} \right] \quad (17)$$

**Proof.**

$$\begin{aligned}
 E_1[V(\hat{\mu}_x)] &= \int_a^b \int_0^1 V(\hat{\mu}_x) f(p, \theta) dp d\theta \\
 &= \frac{1}{n(b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \\
 &\quad \times \int_a^b \int_0^1 [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1} \left[ \sigma_x^2 + \mu_x^2(1+C_x^2) \left\{ \frac{p(1-p)(1-\theta)^2 + (1-p)\gamma^2}{\{p + (1-p)\theta\}^2} \right\} \right] dp d\theta \\
 &= \frac{1}{n(b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \\
 &\quad \times \left[ \int_a^b \int_0^1 \sigma_x^2 [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \right. \\
 &\quad \left. + \int_a^b \int_0^1 \mu_x^2(1+C_x^2) \left\{ p(1-p)(1-\theta)^2 + (1-p)\gamma^2 \right\} p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \right] \\
 &= \frac{1}{n(b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \\
 &\quad \times \left[ \sigma_x^2 \int_a^b \int_0^1 [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \right. \\
 &\quad \left. + \mu_x^2(1+C_x^2) \int_a^b \int_0^1 \left\{ p(1-p)(1-\theta)^2 + (1-p)\gamma^2 \right\} p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \right] \\
 &= \frac{\sigma_x^2 I_1 + \mu_x^2(1+C_x^2) I_2}{n(b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]} \quad (18)
 \end{aligned}$$

So that

$$\begin{aligned}
 I_1 &= \int_a^b \int_0^1 [p + (1-p)\theta]^2 p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \\
 &= \int_a^b \int_0^1 [p^2 + (1-p)^2 \theta^2 + 2\theta p(1-p)] p^{\alpha-1} (1-p)^{\beta-1} dp d\theta
 \end{aligned}$$



$$\begin{aligned}
&= \int_a^b \left[ \int_0^1 p^{(\alpha+2)-1} (1-p)^{\beta-1} dp + \theta^2 \int_0^1 p^{\alpha-1} (1-p)^{(\beta+2)-1} dp + 2\theta \int_0^1 p^{(\alpha+1)-1} (1-p)^{(\beta+1)-1} dp \right] d\theta \\
&= \int_a^b \left[ B(\alpha+2, \beta) + \theta^2 B(\alpha, \beta+2) + 2\theta B(\alpha+1, \beta+1) \right] d\theta \\
&= B(\alpha+2, \beta) \int_a^b d\theta + B(\alpha, \beta+2) \int_a^b \theta^2 d\theta + 2B(\alpha+1, \beta+1) \int_a^b \theta d\theta \\
&= B(\alpha+2, \beta)(b-a) + B(\alpha, \beta+2) \left( \frac{b^3 - a^3}{3} \right) + B(\alpha+1, \beta+1)(b^2 - a^2) \\
&= (b-a) \left[ B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1) \right]
\end{aligned}$$

and

$$\begin{aligned}
I_2 &= \int_a^b \int_0^1 \left[ p(1-p)(1-\theta)^2 + (1-p)\gamma^2 \right] p^{\alpha-1} (1-p)^{\beta-1} dp d\theta \\
&= \int_a^b \int_0^1 p^{(\alpha+1)-1} (1-p)^{(\beta+1)-1} (1-\theta)^2 dp d\theta + \gamma^2 \int_a^b \int_0^1 p^{\alpha-1} (1-p)^{(\beta+1)-1} dp d\theta \\
&= \int_a^b (1-\theta)^2 \left[ \int_0^1 p^{(\alpha+1)-1} (1-p)^{(\beta+1)-1} dp \right] d\theta + \gamma^2 \int_a^b \left[ \int_0^1 p^{\alpha-1} (1-p)^{(\beta+1)-1} dp \right] d\theta \\
&= \int_a^b (1-\theta)^2 B(\alpha+1, \beta+1) d\theta + \gamma^2 \int_a^b B(\alpha, \beta+1) d\theta \\
&= B(\alpha+1, \beta+1) \int_a^b (1+\theta^2 - 2\theta) d\theta + \gamma^2 B(\alpha, \beta+1)(b-a) \\
&= B(\alpha+1, \beta+1) \left[ (b-a) + \frac{1}{3}(b^3 - a^3) - (b^2 - a^2) \right] + \gamma^2 B(\alpha, \beta+1)(b-a) \\
&= (b-a) \left[ B(\alpha+1, \beta+1) \left\{ 1 + \frac{1}{3}(a^2 + ab + b^2) - (b+a) \right\} + \gamma^2 B(\alpha, \beta+1) \right]
\end{aligned}$$

Substituting the values of  $I_1$  and  $I_2$  into (18), proves theorem.

### Simulation Study

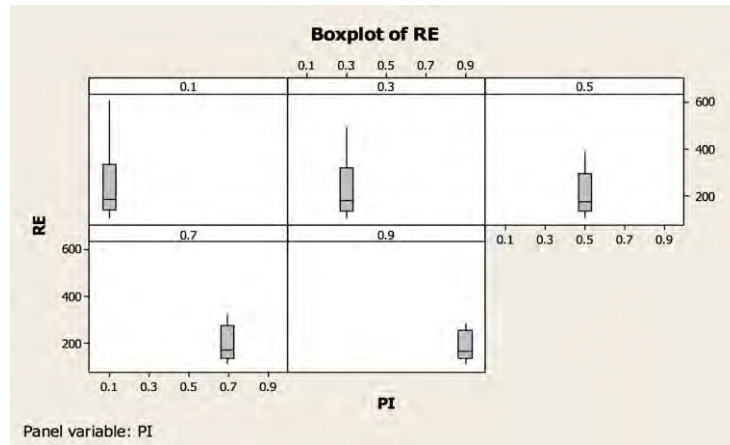
A numerical study was performed to investigate the various choices of parameters where the proposed stochastic randomized response model is more efficient than the non-stochastic BBB model. The percent relative efficiency of the proposed stochastic randomized response model estimator  $\hat{\pi}_{S(BBB)}$  with respect to the non-stochastic BBB model estimator  $\hat{\pi}_{x(BBB)}$  is given by

$$RE = \frac{V(\hat{\pi}_{x(BBB)})}{V(\hat{\pi}_{S(BBB)})} \times 100\%$$

$$= \frac{1 + \frac{(1+C_x^2)}{C_x^2} \left\{ B(\alpha+1, \beta+1) \left( 1 + \frac{1}{3}(a^2 + ab + b^2) - (b+a) \right) + \gamma^2 B(\alpha, \beta+1) \right\}}{B(\alpha+2, \beta) + \frac{1}{3}(a^2 + ab + b^2)B(\alpha, \beta+2) + (b+a)B(\alpha+1, \beta+1)} \times 100\%$$

$$= \frac{1 + \frac{(1+C_x^2)C_s^2(p)}{C_x^2}}{1 + \frac{(1+C_x^2)C_s^2(p)}{C_x^2}} \times 100\%$$

Clearly the relative efficiency depends only on the value of  $P$ ,  $\theta$ ,  $a$ ,  $b$ ,  $C_x$ ,  $C_\gamma$ ,  $\alpha$  and  $\beta$ . Certain parameters were fixed as  $P=0.7$ ,  $\theta=15$ ,  $a=5$ , and  $b=25$ . Note that here  $a=5$ , and  $b=25$  are not the lower limit and upper limit of the scrambling variable, but these are the limits for the mean values  $\theta^*$  of various scrambling variables used in a survey. The value of  $C_x$  was changed from 0.1 to 0.9 with step of 0.2; the value of  $C_\gamma$  was also changed from 0.1 to 0.9 with a step of 0.2; the value of  $\alpha$  was changed from 0.5 to 3.5 with a step of 1.5; and the value of  $\beta$  was changed between 0.5 to 5.0 with a step of 1.5. A box plot showing the magnitude of the RE is for each value of  $\pi$  between 0.1 to 0.9 with a step of 0.2 is given in Figure 2.



**Figure 2.** Relative efficiency plot.

For each combination of parameters, the percent relative efficiency of the proposed estimator was computed. The percent relative efficiency results so obtained are presented in [Table 1](#).

**Table 1.** Percent relative efficiency of the proposed stochastic randomized response model.

$C_x$	$C_y$	$\alpha$	$\beta$	$RE$	$C_x$	$C_y$	$\alpha$	$\beta$	$RE$	$C_x$	$C_y$	$\alpha$	$\beta$	$RE$
0.1	0.1	0.5	0.5	164.9	0.3	0.7	2.0	3.5	149.9	0.7	0.3	2.0	5.0	153.1
0.1	0.1	0.5	2.0	316.6	0.3	0.7	2.0	5.0	181.7	0.7	0.5	0.5	0.5	161.7
0.1	0.1	0.5	3.5	463.2	0.3	0.7	3.5	5.0	123.6	0.7	0.5	0.5	2.0	235.8
0.1	0.1	0.5	5.0	604.5	0.3	0.9	0.5	0.5	190.1	0.7	0.5	0.5	3.5	279.9
0.1	0.1	2.0	3.5	125.5	0.3	0.9	0.5	2.0	266.4	0.7	0.5	0.5	5.0	309.1
0.1	0.1	2.0	5.0	164.4	0.3	0.9	0.5	3.5	309.2	0.7	0.5	2.0	3.5	134.1
0.1	0.3	0.5	0.5	170.0	0.3	0.9	0.5	5.0	336.5	0.7	0.5	2.0	5.0	161.3
0.1	0.3	0.5	2.0	307.4	0.3	0.9	2.0	2.0	122.9	0.7	0.5	3.5	5.0	111.2
0.1	0.3	0.5	3.5	425.8	0.3	0.9	2.0	3.5	159.8	0.7	0.7	0.5	0.5	171.0
0.1	0.3	0.5	5.0	528.6	0.3	0.9	2.0	5.0	189.5	0.7	0.7	0.5	2.0	238.5
0.1	0.3	2.0	3.5	131.5	0.3	0.9	3.5	3.5	110.5	0.7	0.7	0.5	3.5	276.1
0.1	0.3	2.0	5.0	169.5	0.3	0.9	3.5	5.0	134.3	0.7	0.7	0.5	5.0	299.9
0.1	0.5	0.5	0.5	178.0	0.3	0.9	5.0	5.0	105.3	0.7	0.7	2.0	2.0	111.1
0.1	0.5	0.5	2.0	295.0	0.5	0.1	0.5	0.5	154.0	0.7	0.7	2.0	3.5	144.1

# STOCHASTIC RR MODEL FOR A SENSITIVE RANDOM VARIABLE

$C_x$	$C_y$	$\alpha$	$\beta$	$RE$	$C_x$	$C_y$	$\alpha$	$\beta$	$RE$	$C_x$	$C_y$	$\alpha$	$\beta$	$RE$
0.1	0.5	0.5	3.5	381.3	0.5	0.1	0.5	2.0	256.2	0.7	0.7	2.0	5.0	170.6
0.1	0.5	0.5	5.0	447.5	0.5	0.1	0.5	3.5	332.0	0.7	0.7	3.5	5.0	121.3
0.1	0.5	2.0	3.5	141.2	0.5	0.1	0.5	5.0	390.3	0.7	0.9	0.5	0.5	179.8
0.1	0.5	2.0	5.0	177.4	0.5	0.1	2.0	3.5	122.1	0.7	0.9	0.5	2.0	240.9
0.1	0.5	3.5	5.0	113.1	0.5	0.1	2.0	5.0	153.6	0.7	0.9	0.5	3.5	272.9
0.1	0.7	0.5	0.5	186.3	0.5	0.3	0.5	0.5	159.1	0.7	0.9	0.5	5.0	292.5
0.1	0.7	0.5	2.0	284.0	0.5	0.3	0.5	2.0	255.6	0.7	0.9	2.0	2.0	121.2
0.1	0.7	0.5	3.5	346.7	0.5	0.3	0.5	3.5	323.0	0.7	0.9	2.0	3.5	154.0
0.1	0.7	0.5	5.0	390.2	0.5	0.3	0.5	5.0	372.7	0.7	0.9	2.0	5.0	179.3
0.1	0.7	2.0	2.0	112.6	0.5	0.3	2.0	3.5	127.6	0.7	0.9	3.5	3.5	109.8
0.1	0.7	2.0	3.5	151.9	0.5	0.3	2.0	5.0	158.7	0.7	0.9	3.5	5.0	131.4
0.1	0.7	2.0	5.0	185.8	0.5	0.5	0.5	0.5	167.4	0.9	0.1	0.5	0.5	144.3
0.1	0.7	3.5	5.0	124.3	0.5	0.5	0.5	2.0	254.8	0.9	0.1	0.5	2.0	214.6
0.1	0.9	0.5	0.5	193.6	0.5	0.5	0.5	3.5	310.6	0.9	0.1	0.5	3.5	257.8
0.1	0.9	0.5	2.0	275.8	0.5	0.5	0.5	5.0	349.2	0.9	0.1	0.5	5.0	286.9
0.1	0.9	0.5	3.5	323.0	0.5	0.5	2.0	3.5	136.6	0.9	0.1	2.0	3.5	118.9
0.1	0.9	0.5	5.0	353.6	0.5	0.5	2.0	5.0	167.0	0.9	0.1	2.0	5.0	144.1
0.1	0.9	2.0	2.0	123.5	0.5	0.5	3.5	5.0	111.9	0.9	0.3	0.5	0.5	149.2
0.1	0.9	2.0	3.5	161.8	0.5	0.7	0.5	0.5	176.5	0.9	0.3	0.5	2.0	217.5
0.1	0.9	2.0	5.0	193.1	0.5	0.7	0.5	2.0	253.9	0.9	0.3	0.5	3.5	258.1
0.1	0.9	3.5	3.5	110.8	0.5	0.7	0.5	3.5	299.1	0.9	0.3	0.5	5.0	285.0
0.1	0.9	3.5	5.0	135.2	0.5	0.7	0.5	5.0	328.7	0.9	0.3	2.0	3.5	123.7
0.1	0.9	5.0	5.0	105.4	0.5	0.7	2.0	2.0	111.6	0.9	0.3	2.0	5.0	148.9
0.3	0.1	0.5	0.5	160.3	0.5	0.7	2.0	3.5	147.0	0.9	0.5	0.5	0.5	157.3
0.3	0.1	0.5	2.0	288.7	0.5	0.7	2.0	5.0	176.0	0.9	0.5	0.5	2.0	222.0
0.3	0.1	0.5	3.5	398.6	0.5	0.7	3.5	5.0	122.4	0.9	0.5	0.5	3.5	258.6
0.3	0.1	0.5	5.0	493.5	0.5	0.9	0.5	0.5	184.9	0.9	0.5	0.5	5.0	282.1
0.3	0.1	2.0	3.5	124.1	0.5	0.9	0.5	2.0	253.2	0.9	0.5	2.0	3.5	132.0
0.3	0.1	2.0	5.0	159.8	0.5	0.9	0.5	3.5	290.1	0.9	0.5	2.0	5.0	156.9
0.3	0.3	0.5	0.5	165.4	0.5	0.9	0.5	5.0	313.3	0.9	0.5	3.5	5.0	110.6
0.3	0.3	0.5	2.0	284.0	0.5	0.9	2.0	2.0	122.1	0.9	0.7	0.5	0.5	166.6
0.3	0.3	0.5	3.5	377.0	0.5	0.9	2.0	3.5	156.9	0.9	0.7	0.5	2.0	226.8
0.3	0.3	0.5	5.0	451.7	0.5	0.9	2.0	5.0	184.4	0.9	0.7	0.5	3.5	259.1
0.3	0.3	2.0	3.5	129.9	0.5	0.9	3.5	3.5	110.2	0.9	0.7	0.5	5.0	279.2
0.3	0.3	2.0	5.0	164.9	0.5	0.9	3.5	5.0	132.8	0.9	0.7	2.0	2.0	110.6

$C_x$	$C_y$	$\alpha$	$\beta$	RE	$C_x$	$C_y$	$\alpha$	$\beta$	RE	$C_x$	$C_y$	$\alpha$	$\beta$	RE
0.3	0.5	0.5	0.5	173.5	0.5	0.9	5.0	5.0	105.1	0.9	0.7	2.0	3.5	141.8
0.3	0.5	0.5	2.0	277.4	0.7	0.1	0.5	0.5	148.5	0.9	0.7	2.0	5.0	166.2
0.3	0.5	0.5	3.5	349.3	0.7	0.1	0.5	2.0	231.4	0.9	0.7	3.5	5.0	120.3
0.3	0.5	0.5	5.0	402.0	0.7	0.1	0.5	3.5	286.4	0.9	0.9	0.5	0.5	175.6
0.3	0.5	2.0	3.5	139.3	0.7	0.1	0.5	5.0	325.4	0.9	0.9	0.5	2.0	231.2
0.3	0.5	2.0	5.0	173.1	0.7	0.1	2.0	3.5	120.3	0.9	0.9	0.5	3.5	259.5
0.3	0.5	3.5	5.0	112.6	0.7	0.1	2.0	5.0	148.2	0.9	0.9	0.5	5.0	276.7
0.3	0.7	0.5	0.5	182.3	0.7	0.3	0.5	0.5	153.5	0.9	0.9	2.0	2.0	120.4
0.3	0.7	0.5	2.0	271.2	0.7	0.3	0.5	2.0	233.1	0.9	0.9	2.0	3.5	151.6
0.3	0.7	0.5	3.5	326.0	0.7	0.3	0.5	3.5	283.7	0.9	0.9	2.0	5.0	175.2
0.3	0.7	0.5	5.0	363.1	0.7	0.3	0.5	5.0	318.7	0.9	0.9	3.5	3.5	109.5
0.3	0.7	2.0	2.0	112.2	0.7	0.3	2.0	3.5	125.4	0.9	0.9	3.5	5.0	130.2

Table 2 gives the descriptive statistics for the percent relative efficiency values for different values of  $\pi$ .

**Table 2.** Descriptive statistics of the relative efficiency values.

$\pi$	Mean	StDev	Minimum	Median	Maximum
0.1	244.9	131.9	105.4	185.8	604.5
0.3	228.6	109.5	105.3	181.7	493.5
0.5	210.0	87.5	105.1	176.0	390.3
0.7	198.2	71.0	109.8	170.8	325.4
0.9	188.0	61.3	109.5	166.4	286.9

Table 1 shows that overall the minimum RE value is 105.1% and maximum RE value is 604.5%. The average value the RE is 214.2% with a standard deviation of 97.06%. The median value of the percent relative efficiency is 175.61%. Thus, in conclusion, it is possible to make a stochastic randomization device which will remain more efficient than the BBB model and more cooperation could be expected from the respondents.

## References

- Abdelfatah, S., Mazloun, R., & Singh, S. (2013). Efficient use of a two-stage randomized response procedure. *Brazilian Journal of Probability and Statistics*, 7(4): 608-617.
- Bar-Lev, S. K., Bobovitch, E., & Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60(3): 255-260.
- Franklin, L. A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population. *Communications in Statistics - Theory and Methods*, 18(2): 489-505.
- Greenberg, B. G., Abul-El, A. L. A., Simons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64: 520-539.
- Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. *Proceedings of Social Statistics Section. American Statistical Association*, pp. 65-72.
- Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77(2): 436-438.
- Singh, S. (2002). A new stochastic randomized response technique. *Metrika*, 56(2): 130-142.
- Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60: 63-69.

# Comparison of Three Calculation Methods for a Bayesian Inference of Two Poisson Parameters

**Yohei Kawasaki**

Tokyo University of Science  
Tokyo, Japan

**Etsuo Miyaoka**

Tokyo University of Science  
Tokyo, Japan

---

The statistical inference drawn from the difference between two independent Poisson parameters is often discussed in medical literature. Kawasaki and Miyaoka (2012) proposed an index  $\theta = P(\lambda_{1,\text{post}} < \lambda_{2,\text{post}})$ , where  $\lambda_{1,\text{post}}$  and  $\lambda_{2,\text{post}}$  denote Poisson parameters following posterior density. A new calculation method is proposed using MCMC and an approximate expression and exact expression for  $\theta$  are compared.

*Keywords:* Poisson distribution, Bayesian inference, MCMC method, Hypergeometric series

---

## Introduction

The statistical inference drawn from the difference between two independent Poisson parameters is often discussed in terms of the frequentist viewpoint rather than the Bayesian viewpoint. In this article, a Poisson parameter is assumed as the relapse rate of a wrong outcome and an adverse reaction rate. Therefore, a low value of the Poisson parameter is desirable.

Classical statistical analysis of outcomes observed in a randomized controlled clinical trial is based on the frequentist approach. The frequentist approach to hypothesis testing is based on the p-value. The inconvenience of using the p-value is well-known and has been documented by Lindley (1957) and Hwang, et al. (1992) among others.

A few different techniques for hypothesis testing have been developed under the Bayesian approach. Basu (1996) briefly showed the use of the Bayesian approach with respect to hypothesis testing. Let  $y$  be data from the probability density function; it is desired to test the null hypothesis against the alternative

---

*Dr. Kawasaki and Dr. Miyaoka are both Professors in the Department of Mathematics.  
Email Dr. Kawasaki at: [yk\\_sep10@yahoo.co.jp](mailto:yk_sep10@yahoo.co.jp).*

## COMPARISON OF THREE CALCULATION METHODS

hypothesis. One approach computes the posterior probability. Poisson parameters were applied to the posterior probability  $\theta$  that shows the difference between the posterior densities of the two independent Poisson parameters, which are considered as random variables. This index can be used to determine the probability that the Poisson parameter of a study drug is superior to that of a control drug.

Kawasaki and Miyaoka (2012) applied  $\theta$  to a one-side hypothesis based on a two-sample situation. They derived an exact and an approximate expression to determine  $\theta$ .

There are some pending issues with the above-mentioned method. An approximate method and exact method of  $\theta$  were adopted only while using a conjugate prior. The drawback of the approximate method is that it occasionally leads to a rough result in a small sample. The drawback of the exact method is that it is slightly complicated. In addition, the exact method requires extensive computing time with a large sample size. Hence, a Markov Chain Monte Carlo (MCMC) method for  $\theta$  is proposed as a solution to these problems.

### Methodology

If  $X_i$  is the number of events in a population of  $n_i$  patients (or over  $n_i$  units of time), and  $\lambda_i$  is the event rate, then the sampling distribution is

$$f_{X_i}(x_i | \lambda_i) = \frac{e^{-n_i \lambda_i} (n_i \lambda_i)^{x_i}}{x_i!}, \quad (1)$$

where  $i = 1, 2$ . The conjugate prior density of  $\lambda_i$  is the gamma distribution with parameters  $\alpha_i$  and  $\beta_i$ :

$$f_i(\lambda_i | \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i-1} e^{-\beta_i \lambda_i} \quad \lambda_i > 0 \quad (2)$$

where  $\alpha_i > 0$  and  $\beta_i > 0$ . The posterior density for  $\lambda_i$  is given as

$$g_i(\lambda_{i,post}) = \frac{b_i^{a_i}}{\Gamma(a_i)} \lambda_i^{a_i-1} e^{-b_i \lambda_i}, \quad (3)$$



where  $a_i = \alpha_i + x_i$ ,  $b_i = n_i + \beta_i$  and  $\Gamma(a)$  denotes the gamma function. Let  $\lambda_{i,post}$  denote the Poisson parameter in the posterior density.

### Approximate expression for $\theta$

$\theta$  can be calculated via an approximation using the standard normal table; assume that sample sizes,  $n_1$  and  $n_2$ , are large. It is necessary to determine a Z-test statistic. The expected difference in the posterior density and the variance in this difference can be expressed as:

$$E(\lambda_{1,post} - \lambda_{2,post}) = \mu_{1,post} - \mu_{2,post}, \quad (4)$$

$$V(\lambda_{1,post} - \lambda_{2,post}) = a_1 / b_1^2 + a_2 / b_2^2 \quad (5)$$

where  $\mu_{i,post} = a_i / b_i$  denote the posterior mean of  $\lambda_{i,}$ . The  $Z_g$ -test statistic is

$$Z_g = \frac{(\lambda_{1,post} - \lambda_{2,post}) - E(\lambda_{1,post} - \lambda_{2,post})}{\sqrt{V(\lambda_{1,post} - \lambda_{2,post})}} \quad (6)$$

The  $Z_g$ -test statistic is approximately distributed according to the standard normal distribution. Therefore, the approximate probability of the index  $\theta$  is given as

$$\theta = P(\lambda_{1,post} < \lambda_{2,post}) \approx \Phi\left(\frac{-\mu_{1,post} + \mu_{2,post}}{\sqrt{a_1 / b_1^2 + a_2 / b_2^2}}\right) \quad (7)$$

where  $\Phi(\bullet)$  is the cumulative distribution function of the standard normal distribution. From this the approximate probability can be easily calculated.

### Exact method for $\theta$

Kawasaki and Miyaoka (2012) derived the exact expression for  $\theta$  using the posterior density. The exact expression for  $\theta$  is

## COMPARISON OF THREE CALCULATION METHODS

$$\begin{aligned}\theta &= P(\lambda_{1,post} < \lambda_{2,post}) \\ &= 1 - \frac{1}{a_2 \text{Beta}(a_1, a_2)} \left( \frac{b_2}{b_1 + b_2} \right)^{a_2} {}_2F_1\left(a_2, 1 - a_1, 1 + a_2, \frac{b_2}{b_1 + b_2}\right)\end{aligned}\quad (8)$$

where

$${}_2F_1(k_1, k_2; l; u) = \sum_{t=0}^{\infty} \frac{(k_1)_t (k_2)_t}{(l)_t} \frac{u^t}{t!} \quad (9)$$

is the hypergeometric series, and  $(k)_t$  is the Pochhammer symbol.

### MCMC Method for $\theta$

A computational procedure for  $\theta$  can be described using the MCMC method. The MCMC method is a means of sampling from a posterior density. A random-walk Metropolis-Hasting algorithm was used as the MCMC Method. Given that the samples come from two independent populations, the posterior joint distribution of  $\lambda_1$  and  $\lambda_2$  is a product of its marginal distributions. For this reason, samples can be obtained from the posterior distribution of  $\lambda_1 - \lambda_2$  by simulating  $k$  values from the posterior distribution of  $\lambda_1$  and  $\lambda_2$  using MCMC procedure of SAS, e.g.,  $\lambda_{1,post}^1, \lambda_{1,post}^2, \dots, \lambda_{1,post}^k$  and  $\lambda_{2,post}^1, \lambda_{2,post}^2, \dots, \lambda_{2,post}^k$ , respectively. By computing  $\lambda_{1,post}^1 - \lambda_{1,post}^1, \lambda_{1,post}^2 - \lambda_{1,post}^2, \dots, \lambda_{1,post}^k - \lambda_{1,post}^k$ , it is possible to obtain the simulated values from the posterior distribution of  $\lambda_1 - \lambda_2$ . The posterior samples obtained by the MCMC method after the burn-in period are  $\delta_1, \delta_2, \dots, \delta_k$ . Let  $\Delta_1, \Delta_2, \dots, \Delta_k$  be independent identically distributed random variables with distribution function  $F$ . The posterior sample is the observed value of  $\Delta_1, \Delta_2, \dots, \Delta_k$ . Note that  $\theta = P(\lambda_{1,post} < \lambda_{2,post})$  equals  $\theta = P(\lambda_{1,post} - \lambda_{2,post} < 0)$ , thus,  $\theta$  can be expressed as

$$\theta = P(\lambda_{1,post} < \lambda_{2,post}) = P(\lambda_{1,post} - \lambda_{2,post} < 0) \approx 1 - \hat{F}_k(0) \quad (10)$$

where

$$\hat{F}_k(s) = \frac{1}{k} \sum_{i=1}^k I(\Delta_i \leq s) \quad (11)$$

and

$$I(\Delta_i \leq s) = \begin{cases} 1 & \text{if } \Delta_i \leq s \\ 0 & \text{if } \Delta_i > s \end{cases} \quad (12)$$

is the empirical distribution function.

## Results

### Comparison of three methods

To compare the probabilities of the three methods for  $\theta$ , the difference between the sample rates (horizontal axis) where sample rate is calculated as  $X_i/n_i$ , were plotted against the difference between the probabilities of the MCMC and exact methods (vertical axis), as shown in Figures 1, 3, and 5. Similarly, the difference between the sample rates (horizontal axis) were plotted against the difference between the probabilities of the approximate and exact methods (vertical axis), as shown in Figures 2, 4, and 6. Figures 1, and 2 show situations that considered small sample sizes, i.e.,  $n_1 = n_2 = 10, 15, 20$ , and 25; Figures 3 and 4, show larger sample sizes, i.e.,  $n_1 = n_2 = 60, 70, 90$ , and 100. Figures 5 and 6 consider groups of different sample sizes, that is,  $n_1 = 5, n_2 = 15$ ;  $n_1 = 5, n_2 = 25$ ;  $n_1 = 15, n_2 = 5$  and  $n_1 = 25, n_2 = 5$ .

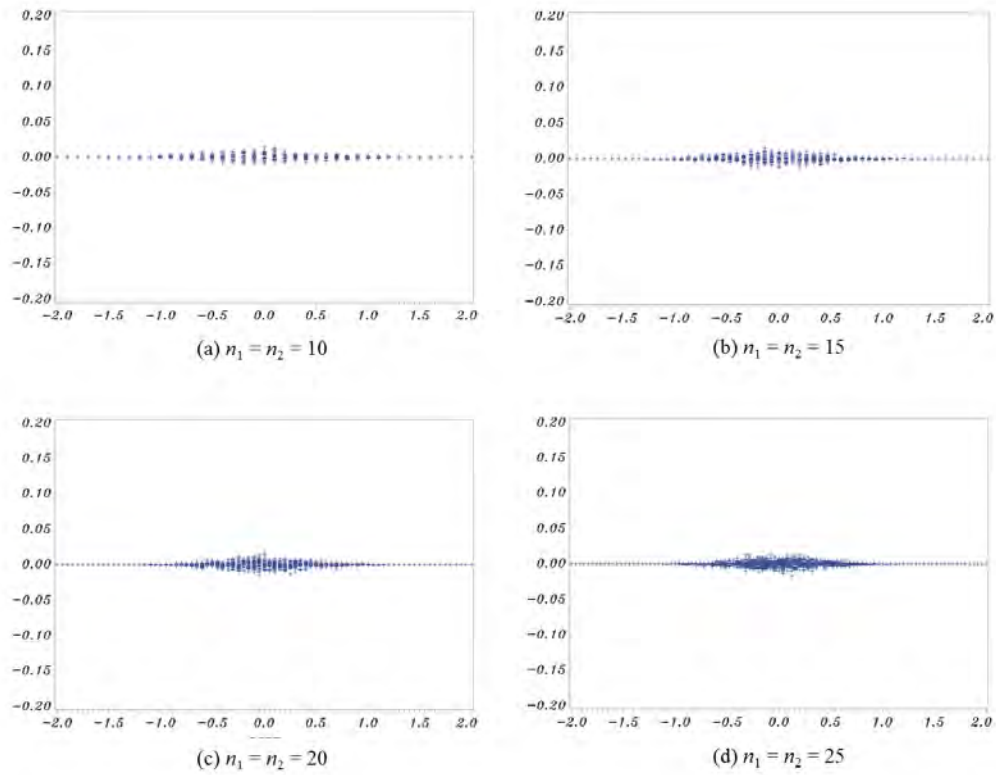
### Relationship between the difference in the probabilities and the difference in the sample rates.

In Figures 1(d) and 3(d), the probability of the MCMC method is approximately equal to that of the exact method when the difference between the sample rates is 1.0. Conversely, the difference between the probabilities of the MCMC and exact methods is around 0.01 when the difference between the sample rates is zero. Overall, when the difference between the sample rates is large, the probabilities of the MCMC and exact methods are roughly equal. By contrast, when the difference between the sample rates is small, the probability of the MCMC method is different from that of the exact method. This general pattern is similar for the difference in the probabilities of the approximation and exact methods.

### **Relationship between the sample size and the difference in the probabilities**

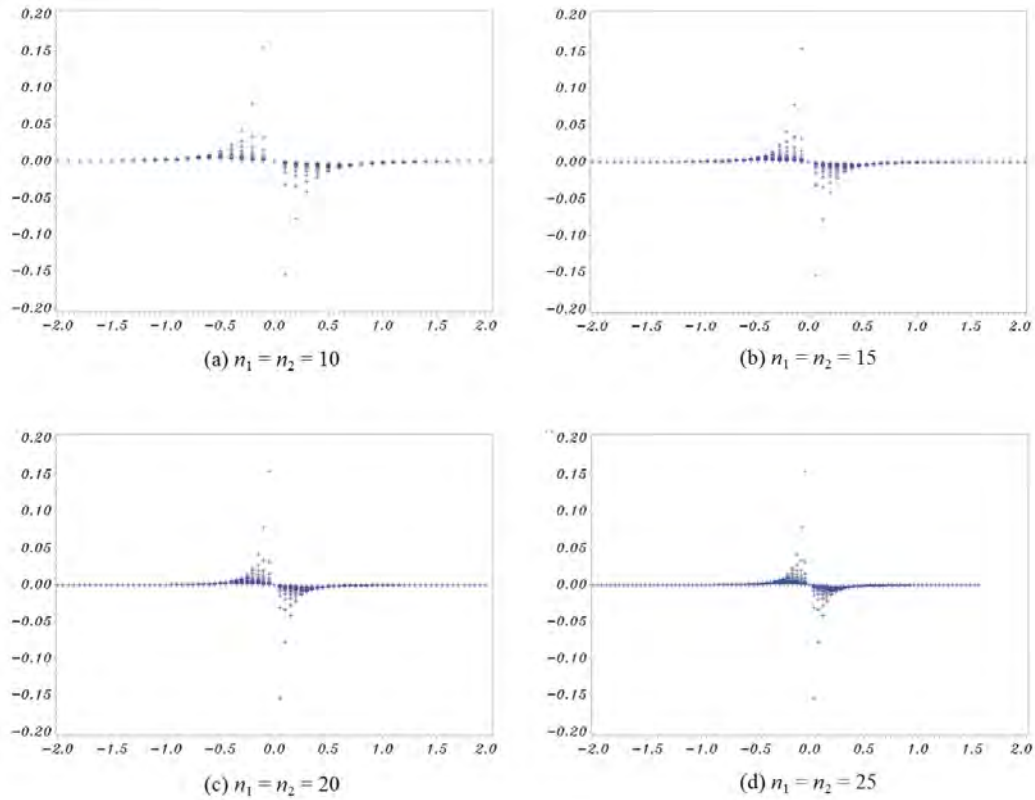
In Figure 2(a), the difference between the probabilities of the approximate and exact methods is around 0.15 when the difference between the sample rates is 0.01. For a slightly larger sample size (see Figure 2(d)), the difference between the probabilities of the approximate and exact methods is around 0.05 for the same difference between the sample rates. In addition, there is virtually no difference between the probabilities of the approximate and exact methods when the sample size is further increased, as shown in Figure 4(d). Thus, the sample size influences the accuracy of the probability of the approximate method. Also shown is the difference in the probabilities of the MCMC and exact methods. In Figure 1(a), the difference between the probabilities of the MCMC and exact methods is around 0.01 when the difference between the sample rates is zero. For a slightly larger sample size (see Figure 3(d)), the difference between the probabilities of the MCMC and exact method is around 0.01 for the same difference between the sample rates. Thus, the accuracy of the probability of the MCMC method always remains high even when the sample sizes are small.

Finally, the difference between the probabilities when groups of different sample sizes are considered was investigated. In Figure 2(a), the difference between the probabilities of the approximate and exact methods is around -0.01 when the difference between the sample rates is 0.5. Conversely, in Figure 6(a), the difference between the probabilities of the approximate and exact methods is around -0.05 for the same difference between the sample rates. In both the cases, the total sample size ( $n_1 + n_2$ ) is the same. However, the difference between the probabilities of the approximate and exact methods is slightly greater in the case of groups with different sample sizes; the case of the MCMC method is also shown. In Figure 1(d), the difference between the probabilities of the MCMC and exact methods is around 0.01 when the difference between the sample rates is zero. Conversely, in Figure 5(d), the difference between the probability of the MCMC and exact methods is around 0.01 for the same difference between the sample rates. Therefore, the difference between the probabilities of the MCMC and exact methods is the same regardless of whether the sample sizes are equal or different.

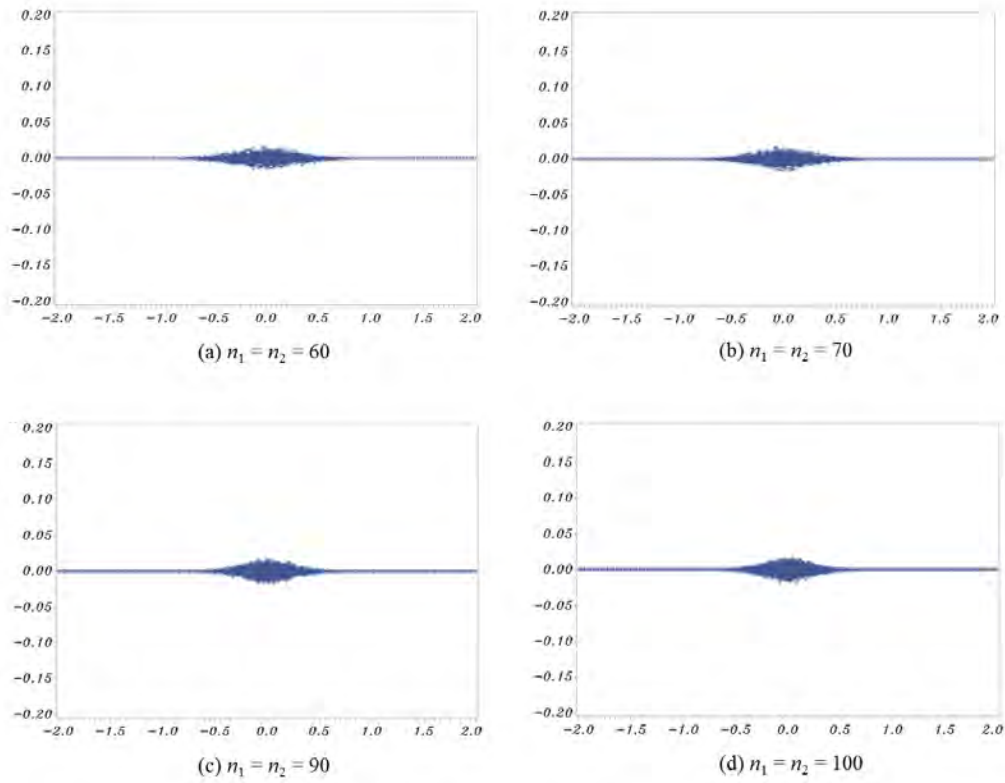


**Figure 1.** Comparison of the Exact and MCMC Method when sample sizes are small. (vertical axis : Differences of  $\theta$  in Exact and MCMC method. Prior distribution is Gamma(0.01,0.01). horizontal axis : Differences of two sample rates.

## COMPARISON OF THREE CALCULATION METHODS

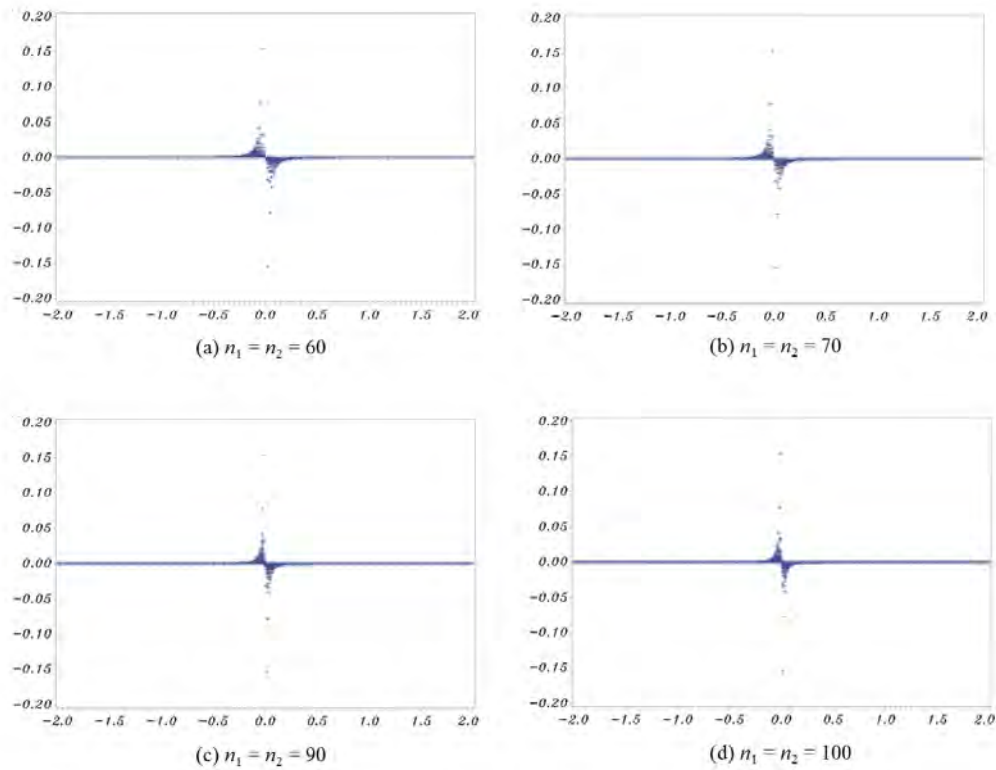


**Figure 2.** Comparison of the Exact and Approximate method when sample sizes are small. (vertical axis : Differences of  $\theta$  in Exact and Approximation method. horizontal axis : Differences of two sample rates.



**Figure 3.** Comparison of the Exact and MCMC Method when sample sizes are large.  
 (vertical axis : Differences of  $\theta$  in Exact and MCMC method. Prior distribution is  $\text{Gamma}(0.01, 0.01)$ . horizontal axis : Differences of two sample rates.

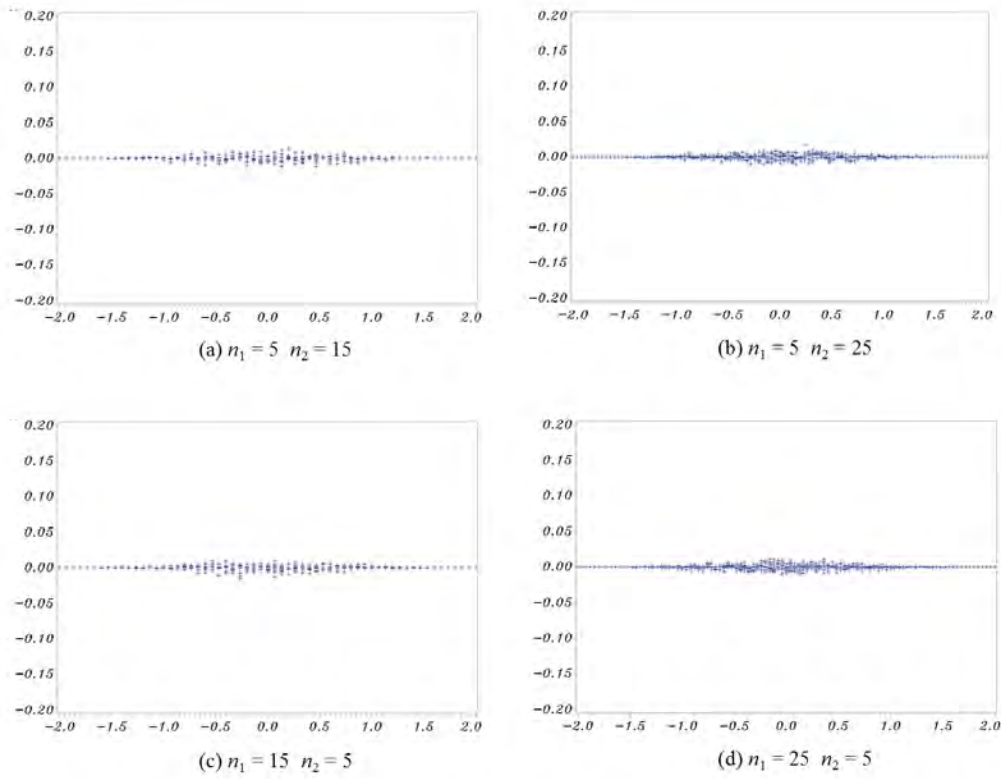
## COMPARISON OF THREE CALCULATION METHODS



**Figure 4.** Comparison of the Exact and Approximate method when sample sizes are large. (vertical axis : Differences of  $\theta$  in Exact and Approximation method. horizontal axis : Differences of two sample rates.

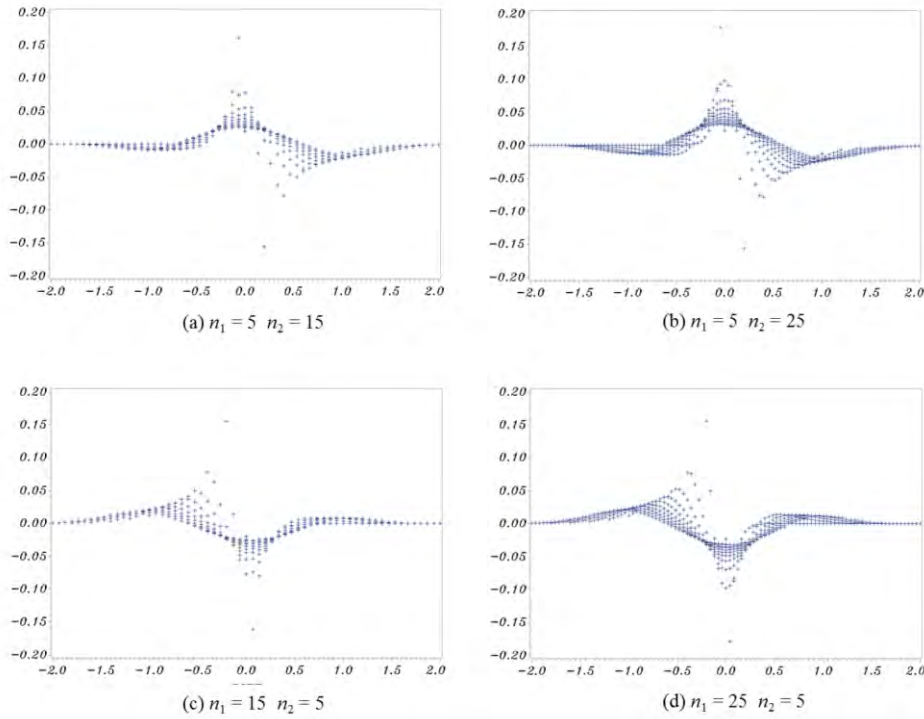
---





**Figure 5:** Comparison of the Exact and MCMC Method when sample sizes are unbalanced. (vertical axis : Differences of  $\theta$  in Exact and MCMC method. Prior distribution is Gamma(0.01,0.01). horizontal axis : Differences of two sample rates.

## COMPARISON OF THREE CALCULATION METHODS



**Figure 6:** Comparison of the Exact and Approximate method when sample sizes are unbalanced. (vertical axis : Differences of  $\theta$  in Exact and Approximation method. horizontal axis : Differences of two sample rates.

## Conclusion

Three calculation methods were presented for the index  $\theta = P(\lambda_{1,post} < \lambda_{2,post})$ . A new procedure was described based on the MCMC method. The probabilities of these three methods were compared in order to test the relative effectiveness of each.

The expression for the exact method was presented, which includes a hypergeometric series, and it was speculated that this series causes the decrease in calculation efficiency when the sample size is very large. In addition, hypergeometric series are not built into SAS, which is a statistical software program frequently used in pharmaceutical development. Therefore, if SAS is used, a calculation program for hypergeometric series must be developed.

It is easy to calculate the probability for using the approximation method. This is an advantage when the approximate probability is used. Conversely, when

the difference in the sample rates is small and the sample sizes are unbalanced, the accuracy the approximation method is poor. That is, the accuracy of the probability of the approximation method depends on the sample size.

This study showed that the accuracy of the MCMC method was greater than that of the approximation method. Moreover, the probability of the MCMC method can be easily calculated using SAS. In addition, it is possible to use the non-conjugate prior for the prior distribution in the MCMC method. This is considered to be one of the advantages of the MCMC method.

## References

- Basu, S. (1996). Bayesian hypotheses testing using posterior density ratios. *Statistics and Probability Letters*, 30(1): 79-86.
- Hwang, J. T., Casella, G., Robert, C., Wells, M. T., & Farrell, R. H. (1992). Estimation of accuracy in testing. *The Annals of Statistics*, 20(1): 490-509.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2): 187-192.
- Kawasaki, Y., & Miyaoka, E. (2012). A Bayesian inference of  $P(\lambda_1 < \lambda_2)$  for two Poisson parameters. *Journal of Applied Statistics*, 39(10): 2141-2152.

# Specifying Asymmetric STAR models with Linear and Nonlinear GARCH Innovations: Monte Carlo Approach

**OlaOluwa S. Yaya**

University of Ibadan  
Ibadan, Nigeria

**Olanrewaju I. Shittu**

University of Ibadan  
Ibadan, Nigeria

---

Economic and finance time series are typically asymmetric and are expected to be modeled using asymmetrical nonlinear time series models. Smooth Transition Autoregressive (STAR) models: Logistic (LSTAR) and Exponential (ESTAR) are known to be asymmetric and symmetric respectively. Under non-normal and heteroscedastic innovations, the residuals of these models are estimated using Generalized Autoregressive Conditionally Heteroscedastic (GARCH) models with variants which include linear and nonlinear forms. The small sample properties of STAR-GARCH variants are yet to be established but these properties are investigated using Monte Carlo (MC) simulation. An MC investigation was conducted to investigate the performance of selections of STAR-GARCH models by classical nonlinear selection approaches. The ARCH(1) and GARCH(1,1) models were the linear GARCH specifications while the Logistic Smooth Transition-ARCH (LST-ARCH(1,1)), Logistic Smooth Transition-GARCH (LST-GARCH(1,1)) and Asymmetric Nonlinear Smooth Transition-GARCH (ANST-GARCH(1,1)) models were the nonlinear GARCH specifications. The nonlinearity parameter in the variance equations and Autoregressive (AR) parameters were varied along with different sample sizes. With the assumption of normality, the results showed that the selection of LSTAR models were actually affected by the structure of the innovations and this improved as sample size increased. Misspecification tests showed that these models cannot be misrepresented in the real sense.

*Keywords:* Asymmetry, Monte Carlo simulations, nonlinear GARCH, Smooth transition autoregression, specification

---

## Introduction

Smooth Transition Autoregressive (STAR) and Generalized Conditionally Heteroscedastic (GARCH) models are gaining their popularities in economics and

---

*OlaOluwa S. Yaya is a Lecturer in the Department of Statistics. Email at: [os.yaya@mail.ui.edu.ng](mailto:os.yaya@mail.ui.edu.ng). Dr. Shittu is a Lecturer in the Department of Statistics. Email at [oi.shittu@ui.edu.ng](mailto:oi.shittu@ui.edu.ng).*

finance. STAR models of Granger and Teräsvirta (1993) classify market into two phases of contraction and expansion, whereas GARCH model of Bollerslev (1986) is often used to study the behavior of asset returns or innovations of the ‘parent’ model. In that case, such a ‘parent’ model is the mean equation and the (GARCH) model is the variance equation. The innovations of the STAR model are expected follow normal distribution (homoscedasticity) but in case this is not true, the innovations are said to possess heteroscedasticity, which can be of various forms (Pavlidis, Paya and Peel, 2010). The mean and variance equations are then compounded as STAR-GARCH model.

Maximum Likelihood Estimation (MLE) of STAR-GARCH model was examined in Chan and McAleer (2002). The structural and statistical properties of the model were also established in the paper, even though the asymptotic normality and finite sample properties are still examined using Monte Carlo simulation approach. Chan and McAleer (2002) also considered the effects of misspecifying the transition functions (logistic or exponential) in the STAR model and the results obtained showed that greater bias will be induced in the GARCH estimates for the STAR-GARCH model whenever STAR mode is misspecified. Their results further showed that Logistic STAR model can easily be substituted for Exponential STAR model.

In the study of financial returns, negative returns tend to be followed by periods of higher volatility than positive returns of the same magnitude, that is negative and positive shocks exert different values for the leverage of a firm which on the long run realize different volatilities (Black, 1976). This property has therefore led to the development of GARCH variants that are robust to asymmetry. These variants are nonlinear in their structures due to the fact that the conditional variance is no longer specified as a linear function of lagged squared error and lagged variance. These common asymmetric variants are the Exponential GARCH (EGARCH) (Nelson, 1991), Asymmetric Power ARCH (APARCH) (Ding, *et al.*, 1993) and Glosten Jaganathan and Runkle (GJR-GARCH) (Glosten, *et al.*, 1993) models, but in this work we investigate GARCH variants which display regime switching dynamics.

This study is motivated by the work of Chan and McAleer (2002). We applied the linear GARCH and Smooth Transition specification of ARCH/GARCH models in a Monte Carlo simulation approach. Nonlinearities were first introduced in the ARCH functional form in Engle and Bollerslev (1986). They proposed in their model the dynamics of conditional variance,  $\sigma_t^2$  as it changes with the squared residuals and the transition between different conditional variance determined by normal cumulative distribution function. A

few years later, Higgins and Bera (1992) developed a Nonlinear ARCH (NARCH) model which accommodated different functional forms to predict the conditional variance. Apart from the classical ARCH and GARCH models of Engle (1982) and Bollerslev (1986), Smooth Transition ARCH (STARCH), Smooth Transition GARCH (ST-GARCH) and Asymmetric Nonlinear Smooth Transition GARCH (ANST-GARCH) models of Hagerud (1996; 1997), González-Rivera (1998) and Anderson *et al.* (1999) respectively are also considered. The ST-ARCH model was applied on the Nordic and Stockholm stock returns and found the model better than the linear GARCH model. González-Rivera (1998) used MC simulation experiment to study the model and applied the models on stock returns and exchange-rate data.

### **The STAR-GARCH and STAR-STGARCH Models**

This article presents compounded regime switching and volatility models, with the regime switching model as the mean equation and volatility models as the variance equation. For a time series  $y_t, t=1, \dots, N$  with  $y_t \sim N(\mu, \sigma^2)$  in the structural model,

$$\hat{y}_t = f(.) + \varepsilon_t \quad (1)$$

where  $f(.)$  is the function of  $y_t$  and  $\varepsilon_t$  is the innovation process, expected to be independently and identically distributed with mean 0 and variance 1 that is homoscedasticity case. In the case where this assumption of normal distribution fails, the innovations are estimated with volatility models.

### **The Mean Equation: STAR model**

The Smooth Transition Autoregressive (STAR) model is introduced in Granger and Teräsvirta (1993) and the specification, estimation and evaluation of the model are itemized following standard procedures in Teräsvirta (1994). Since then, the model has been applied to study nonlinearity in business cycle (Teräsvirta and Anderson, 1992); Skalin and Teräsvirta 1996; 1998) and real exchange rates (Baum *et al.*, 1998; Liew *et al.*, 2002). The connection between business cycle-regimes and nonlinearity in the UK labour market is studied in Acemoglu and Scotts (1994). Öcal (2000) applied STAR model on the nonlinearities in growth rates of some selected UK macroeconomic time series

and suggest either two-regime or three-regime model for UK economy. Mourelle, Cuestas and Gil-Alana (2011), Shittu and Yaya (2011) and Yaya (2013) considered STAR model for Nigerian inflation series.

Apart from real life time series data that have been considered for the STAR model, Escribano and Jordá (2001) and Yaya and Shittu (2011) investigated the selection of STAR model by varying some of the parameters and conditions in the models and obtained results that serve as guide for nonlinear time series modelers; then, there is need to study, and if possible develop the structural and small sample properties of the STAR model.

The STAR model of order  $p$  is given as,

$$y_t = \phi_{10} + \sum_{i=1}^p \phi_{1i} y_{t-i} + \left( \phi_{20} + \sum_{i=1}^p \phi_{2i} y_{t-i} \right) F(y_{t-d}; \gamma, c) + \varepsilon_t \quad (2)$$

where  $\phi_{10}, \phi_{20}$  are the constants and  $\phi_{1i}, \phi_{2i}$  ( $i=1, \dots, p$ ) are the autoregressive parameters of order  $p$ . The transition function,  $F(y_{t-d}; \gamma, c)$  causes the nonlinear dynamics in the model, and this are of logistic and exponential forms as given as,

$$F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(y_{t-d} - c)]} \quad (3)$$

and

$$F(y_{t-d}; \gamma, c) = 1 - \exp[-\gamma(y_{t-d} - c)^2] \quad (4)$$

respectively, with  $\gamma > 0$  in both cases. The logistic type is known to be asymmetric whereas the exponential type is symmetric. Economic and finance series often exhibit forms of asymmetries, and therefore Logistic STAR (LSTAR) model is often applied to model nonlinear dynamics in the series. In the transition functions, the transition variable is  $y_{t-d}$  with  $d$  assuming values 1, 2, ...,  $p$ . the value of  $d$  is varied in order to improve nonlinearity in the system when it is not known prior to model estimation. The slope,  $\gamma$  and intercept,  $c$  are parts of the nonlinearity parameters in the transition function. As  $\gamma$  assumes values from 1 to say 100, the nonlinearity becomes sharper, and the dynamics shift from lower linear region to upper linear region at faster rate, after being in the nonlinear state

for some period. At  $\gamma = 1$ , depending on the variance of  $y_t$  and size of  $c$ , discrimination between the nonlinear and linear series may not be significant. (Yaya and Shittu, 2011). The transition functions in (3) and (4) are bounded between 0 and 1, and this makes the STAR modelling of interesting application. When the transition function is at zero state, the entire system in (2) becomes linear, and at unity state, it is also linear. Most of the time, the transition function is such that  $0 < F(y_{t-d}; \gamma, c) < 1$ , which is a nonlinear state.

Specification between the asymmetric and symmetric transition function is often carried out using the approach outlined in Teräsvirta (1994). Though there is a newer specification approach proposed in Escribano and Jordá (2001), the approach of Teräsvirta (1994) is not dominated by that of Escribano and Jordá (2001). Further readings on the specification of STAR models are referred to the two articles as well as Luukkonen, Saikkonen and Teräsvirta (1988).

### The Variance Equation: GARCH and ST-GARCH models

Apart from the issue of nonlinearity of the time series  $y_t$ , the innovations of the estimated model (mean equation) is often heteroscedastic for economic and finance series to be specific. Engle (1982) proposed the Autoregressive Conditionally Heteroscedastic (ARCH) model of order  $q$  for UK inflation.

$$\sigma_t^2 = w + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 \quad (5)$$

where  $\sigma_t^2$  is the conditional variance,  $w$  is the constant and  $\alpha_i (i=1, \dots, q)$  are the parameters in the ARCH model. The  $\varepsilon_{t-i}$  are the residuals from the mean equation which are assumed to be heteroscedastic.

Bollerslev (1986) proposed the generalized version of Engle's model which is named the Generalized Autoregressive Conditionally Heteroscedastic (GARCH) model of order  $(p, q)$  given as,

$$\sigma_t^2 = w + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^r \beta_j \sigma_{t-j}^2 \quad (6)$$

where  $\beta_j (j=1, \dots, r)$  are the parameters in the GARCH term. In the ARCH( $q$ ) and GARCH( $q, r$ ) models in (5) and (6),  $w > 0$ ,  $\alpha_i \geq 0$  and  $\beta_j \geq 0$  and the



existence of covariance-stationarity is  $\sum_{i=1}^q \alpha_i < 1$  for ARCH( $q$ ) and

$\sum_{i=1}^q \alpha_i + \sum_{j=1}^r \beta_j < 1$  for GARCH( $q, r$ ) model.

Hagerud (1996; 1997) and González-Rivera (1998) considered introducing regime switching functional forms in the ARCH/GARCH systems. Their propositions are further developed in Lundbergh and Teräsvirta (1999). Hagerud (1996) proposed Smooth Transition-ARCH ( $q$ ) (STARARCH) model,

$$\sigma_t^2 = w + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 [1 - F(\varepsilon_{t-i})] + \sum_{i=1}^q \delta_i \varepsilon_{t-i}^2 F(\varepsilon_{t-i}) \quad (7)$$

where  $w$  and  $\alpha_i$  are as defined in ARCH model. The additional parameter,  $\delta_i$  ( $i=1, \dots, q$ ) defines the model in two-regimes. The transition function, with the transition variable  $\varepsilon_{t-i}$  is of logistic and exponential as well. These are given as,

$$F(\varepsilon_{t-i}) = \frac{1}{1 + \exp(-\theta \varepsilon_{t-i})} \quad (8)$$

and

$$F(\varepsilon_{t-i}) = 1 - \exp(-\theta \varepsilon_{t-i}^2) \quad (9)$$

for the two forms respectively with  $\theta > 0$  in both cases. The two transition functions in (8) and (9) will generate different data dynamics for the conditional variance. The logistic form in (8) will produce a return process where the dynamics of the conditional variance differ depending on the signs of the innovations (Hagerud, 1997). As  $\varepsilon_{t-j} \rightarrow -\infty$ , the logistic function equals to  $-1/2$  and as  $\varepsilon_{t-j} \rightarrow +\infty$ , the function equals to  $1/2$ . The exponential function in (9) is symmetric with respect to the sign of the error term, hence it generates data for which the dynamics of the conditional variance depends only on the magnitude of the innovations. As  $|\varepsilon_{t-j}| \rightarrow \infty$ , the impact of  $\varepsilon_{t-1}^2$  on  $\sigma_t^2$  changes smoothly from  $\alpha_i$  to  $\delta_i$  in both logistic ST-ARCH( $q$ ) and ST-GARCH( $q, r$ ) when the function equals 1, and as  $\varepsilon_{t-j} = 0$ , the logistic function equals 0. Also, as the parameter  $\theta$

becomes larger, both the logistic ST-ARCH and ST-GARCH functions approach step functions which equal 0 for negative  $\varepsilon_{t-1}$  and 1 for positive  $\varepsilon_{t-1}$ , therefore, for logistic function,  $-1/2 \leq F(\cdot) \leq 1/2$  and for exponential function,  $0 \leq F(\cdot) \leq 1$ .

For positive conditional variance in the logistic ST-ARCH model, the condition  $\alpha_i \geq \frac{1}{2}|\delta_i|$  and for stationarity of the innovations  $\varepsilon_t$ ,

$\sum_{i=1}^q \left[ \alpha_i - \frac{1}{2}|\delta_i| + \max(\delta_i, 0) \right] < 1$ . For the positive conditional mean in the exponential ST-ARCH,  $\alpha_i + \delta_i \geq 0$  and for stationarity of the innovations  $\varepsilon_t$ ,

$\sum_{i=1}^q [\alpha_i + \max(\delta_i, 0)] < 1$  (Hagerud, 1997).

The generalized form of the model called Smooth Transition-GARCH ( $q, r$ ) (ST-GARCH) is proposed in Hagerud (1997) and González-Rivera (1998) as,

$$\sigma_t^2 = w + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 [1 - F(\varepsilon_{t-i})] + \sum_{i=1}^q \delta_i \varepsilon_{t-i}^2 F(\varepsilon_{t-i}) + \sum_{j=1}^r \beta_j \sigma_{t-j}^2 \quad (10)$$

with the transition functions in (8) and (9) for the logistic and exponential cases respectively. The ST-GARCH model only included the GARCH term,  $\sigma_{t-j}^2$ .

For positive conditional variance in the logistic ST-GARCH model, all the covariance stationarity condition of GARCH( $p, q$ ) model hold here in ST-GARCH, and apart from these,  $\alpha_i \geq \frac{1}{2}|\delta_i|$  for the logistic case and for the

stationarity of the innovations  $\varepsilon_t$ ,  $\sum_{i=1}^q \left[ \alpha_i - \frac{1}{2}|\delta_i| + \max(\delta_i, 0) \right] + \sum_{j=1}^r \beta_j < 1$ . For the positive conditional mean in the exponential ST-GARCH,  $\alpha_i + \delta_i \geq 0$  and for stationarity of the innovations  $\varepsilon_t$ ,  $\sum_{i=1}^q [\alpha_i + \max(\delta_i, 0)] < 1$  (Hagerud, 1997).

A similar ST-GARCH ( $p, q$ ) is proposed in Anderson, et al. (1999) and applied recently in Nam, et al. (2002). This is given as,

$$\sigma_t^2 = \left( w_{10} + \sum_{i=1}^q \alpha_{1i} \varepsilon_{t-i}^2 + \sum_{j=1}^r \beta_{1j} \sigma_{t-j}^2 \right) [1 - F(\varepsilon_{t-i})] + \left( w_{20} + \sum_{i=1}^q \alpha_{2i} \varepsilon_{t-i}^2 + \sum_{j=1}^r \beta_{2j} \sigma_{t-j}^2 \right) F(\varepsilon_{t-i}) \quad (11)$$

This is a variant of GARCH model in regime switching functional form. The parameters and the conditions of existence of GARCH as defined for the GARCH specification in (6) holds for the ST-GARCH model. The model in (11) is defined only for the asymmetric function (8), and therefore, the ST-GARCH model is otherwise known as Asymmetric nonlinear Smooth Transition-GARCH (ANST-GARCH) model (Nam, *et al.*, 2002). Franses and van Dijk (2003) showed that there is similarity between the ST-GARCH ( $q, r$ ) model of Hagerud (1997), even in the conditions of existence of conditional volatility and stationarity. Our selection of asymmetric variants of GARCH in this paper is based on similarity with STAR model and their abilities to realize smooth changing dynamics.

## Structure of the Data Generating Process and Nonlinearity Tests

The structure of the Data Generating Process (DGP) model used in the simulation is first explained analytically using a particular STAR model used in Granger and Teräsvirta (1993), Teräsvirta, Lin and Granger (1993), Teräsvirta (1994), Escribano, Franses and van Dijk (1998), Escribano and Jordá (2001) and Lopes and Salazar (2006). The DGP is examined by varying the nonlinearity parameters in the models. From the results, nonlinearity tests are described. The DGP is,

$$y_t = 1.8y_{t-1} - 1.06y_{t-2} + (\phi_{20} - 0.9y_{t-1} + 0.795y_{t-2})F(y_{t-d}; \gamma, c) + \varepsilon_t \quad (12)$$

where  $\varepsilon_t \sim N(0, 0.1\sigma_t)$  and  $F(y_{t-d}; \gamma, c)$  is either the logistic or exponential transition function as given in (3) and (4) respectively. The  $\phi_{20}$  is the intercept in the nonlinear part of the Autoregressive model.

Following Teräsvirta (1994), the LSTAR transition function in (3) is approximated by the third order Taylor's series expansion as,

## SPECIFYING ASYMMETRIC STAR MODELS

$$F(y_{t-d}; \gamma, c) \approx -\left(\frac{1}{4}c\gamma + \frac{1}{48}\gamma^3c^3\right) + \left(\frac{1}{4}\gamma + \frac{c^2\gamma^3}{16}\right)y_{t-d} - \frac{1}{16}c\gamma^3y_{t-d}^2 + \frac{1}{48}\gamma^3y_{t-d}^3 + R. \quad (13)$$

where  $R$  is the remainder series.

Substituting  $\gamma = 100$  and  $c = 0.2$ , then (13) becomes

$$F(y_{t-d}; \gamma, c) \approx -20838.33 + 2525y_{t-d} - 12500y_{t-d}^2 + 20833.33y_{t-d}^3 \quad (14)$$

this is then substituted in (12) to obtain

$$\begin{aligned} y_t = & (-20838.3\phi_{20} + 18751.8y_{t-1} - 16567.5y_{t-2}) \\ & + (2525\phi_{20} - 2272.5y_{t-1} + 2007.4y_{t-2})y_{t-d} \\ & + (-12500\phi_{20} - 11250y_{t-1} + 9937.5y_{t-2})y_{t-d}^2 \\ & + (20833.3\phi_{20} - 18750y_{t-1} + 16562.5y_{t-2})y_{t-d}^3 + R^* \end{aligned} \quad (15)$$

The expansion in (15) can be generalized as,

$$y_t = \phi' y_t^{(2)} + \beta_1' y_t^{(2)} y_{t-d} + \beta_2' y_t^{(2)} y_{t-d}^2 + \beta_3' y_t^{(2)} y_{t-d}^3 + \mathcal{G}_t \quad (16)$$

where  $\mathcal{G}_t$  is some noise process and  $y_t^{(2)}$  is the AR process of order 2 and  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  are the parameters of the nonlinear regression model. From (16), the LSTAR model is specified if the parameter  $\beta_2$  is not significant at  $\alpha$ -level or if it is the least significant among the three betas. Otherwise, ESTAR is specified.

Similar nonlinearity test to the above is developed in Escribano and Jordá (2001). Here there is suggestion to apply second order Taylor's series expansion of the ESTAR function in (4) to approximate the transition function. The approximation is given as,

$$\begin{aligned} F(y_{t-d}; \gamma, c) \approx & \left(\gamma c^2 - \frac{1}{2}\gamma^2 c^4\right) + (2c^3\gamma^2 - 2c\gamma)y_{t-d} \\ & + \left(\gamma - 3c^2\gamma^2\right)y_{t-d}^2 + 2c\gamma^2y_{t-d}^3 - \frac{1}{2}\gamma^2y_{t-d}^4 + R. \end{aligned} \quad (17)$$

Substituting  $\gamma = 100$  and  $c = 0.2$ , then (17) becomes

$$F(y_{t-d}; \gamma, c) \approx -12 + 120y_{t-d} - 1100y_{t-d}^2 + 4000y_{t-d}^3 + 5000y_{t-d}^4 \quad (18)$$

this is then substituted in (12) to obtain

$$\begin{aligned} y_t = & (-12\phi_{20} + 12.4y_{t-1} - 11.6y_{t-2}) \\ & + (120\phi_{20} - 108y_{t-1} + 95.4y_{t-2})y_{t-d} \\ & + (-1100\phi_{20} - 990y_{t-1} + 874.5y_{t-2})y_{t-d}^2 \\ & + (4000\phi_{20} - 3600y_{t-1} + 3180y_{t-2})y_{t-d}^3 \\ & + (5000\phi_{20} - 4500y_{t-1} + 3975y_{t-2})y_{t-d}^4 + R^* \end{aligned} \quad (19)$$

The expansion in (19) can be generalized as,

$$y_t = \phi' y_t^{(2)} + \beta_1' y_t^{(2)} y_{t-d} + \beta_2' y_t^{(2)} y_{t-d}^2 + \beta_3' y_t^{(2)} y_{t-d}^3 + \beta_4' y_t^{(2)} y_{t-d}^4 + \mathcal{G}_t \quad (20)$$

Here, the parameters are of order 2 and LSTAR is specified once the odd parameters  $\beta_1$  and  $\beta_3$  are most significant. Otherwise, ESTAR is specified if the parameters  $\beta_2$  and  $\beta_4$  are most significant.

## Monte Carlo Simulation Experiment

The Data Generating Process (DGP) defined as,

$$y_t = 1.8y_{t-1} - 1.06y_{t-2} + (\phi_{20} - 0.9y_{t-1} + 0.795y_{t-2})F(y_{t-d}; \gamma, c) + \varepsilon_t \quad (21)$$

with the nonlinear transition functions  $F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-100(y_{t-1} - 0.2)]}$  and  $F(y_{t-d}; \gamma, c) = 1 - \exp[-100(y_{t-1} - 0.2)^2]$  for Logistic STAR and Exponential STAR respectively. In the DGP, the autoregressive parameter  $\phi_{20}$  is varied as  $\phi_{20} = \{0, 0.2, 0.5\}$  and the innovations are assumed to have non-constant variance, that is  $\varepsilon_t \sim N(0, 0.1\sigma_t)$ . The values of  $\phi_{20}$  are chosen such that the DGP will

## SPECIFYING ASYMMETRIC STAR MODELS

realized stationary series. At  $F(s_i; \gamma, c) = 0$ , the resulting linear model has complex roots that are less than unity in absolute term, hence the process becomes nonstationary and there is possibility of explosion. At  $F(s_i; \gamma, c) = 1$ , the behavior of the process is influenced by the values of  $\phi_{20}$ . For example, when  $\phi_{20} = \{0, 0.2\}$ , the resulting characteristic equation has complex roots that are less than unity in absolute terms, hence the system reverts back to stationary region. At  $\phi_{20} = 0.5$ , the roots of the characteristic equations are real and the system realize nonstationary series.

The variance equations used in the simulations are the ARCH (1), GARCH (1,1), STARCH (1), ST-GARCH (1,1) and ANST-GARCH (1,1) are:

$$\sigma_t^2 = 0.02 + 0.3\varepsilon_{t-1}^2 \quad (22)$$

$$\sigma_t^2 = 0.02 + 0.3\varepsilon_{t-1}^2 + 0.6\sigma_{t-1}^2 \quad (23)$$

$$\sigma_t^2 = 0.02 + 0.3\varepsilon_{t-1}^2 [1 - F(\varepsilon_{t-1})] + 0.5\varepsilon_{t-1}^2 \quad (24)$$

$$\sigma_t^2 = 0.02 + 0.3\varepsilon_{t-1}^2 [1 - F(\varepsilon_{t-1})] + 0.5F(\varepsilon_{t-1}) + 0.6\sigma_{t-1}^2 \quad (25)$$

$$\begin{aligned} \sigma_t^2 = & 0.05 + 0.5\varepsilon_{t-1}^2 + 0.3\sigma_{t-1}^2 [1 - F(\varepsilon_{t-1})] \\ & + (0.02 + 0.3\varepsilon_{t-1}^2 + 0.6\sigma_{t-1}^2)F(\varepsilon_{t-1}) \end{aligned} \quad (26)$$

The logistic and exponential functions for the innovations  $\varepsilon_t$  are  $F(\varepsilon_{t-1}) = \frac{1}{1 + \exp(-\theta\varepsilon_{t-1})}$  and  $F(\varepsilon_{t-1}) = 1 - \exp(-\theta\varepsilon_{t-1}^2)$  respectively. In each case, the nonlinear parameter in the variance equations in (14) to (18) is varied as  $\theta = \{1, 5, 10\}$ . The experiment is carried out over 1,000 replications with sample sizes  $N = \{50, 100, 200, 500, 1000\}$ . Initialization problem is catered for by discarding the first 100 observations in each replication.

The experiment was carried out in two scenarios:

1. When the LSTAR DGP was used to realize LSTAR series with the specifications of the variance equations (Tables 1-3). When ESTAR model was misspecified for LSTAR model (Tables 4-6).
2. The relative frequencies of selecting an asymmetric STAR model with a particular variance equation are computed on every 1,000 replications at 5% nominal significant level.

The relative frequencies of selecting an asymmetric STAR model with a particular variance equation are computed on every 1,000 replications at 5% nominal significant level.

# SPECIFYING ASYMMETRIC STAR MODELS

**When the LSTAR DGP is used to realise LSTAR series.**

**Table 1.** Selection Frequencies of models at different  $\phi_{20}$  with fixed  $\theta = 1$

$\phi_{20} = 0, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	0.522	0.934	0.692	0.879	0.551	0.918	0.683	0.875	0.697	0.875
100	0.627	0.988	0.772	0.961	0.643	0.980	0.777	0.963	0.803	0.956
200	0.707	0.999	0.903	0.996	0.741	0.998	0.902	0.996	0.917	0.994
500	0.869	1.000	0.990	1.000	0.872	1.000	0.992	1.000	0.994	1.000
1000	0.949	1.000	1.000	1.000	0.966	1.000	1.000	1.000	1.000	1.000

$\phi_{20} = 0.2, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.598	0.500	-	-	0.623	0.500	0.590	0.494
1000	-	-	0.672	0.554	-	-	0.630	0.507	0.670	0.557

$\phi_{20} = 0.5, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** Table 1 presents the results of the selections of LSTAR models with forms of heteroscedastic innovation processes. EJP performed better than TP in selecting the LSTAR models at zero intercept,  $\phi_{20} = 0$  of the DGP. Both LSTAR-GARCH and LSTAR-ANLSTGARCH models were detected at frequencies higher than that of other model variants. As  $\phi_{20}$  increased beyond 0, there was failure in model specifications as a result of matrix inversion problems encountered by the simulator. The results were worse when computed at the nonstationary region ( $\phi_{20} = \{0.2, 0.5\}$ ) of the DGP.



**Table 2.** Selection Frequencies of models at different  $\phi_{20}$  with fixed  $\theta = 5$ 

$\phi_{20} = 0, \theta = 5$										
$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	0.522	0.934	0.692	0.879	0.549	0.921	0.682	0.868	0.697	0.878
100	0.627	0.988	0.772	0.961	0.647	0.979	0.789	0.963	0.811	0.954
200	0.707	0.999	0.903	0.996	0.751	0.997	0.907	0.997	0.918	0.995
500	0.869	1.000	0.990	1.000	0.878	1.000	0.993	1.000	0.992	1.000
1000	0.949	1.000	1.000	1.000	0.962	1.000	1.000	1.000	1.000	1.000

$\phi_{20} = 0.2, \theta = 5$										
$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.598	0.500	-	-	0.620	0.496	0.605	0.488
1000	-	-	0.672	0.554	-	-	0.631	0.483	0.687	0.556

$\phi_{20} = 0.5, \theta = 5$										
$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** Increasing  $\theta$  as 2 in Table 2, similar results to that of Table 2 were obtained. This implies that little increase in the nonlinearity of the residuals may not have significant effect on the specification of STAR models with Smooth Transition GARCH. The results were also worse at  $\theta = 5$

# SPECIFYING ASYMMETRIC STAR MODELS

**Table 3.** Selection Frequencies of models at different  $\phi_{20}$  with fixed 10

$\phi_{20} = 0, \theta = 10$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	0.522	0.934	0.692	0.879	0.551	0.917	0.694	0.871	0.712	0.881
100	0.627	0.988	0.772	0.961	0.641	0.780	0.790	0.962	0.823	0.958
200	0.707	0.999	0.903	0.996	0.753	0.998	0.912	0.996	0.923	0.999
500	0.869	1.000	0.990	1.000	0.888	1.000	0.991	1.000	0.988	1.000
1000	0.949	1.000	1.000	1.000	0.961	1.000	0.999	1.000	1.000	1.000

$\phi_{20} = 0.2, \theta = 10$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.598	0.500	-	-	0.616	0.496	0.605	0.477
1000	-	-	0.672	0.554	-	-	0.644	0.450	0.690	0.530

$\phi_{20} = 0.5, \theta = 10$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** Table 3 gives similar results to Tables 1 and 2. As we see in the previous results that correct model specifications were carried out at intercept  $\phi_{20} = 0$  and at this point, the process realized stationary time series.

**When ESTAR model is misspecified for LSTAR model****Table 4.** Selection Frequencies of models at different  $\phi_{20}$  with fixed  $\theta = 1$ 

$\phi_{20} = 0, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.750	0.569	-	-	0.752	0.570	0.767	0.607
1000	-	-	0.784	0.638	-	-	0.785	0.630	0.790	0.631

$\phi_{20} = 0.2, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.614	0.470	-	-	0.614	0.470	0.656	0.522
1000	-	-	0.673	0.554	-	-	0.667	0.543	0.718	0.563

$\phi_{20} = 0.5, \theta = 1$										
$N$	LSTAR-ARCH		LSTAR-GARCH		LSTAR-LSTARCH		LSTAR-LSTGARCH		LSTAR-ANLSTGARCH	
	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** Tables 4-6 give the results of specifying ESTAR for LSTAR in the DGP in (12). At  $\phi_{20} = 0$ , the simulator could not specify LSTAR and it reported matrix inversion problems. Also, TP performed better than EJP in selecting LSTAR from ESTAR DGP

# SPECIFYING ASYMMETRIC STAR MODELS

**Table 5.** Selection Frequencies of models at different  $\phi_{20}$  with fixed  $\theta = 5$

$$\phi_{20} = 0, \theta = 5$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.750	0.569	-	-	0.755	0.579	0.769	0.604
1000	-	-	0.784	0.638	-	-	0.786	0.641	0.790	0.629

$$\phi_{20} = 0.2, \theta = 5$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.614	0.470	-	-	0.612	0.482	0.659	0.527
1000	-	-	0.673	0.554	-	-	0.657	0.546	0.718	0.563

$$\phi_{20} = 0.5, \theta = 5$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** The results obtained here are similar to that of Table 4.

**Table 6.** Selection Frequencies of models at different  $\phi_{20}$  with fixed  $\theta = 10$ 

$$\phi_{20} = 0, \theta = 10$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.750	0.569	-	-	0.760	0.578	0.768	0.609
1000	-	-	0.784	0.638	-	-	0.789	0.644	0.788	0.625

$$\phi_{20} = 0.2, \theta = 10$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	0.614	0.470	-	-	0.612	0.482	0.656	0.527
1000	-	-	0.673	0.554	-	-	0.655	0.527	0.714	0.552

$$\phi_{20} = 0.5, \theta = 10$$

$N$	<i>LSTAR-ARCH</i>		<i>LSTAR-GARCH</i>		<i>LSTAR-LSTARARCH</i>		<i>LSTAR-LSTGARCH</i>		<i>LSTAR-ANLSTGARCH</i>	
	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>	<i>TP</i>	<i>EJP</i>
50	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-
200	-	-	-	-	-	-	-	-	-	-
500	-	-	-	-	-	-	-	-	-	-
1000	-	-	-	-	-	-	-	-	-	-

**Note:** The results obtained here seem to improve insignificantly over that of Table 5.

## Conclusion

This study considered the specification of asymmetric Smooth Transition Autoregressive (STAR) models with linear and nonlinear GARCH innovations. The GARCH error specifications are those proposed already in the literature.

## SPECIFYING ASYMMETRIC STAR MODELS

Specifications of the Logistic STAR-GARCH (LSTAR-GARCH) variants were carried out using the usual STAR specification procedures. The empirical results showed strong support for modelling STAR models with different GARCH error specifications. The results further showed that STAR model in STAR-GARCH model cannot be misrepresented in the real sense.

### References

- Acemoglu, D. & Scott, A. (1994). Asymmetries in the cyclical behaviour of UK labour markets. *Economic Journal*, 104: 1303-1323.
- Anderson, H. M., Nam, K. & Vahid, F. 1999. Asymmetric nonlinear smooth transition GARCH models. Rothman, P., (editor) *Nonlinear time series analysis of economic and financial data* (pp. 191-207). Boston, MA: Kluwer.
- Baum, C. F., Caglayan, M. & Barkoulas, J. T. (1998). Nonlinear adjustment to purchasing power parity in the post-Bretton woods era. Working paper No. 404, Department of Economics, Boston College.
- Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics*, 3: 167-179.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, 31: 307-27.
- Chan, K. & McAleer, M. (2002). Maximum Likelihood Estimation of STAR and STAR-GARCH Models: Theory and Monte Carlo Evidence. *Journal of applied Econometrics*, 17: 509-534.
- 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: 83-106.
- Engle R. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987-1007.
- Engle, R. F. & Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, 5(1): 1-50.
- Escribano, Á. & Jordá, O. (2001). Testing nonlinearity: Decision Rules for Selecting between Logistic and Exponential STAR models. *Spanish Economic review*, 3: 193-209.

Escribano, A., Franses, P. H. & van Dijk, D. (1998). Nonlinearities and Outliers: Robust Specification of STAR Models. *Econometric Institute Research Report 9832/A*.

Franses, P. H. & van Dijk, D. 2003. *Nonlinear Time Series Models in Empirical Finance*. Cambridge: Cambridge University Press.

Glosten, L. W., Jaganathan, R. & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48: 1779–801.

González-Rivera, G. (1998). Smooth Transition GARCH models. *Studies in nonlinear Dynamics and Econometrics*, 3: 61-78.

Granger, C. W. J. & Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press. Chinese edition 2006: Shanghai University of Finance and Economics Press.

Hagerud, G. E. (1996). A smooth transition Arch model for asset returns. *Working Paper Series in Economics and Finance 162*, Department of Finance, Stockholm School of Economics.

Hagerud, G. E. (1997). *Specification tests for Asymmetric ARCH*. Working paper. Stockholm School of Economics.

Higgins, M. L. & Bera, A. K. (1992). A class of nonlinear ARCH models. *International Economic Review*, 33: 137–158.

Liew, V. K. S., Ahmad, Z. B. & Sie-Hoe, L. 2002. Forecasting Performance of Logistic Smooth Transition Autoregressive Exchange Rate model: The Original and Re-parameterised versions. *MPRA Paper 511*: pp 1-20.

Lopes, H. F. & Salazar, E. (2006). Time Series Mean Level and Stochastic Volatility Modelling by Smooth Transition Autoregressions: A Bayesian Approach. *Advances in Econometrics*, 20: 225-238.

Lundbergh, S. & Teräsvirta, T. (1999). Modelling Economic high Frequency Time series with STAR-STGARCH models. *SSE/EFI Working Paper Series in Economics and Finance*, No. 291.

Luukkonen, R., Saikkonen, P. & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika* 75: 491-499.

Mourelle, E., Cuestas, J. C. & Gil-Alana, L. A. (2011). Is there an Asymmetric Behaviour in African Inflation? A Non-linear Approach. *South African Journal of Economics*, 79: 68-88.

## SPECIFYING ASYMMETRIC STAR MODELS

- Nam, K., Pyun, C. S. & Arize, A. C. (2002). Asymmetric mean reversion and contrarian profits: ANST-GARCH approach. *Journal of Empirical Finance*, 9: 563-588.
- Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns: A new approach. *Econometrica*, 59: 347-370.
- Öcal, N. (2000). Nonlinear Models for U.K. Macroeconomic Time Series. *Studies in Nonlinear Dynamics and Econometrics*, 43: 123-135.
- Pavlidis, E. G., Paya, I. & Peel, D. A. (2010). Specifying Smooth Transition Regression Models in the Presence of Conditional Heteroscedasticity of Unknown form. *Studies in Nonlinear Dynamics and Econometrics*, 143: 1-37.
- Skalin, J. & Terasvirta, T. (1996). Another look at Swedish business cycles. *Working paper series in Economic and Finance. No. 130*, Stockholm School of Economics.
- Skalin, J. & Teräsvirta, T. (1998). *Modelling Asymmetries in Unemployment rates*, Mimeo, Stockholm School of Economics.
- Shittu, O. I. & Yaya, O. S. (2011). On Fractionally Integrated Logistic Smooth Transitions in Time Series. *Central Bank of Nigeria Journal of Applied Statistics*, 2(1): 1-13.
- Teräsvirta, T. (1994). Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association*, 89: 208-218.
- Teräsvirta, T. & Anderson, H. M. (1992). Characterizing nonlinearities in Business cycles using Smooth Transition Autoregressive models. *Journal of Applied Econometrics*, 7: S119- S136.
- Teräsvirta, T., Lin, C-F. J. & Granger, C. W. J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, 14: 209-220.
- Yaya, O. S. (2013). On the variants of Nonlinear models. PhD Thesis. Department of Statistics, University of Ibadan.
- Yaya, O. S. & Shittu, O. I. (2011). On Misspecification of Exponential Transition Models with GARCH Error Terms: The Monte Carlo Evidence. *Proceedings of the 58<sup>th</sup> World Statistical Congress of the International Statistical Institute, Dublin. Section CPS001*: 5907-5912.



# On the Exponentiated Weibull Distribution for Modeling Wind Speed in South Western Nigeria

**Olanrewaju I. Shittu**

University of Ibadan  
Ibadan, Nigeria

**K. A. Adepoju**

University of Ibadan  
Ibadan, Nigeria

---

One of the bases for assessment of wind energy potential for a specified region is the probability distribution of wind speed. Thus, appropriate and adequate specification of the probability distribution of wind speed becomes increasingly important. Several distributions have been proposed for describing wind distribution. Among the most popular distributions is the Weibull whose choice is due to its flexibility. An exponentiated Weibull distribution is proposed as an alternative to model wind speed data with a view to comparing it with the existing Weibull distribution. Results indicate that the proposed distribution outperforms the existing Weibull distribution for modeling wind speed data in terms of minimum Akaike information criterion (AIC) and likelihood function. Thus, the exponentiated Weibull can be used as an alternative distribution that adequately describe the wind speed and thereby provide better representation of the potentials of wind energy.

*Keywords:* Wind power, Weibull, exponentiated Weibull, model selection criteria, maximum likelihood estimation

---

## Introduction

Energy demand increases proportionally as world population grows rapidly. Governments and societies become interested to renewable energies. Wind energy is considered the most attractive as it ensures high output power compared to other renewable energies. Nevertheless, the assessment of the wind energy potential is complicated since the wind speed availability is probabilistic. Several statistical distributions have been used for the description of the wind speed distribution. The two-parameter Weibull distribution function has been commonly

---

*Dr. Shittu and Mr. Adepoju are lecturers in the Department of Statistics. Email Dr. Shittu at [oi.shittu@ui.edu.ng](mailto:oi.shittu@ui.edu.ng).*

used in many fields including wind energy assessment (Rehman et al., 1994; Bivona et al., 2003).

Silva and Cordeiro (2012) were among the first among researchers to use compound distributions to model wind speed. They showed that Burr type XII distribution outperformed the commonly used Weibull distribution. Therefore, this article received its motivation from this and attempts to model wind speed using exponentiated Weibull distribution, which is a generalization of the Weibull distribution for increased and improved modeling potential.

## Weibull Distribution

The Weibull distribution is characterized by two parameters  $K$  and  $S$ , the shape and scale respectively. A random variable  $V$  (wind speed) is distributed as Weibull if it satisfies the following probability density function.

$$f_{(V)} = \frac{K}{C} \left( \frac{V}{C} \right)^{K-1} \exp \left[ - \left( \frac{V}{C} \right)^K \right]. \quad (1)$$

The corresponding distribution function is

$$F_{(V)} = 1 - \exp \left[ - \left( \frac{V}{C} \right)^K \right]. \quad (2)$$

If  $V$  denotes the wind speed, then the average wind speed is expressed as

$$E_{(V)} = \int_0^{\infty} V f(V) dv = \int_0^{\infty} V \frac{K}{C} \left( \frac{V}{C} \right)^{K-1} \exp \left[ - \left( \frac{V}{C} \right)^K \right] dv \quad (3)$$

giving rise to

$$E_{(V)} = \bar{V} = C \Gamma \left( 1 + \frac{1}{K} \right). \quad (4)$$

The variance of  $V$  is

$$Var(V) = \int_0^{\infty} (V - \bar{V})^2 f(v) dv \quad (5)$$

which on simplification gives

$$\sigma^2 = Var(V) = C^2 \left[ \Gamma\left(1 + \frac{2}{K}\right) - \Gamma^2\left(1 + \frac{1}{K}\right) \right] \quad (6)$$

from which

$$\sigma = C \left[ \Gamma\left(1 + \frac{2}{K}\right) - \Gamma^2\left(1 + \frac{1}{K}\right) \right]^{\frac{1}{2}}. \quad (7)$$

## Method of Estimating the Weibull Parameters

Commonly used methods known as graphical and maximum likelihood methods are now considered.

### Graphical Method

From (2)

$$\frac{1}{1 - F_{(V)}} = \exp \left[ \left( \frac{V}{C} \right)^K \right] \quad (8)$$

Introducing  $\ln$  to both sides results in

$$\ln \left[ \frac{1}{1 - F_{(V)}} \right] = \left( \frac{V}{C} \right)^K \quad (9)$$

and further introduction of  $\ln$  results in

$$\ln \ln \left[ \frac{1}{1 - F_{(V)}} \right] = K \ln V - K \ln C \quad (10)$$

Equation (10) can be expressed as  $Y = aX + b$  where

$$Y = \ln \ln \left[ \frac{1}{1 - F(v)} \right] \quad X = \ln v, a = K \text{ and } b = -K \ln C$$

### Maximum Likelihood Method

Harter and Moore (1965) were the earliest statisticians to use the maximum likelihood procedure because of its desirable characteristics. Given a random sample of size  $n$  wind speed drawn from a probability density function in (1), then the likelihood function will be

$$L(V_1, V_2, \dots, V_n, K, C) = \prod_{i=1}^n \left( \frac{K}{C} \right) \left( \frac{V_i}{C} \right)^{K-1} \exp \left[ - \left( \frac{V_i}{C} \right)^K \right] \quad (11)$$

The logarithm of (11) becomes

$$l = n \log \frac{K}{C} + (K-1) \sum_{i=1}^n \log \left( \frac{V_i}{C} \right) - \sum_{i=1}^n \left( \frac{V_i}{C} \right)^K \quad (12)$$

by differentiating (12) with respect to  $K$  and  $C$  in turn and equating to zero, the following are obtained

$$\frac{\partial l}{\partial C} \Rightarrow \frac{-nK}{C} + \frac{K}{C} \sum_{i=1}^n \left( \frac{V_i}{C} \right)^K = 0 \quad (13)$$

$$\frac{\partial l}{\partial K} = \frac{n}{K} + \sum_{i=1}^n \log \left( \frac{V_i}{C} \right)^K - \log \left( \frac{V_i}{C} \right) = 0 \quad (14)$$

Equations (13) and (14) are termed normal equations and can be solved numerically to obtain the maximum likelihood estimates of  $K$  and  $C$ .

### Exponentiated Weibull Distributions

According to Mudhokar, et al., (1995), the exponentiated Weibull density function is defined as

$$g(V) = \frac{K\delta}{C} \left[ 1 - \exp\left(-\frac{V}{C}\right)^K \right]^{\delta-1} \left(\frac{V}{C}\right)^{K-1} \exp\left(-\frac{V}{C}\right)^K \quad (15)$$

where  $K, C$  and  $d > 0$   $V_i > 0$ .

This distribution is proposed to model wind speed for the first time. For adequate determination of wind speed, the parameters in equation (15) need to be estimated. For this, we adopt the use of maximum likelihood method.

As before if  $V_1, V_2, \dots, V_n$  is a random sample of size  $n$  wind speed drawn from the density function in (15), then the likelihood function is

$$L(V_1, V_2, \dots, V_n, K, C, \delta) = \frac{K^n \delta^n}{C^n} \prod_{i=1}^n \left[ 1 - \exp\left(-\frac{V_i}{C}\right)^K \right]^{\delta-1} \prod_{i=1}^n \left(\frac{V_i}{C}\right)^{K-1} \exp\left(-\frac{V_i}{C}\right)^K. \quad (16)$$

The corresponding log-likelihood function is obtained by finding the logarithm of (16) is

$$l = n \log K + n \log \delta - n \log C + (\delta - 1) \sum_{i=1}^n \log \left( 1 - \exp\left(-\frac{V_i}{C}\right)^K \right) + (K - 1) \sum_{i=1}^n \log \left( \frac{V_i}{C} \right) - \sum_{i=1}^n \log \left( \frac{V_i}{C} \right)^K. \quad (17)$$

Taking the derivative of (17) with respect to  $K$ ,  $C$  and  $\delta$ , results in

$$\begin{aligned} \frac{\delta l}{\delta K} &= \frac{n}{K} + (\delta - 1) \sum_{i=1}^n \left[ \frac{\left(\frac{V_i}{C}\right)^K \exp\left(-\frac{V_i}{C}\right)^K \log\left(\frac{V_i}{C}\right)}{\left(1 - \exp\left(-\frac{V_i}{C}\right)^K\right)} \right] \\ &+ \sum_{i=1}^n \log\left(\frac{V_i}{C}\right)^K - \log\left(\frac{V_i}{C}\right) = 0 \end{aligned} \quad (18)$$

$$\frac{\delta l}{\delta C} = \frac{-n}{C} + \frac{K(\delta-1)}{C} \sum_{i=1}^n \left[ \frac{\left(\frac{V_i}{C}\right)^K \exp\left(-\frac{V_i}{C}\right)^K}{\left(1 - \exp\left(-\frac{V_i}{C}\right)^K\right)} \right] - \frac{n(K-1)}{C} - \frac{K}{C} \sum_{i=1}^n \left(\frac{V_i}{C}\right)^K \quad (19)$$

$$\frac{\delta l}{\delta d} = \frac{n}{d} + \sum_{i=1}^n \log \left( 1 - \exp\left(-\frac{V_i}{C}\right)^K \right). \quad (20)$$

Equations (18), (19) and (20) are solved iteratively to obtain the maximum likelihood estimates of the parameters  $K$ ,  $C$  and  $d$ .

### Moments of the Exponentiated Weibull Distribution

Following the density function in (15), its  $r^{\text{th}}$  moment can be obtained as:

$$\mu'_r = E(X^r) = \int_0^\infty \frac{Kd}{C} V^r \left( 1 - \exp\left(-\frac{V}{C}\right)^K \right)^{d-1} \left(\frac{V}{C}\right)^{K-1} \exp\left(-\frac{V}{C}\right)^K dy$$

$$\text{If } y = \left(\frac{V}{C}\right)^K \Rightarrow V = y^{\frac{1}{K}C} \text{ and } dv = \frac{C^K}{KV^{K-1}} dy,$$

which reduces to

$$\mu'_r = E(X^r) = \int \left(y^{\frac{1}{K}C}\right)^r \exp(-y) (1 - \exp(-y))^{d-1} dy. \quad (21)$$

Note from binomial series expansion that

$$(1-m)^b = \sum_{j=0}^{\infty} (-1)^j \binom{b}{j} m^j, \text{ then } (1 - \exp(-y))^{d-1} = \sum_{j=0}^{\infty} (-1)^j \binom{d-1}{j} \exp(-y_j)$$

thus, equation (21) becomes

$$\mu_r' = C^r d \sum_{j=0}^{\infty} (-1)^j \binom{d-1}{j} \int_0^{\infty} y^{r/K} \exp(-y(1+j)) dy. \quad (22)$$

If  $P = y(1+j) \Rightarrow y = \frac{P}{1+j}$  and  $dy = \frac{dP}{1+j}$ , then

$$\mu_r' = E(X^r) = C^r d \sum_{j=0}^{\infty} \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{r/K+1}} \int_0^{\infty} P^{r/K} \exp(-P) dy, \quad (23)$$

therefore, the  $r^{\text{th}}$  moment of the exponentiated Weibull distribution is

$$E(X^r) = C^r d \sum_{j=0}^{\infty} \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{r/K+1}} \Gamma\left(\frac{r}{K} + 1\right). \quad (24)$$

For simplicity let  $w_j = \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{r/K+1}}$

$$E(X^r) = C^r d \sum_{j=0}^{\infty} w_j \Gamma\left(\frac{r}{K} + 1\right) \quad (25)$$

If  $r=1$  and  $d=1 \Rightarrow \sum w_j = 1$ , then this reduces to the mean of the Weibull distribution and the moments, such as the Mean, Variance, Skewness and Kurtosis, can be obtained from (24).

The mean and variance are respectively

$$E(V_i) = C d \Gamma\left(\frac{1}{K} + 1\right) \sum_{j=0}^{\infty} \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{\frac{1}{K}+1}}$$

## ON THE EXPONENTIATED WEIBULL DISTRIBUTION

$$Var(V_i) = C^2 d \Gamma\left(\frac{2}{K} + 1\right) \sum_{j=0}^{\infty} \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{\frac{1}{K}+1}} - C^2 d^2 \Gamma^2\left(\frac{1}{K} + 1\right) \left( \sum_{j=0}^{\infty} \frac{(-1)^j \binom{d-1}{j}}{(1+j)^{\frac{1}{K}+1}} \right)^2$$

### Application

The fitting of monthly wind data collected across regions in the south western part of Nigeria was considered using data from the period between 1992 and 2012. Using the R-Package, the following results were obtained.

#### Estimates and Goodness-of-Fit for the Wind Speed Data

##### January

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2276	0.0002		1	19.3754	23.3754
Exponentiated Weibull	0.6678	1		10.2721	0.9916	4.9916

##### February

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2168	0.0001		1	21.32016	25.32016
Exponentiated Weibull	0.6598	1		13.5975	3.15927	7.15927

##### March

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.20399	0.00006		1	23.75956	27.75956
Exponentiated Weibull	0.641638	1		15.335852	6.31854	10.31854



**April**

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.19996	0.00005		1	24.55724	28.55724
Exponentiated Weibull	0.6896499	1		41.4258026	15.38302	19.38302

**May**

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.209261	0.000072		1	22.73786	26.73786
Exponentiated Weibull	0.6710878	1		20.7852047	8.297559	12.297589

**June**

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2096060	0.000073		1	22.67258	26.67258
Exponentiated Weibull	0.6915926	1		25.247869	10.20779	14.20779

**July**

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.21815	0.000106		1	21.0735	25.0735
Exponentiated Weibull	0.677590	1		15.3195704	5.094789	9.094789

**August**

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.211048	0.0000782		1	22.39852	26.39852
Exponentiated Weibull	0.692516	1		23.699390	9.502349	13.502349

## ON THE EXPONENTIATED WEIBULL DISTRIBUTION

### September

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2077786	0.00067		1	23.0227	27.0227
Exponentiated Weibull	0.6699132	1		21.8338828	8.774993	12.774993

### October

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.222508	0.000127		1	20.28256	24.28256
Exponentiated Weibull	0.6600494	1		11.9804517	2.943342	6.943342

### November

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2262799	0.0001477		1	19.61016	23.61016
Exponentiated Weibull	0.5949411	1		4.3888947	21.41094	25.41094

### December

Distributions	K	MLE		$\delta$	-2log l	AIC
		C				
Weibull	0.2426764	0.0002682		1	16.81126	20.81126
Exponentiated Weibull	0.6716818	1		7.055548	5.285842	9.285842

### Summary Statistics

Min	1 <sup>st</sup> Quarter	Median	Mean	3 <sup>rd</sup> Quarter	Mae
1.54	2.94	4.06	4.578	5.88	9.81

Note: Kurtosis = 2.502187, Skewness = 0.6333066

### Conclusion

The performance of Exponentiated Weibull and Weibull distribution functions to model wind energy was systematically compared. It was observed that the log

likelihood values and the Akaike information criterion (AIC) for the Exponentiated Weibull was always smaller for the Weibull distribution for each month except the month of November. This indicates that the proposed Exponentiated Weibull distribution outperformed the existing Weibull distribution for wind speed data in terms of minimum AIC and likelihood function over the months of the years under review. Thus, the exponentiated Weibull can be used as an alternative distribution that adequately describes wind speed, and may provide better representation of the potentials of wind energy.

## References

- Akpinar, S. & Akpinar, E. K. (2009). Estimation of wind energy potential using finite mixture distribution models. *Energy Conversion and Management*, 50(4): 877–84.
- Bivona, S., Burlon, R., & Leone, C.. (2003). Hourly wind speed analysis in Sicily. *Renewable Energy*, 28(9): 1371-1385.
- Burton, T., Sharpe, D., Jenkins, N., & Bossanyi, E. (2001). *Wind energy handbook*. Wiley.
- Carta, J. A., Ramirez, P., & Velazquez, S. (2009). A review of wind speed probability distributions used in wind energy analysis Case studies in the Canary Islands. *Renewable and Sustainable Energy Reviews*, 13(5): 933–55.
- Carta, J. A., Ramirez, P., & Velazquez, S. (2008) Influence of the level of fit of a density probability function to wind-speed data on the WECS means power output estimation. *Energy Conversion and Management*, 49(10):2647–55.
- Carta, J. A. & Ramirez, P. (2007). Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions. *Renewable Energy*, 32(3): 518–31.
- Carta, J. A. & Ramirez, P. (2007). Use of finite mixture distribution models in the analysis of wind energy in the Canarian Archipelago. *Energy Conversion and Management*, 48(1): 281–91.
- Celik, A. N. (2004). A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey. *Renewable Energy*, 29(4): 593–604.
- Chang, T-J & Tu, Y-L. (2007). Evaluation of monthly capacity factor of WECS using Chronological and probabilistic wind speed data: a case study of Taiwan. *Renewable Energy*, 32(12): 1999–2010.

## ON THE EXPONENTIATED WEIBULL DISTRIBUTION

Garcia, A., Torres, J. L., Prieto, E., & Francisco, A. D. (1998). Fitting wind speed distributions: a case study. *Solar Energy*, 62(2): 139–44.

Harris, R. I. (2006). Errors in gev analysis of wind epoch maxima from Weibull parents. *Wind and Structures, An International Journal*, 9(3): 179–91.

Harris, R. I. (2005). Generalized Pareto methods for wind extremes. Useful tool or mathematical mirage? *Journal of Wind Engineering and Industrial Aerodynamics*, 93(5): 341–60. doi:10.1016/j.jweia.2005.02.004.

Harter, H. L. & Moore, A. H. (1965). Maximum-Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and from Censored Samples. *Technometrics*, 7: 639 - 643.

Hennessey, J. P. (1977). Some aspects of wind power statistics. *Journal of Applied Meteorology and Climatology*, 16(2): 119–28.

International Electrotechnical Commission. *Wind turbines – part 1: design requirements. Tech. rep. 61400-1*, (Ed.3). International Electrotechnical Commission; 2005.

Jaramillo, O. A. & Borja, M. A. (2004). Wind speed analysis in La Ventosa, Mexico: a bimodal probability distribution case. *Renewable Energy*, 29(10): 1613–30.

Kiss, P. & Janosi, I. M. (2008). Comprehensive empirical analysis of ERA-40 surface wind speed distribution over Europe. *Energy Conversion and Management*, 49(8): 2142–51. doi:10.1016/j.enconman.2008.02.003.

Krishnamurthy, K. (2006). *Handbook of statistical distributions with applications*. New York: Chapman & Hall.

Lackner, M. A., Rogers, A. L., & Maxwell, J. F. (2008). Uncertainty Analysis in MCP-Based Wind Resource Assessment and Energy Production Estimation. *Journal of solar energy engineering*, 130(3): 31006.

Maxwell, J. F., McGowan, J. G., Rogers, A. L. (2002). *Wind energy explained: theory, design and application*. New York: Wiley.

Mudholkar, G. S., Srivastava, D. K. & Friemer, M. (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37: 436-445.

NOAA's National Data Buoy Center. (2009) National Data Buoy Center. Retrieved July 1<sup>st</sup>, 2009, from <http://www.ndbc.noaa.gov/>.

Ramirez, P. & Carta, J. A. (2005). Influence of the data sampling interval in the estimation of the parameters of the Weibull wind speed probability density

distribution: a case study. *Energy Conversion and Management*, 46(15–16): 2419–38.

Rehman, S., Halawani, T. O., Husain, T. (1994). Weibull parameters for wind speed distribution in Saudi Arabia. *Solar Energy*, 53(6): 473-479.

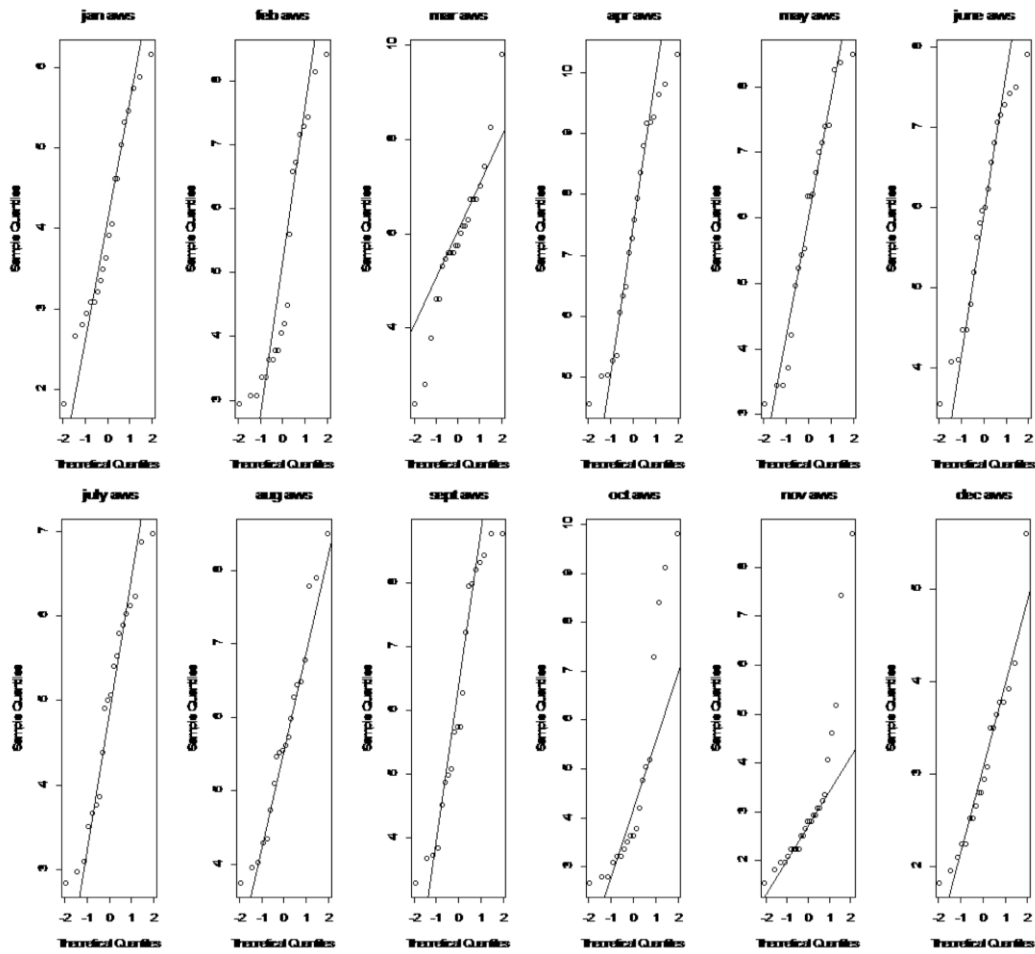
Silva, B. R. and Cordeiro, G. M. (2012). The Burr XII power series distribution: a new compounding family. In press. *Brazilian Journal of Probability and Statistics*. <http://imstat.org/bjps/papers/BJPS234.pdf>

Simiu, E., Heckert, N., Filliben, J., & Johnson, S. (2001). Extreme wind load estimates based on the Gumbel distribution of dynamic pressures: an assessment. *Structural Safety*, 23(3): 221–9.

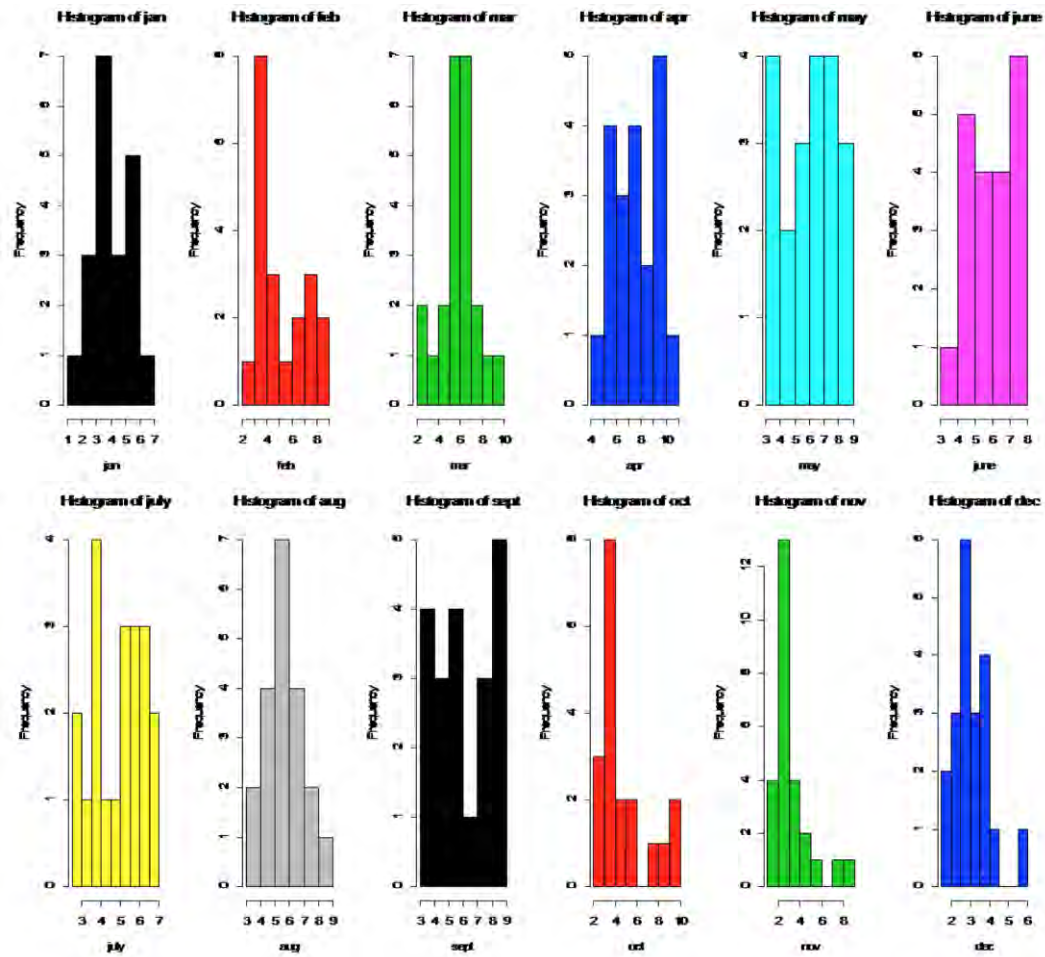
*Wind powering America*. (2009). US Dept. of Energy. Retrieved September, 2009, from [http:// www.windpoweringamerica.gov](http://www.windpoweringamerica.gov)

## Appendix

### QQ Plots



## Histograms



# Robust Regression Analysis for Non-Normal Situations under Symmetric Distributions Arising In Medical Research

**S. S. Ganguly**  
Sultan Qaboos University  
Muscat, Oman

---

In medical research, while carrying out regression analysis, it is usually assumed that the independent (covariates) and dependent (response) variables follow a multivariate normal distribution. In some situations, the covariates may not have normal distribution and instead may have some symmetric distribution. In such a situation, the estimation of the regression parameters using Tiku's Modified Maximum Likelihood (MML) method may be more appropriate. The method of estimating the parameters is discussed and the applications of the method are illustrated using real sets of data from the field of public health.

*Keywords:* Maximum likelihood, modified maximum likelihood, student's t-distribution, order statistics, delta method

---

## Introduction

Often in medicine, a relationship is established between a response variable  $y$ , which depends on the  $r$  covariates  $x_1, x_2, \dots, x_r$ , which are independent of each other, so that, in total, there may be  $(r + 1)$  variables. In classical regression model, the response variable  $y$  is treated as a random variable whose mean depends upon fixed variables of the  $x_i$ 's. The mean is assumed to be linear function of the regression coefficients  $\alpha, \beta_1, \beta_2, \dots, \beta_r$ .

The linear regression model also arises in a different setting. Suppose all the variables  $y, x_1, x_2, \dots, x_r$  are random and have a joint distribution

$$f(y, x_1, x_2, \dots, x_r),$$

which is not necessarily normal so that

---

*S. S. Ganguly is a Professor in the Department of Family Medicine and Public Health.  
Email at: [ganguly@squ.edu.om](mailto:ganguly@squ.edu.om).*



$$f(y, x_1, x_2, \dots, x_r) = g(y|x_1, x_2, \dots, x_r) \prod_{i=1}^r h(x_i). \quad (1)$$

It is assumed herein that the conditional distribution of  $y$  given  $x_1, x_2, \dots, x_r$  is normal and is given by

$$g(y|x_1, x_2, \dots, x_r) \approx \left\{ \sigma_o^2 (1 - \sum_{i=1}^r \rho_{oi}^2) \right\}^{-\frac{1}{2}} \times \exp \left[ -\frac{1}{2\sigma_o^2 (1 - \sum_{i=1}^r \rho_{oi}^2)} \left\{ y - \mu_o - \sum_{i=1}^r \rho_{oi} \left( \frac{\sigma_o}{\sigma_i} \right) (x_i - \mu_i) \right\}^2 \right] \quad (2)$$

with mean

$$E(y|\underline{x}) = \mu_o - \sum_{i=1}^r \rho_{oi} \left( \frac{\sigma_o}{\sigma_i} \right) (x_i - \mu_i) \quad (3)$$

and variance

$$V(y|\underline{x}) = \sigma_o^2 \left( 1 - \sum_{i=1}^r \rho_{oi}^2 \right). \quad (4)$$

The marginal density corresponding to the covariate  $x_i$  is assumed to be symmetric about mean of the form:

$$\frac{1}{\sigma_i} f \left( \frac{x_i - \mu_i}{\sigma_i} \right) \quad (5)$$

Here  $\mu_i = E(x_i)$ ,  $\sigma_i^2 = V(x_i)$  and  $\rho_{oi} (i=1, 2, \dots, r)$  is the correlation coefficient between  $y$  and  $x_i$ . Relation (2) provides for the measurement of dependency of the response random variable on the random covariates  $x_i (i=1, 2, \dots, r)$ .

The linear relationship may also be written in the form of classical regression model as

$$E(y|\underline{x}) = \alpha + \sum_{i=1}^r \beta_i x_i \quad (6)$$

where

$$\alpha = \mu_o - \sum_{i=1}^r \rho_{oi} \left( \frac{\sigma_o}{\sigma_i} \right) \mu_i \quad (7)$$

and

$$\beta_i = \rho_{oi} \left( \frac{\sigma_o}{\sigma_i} \right), \quad i = 1, 2, \dots, r \quad (8)$$

are the regression coefficients. It may be noted that  $E(y|\underline{x})$  is the best linear predictor of the response variable  $y$  where the population is  $N_{r+1}(\underline{\mu}, \Lambda)$ .

In medical epidemiology, one often encounters situations where some (if not all) covariates  $x_i$  have non-normal symmetric distributions. This article is restricted to a situation where the covariates have non-normal symmetric distributions. The objective, therefore, is to estimate the parameters  $(\alpha, \underline{\beta})^T$  from  $n$  sample values  $(y_i, \underline{x}_i)$ ,  $1 \leq i \leq n$ . For this, consider the family of student's  $t$ -distributions. The method, which has been developed here, is, of course, general and can be used for other families of location-scale distributions of the type (5).

## Likelihood equations

Suppose that the covariate  $x_i$  ( $i=1, 2, \dots, r$ ) has the symmetric distribution with the density given by

$$h(x_i) \approx \left( k_i \sigma_i^2 \right)^{-\frac{1}{2}} \left\{ 1 + \frac{(x_i - \mu_i)^2}{k_i \sigma_i^2} \right\}^{-p_i}, \quad -\infty < x_i < \infty \quad (9)$$

where  $k_i = 2p_i - 3$ ,  $p_i \geq 2$ ;  $E(x_i) = \mu_i$  and  $v(x_i) = \sigma_i^2$ . Assume that  $p_i$  is known. For  $p_i = 5$ , (9) is almost indistinguishable from logistic distribution, because the two distributions are both symmetric and have first four moments common (Pearson, 1963). If the two distributions are plotted, it will be seen that one sits almost on top of the other. It may be noted that

$$t = \frac{\{v_i(x_i - \mu_i)\}^{\frac{1}{2}}}{\sigma_i(k_i)^{\frac{1}{2}}}$$

has Student's  $t$  – distribution with  $(2p_i-1)$  degrees of freedom. For  $1 \leq p_i \leq 2$ ,  $k$  is equal to 1 in which case  $\sigma$  in (9) is simply a scale parameter.

Given the data matrix  $(n > r+1)$  of the form

$$(y_j; x_{j1}, \dots, x_{ji}, \dots, x_{jk}), j = 1, 2, \dots, n \quad (10)$$

where  $y$  is the response variable and the  $x$  terms as explanatory variables or covariates. Then the likelihood function  $L$  based on relation (1) can be written as usual and is given by

$$\begin{aligned} L \approx & \left\{ \left( \prod_{i=1}^r \sigma_i^2 \right) \left( 1 - \sum_{i=1}^r \rho_{oi}^2 \right) \right\}^{\frac{n}{2}} \\ & * \exp \left[ - \frac{1}{2 \sigma_o^2 \left( 1 - \sum_{i=1}^r \rho_{oi}^2 \right)} \sum_{j=1}^n \left\{ y_{[j]} - \mu_o - \sum_{i=1}^r \rho_{oi} \left( \frac{\sigma_o}{\sigma_i} \right) (x_{i(j)} - \mu_i) \right\}^2 \right] \quad (11) \\ & * \prod_{j=1}^n \prod_{i=1}^r \left\{ 1 + \frac{(x_i - \mu_i)^2}{k_i \sigma_i^2} \right\}^{-p_i} \end{aligned}$$

where  $x_{i(j)}$ ,  $(i = 1, 2, \dots, r; j = 1, 2, \dots, n)$  are the order statistics of  $x_i$  observations, and  $y_{[j]}$  ( $j = 1, \dots, n$ ) are the corresponding concomitant  $y$  observations. The maximum likelihood estimators are the solutions of the likelihood equations, i.e, of the derivatives of  $\ln L$ . These equations are, however, intractable. Solving them by iterative procedures may be problematic, for example, one may encounter multiple roots, slow convergence, or convergence to wrong values (see specifically Barnett, 1966; Lee et al., 1980; Tiku and Suresh, 1992; Vaughan, 1992). Instead the Tikus method of modified likelihood (MML) estimation was employed, which gives explicit estimators and involves replacing intractable terms by linear approximations. Because this method is already well established

and is known to produce estimators which are fully efficient for large  $n$  (Tiku, 1970; Bhattacharyya, 1985) and almost fully efficient for small  $n$  (Tiku et al, 1986; Tiku and Suresh, 1992; Vaughan, 1992, 1994).

## Modified Maximum Likelihood

Consider the  $i^{\text{th}}$  covariate of a random sample of size  $n$  denoted by  $x_{1i}, x_{2i}, \dots, x_{ni}$  from any location-scale distribution with density given by

$$\frac{1}{\sigma_i} f\left[\frac{x_{ji} - \mu_i}{\sigma_i}\right], \quad i = 1, 2, \dots, r.$$

For simplicity of notation, suppress the suffix  $i$  and consider  $f$  to be a student  $t$  density. Then the likelihood equations for estimating  $\mu$  and  $\sigma$  corresponding to each covariate are

$$\frac{\partial \ell n L}{\partial \mu} = \frac{2p}{k\sigma} \sum_{j=1}^n g(z_j) = 0$$

and

$$\frac{\partial \ell n L}{\partial \sigma} = \frac{-n}{\sigma} + \frac{2p}{k\sigma} \sum_{j=1}^n z_j g(z_j) = 0$$

where

(12)

$$z_j = \frac{(x_j - \mu)}{\sigma}$$

and

$$g(z_j) = \frac{z_j}{\{1 + (1/k)z_j^2\}}.$$

Equations (12) do not provide explicit solutions. Following Tiku-Suresh (1992); Vaughan and Tiku (2000), the first step is to express these equations in terms of order statistics  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Because complete sums are invariant to ordering

$$\frac{\partial \ell n L}{\partial \mu} = \frac{2p}{k\sigma} \sum_{j=1}^n g(z_{(j)}) = 0$$

and

$$\frac{\partial \ell n L}{\partial \sigma} = \frac{-n}{\sigma} + \frac{2p}{k\sigma} \sum_{j=1}^n z_{(j)} g(z_{(j)}) = 0 \quad (13)$$

where

$$z_{(j)} = \frac{(x_{(j)} - \mu)}{\sigma}, j = 1, 2, \dots, n.$$

Under appropriate regularity considerations which are very general in nature,  $g(z_{(j)})$  can be replaced by linear approximations given by the first two terms of Taylor series expansions (Tiku, 1967, 1968; Tiku and Suresh, 1992; Tiku and Kambo, 1992, Vaughan, 1992; Vaughan and Tiku, 2000), so that

$$\begin{aligned} g\{z_{(j)}\} &\approx g\{t_{(j)}\} + [z_{(j)} - t_{(j)}] \left\{ \frac{d}{dz} g(z) \right\}_{z=t_{(j)}} \\ &= \alpha_j + \beta_j z_{(j)}, j = 1, 2, \dots, n \end{aligned} \quad (14)$$

where

$$t_{(j)} = E\{z_{(j)}\}.$$

Thus, the modified equations are obtained, i.e.

$$\frac{\partial \ell n L}{\partial \mu} \approx \frac{\partial \ell n L^*}{\partial \mu} = \frac{2p}{k\sigma} \sum_{j=1}^n \{\alpha_j + \beta_j z_{(j)}\} = 0$$

(15)

and

$$\frac{\partial \ell n L}{\partial \sigma} \approx \frac{\partial \ell n L^*}{\partial \sigma} = \frac{-n}{\sigma} + \frac{2p}{k\sigma} \sum_{j=1}^n z_{(j)} \{\alpha_j + \beta_j z_{(j)}\} = 0$$

Equations (15) have explicit solutions, which are called modified maximum likelihood (MML) estimators. Note that the ML and MML estimators are asymptotically equivalent.

For distribution  $(p \geq 2, k = 2p - 3)$

$$h(x_j) \propto (k\sigma)^{-\frac{1}{2}} \left\{ 1 + \frac{x_{(j)} - \mu}{k\sigma^2} \right\}^{-p}, -\infty < x_j < \infty \quad (16)$$

This method gives the following MML estimators (see Tiku and Suresh, 1992; Tiku and Kambo, 1992; Vaughan, 1992; Vaughan and Tiku, 2000; Tiku et al, 2008)

$$\hat{\mu} = \frac{1}{m} \sum_{j=1}^n \beta_j x_{(j)} \quad (m = \sum_{j=1}^n \beta_j) \quad (17)$$

and

$$\hat{\sigma} = \frac{\left\{ B + (B^2 + 4nc)^{\frac{1}{2}} \right\}}{2\{n(n-1)\}} \quad (18)$$

where

$$B = \frac{2p}{k} \sum_{j=1}^n \alpha_j x_{(j)} \text{ and } C = \frac{2p}{k} \left\{ \sum_{j=1}^n \beta_j y_{(j)}^2 - m \hat{\mu}^2 \right\} \quad (19)$$

The coefficients  $\alpha_j$  and  $\beta_j$  are obtained from the equations

$$\alpha_j = \frac{(2/k) t_{(j)}^3}{\left\{ 1 + \left( \frac{1}{k} \right) t_{(j)}^2 \right\}^2} \text{ and } \beta_j = \frac{1 - (1/k) t_{(j)}^2}{\left\{ 1 + (1/k) t_{(j)}^2 \right\}^2}, j = 1, 2, \dots, n \quad (20)$$

For  $p = \infty$  (i.e. for normal distribution),  $\alpha_j = 0$  and  $\beta_j = 1$ , because  $k=2p-3$ . Note that  $\alpha_j = -\alpha_{n-j+1}$ ,  $\beta_j = \beta_{n-j+1}$  ( $1 \leq j \leq n$ ) and  $\sum_{j=1}^n \alpha_j = 0$ . Tables of the value of  $t_{(j)}$  are available for  $p=2(5) 10$  and  $n \leq 20$  (Tiku and Kumra, 1985). For  $n > 20$ ,  $t_{(j)}$  are obtained from the equation

$$\int_{-\infty}^{t_{(j)}} f(z)dz = \frac{j}{n+1} \quad (1 \leq j \leq n). \quad (21)$$

In evaluating (21), it should be noted that  $\{(k/\nu)z\}^{1/2}$  has student's t-distribution with  $\nu = 2p - 1$  degrees of freedom.

It may be of interest to note that in deriving the estimators  $\mu$  and  $\sigma$  given by the equations (17)-(20), the method of MML estimation for  $p < \infty$  automatically gives small weights to extreme order statistics close to the center. It is precisely due to this fact these estimators are robust to reasonable departures from the true value of  $p$  in (16). In most applications, therefore, it is not very important to pinpoint the true value of  $p$  and use it in all derivatives. Any reasonable value of  $p$  gives almost optimal results.

A Q-Q plot can be employed to give a reasonable value closure (if not exactly) the true value of  $p$  corresponding to covariate  $x$  (Tiku et al, 1986, p.277). The order statistic  $x_{(j)}$  is plotted against the values  $t_{(j)} = E(z_{(j)})$ ,  $z_j = (x_{(j)} - \mu) / \sigma$ ,  $j = 1, 2, \dots, n$ , under the assumed model, i.e. for a particular value of  $p$  in (16). If the plot gives a straight line (or nearly so), the model is taken to be valid for the MML estimation.

Following the above procedure, the parameters  $\mu_i$  and  $\sigma_i (i = 1, 2, \dots, r)$  are estimated. In order to estimate the remaining parameters viz.,  $\mu_o, \sigma_o, \rho_{oi} (i = 1, 2, \dots, r)$ , the likelihood function (11) is considered. Because  $\frac{\partial \ln L}{\partial \mu_i}$  and  $\frac{\partial \ln L}{\partial \sigma_i}$ ,  $(i = 1, 2, \dots, r)$  are expressed in terms of  $g\{z_{i(j)}\}$ , the likelihood equations  $\frac{\partial \ln L}{\partial \mu_i} = 0$ ,  $\frac{\partial \ln L}{\partial \sigma_i} = 0$  ( $i = 1, 2, \dots, r$ ) and  $\frac{\partial \ln L}{\partial \rho_{oi}} = 0$  ( $i = 1, 2, \dots, r$ ) have no explicit solutions. The modified likelihood equations are  $\frac{\partial \ln L^*}{\partial \mu_i} = 0$ ,  $\frac{\partial \ln L^*}{\partial \sigma_i} = 0$ , ( $i = 0, 1, \dots, r$ ) and  $\frac{\partial \ln L^*}{\partial \rho_{oi}} = 0$  ( $i = 1, 2, \dots, r$ ), and are obtained by replacing  $g\{z_{i(j)}\}$  with the linear approximations given by (14). The solutions of these equations are the following MML estimators:

$$\hat{\mu}_o = \bar{y} - \sum_{i=1}^r \hat{\rho}_{oi} \left( \frac{\hat{\sigma}_o}{\hat{\sigma}_i} \right) (\bar{x}_i - \hat{\mu}_i) \quad (22)$$

$$\hat{\sigma}_o = \left[ S_o^2 + \sum_{i=1}^r \left\{ \frac{S_{oi}^2}{S_i^2} \left( \frac{\sigma_i^2}{S_i^2} - 1 \right) \right\} \right]^{1/2} \quad (23)$$

$$\hat{\rho}_{oi} = \frac{S_{oi}^2}{S_i^2} \left[ \frac{\hat{\sigma}_1}{\hat{\sigma}_o} \right], \quad i = 1, 2, \dots, r \quad (24)$$

Here,

$$n\bar{y} = \sum_{j=1}^n y_{[j]} = \sum_{j=1}^n y_j, \quad n\bar{x} = \sum_{j=1}^n x_{i(j)} = \sum_{j=1}^n x_{ij} \quad (25)$$

$$(n-1)S_o^2 = \sum_{j=1}^n [y_{[j]} - \bar{y}]^2 = \sum_{j=1}^n (y_j - \bar{y})^2 \quad (26)$$

$$(n-1)S_i^2 = \sum_{j=1}^n [x_{i(j)} - \bar{x}_i]^2 = \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \quad (27)$$

and

$$\begin{aligned} (n-1)S_{oi}^2 &= \sum_{j=1}^n [y_{[j]} - \bar{y}] [x_{i(j)} - \bar{x}_i] \\ &= \sum_{j=1}^n [y_j - \bar{y}] [x_{ij} - \bar{x}_i], \quad i = 1, 2, \dots, r. \end{aligned} \quad (28)$$

Relation (22) provides for the measurement of dependency of the response random variable on the random covariates  $x_i$  ( $i = 1, 2, \dots, r$ ). The linear relationship is also represented in the form of classical model (6).

The asymptotic variances and covariances of the estimators  $\hat{\mu}_o, \hat{\mu}_1, \hat{\sigma}_o, \hat{\sigma}_i$  and  $\hat{\rho}_{oi}$  ( $i = 1, 2, \dots, r$ ) are obtained with the use of the second partial derivatives of the likelihood function (11). The matrix formed by the negative of the expected values of the second partial derivatives gives the information matrix, which may be expressed as the partitioned matrix



$$\underline{V} = \begin{pmatrix} \underline{V}_1 & \underline{O} \\ \underline{O} & \underline{V}_2 \end{pmatrix} \quad (29)$$

where the matrix is of the order  $(3r+2) \times (3r+2)$  and

$$\underline{V}_1 = E \left[ -\frac{\partial^2 \ln L^*}{\partial \mu_i \partial \mu_{i'}} \right]$$

of order  $(r+1) \times (r+1)$  and

$$\underline{V}_2 = E \left[ -\frac{\partial^2 \ln L^*}{\partial \theta_i \partial \theta_{i'}} \right], \quad i, i' = 1, 2, \dots, r$$

of order  $(2r+1) \times (2r+1)$  with  $(\theta_1 = \sigma_o, \theta_2 = \sigma_1, \dots, \theta_{2r+1} = \rho_{ok})$ .

The inverse of  $\underline{V}_1$  and  $\underline{V}_2$  matrices provides the elements of the precision and covariance structure of the estimated coefficients.

The estimated values of the parameters obtained above are used in relation (7) and (8) which give the estimated values of the regression coefficients  $\alpha$  and  $\beta_i$  ( $i=1, 2, \dots, r$ ) of the model (6). The asymptotic covariance structure of the estimated regression coefficients  $\hat{\alpha}$  and  $\hat{\beta}_i$  ( $i=1, 2, \dots, r$ ) are obtained using delta method (Serfling, 1980) as:

Let  $g(\alpha, \beta)$  and  $\underline{\theta} = (\underline{\mu}, \underline{\sigma}, \underline{\rho})$ , then

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \cong N_{(r+1)} \left[ \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, n^{-1} G^T \hat{\Sigma} G \right], \quad (30)$$

where

$$G = \left[ \frac{\partial \underline{g}}{\partial \underline{\theta}} \right]_{\mu_o, \dots, \rho_{ok}}$$

of order  $(3r+2) \times (r+1)$  and

$$\hat{\Sigma} = \begin{pmatrix} \hat{\underline{V}}_1^{-1} & \underline{O} \\ \underline{O} & \hat{\underline{V}}_2^{-1} \end{pmatrix}$$

of order  $(3r+2) \times (3r+2)$ . Note that when  $p = \infty$  the distribution (16) reduces to the ideal normal distribution in which case  $\hat{\mu} = \bar{x}$  (sample mean) and  $\hat{\sigma}^2 = s^2$  (sample variance),  $\bar{x}$  and  $s^2$  being optimal under the assumption of normality.

## Examples

### Example 1

Consider the part of the data set pertaining to 20 male insulin-dependent diabetic patients as provided in Dobson (1990, p. 69), which is reproduced in Table 1.

**Table 1.** Carbohydrate, age and weight for twenty insulin-dependent diabetics

$y = \text{Carb. (gm)}$	$x_1 = \text{Age (yrs)}$	$x_2 = \text{Wgt (kg)}$	$y = \text{Carb. (gm)}$	$x_1 = \text{Age (yrs)}$	$x_2 = \text{Wgt (kg)}$
33	33	100	50	31	108
40	47	92	51	61	85
37	49	135	30	63	130
27	35	144	36	40	127
30	46	140	41	50	109
43	52	101	42	64	107
34	62	95	46	56	117
48	23	101	24	61	100
30	32	98	35	48	118
38	42	105	37	28	102

In this sample, the goal is to establish the relationship between the response variable  $y$  (amount of carbohydrate) and the two covariates  $x_1$  (age) and  $x_2$  (body weight, relative to “ideal” weight for height) using the linear regression model (6) which takes the form

$$E(y|x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 \quad (31)$$

Here, it is assumed that, in relation (1), the conditional distribution of the response random variable  $y$  is normal; however, the covariates follow

independently non-normal symmetric distribution. The model (31) is fitted using above described modified maximum likelihood method.

First obtain the values of  $p_1$  and  $p_2$  corresponding to the two covariates  $x_1$  and  $x_2$  using Q-Q plots, where the order statistics  $x_{1(j)}$  and  $x_{2(j)}$  were plotted separately against  $t_{1(j)}$  and  $t_{2(j)}$  respectively,  $j = 1, \dots, n$  for different values of  $p$  as given in Tiku and Kumra (1985). The values of  $p_1 = 5$  and  $p_2 = 7$  provided an approximate straight line patterns which determined the appropriate types of densities in (16). Once  $p_1$  and  $p_2$  are known, then using the equations (17)-(20), the MML estimates of the parameters  $\mu_1, \sigma_1$  and  $\mu_2, \sigma_2$  are obtained. Using these values in equations (22)-(28) the rest of the parameters  $\mu_o, \sigma_o, \rho_{o1}$  and  $\rho_{o2}$  are estimated. Solutions of the information matrix (29) provided the elements of the precision and covariance structure of the estimated parameters. The estimated values and their standard errors are presented in Table 2.

**Table 2.** MML estimates of the parameters and their standard errors for the data set in Table 1

Param.	Est.	Std. Err.
$\mu_o$	37.732	1.848
$\mu_1$	46.437	3.008
$\mu_2$	109.936	3.776
$\sigma_o$	7.635	1.411
$\sigma_1$	13.989	1.789
$\sigma_2$	17.265	2.351
$\rho_{o1}$	-0.064	0.228
$\rho_{o2}$	-0.420	0.243

**Table 3.** MML and ML estimates of the parameters and their standard errors for the data set in Table 1

Param.	Est.	Std. Err.	W
Constant ( $\alpha$ )	59.783	12.469	
<b>MML</b> Coefficient ( $\beta_1$ )	-0.035	0.124	-0.282
Coefficient ( $\beta_2$ )	-0.186	0.099	-1.879
Constant ( $\alpha$ )	60.432	13.017	
<b>ML</b> Coefficient ( $\beta_1$ )	-0.046	0.131	-0.351
Coefficient ( $\beta_2$ )	-0.187	0.101	-1.851

Using the estimated values in Table 2 in relation (7) and (8), obtain MML estimates of the regression parameters  $\alpha, \beta_1$  and  $\beta_2$ . Use of delta method as described in (30) provided the asymptotic standard errors; also these parameters based on usual maximum likelihood method were estimated. The results, obtained under the two methods are summarized in Table 3.

The analysis in Table 3 reveals that the MML estimates of the regression parameters for the data set in Table 1 are very close to the values obtained using maximum likelihood method, as expected. Moreover, the two methods gave approximately the same results for the Wald statistics  $W$ , which permits to test the

null hypothesis  $H_o : \beta_1 = 0$  and  $\beta_2 = 0$ . For large  $n$ , the null distribution of  $W$  is referred to a standard normal distribution.

### Example 2

Consider another data set from Murray (1937), reproduced in El-Saidi (1995, p. 214) as shown in Table 4. The data provides 11 observations on the number of male flies died after twenty minutes exposure to pyrethrum at various concentrations.

The main objective is to describe the probability of success  $p_j$  as a function of dose  $x_j$ . In literature, such type of analysis are carried out usually considering either probit or logit models (Cox, 1970). However, the logit model is preferred to a probit model due to two primary reasons (Hosmer and Lameshow, 1989): from mathematical point of view, it is an easily used function, and it leads to itself to a biological meaningful interpretation.

**Table 4.** Mortality of male flies after twenty minutes exposure to pyrethrum

Concentration (log <sub>10</sub> )	Number of flies		Proportions Died
	Exposed	Died	
1.6020	462	109	0.2359
1.7782	500	199	0.3980
1.9031	467	298	0.6381
2.0000	515	370	0.7185
2.0792	561	459	0.8182
2.1461	469	400	0.8529
2.2041	550	495	0.9000
2.2553	542	499	0.9207
2.3010	479	450	0.9395
2.3979	497	476	0.9577
2.4771	453	442	0.9757

The logit model is a family of Generalized Linear Models (GLMs) with link function  $g(p_j)$  as  $\ln\left(\frac{p_j}{1-p_j}\right)$  (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). The link function  $g(p_j)$  is continuous and maps the  $[0,1]$  range of probabilities onto  $(-\infty, \infty)$  and is represented by

$$g(p_j) = \ln \left( \frac{p_j}{1-p_j} \right) = \alpha + \beta x_j, j = 1, 2, \dots, n \quad (32)$$

so that

$$p_j = \frac{\exp(\alpha + \beta x_j)}{1 + \exp(\alpha + \beta x_j)}, j = 1, 2, \dots, n \quad (33)$$

The relation (33) is known as binary logistic model with probability of success  $p_j$ , this belongs to the standardized logistic distribution which is symmetric in nature (Rao and Toutenburg, 1995, p. 263).

In order to estimate the unknown parameters  $\alpha$  and  $\beta$  in (32), usually ML method is used. The technique involves the solution of the likelihood equations, which have no explicit solutions and have to be solved by interactive procedures. Solving these equations is, therefore, tedious and time consuming. Therefore, these parameters are estimated using MML method.

For this, consider the link function i.e. log odds as a response variable and  $x_j$  as a covariate. First estimate  $\mu_1$  and  $\sigma_1$  for  $p=5$  in distribution (16). Using these values in equations (22)-(28), the rest of the parameters  $\mu_o, \sigma_o$  and  $\rho_{o1}$  involved in the likelihood function (11) were obtained. The estimated values of the variances and co-variances were obtained using these values in second partial derivatives of the likelihood function (11) and solving for the inverse of the information matrix (29). The estimated values of the parameters and their standard errors involved in the likelihood function (11) with  $p=5$  for the data set in Table 4 are shown in Table 5.

Using these estimated values of the parameters in relation (7) and (8), obtain the MML estimates of the parameters  $\hat{\alpha}$  and  $\hat{\beta}$  of the logistic model (33). The use of delta method (30) gave the asymptotic variances of  $\hat{\alpha}$  and  $\hat{\beta}$ . The ML estimates of these parameters and their variances under the logit model (32) were also obtained using iterative procedures viz; Newton-Raphson method (Cox, 1970, Chapter 2). The results obtained under the two procedures are summarized in Table 6.

These analyses also reveal that the MML estimates of the regression parameters  $\alpha$  and  $\beta$  for the data set in Table 4 are very close to the values obtained using maximum likelihood method, as expected.

## ROBUST REGRESSION ANALYSIS FOR NON-NORMAL SITUATIONS

**Table 5.** MML estimates of the parameters and their standard errors for the data set in Table 4

Param.	Est.	Std. Err.
$\mu_0$	1.642	0.459
$\mu_1$	2.115	0.082
$\sigma_0$	1.593	0.198
$\sigma_1$	0.284	0.035
$\rho_{01}$	0.999	0.003

**Table 6.** MML and ML estimates of the parameters and their standard errors for logit model (32)

	Param.	Est.	Std. Err.
<b>MML</b>	Constant ( $\alpha$ )	-10.219	0.186
	Coefficient ( $\beta$ )	5.608	0.087
<b>ML</b>	Constant ( $\alpha$ )	-10.329	0.343
	Coefficient ( $\beta$ )	5.661	0.172

This study used Tiku's modified maximum likelihood method for carrying out regression analysis when the underlying distributions of the data set have non-normal symmetric distributions. The method yields estimators which are explicit functions of sample observations and are numerically very close to the maximum likelihood estimators and equally efficient.

## References

- Barnet, V. D. (1966). Order statistics estimators of the location of the Cauchy distribution. *Journal of America Statistical Association*, 61(316): 1205-1218.
- Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. *Journal of American Statistical Association*, 80(390): 398-404.
- Cox, D. R. (1970). *The analysis of binary data*. Methuen: London.
- Dobson, A. J. (1990). *An introduction to generalized linear models*. Chapman and Hall: New York.
- El-Saidi, M. A. (1995). A symmetric extended logistic model with applications to experimental toxicity data. *Biometrical Journal*, 37(2), 205-216.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression*. John Wiley: New York.
- Lee, K. R., Kapadia, C. H. and Dwight, B. B. (1980). On estimating the scale parameters of the Rayleigh distribution from doubly censored samples. *Statistische Hefte*, 21(1): 14-29.
- McCullagh, P and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall: London.

- Murray, C.A. (1937). A statistical analysis of fly mortality data. *Soap*, 13(8): 89-105.
- Nelder, J. A. and Wedderburn, R. W. N. (1972). Generalized linear models. *Journal of Royal Statistical Society, Series A*, 135(3): 370-384.
- Pearson, E. S. (1963). Some problems arising in approximating to probability distributions using moments. *Biometrika*, 50, 95-112.
- Rao, C. R. and Toutenburg, H. (1995). *Linear models: least squares and alternatives*. Springer-Verlag: New York.
- Serfling, R. J. (1980). *Approximation theorems of mathematical studies*. Wiley: New York.
- Tiku, M. L. (1967). Estimating the mean and standard deviation from a censored normal sample. *Biometrika*, 54(1/2): 155-165.
- Tiku, M. L. (1968). Estimating the parameters of normal and logistic distributions from censored samples. *Australian Journal of Statistics*, 10(2): 64-74.
- Tiku, M. L., Islam, M. Q. and Sazak, H.S. (2008). Estimation in bivariate non-normal distributions with stochastic variance functions. *Computational Statistics & Data Analysis*, 52(3): 1728-1745.
- Tiku, M. L. (1970) Monte Carlo study of some simple estimators in censored normal samples. *Biometrika*, 57(1): 207-211.
- Tiku, M. L. and Kambo, N. S. (1992) Estimation and hypothesis testing for a new family of bivariate non normal distributions. *Communications in Statistics – Theory and Methods*, 21(6): 1683-1705.
- Tiku, M. L. and Kumra, S. (1985). Expected values and variances and covariances of order statistics for a family of symmetric distributions (Student's  $t$ ). In B. J. Trawinski, R. E. Bechhofer, S. Kumra, M. L. Tiku, & A. C. Tahmane (Eds.) *Selected tables in mathematical statistics, Vol. 8*. Providence, R.I.: American Mathematical Society: pp. 141-270.
- Tiku, M. L. and Suresh, R. P. (1992). A new method of estimation for location and scale parameters. *Journal of Statistical Planning and Inference*, 30(2): 281-292.
- Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). *Robust Inference*. Marcel Dekker : New York.
- Vaughan, D. C. (1992). On the Tiku-Suresh method of estimation. *Communications in Statistics – Theory and Methods*, 21(2): 451-469.

## ROBUST REGRESSION ANALYSIS FOR NON-NORMAL SITUATIONS

Vaughan, D. C. (1994). The exact values of the expected values, variances and covariances of the order statistics from the Cauchy distribution. *Journal of Statistical Computation and Simulation*, 49(1-2): 21-32.

Vaughan, D. C. & Tiku, M. L. (2000). Estimation and hypothesis testing for a non-normal bivariate distribution with applications. *Mathematical and Computer Modelling*, 32: 53-67.



## **JMASM Algorithms and Code: A Flexible Method for Conducting Power Analysis for Two- and Three-Level Hierarchical Linear Models in R**

**Yi Pan**

University of North Carolina – Chapel Hill  
Chapel Hill, NC

**Matthew T. McBee**

East Tennessee State University  
Johnson City, Tennessee

---

A general approach for conducting power analysis in two- and three-level hierarchical linear models (HLMs) is described. The method can be used to perform power analysis to detect fixed effects at any level of a HLM with dichotomous or continuous covariates. It can easily be extended to perform power analysis for functions of parameters. Important steps in the derivation of this approach are illustrated and numerical examples are provided. Sample code implementing this approach is provided using the free program *R*.

*Keywords:* power analysis, hierarchical linear model, mixed model, *R*, power analysis for hierarchical linear model

---

Hierarchical linear modeling (HLM) is widely used in various areas of social science (Singer, 1998; Raudenbush & Bryk, 2002). As with any quantitative method, it is frequently important to perform power analysis in order to determine the necessary sample size to achieve a given level of power, to describe the minimum detectable effect size, or to describe the level of precision in the estimation of effects that is achievable by a given study design and sample size.

Power analysis in the general linear model context is straightforward. Many empirical researchers are trained in the methods of performing power analysis for linear models and several excellent pieces of software, such as *GPower* and *SAS PROC GLMPOWER*, are widely available (Thomas & Krebs, 1997; Lewis, 2006). The penetration of HLM into the mainstream of a variety of social science disciplines has created a need for convenient tools to perform power analysis for HLMs. Several software applications are currently available for HLM power

---

*Dr. Pan is a statistician at UNC-Chapel Hill's Frank Porter Graham Child Development Institute. Email him at [yi.pan@unc.edu](mailto:yi.pan@unc.edu). Dr. McBee is an Assistant Professor in the Department of Psychology. Email him at: [mcbeem@etsu.edu](mailto:mcbeem@etsu.edu).*

analysis. *Optimal Design* (Raudenbush, et al., 2004) is a widely used HLM power analysis software in social sciences, and allows researchers conduct power analysis on difference between treatment and control group in a number of cluster data analysis scenarios. However, it lacks the functionality of conducting power analysis for continuous predictors. *Power Analysis in Two-Level Designs* (PinT; Snijders & Bosker, 1993; Bosker, Snijders, & Guldemon, 1999) accommodates power analysis for continuous variables, but is limited to 2-level HLM's. Simulation-based power analysis software, like *MLPowSim* (Browne, Golarizadeh & Parker, 2009) and *ML-Des* (Cools, Van den Noortgate & Onghena, 2008), offer more flexibility, but it takes a much longer time to conduct simulation-based power analysis, and they do not allow unbalanced design.

This article provides insights about how to conduct power analysis in HLM studies and introduce ways to increase flexibility in power analysis previously mentioned pre-packaged software are lacking. Some reader familiarity with the basics of power analysis in a linear models framework is assumed; readers are referred to Cohen (Cohen, 1988, 1992) for a review of the fundamentals. A general strategy is put forth for performing power analysis in HLMs and the calculation of the covariance matrix of parameter estimators for models of various complexities, which is the critical component to calculate power, is illustrated. Also illustrated is how to use the equations derived to perform power calculations using *R*, although they could be performed in any software that performs matrix calculations. The goal is to provide a flexible and general approach that can be used for different scenarios, many of which may not be implemented in existing software.

### Review of Power Analysis

Performing a power analysis involves calculating standard errors for estimators of parameters of interest. Once armed with an effect size and a standard error, a researcher can produce a test statistic that may then be compared against a chi-square, *T*, or *F* distribution (Cohen, 1998) to estimate approximate power. This paper focuses on the process of appropriately obtaining the standard error of a parameter estimator in HLM, which is the square root of the variance estimate of the parameter estimator. The actual power calculation using an assumed effect size and standard error is shown in numeric examples.

## Statistical Power in HLM

The process of power analysis for multilevel models differs depending on whether one wishes to calculate power for a continuous variable or a dichotomous variable. This article will show that the dichotomous case is much simpler. In fact, an explicit analytical result is derived; however, the starting point is the more general continuous case. In the continuous case, the variance of the parameter estimator of interest depends on the sample data which researchers may not have when they conduct their power analysis. Therefore, additional information about unknown sample data must be assumed. In addition, the inclusion of covariates as well as whether the model contains random slopes will impact the power analysis. Although analytical solutions could be derived for some special cases, slightly different models could end up with very different analytical forms. Therefore a general numerical approach that may be used with a variety of models will be illustrated.

The goal is to calculate a test statistic, whose approximate distribution is known, that can be used to estimate the power of a statistical test of a parameter. Given certain assumptions regarding the model, parameter values, and sample data, the variance-covariance matrix of all the fixed effect parameters in the model can be approximated. This implies that the power to detect any fixed effect can be easily calculated. Furthermore, it will be demonstrated that the power to detect functions of parameters (e.g., contrasts) can also be calculated once the variance-covariance matrix is obtained. For maximum generality matrix notation is used to describe the model.

## Power Analysis for Two-Level Models with Continuous Variables

According to the Gauss-Markov theorem, when errors are independently identically normally distributed with mean of zero and a constant variance in a simple linear regression model, the ordinary least squares estimator (OLSE) is the best linear unbiased estimator (BLUE; Hayashi, 2000). However, the assumption of independently identically distributed (i.i.d.) errors is not realistic for multilevel data. The variance-covariance matrix of random errors in response variables can be assumed to be  $\sigma^2\Omega$  as opposed to  $\sigma^2I$ , where  $I$  is an identity matrix according to the conditions specified in the Gauss-Markov theorem. As a result, the OLSE can be generalized to obtain a generalized least squares estimator

(GLSE)  $(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ . Note that when  $\Omega = I$ , GLSE is OLSE. Under the assumption that  $\sigma^2\Omega$  is specified correctly, the GLSE is also BLUE (Aitken, 1935). Suppose a researcher conducts a study in which she enrolls  $J$  groups of participants and each group consists of  $n$  individuals. There are all together  $m$  level-one predictors. The level-one equation in matrix form is  $Y = X\beta + e$ , where  $Y$  is a  $nJ \times 1$  vector,  $X$  is a  $nJ \times mJ$  diagonal block matrix,  $\beta$  is  $mJ \times 1$ , and  $e$  is a  $nJ \times 1$  vector.

The level-one equation in matrix form is

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{n1} \\ y_{12} \\ \vdots \\ y_{nJ} \end{pmatrix} = \begin{pmatrix} x_{111} & x_{112} & \cdots & x_{11m} & & & & & \\ x_{211} & x_{212} & \cdots & x_{21m} & & & & & \\ \vdots & \vdots & \ddots & \vdots & & & & & \\ x_{n11} & x_{n12} & \cdots & x_{n1m} & & & & & \\ & & & & x_{121} & x_{122} & \cdots & x_{12m} & \\ & & & & x_{221} & x_{222} & \cdots & x_{22m} & \\ & & & & \vdots & \vdots & \ddots & \vdots & \\ & & & & x_{n21} & x_{n22} & \cdots & x_{n2m} & \\ & & & & & & & & \ddots & \\ & & & & & & & & & x_{1J1} & x_{1J2} & \cdots & x_{1Jm} \\ & & & & & & & & & x_{2J1} & x_{2J2} & \cdots & x_{2Jm} \\ & & & & & & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & & & & & & & & & & x_{nJ1} & x_{nJ2} & \cdots & x_{nJm} \end{pmatrix} \begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1m} \\ \beta_{21} \\ \vdots \\ \beta_{Jm} \end{pmatrix} + \begin{pmatrix} e_{11} \\ \vdots \\ e_{n1} \\ e_{12} \\ \vdots \\ e_{nJ} \end{pmatrix}$$

The elements of one column in each block in  $X$  are all 1 if the level-one model has an intercept. The intercept can also be considered as a slope when  $x$  is always equal to 1; no further distinction will be given between intercepts and slopes in the remainder of this article.

The level-two equation is

$$\begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1m} \\ \beta_{21} \\ \vdots \\ \beta_{jm} \end{pmatrix} = \begin{pmatrix} z_{111} & \cdots & z_{11p} & \mathbf{0} & \cdots & \mathbf{0} \\ & \mathbf{0} & & z_{211} & \cdots & z_{21p} & \cdots & \mathbf{0} \\ & \vdots & & \vdots & & \ddots & & \vdots \\ & \mathbf{0} & & \cdots & & \mathbf{0} & z_{m11} & \cdots & z_{m1p} \\ z_{121} & \cdots & z_{12p} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ & \mathbf{0} & & z_{221} & \cdots & z_{22p} & \cdots & \mathbf{0} \\ & \vdots & & \vdots & & \ddots & & \vdots \\ & \mathbf{0} & & \cdots & & \mathbf{0} & z_{m21} & \cdots & z_{m2p} \\ & \vdots & & \vdots & & \ddots & & \vdots \\ z_{1J1} & \cdots & z_{1Jp} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ & \mathbf{0} & & z_{2J1} & \cdots & z_{2Jp} & \cdots & \mathbf{0} \\ & \vdots & & \vdots & & \ddots & & \vdots \\ & \mathbf{0} & & \cdots & & \mathbf{0} & z_{mJ1} & \cdots & z_{mJp} \end{pmatrix} \begin{pmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1p} \\ \gamma_{21} \\ \vdots \\ \gamma_{mp} \end{pmatrix} + \begin{pmatrix} u_{11} \\ \vdots \\ u_{1m} \\ u_{21} \\ \vdots \\ u_{jm} \end{pmatrix}$$

where  $z_{ijk}$  indicates the  $k^{\text{th}}$  second level predictor that is for the  $j^{\text{th}}$  group and has an effect on  $i^{\text{th}}$  level-one variable.

The size of the level-two predictor design matrix  $Z$  is  $mJ * mp$ . The size of the level-two parameter vector  $\gamma$  is  $mp * 1$ , and the random slope  $U$  is a  $mj * 1$  vector. Note that the above equation assumes that all level two predictors have effects on all  $\beta$ 's. In practice, the design matrix of level two predictors should be constructed according to the actual model of interest. Also, researchers may specify some level-one parameters to have random effects. A level-two HLM equation can also be written in the following fashion:

$$Y = X(Z\gamma + U) + e \quad (1)$$

By distributing  $X$  :

$$Y = XZ\gamma + XU + e \quad (2)$$

Because  $\gamma$  can be considered as a vector of fixed effects, only  $(XU + e)$  is random in  $Y$ .

$$\begin{aligned} V &= Var(Y) = Var(XU + e) \\ &= X\tau_k^2 X' + \pi^2 I = \sigma^2 \Omega \end{aligned} \quad (3)$$

As can be observed, the variance components are divided into multiple parts, and the number depends on how many level-one predictors have random effects. Directly following generalized linear model theory, results in

$$Var(\hat{\gamma}) = \left( (XZ)' (\sigma^2 \Omega)^{-1} (XZ) \right)^{-1} \quad (4)$$

(Snijders & Bosker, 1993). Using a conclusion from De Leeuw and Kreft (1986, p. 25) that

$$(\sigma^2\Omega)^{-1} = \begin{pmatrix} \pi^{-2}I - \pi^{-2}X(X'X)^{-1}X' \\ +X(X'X)^{-1}(\pi^2(X'X)^{-1} + \tau^2)^{-1}(X'X)^{-1}X' \end{pmatrix} \quad (5)$$

As a result, if the  $\sigma^2, \tau^2, X, Z$ , and  $V$  matrices are known, the variance-covariance matrix of  $\hat{\gamma}$  can be calculated. If this information and the assumed effect size of  $\hat{\gamma}$  are determined before conducting power analysis, power can be estimated using the  $\chi^2$  distribution. However, one problem remains: Prospective power analysis takes place before the study begins, so some of the needed information may be unavailable. In order to proceed with the power calculation, it is necessary to make some assumptions about the values of level one and level two covariates. One obvious option is to gather information on the distribution of the covariates, draw random variates from the distributions, and use that information in the calculations.

A general strategy to estimate power for the effects of covariates in two-level HLM is now presented; the procedure is as follows: First, assume values for the following: the effect size for the parameter of interest, the level one residual variance  $\pi^2$ , the level-two random effects' variance-covariance matrix  $\tau^2$ , means and variance-covariance matrix of the level-one covariates,  $X$ , and the means of the level-two covariates,  $Z$ . Second, write down their specific models in matrix form and get detailed expressions for  $X, Z$ , and  $\Omega$ . Third, perform the matrix calculation and describe and obtain an estimation of the variance-covariance matrix of all the fixed effects' parameter estimators. Finally, the assumed effect size and the variance of the parameter estimator of interest can be used to construct a  $\chi^2$  statistic to obtain the estimated power.

### **Example Power Analysis for Two-Level Model Where Covariate Values are Known**

Consider an example of a growth model with ten time points, a random intercept and a random slope for time. The model may be written as:

$$\begin{aligned}
 y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + \varepsilon_{ij} \\
 \beta_{0j} &= \gamma_{00} + \gamma_{01}z_j + u_{0j} \\
 \beta_{1j} &= \gamma_{10} + \gamma_{11}z_j + u_{1j} \\
 y_{ij} &= \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}z_j + u_{0j} + u_{1j}t_{ij} + \varepsilon_{ij} \\
 u_j &\sim N(0, \tau^2), \varepsilon_{ij} \sim N(0, \sigma^2)
 \end{aligned} \tag{6}$$

When considering the values of the covariates  $x_{ij}$  and  $z_j$ , researchers may face two situations. One is that  $x_{ij}$  and  $z_j$  are completely or partially unknown prior to data collection. In this case, to conduct power analysis, the researcher will have to assume the first and second moments of the covariates. The second situation is that  $x_{ij}$  and  $z_j$  are known. For example, if  $z_j$  represents different levels of treatment, the number of levels and the number of individuals assigned to each is known in advance of data collection. For this example, assume that  $x_{ij}$  represents the coding of ten equally-spaced time points,  $z_j$  represents five levels of treatment, and the model assumes linear effects of  $x_{ij}$  and  $z_j$ .

### Step one: Assume necessary values.

Assume that the effect size,  $\delta$ , of  $\gamma_{01}$  is 1.0, The level-one error variance,  $\pi^2$ , is 10. The variance-covariance matrix of the level-two random components  $u_j$  is  $\begin{pmatrix} 5.0 & 1.0 \\ 1.0 & 4.0 \end{pmatrix}$ . The number of clusters,  $j$ , is 50. The number of repeated measures per cluster,  $n$ , is 10. The total sample size is 500; input these values into *R* by creating variables to hold them.

```

pisq <- 10
tausq <- array(c(5,1,1,4), dim=c(2,2))
delta <- 1
n <- 10
j <- 50

```

### Step two: Write out matrix forms of $X$ and $Z$

The matrix format of the reduced form equation is:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{n1} \\ y_{12} \\ \vdots \\ y_{nJ} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & & \\ 1 & x_{21} & & & \\ \vdots & \vdots & \mathbf{0} & \cdots & \mathbf{0} \\ 1 & x_{n1} & & & \\ & & 1 & x_{12} & \\ & & 1 & x_{22} & \cdots & \mathbf{0} \\ \mathbf{0} & & \vdots & \vdots & & \\ & & 1 & x_{n2} & & \\ \vdots & & \vdots & \ddots & \vdots & \\ & & & & 1 & x_{1J} \\ & & & & 1 & x_{2J} \\ \mathbf{0} & \cdots & \mathbf{0} & & \vdots & \\ & & & & 1 & x_{nJ} \end{pmatrix} \left[ \begin{pmatrix} 1 & z_1 & \mathbf{0} \\ \mathbf{0} & 1 & z_1 \\ 1 & z_2 & \mathbf{0} \\ \mathbf{0} & 1 & z_2 \\ \vdots & \vdots & \\ 1 & z_j & \mathbf{0} \\ \mathbf{0} & 1 & z_j \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix} + \begin{pmatrix} u_{01} \\ u_{11} \\ u_{02} \\ u_{12} \\ \vdots \\ u_{0J} \\ u_{1J} \end{pmatrix} \right] + \begin{pmatrix} e_{11} \\ \vdots \\ e_{n1} \\ e_{12} \\ \vdots \\ e_{nJ} \end{pmatrix} \quad (7)$$

In this case,  $X$  is a block diagonal matrix. Each block contains a vector of ones to define the intercept and a second vector coding the time points according to the model.  $Z$  is stack of block diagonal matrices. Within each submatrix, the first row describes how the level-one intercept is a function of the level-two parameters,  $\gamma_{00}, \gamma_{01}, \gamma_{10}$ , and  $\gamma_{11}$ . The second row describes how the slope for the time parameters is a function of the same parameters. These matrices are specified in  $R$  by creating two operation matrices,  $A$  and  $B$ , and then their Kronecker product is calculated to obtain  $X$ .  $A$  is a  $j$  by  $j$  identity matrix and  $B$  is an  $n$  by 2 matrix containing a column vector of ones and column vector containing the coding of time.

```
A <- diag(j)
B <- array(c(1,1,1,1,1,1,1,1,1,1,1,2,3,4,5,6,7,8,9,10), dim=c(10,2))
X <- kronecker(A,B)
```

The  $Z$  matrix is created by the following code:

```
Zmean1 <- 1
Zmean2 <- 2
Zmean3 <- 3
Zmean4 <- 4
Zmean5 <- 5
B1 <- matrix(data=c(1,0,Zmean1,0,0,1,0,Zmean1), nrow=2, ncol=4)
A1 <- array(1, dim=c(j/5,1))
Z1 <- kronecker(A1,B1)
B2 <- matrix(data=c(1,0,Zmean2,0,0,1,0,Zmean2), nrow=2, ncol=4)
A2 <- array(1, dim=c(j/5,1))
Z2 <- kronecker(A2,B2)
B3 <- matrix(data=c(1,0,Zmean3,0,0,1,0,Zmean3), nrow=2, ncol=4)
A3 <- array(1, dim=c(j/5,1))
Z3 <- kronecker(A3,B3)
```



```

B4 <- matrix(data=c(1,0,Zmean4,0,0,1,0,Zmean4), nrow=2, ncol=4)
A4 <- array(1, dim=c(j/5,1))
Z4 <- kronecker(A4,B4)
B5 <- matrix(data=c(1,0,Zmean5,0,0,1,0,Zmean5), nrow=2, ncol=4)
A5 <- array(1, dim=c(j/5,1))
Z5 <- kronecker(A5,B5)
Z <- rbind(Z1,Z2,Z3,Z4,Z5)

```

In this example, a balanced design is assumed because there are equal numbers of time points assigned to each individual and equal number of individuals assigned to each level of treatment. However, researchers can conduct power analysis for unbalanced designs using this method by assigning varying numbers of time points to individuals or varying numbers of individuals across levels of treatment.

### Step three: Obtain the approximate variance-covariance matrix

In order to simplify the syntax for calculating  $(\sigma^2 * \Omega)^{-1}$ , as shown in Equation 5, pre-define the identity matrix  $I$  and perform a calculation to obtain  $\psi$ , the block-diagonal matrix with  $J$  blocks of the 2 by 2  $\tau^2$  matrix of variance components:

```

I <- diag(n*j)
I1 <- diag(j)
psi <- kronecker(I1,tausq)

```

Now  $(\sigma^2 * \Omega)^{-1}$  can be obtained using the following code. The  $a$ ,  $b$ ,  $c$ , and  $d$  matrices correspond with the components of Equation 5. Use the solve command to perform matrix inversion.

```

a <- (pisq^-1)*I
b <- (pisq^-1)*(X %*%solve(t(X) %*% X)%*% t(X))
c <- (pisq)*(solve(t(X)%*%X))+ psi
d <- X%*%(solve(t(X)%*%X))%*%(solve(c))%*%(solve(t(X)%*%X))%*%t(X)
OmegaInv <- (a-b+d)

```

With this information, obtain the variance-covariance matrix of the parameter estimates using Equation 4.

```

e <- t(Z) %*% t(X) %*% OmegaInv %*% X %*% Z
Var_gamma <- solve(e)

```

After the covariance matrix has been obtained, the power estimate may be calculated by using the chi-square approximation. In the following code, compute

the non-centrality parameter  $Z1$  by dividing the squared effect size by the relevant element from the covariance matrix and then obtaining the probability from the chi-square distribution with 1 degree of freedom. Interest lies in the power to detect the effect of  $\gamma_{01}$ , the parameter describing the effect of  $z_j$  on the outcome, which is the second of the four fixed effects. Its variance is represented by the (2, 2) entry of the variance-covariance matrix. It turns out the power to detect  $\gamma_{01}$  for an effect size of 1.0 under all the above assumptions is about 0.9. The following *R* code produces the power estimate:

```
Z1 <- (delta^2)/Var_gamma[2,2]
pchisq(3.841459, 1, Z1, lower.tail=FALSE)
```

## Example Power Analysis for Two-Level Model where Covariate Values are Unknown

In the previous example, the values of the level-one and level-two covariates were known prior to data collection. The level-one covariate  $x$  in the growth model represented ten time points while the level-two covariate  $z$  represented five levels of treatment. Because the values were known, the  $X$  and  $Z$  design matrices could be constructed with the known values. However, in many cases the values of covariates are unknown prior to data collection. In this situation researchers will need to assume values for the means, variances, and covariances of the covariates in  $X$  and  $Z$ . The design matrices may then be constructed with values obtained from taking random draws from the appropriate univariate or multivariate distributions. In this section, power analysis for the model considered in the first example will be performed, but this time  $X$  and  $Z$  will contain continuous covariates with unknown values.

Step one: Assume necessary values. The assumed values for all model parameters will be the same as the previous example, except  $x \sim N(0,1)$  and  $z \sim N(0,1)$ . The effect size of  $\gamma_{01}$  is 1.0, the level-one error variance,  $\pi^2$ , is 10. The variance-covariance matrix of the level-two random components  $u_j$  is  $\begin{pmatrix} 5.0 & 1.0 \\ 1.0 & 4.0 \end{pmatrix}$  as before. The number of clusters,  $j$ , is 50. The number of repeated measures per cluster,  $n$ , is 10. The total sample size is 500. The *R* code is identical to that provided for the first example, with the exception of the creation of the  $X$  and  $Z$  matrices. New variables, however, will be added to hold the means and standard deviations of the covariates. If the model included more than

one covariate in  $X$  or  $Z$ , additional variables would be needed to contain their pairwise covariances.

```
meanx <- 0
sdx <- 1
meanz <- 0
sdz <- 1
```

### Step two: Write out matrix forms of $X$ and $Z$

Readers are referred to Equations 6 and 7 for scalar and matrix representations of the model. The status of covariates as known or unknown does not affect the representation of the model. The issue is the creation of  $X$  and  $Z$  with randomly drawn values. The following code will perform this task:

```
library(Matrix)
B <- list()
set.seed(1234)
for (i in 1:j) {
  Bx1 <- rep(1, times=n)
  Bx2 <- rnorm(n, mean=meanx, sd=sdx)
  B[[i]] <- cbind(Bx1, Bx2)
}
```

This code loads the Matrix library and defines the object  $B$  as a list. A loop creates a design matrix for each  $j$  by creating a vector of ones to code the intercept and then making  $n$  draws from the normal distribution to determine plausible values in  $X$ . These blocks are stored in objects named  $B[[1]]$  to  $B[[j]]$ . Now these must be assembled into the overall design matrix  $X$  which has a block-diagonal structure as shown in Equation 7. The random number seed ensures that repeated runs of the code will produce identical pseudo-random draws for  $x$ .

```
C <- list()
for (i in 1:j) {
  if (i == 1) {C[[i]] <- B[[1]]}
  else {C[[i]] <- bdiag(C[[i-1]], B[[i]])}
}
X <- C[[j]]
```

This code assembles the  $X$  matrix by adding one block at a time using the `bdiag` command from the Matrix package. A similar procedure will be used to create the  $Z$  matrix.

```
D <- list()
set.seed(4321)
```

## A FLEXIBLE METHOD FOR CONDUCTING POWER ANALYSIS

```

for (i in 1:j) {
  zj <- rnorm(1, mean=meanz, sd=sdz)
  Dz1 <- c(1, zj, 0, 0)
  Dz2 <- c(0, 0, 1, zj)
  D[[i]] <- rbind(Dz1, Dz2)
}
E <- list()
for (i in 1:j) {
  if (i == 1) {E[[i]] <- D[[1]]}
  else {E[[i]] <- rbind(E[[i-1]], D[[i]])}
}
Z <- E[[j]]

```

The first loop creates  $j$  design matrices, stored in objects  $D[[1]]$  to  $D[[j]]$ . Because the both the level-one intercept and slope are regressed on the same the same variable, a single draw for  $z$  is used in both rows of the “ $D$ ” matrix. The second loop binds all  $j$  matrices together into the complete  $Z$ . A different random number seed should be specified here so the random draws that provide values for  $z$  are not identical to the first  $j$  draws of  $x$ .

### Step three: Obtain the approximate variance-covariance matrix

After  $X$  and  $Z$  are specified the variance-covariance matrix of fixed effects parameter estimates may be obtained using the same code used in the previous example. The (2, 2) entry of this matrix provides the approximate variance of the fixed effect  $\gamma_{01}$ . The following code performs this calculation:

```

I <- diag(n*j)
I1 <- diag(j)
psi <- kronecker(I1,tausq)
a <- (pisq^-1)*I
b <- (pisq^-1)*(X %>%solve(t(X) %>% X)%% t(X))
c <- (pisq)*(solve(t(X)%%X))+ psi
d <- X%%(solve(t(X)%%X))%%(solve(c))%%(solve(t(X)%%X))%*t(X)
OmegaInv <- (a-b+d)
e <- t(Z) %>% t(X) %>% OmegaInv %>% X %>% Z
Var_gamma <- solve(e)

```

If the assumed effect size is 1.0, then the power estimate is obtained by the following code:

```

Z1 <- (delta^2)/Var_gamma[2,2]
pchisq(3.841459, 1, Z1, lower.tail=FALSE)

```

The power estimate is about 0.60. It is important to note that when this approach is used there may be considerable sampling variation across runs in the

draws of  $x$  and  $z$ . The amount of sampling variability in  $x$  is much smaller than in  $z$  because there are  $nj$   $x$ 's but only  $j$   $z$ 's. This may lead to some between-run variability in the power estimate. It is recommended that researchers run the program several times with different random number seeds and average the power estimates across runs.

## Power Analysis for Three-Level Models with Continuous Variables

Next is an outline of how to perform power analysis for a three-level model using the same method. First a general matrix formulation of a three-level HLM is provided.

$$Y = X[Z(W\omega + V) + U] + e \quad (8)$$

$$Y = XZW\omega + XZV + XU + e \quad (9)$$

Only  $(XZV + XU + e)$  is random in  $Y$ .

$$\begin{aligned} Var(Y) &= Var(XZV + XU + e) \\ &= XZv^2Z'X' + X\tau^2X' + \pi^2I = \sigma^2\Omega \end{aligned} \quad (10)$$

In Equation 10,  $v^2$  is the variance-covariance matrix of level-three random components; the remaining terms are defined as previously. The variance-covariance matrix of  $\hat{\gamma}$  can be calculated using:

$$Var(\hat{\gamma}) = \left( (XZW)'(\sigma^2\Omega)^{-1}(XZW) \right)^{-1} \quad (11)$$

Through simple derivation:

$$(\sigma^2\Omega)^{-1} = \begin{pmatrix} \pi^{-2}I - \pi^{-2}X(X'X)^{-1}X' \\ + X(X'X)^{-1}[\pi^2(X'X)^{-1} + Zv^2Z' + \tau^2]^{-1}(X'X)^{-1}X' \end{pmatrix} \quad (12)$$

As the following example will illustrate, the remainder of the procedure for the power analysis in a three-level HLM follows the same logic as that in a two-level HLM.

### Example Power Analysis for Three-Level Model

An example is provided to perform power analysis for a simple three-level model.

$$\begin{aligned}
 y_{ij} &= \beta_{0jk} + \varepsilon_{ij} \\
 \beta_{0jk} &= \gamma_{00} + u_{0jk} \\
 \gamma_{00k} &= \omega_{000} + \omega_{001}\omega_{1k} + v_{00k} \\
 y_{ijk} &= \omega_{000} + \omega_{001}\omega_{1k} + u_{0jk} + v_{00k} + \varepsilon_{ijk} \\
 v_{00k} &\sim N(0, v_{00}^2), u_{0jk} \sim N(0, \tau_{00}^2), \varepsilon_{ij} \sim N(0, \sigma^2)
 \end{aligned} \tag{13}$$

The model could represent students clustered within classrooms and classrooms clustered within schools. The model contains two fixed effects, a grand-mean intercept and a single level-three covariate, presumed to be continuous,  $w_{1k}$ . Like the previous example, assume that the levels of  $w_{1k}$  are known prior to data collection. Sample code is provided only where there are marked differences from the previous example.

#### Step one: Assume necessary values

Assume that the effect size of  $\omega_{001}$  is .20. The outcome is standardized with a total variance,  $\sigma^2$ , of 1.0. The within-cluster variance,  $\pi^2$  is .80. The level-two variance,  $\tau_{00}^2$  is .10. The level-three variance,  $v_{00}^2$ , is also .10. The number of level-two units per level-three unit,  $J$ , is 5. The number of level-three units,  $K$ , is 30. The number of individuals per level-two unit,  $n$ , is 10, yielding a total sample size of 1,500.

### Step two: $X$ , $Z$ , and $W$

The structures of  $X$ ,  $Z$ , and  $W$ , based on the model equations, follow similar logic to the previous example. They are structured as follows:

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{nJK} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & 1 & & \\ 1 & 1 & & \\ 0 & \vdots & \cdots & 0 \\ & 1 & & \\ \vdots & \vdots & \ddots & \vdots \\ & & & 1 \\ 0 & \cdots & 0 & \vdots \\ & & & 1 \end{pmatrix} \begin{bmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & 1 & & \\ 1 & 1 & & \\ 0 & \vdots & \cdots & 0 \\ & 1 & & \\ \vdots & \vdots & \ddots & \vdots \\ & & & 1 \\ 0 & \cdots & 0 & \vdots \\ & & & 1 \end{pmatrix} \left[ \begin{pmatrix} 1 & w_{11} \\ \vdots & \vdots \\ 1 & w_{1k} \end{pmatrix} \begin{pmatrix} \omega_{000} \\ \omega_{001} \end{pmatrix} + \begin{pmatrix} v_{001} \\ \vdots \\ v_{00K} \end{pmatrix} \right] + \begin{pmatrix} u_{011} \\ u_{021} \\ \vdots \\ u_{0JK} \end{pmatrix} \end{bmatrix} + \begin{pmatrix} e_{111} \\ \vdots \\ e_{nJK} \end{pmatrix} \quad (14)$$

### Step three: Obtain the approximate variance-covariance matrix

In order to calculate  $(\sigma^2 * \Omega)^{-1}$ , as shown in Equation 12, use the code:

```
I <- diag(n*j*k)
a <- (pisq^-1)*I
b <- (pisq^-1)*(X %>%solve(t(X) %>% X)%% t(X))
c <- (pisq)*(solve(t(X)%%X))+(Z%>%Tausqv%>%t(Z)+Tausqv)
d <- X%%(solve(t(X)%%X))%%(solve(c))%%(solve(t(X)%%X))%t(X)
OmegaInv <- (sigmasq*(a-b+d))
```

The variance-covariance matrix of the parameter estimates is obtained using Equation 11.

```
e <- t(W) %>% t(Z) %>% t(X) %>% OmegaInv %>% X %>% Z %>% W
VarW <- solve(e) * sigmasq
```

Now that the covariance matrix is obtained, the power estimate may be calculated by using the chi-square approximation. Interest lies in the power to detect the effect of  $\omega_{001}$ . The final power estimate result is 0.68; the power estimate is obtained using:

```
Z1 <- (delta^2)/VarW[2,2]
pchisq(3.841459, 1, Z1, lower.tail=FALSE)
```

## Power Analysis for Models with Dichotomous Predictors

Raudenbush & Liu (2000) described a simplified method of calculating the power to detect the effect of a dichotomous predictor. For example, this method would conveniently apply to intervention studies with two levels of treatment. Consider the following simple multilevel model with only one dichotomous fixed level-two variable and a random intercept to illustrate some of the issues involved in power calculation. Suppose a researcher is interested in whether an intervention helps participants improve their outcome scores ( $y_{ij}$ ).  $J$  groups are randomly enrolled to have the intervention as the experimental group and  $J$  groups are randomly chosen to be the control group ( $z_{0j}=1$  if  $j^{th}$  group receive the intervention, otherwise the value is 0). There are  $n$  students from each group enrolled in the study. The researcher is interested in estimating the main effect of intervention ( $\gamma_{01}$ ) on participants' outcome scores. The model is:

$$\begin{aligned} y_{ij} &= \beta_0 + \varepsilon_{ij} \\ \beta_0 &= \gamma_{00} + \gamma_{01} * z_{0j} + u_{0j} \\ y_{ij} &= \gamma_{00} + \gamma_{01} * z_{0j} + u_{0j} + \varepsilon_{ij} \\ u_{0j} &\sim N(0, \tau_{00}^2), \varepsilon_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{15}$$

Because  $z_{0j}$  is dichotomous ( $z_{0j} = 0$  for all participants in the control group and  $z_{0j}=1$  for the treatment group), all observations in treatment and control groups can be summed respectively to

$$\bar{y}_{ij \text{ treat}} = \gamma_{00} + \gamma_{01} + \frac{\sum_{j=1}^J n * u_{0j \text{ treat}}}{J * n} + \frac{\sum_{j=1}^J \sum_{i=1}^n \varepsilon_{ij \text{ treat}}}{J * n} \tag{16}$$

$$\bar{y}_{ij \text{ control}} = \gamma_{00} + \frac{\sum_{j=1}^J n * u_{0j \text{ control}}}{J * n} + \frac{\sum_{j=1}^J \sum_{i=1}^n \varepsilon_{ij \text{ control}}}{J * n} \tag{17}$$

Subtracting Equation 17 from Equation 16, it is possible to cancel out  $\gamma_{00}$  and get



$$\bar{y}_{ij \text{ treat}} - \bar{y}_{ij \text{ control}} = \gamma_{00} + \frac{\sum_{j=1}^J n^*(u_{ij \text{ treat}} - u_{ij \text{ control}})}{J} + \frac{\sum_{j=1}^J \sum_{i=1}^n (u_{ij \text{ treat}} - u_{ij \text{ control}})}{J * n} \quad (18)$$

If the expectation of  $\bar{y}_{ij \text{ treat}} - \bar{y}_{ij \text{ control}}$  is taken, all random intercepts and residuals drop out because their expectations are all 0 according to the assumption. Finally this results in

$$E(\bar{y}_{ij \text{ treat}} - \bar{y}_{ij \text{ control}}) = E(\gamma_{01}) = \gamma_{01} \quad (19)$$

The fact that  $z_{01}$  is either 0 or 1, and ordinary assumptions about random slopes and residuals allow a simple unbiased estimator of  $\gamma_{01}$  to be derived. Because observations from treatment and control group are independent of each other, the property:

$$\begin{aligned} Var(\hat{\gamma}_{01}) &= Var(\bar{y}_{ij \text{ treat}} - \bar{y}_{ij \text{ control}}) \\ &= Var(\bar{y}_{ij \text{ treat}}) + Var(\bar{y}_{ij \text{ control}}) \end{aligned} \quad (20)$$

is observed. Therefore, using [Equations 18 and 20](#), results in

$$Var(\hat{\gamma}_{01}) = \left[ \frac{\sum_{j=1}^J n^*(u_{0j \text{ treat}} - u_{0j \text{ control}})}{J} + \frac{\sum_{j=1}^J \sum_{i=1}^n (\varepsilon_{ij \text{ treat}} - \varepsilon_{ij \text{ control}})}{J * n} \right] \quad (21)$$

According to the assumptions,  $Var(u_{0j}) = \tau_{00}^2$  for all  $u_{0j}$  's, and  $Var(\varepsilon_{ij} = \sigma^2)$  for all  $\varepsilon_{ij}$  's, [Equation 21](#) can be expressed as:

$$\begin{aligned}
 Var(\hat{\gamma}_{01}) &= \frac{\sum_{j=1}^J n^* (\tau_{0j}^2_{treat} - \tau_{0j}^2_{control})}{J^2} + \frac{\sum_{j=1}^J \sum_{i=1}^n (\sigma_{ij}^2_{treat} - \sigma_{ij}^2_{control})}{n^2 * J^2} \\
 &= \frac{2 * \left( \tau_{00}^2 + \frac{\sigma^2}{n} \right)}{J}
 \end{aligned} \tag{22}$$

Given the derivation in Equation 22, the non-centrality chi-square distribution parameter of random quantity  $\frac{(\hat{\gamma}_{01} - \gamma_0)^2}{var(\hat{\gamma}_{01})}$  can be estimated, where  $\gamma_0$  is the parameter under the null hypothesis. For A two-level HLM with only a random intercept, the intra-class correlation  $p$  is defined as  $\frac{\tau_{00}^2}{\tau_{00}^2 + \sigma^2}$ . As a result,

$$\lambda \approx \frac{(\hat{\gamma}_{01} - \gamma_0)^2}{var(\hat{\gamma}_{01})} = \frac{\frac{(\hat{\gamma}_{01} - \gamma_0)^2}{\tau_{00}^2 + \sigma^2}}{\frac{2 * \left( \tau_{00}^2 + \frac{\sigma^2}{n} \right)}{J * (\tau_{00}^2 + \sigma^2)}} = \frac{J * \delta^2}{2 * \left( p + \frac{1-p}{n} \right)} \tag{23}$$

Where  $\delta^2 = \frac{(\hat{\gamma}_{01} - \gamma_0)^2}{\tau_{00}^2 + \sigma^2}$ , which is the standardized effect size. The same result is provided in Raudenbush, et al. (2004). Because researchers will need to use results from similar previous studies to obtain an assumed effect size and measurements in various studies may be measured on different scales, it makes sense to consider a standardized response variable. After standardization, the variance of the response variable is 1, which means  $\tau_{00}^2 + \sigma^2 = 1$ .

Therefore,  $\delta = \hat{\gamma}_{01} - \gamma_0$ . Equation 23 can be used to generate the proposed test statistic for power calculation.

## Power Analysis for Functions of Parameters in HLM

Sometimes researchers are interested in the power to detect functions of parameters. For example, if a study considers three levels of treatment, detecting

differences between each pair of treatments may be the primary research question. Because the entire variance-covariance matrix will result for the parameter estimators, the power of detecting linear combinations of parameters can be easily calculated. For example, to calculate power of detecting the effect of  $a\beta_1 + b\beta_2$  for constant scalars  $a$  and  $b$ . It is easy calculate the standard error of the linear combination.

$$S.E.(a\hat{\beta}_1 + b\hat{\beta}_2) = \sqrt{a^2\hat{Var}(\hat{\beta}_1) + b^2\hat{Var}(\hat{\beta}_2) + 2ab\hat{Cov}(\hat{\beta}_1, \hat{\beta}_2)} \quad (24)$$

Then the assumed effect size of the two parameters and the calculated standard error of the linear combination of interested parameter estimators can be used to estimate power. Unfortunately this procedure does not apply to the simplified analytical method for dichotomous predictors because the whole variance-covariance matrix is not obtained. Referring back to the first example of the two-level growth model, if it is desirable to conduct power analysis for when  $\gamma_{00} + \gamma_{01}z_j$  when  $z_j = 3$  with additional assumption that the estimated effect size of  $\gamma_{00}$  is 0.50, then by substituting each term in Equation 24 by its corresponding numeric value and results in the power to detect the effect of the linear combination is almost 1. The following code illustrates this calculation:

```
Z2 <- (3.5*3.5)/(1.063333+9*.09666667-6*.29)
pchisq(3.841459, 1, Z2, lower.tail=FALSE)
```

In the cases where interest lies in power of detecting a nonlinear function of parameters, the Taylor expansion can be used to obtain an approximation of the variance of the nonlinear function of parameter estimators.

## Discussion

This article outlined a method for approximating power for a wide variety of HLMs. A theoretical foundation for performing power analysis in two- and three-level models was presented. Examples including *R* code were provided, though any software that can carry out matrix computations and generate random variates may be used. This approach is very flexible and can easily be carried out for models whose power cannot be estimated (or is inconvenient to estimate) using currently available software. The method outlined can perform power analysis for three-level models with many different types of covariates.

## A FLEXIBLE METHOD FOR CONDUCTING POWER ANALYSIS

One limitation of the usability of this approach for applied researchers is the requirement of writing out their models in matrix form. This may be unfamiliar to many researchers and could prevent widespread adoption of this method. In the future, it is hoped that software will be created to automate this process to simplify the implementation of this method and to broaden its appeal.

A second limitation of the approach is the sensitivity to sampling variation when covariate values are unknown. This approach may be thought of as a hybrid of numerical approximation and simulation. The sensitivity increases as the projected sample sizes decrease, which becomes more severe at higher levels of the model. To obtain power estimates robust to sampling variability, it will perhaps be necessary to perform many repetitions of the procedure and obtain a power estimate averaged across repetitions. In most software it is easy to automate multiple repetitions of the procedure to produce the desired stability in the power estimate.

A final limitation is that this method calculates power assuming values for all relevant parameters, such as sample sizes and effect sizes. However, when planning studies, researchers are often interested in determining either the sample size required to reach a given level of power or in the minimum detectable effect size given required power and a fixed sample size. In this case it would be set power equal to some value and solve for the parameter of interest (i.e., sample size or minimum detectable effect size). This article did not directly address these scenarios, as focus was placed on the calculation of power given all other parameter values. However, it is easy to repeat the procedure described in this paper multiple times, specifying a range of values for the parameter of interest in order to find the value of the parameter leading to the desired power.

## References

- Aitken, A. (1935). On least squares and linear combinations of observations. In *Proceedings of the Royal Society of Edinburgh*, 55: 42-28. Farmington, Pennsylvania, USA.
- Bosker, R. J., Snijders, T. A. B., & Guldemon, H. (1999). *PINT (Power IN Two-level Designs): Estimating Standard Errors of Regression Coefficients in Hierarchical Linear Models for Power Calculations: User's Manual*. Retrieved from: [http://www.stats.ox.ac.uk/~snijders/Pint21\\_UsersManual.pdf](http://www.stats.ox.ac.uk/~snijders/Pint21_UsersManual.pdf).

- Browne, W. J., Golalizadeh, M., & Parker, R. (2009). *A guide to sample size calculations for random effect models via simulation and the mlpowsim software package*. University of Bristol.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1): 155-159.
- Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behavioral Research Methods*, 40(1): 236-249.
- De Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, 11(1): 57-86.
- Hayashi, F. (2000). *Econometrics*. Princeton, NJ: Princeton University Press Princeton.
- Lewis, K. (2006). Statistical power, sample sizes, and software to calculate them easily. *BioScience*, 56(7): 607-612.
- Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2<sup>nd</sup> ed.). Thousand Oakes, CA: SAGE Publications Inc.
- Raudenbush, S. & Liu, X. (2000). Statistical power optimal design for multisite randomized trials. *Psychological Methods*, 5(2): 199-213.
- Raudenbush, S., Spybrook, J., Liu X., Congdon, R. (2004). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software*. <http://hlmssoft.net/od/>.
- Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23(4): 323-355.
- Snijders, T., & Bosker, R., (1993). Standard errors and sample sizes for two-levels research. *Journal of Educational and Behavioral Statistics*, 18(3): 237-259.
- Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America*, 78(2): 126-138.

# Investigating the Feasibility of Using *Mplus* in the Estimation of Growth Mixture Models

**Ming Li**  
University of Maryland  
College Park, MD

**Jeffrey R. Harring**  
University of Maryland  
College Park, MD

**George B. Macready**  
University of Maryland  
College Park, MD

---

Hipp and Bauer (2006) investigated the issues of singularities and local maximum solutions within growth mixture models (GMMs) and made recommendations regarding the use of multiple starting values. Building on their work, this simulation study investigates the feasibility of estimating GMMs within *Mplus* as measured by convergence to proper, but local solutions.

*Keywords:* Local maximum solution, convergence, growth mixture modeling, EM algorithm

---

## Introduction

There continues to be growing interest in applying finite mixture models to established statistical methods with the primary goal of accounting for population heterogeneity in model parameters where group membership is latent. One such hybrid is growth mixture modeling (GMM; Muthén, 2001; Muthén & Shedden, 1999) which combines latent growth modeling for the analysis of repeated measures data and latent class analysis (Muthén, 2004). Though GMMs have the advantage of determining possible presence of latent subpopulations with qualitatively distinct patterns of development over time, like many other mixture model applications, GMMs present particular estimation difficulties such as reaching local rather than global optima and, in the case of mixtures of normal distributions, singularities on the likelihood surface (see, e.g., Böhning, 1999). Estimation algorithms for GMMs will impact the convergence rate to proper, global solutions as alternative strategies will likely interact differently with the likelihood surface. Direct maximization of the loglikelihood using gradient

---

*Ming Li is a fifth year PhD candidate. Email at: [liming@umd.edu](mailto:liming@umd.edu). Dr. Jeffrey R. Harring is Associate Professor. Email at: [harring@umd.edu](mailto:harring@umd.edu). Dr. George B. Macready is Professor. Email at: [macready@umd.edu](mailto:macready@umd.edu). All are part of Measurement, Statistics and Evaluation in the Department of Human Development and Quantitative Methodology.*

methods like Newton-Raphson although cumbersome, can be very efficient especially if the intermediate solution is near the maximum (Hsu, 2011). The expectation-maximization (EM) algorithm (Dempster, Laird & Rubin, 1977), on the other hand, provides an indirect approach to obtain maximum likelihood (ML) estimates and is well-suited for estimating GMMs (Muthén & Shedden, 1999). The EM algorithm is an iterative optimization strategy motivated by configuring the statistical model or method as a missing data problem by considering the conditional distribution of what is missing given what has been observed. However, a known deficit of the algorithm is its relatively slow speed to converge (or lack thereof). Yet, the popularity and usefulness of the EM algorithm for GMM applications stems from its seemingly simple implementation and how reliably it can ascertain global optima through stable, uphill steps. This is the primary estimation strategy used in *Mplus*.

The preponderance of methodological studies investigating GMMs has focused on correct class enumeration and parameter recovery across a variety conditions thought to directly influence the hypothesized mixture of latent growth models. To discriminate between local and global solutions and, in general, to avoid likelihood surface irregularities it has been recommended that multiple sets of starting values be used when estimating parameters for finite mixture models (McLachlan & Peel, 2000; Muthén, 2001). The question arises then of how these starting values should be chosen so that they satisfactorily span the parameter space (Hipp & Bauer, 2006) and at the same time do not skirt too close to its space boundaries where divergence is more likely to occur (McLachlan & Basford, 1988). The default in *Mplus* is to generate 10 random sets of starting values although a number of recent studies have encouraged increasing this number in the face of greater model complexity and minimal class separation (Hamilton, 2009; Kohli, 2010; Tolvanen, 2008). In general, these studies have utilized maximum likelihood estimation vis-à-vis the EM algorithm to fit a particular growth mixture model using the mixture modeling module in *Mplus*. However, only the study by Hipp and Bauer (2006) has attempted to qualify the conditions under which estimation of GMMs fails in terms of computational machinery in this modeling and software context.

Building on this work, the primary objectives of this research project are: (1) to empirically investigate the feasibility of the estimation of GMMs within *Mplus* as measured by convergence to a proper, global solution under increasing model complexity and realistic data analytic conditions; and (2) to provide recommendations to practitioners as to what can be expected from the algorithm when applying these models in practical research settings. Issues are examined

related to combating local solutions and nonconvergence including quality of starting values, random perturbations of those values, the number of sets of those starting values, and manipulated arguments in the mixture module related to the EM algorithm on the fitting of GMMs under increased model complexity.

## Methodology

### GMM Specification

The standard latent growth model can be written as

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

where  $\mathbf{y}_i$  is a  $p \times 1$  vector of observed continuous repeated measures for individual  $i$ , where  $p$  denotes the number of waves of data,  $\boldsymbol{\eta}_i$  is a  $q \times 1$  vector of latent growth factors defining the trajectory where  $q$  is the number of latent growth factors ( $q = 2$  for a linear trajectory with intercept and slope), and where  $\boldsymbol{\varepsilon}_i$  is a  $p \times 1$  vector of time-specific residuals for individual  $i$ , and is typically assumed to be distributed normally,  $\boldsymbol{\varepsilon}_i \sim N_p(\mathbf{0}, \Theta_i)$ . The functional form of the individual trajectories is defined by basis functions (columns of  $\mathbf{\Lambda}$ ) whose elements may be constants or parameters to be estimated. For a linear trajectory with latent intercept and slope factors for  $p$  equally spaced repeated measures,  $\mathbf{\Lambda}$  would be set to  $\mathbf{\Lambda} = (\mathbf{1}, \mathbf{t})$ , where  $\mathbf{1}$  is a  $p$ -dimension vector of ones and  $\mathbf{t} = (0, 1, \dots, p-1)'$ .

The joint density of  $\boldsymbol{\varepsilon}_i$  and  $\boldsymbol{\eta}_i$  is assumed to be multivariate normally distributed as

$$\begin{bmatrix} \boldsymbol{\varepsilon}_i \\ \boldsymbol{\eta}_i \end{bmatrix} \sim MVN \left( \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \Theta_i & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix} \right),$$

where  $\boldsymbol{\alpha}$  is a  $q \times 1$  vector of growth factor means,  $\boldsymbol{\Psi}$  is the  $q \times q$  variance-covariance matrix of the growth factors. When coupled with random effects, the time-specific residuals often follow a simple structure like a mutually independent homogenous error structure (i.e.,  $\Theta = \sigma^2 \mathbf{I}_p$  - used throughout the remainder of the study), although any number of other structures could be specified (see, e.g.,



Jennrich & Schluchter, 1986). Because of the normality assumption for the residuals and growth factors the probability density of  $\mathbf{y}_i$  is also multivariate normal

$$f(\mathbf{y}_i) = \phi[\mathbf{y}_i \mid \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})],$$

where the mean vector and covariance structure follow the latent growth model such that,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\alpha}_k \quad \text{and} \quad \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Lambda}\boldsymbol{\Psi}_k\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_k,$$

and  $\boldsymbol{\theta}$  is the vector of parameters from all model matrices.

Muthén (2001) extended the traditional latent growth model to include finite mixtures by permitting the estimation of  $K$  classes each having its own latent growth model with class-specific parameters. The density of  $\mathbf{y}_i$  would then follow a finite mixture of normal distributions of the form

$$f(\mathbf{y}_i) = \pi_k \phi_k[\mathbf{y}_i \mid \boldsymbol{\mu}_k(\boldsymbol{\theta}_k), \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k)],$$

where  $\pi_k$  is the proportion of observations arising from latent class  $k$ . The model-implied mean vector and covariance matrix of a latent growth model again govern each class distribution (Bauer, 2007):

$$\boldsymbol{\mu}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Lambda}\boldsymbol{\alpha}_k \quad \text{and} \quad \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Lambda}\boldsymbol{\Psi}_k\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_k.$$

The growth factor covariance matrices and residual covariance matrices are often presumed to be invariant over classes (i.e.,  $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$  and  $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$  for all  $k$ ). Thus, the only differences between classes are in the model-implied means of the repeated measures as determined by the class-varying growth factor means,  $\boldsymbol{\alpha}_k$ . As Hipp and Bauer (2006) pointed out, an advantage of making the within-class covariance matrices invariant is that ensures the absence of singularities, and ensures the existence of a global solution.

## The EM Algorithm

The EM algorithm is an iterative optimization strategy for finding ML parameter estimates by reformulating the given incomplete data or missing data problem as a complete-data problem (McLachlan & Krishnan, 2008). The algorithm iterates between two steps – an expectation (E) step and a maximization (M) step and then iteratively repeats this sequence until some convergence criteria is met (see, e.g., Harring, 2012; Liu, 2012; Muthén & Shedden, 1999 for a complete description of the algorithm).

## Simulation Design

In this study two Monte Carlo simulations were conducted to help understand the boundaries under which GMM parameters might be successfully estimated within *Mplus*. The conditions manipulated under Simulation 1, which are displayed in Table 1, include: starting value quality (SVQ), number of random starts (RS), number of final optimizations (FO), perturbation level of the starting values (PL), convergence criterion for the EM algorithm (MCONV), and model complexity (MC). A second smaller simulation study was conducted where the data generation model and the estimation model were identical and only the population values for model parameters were used as starting values. Also, only a three-class model (the correct model) was fitted under Simulation 2. Therefore, the main differences between Simulations 1 and 2 are the model complexities considered and the starting values that were used. Although this second simulation design was thought to be unrealistic in practice (because the true number of mixing distributions is unknown nor are the parameter values), it provided a “best case scenario” from which to compare all other non-optimal conditions.

**Table 1.** Conditions and Levels of Manipulated Factors for Simulation 1

Conditions	Levels
Starting value quality (SVQ)	<i>Mplus</i> default values, LGM <sup>a</sup> and LCGM <sup>b</sup> output values
Number of random starts (RS)	25, 50, 100, 200
Number of final optimizations (FO)	5, 10, 25
Perturbation level (PL)	1, 3, 5, 7, 9
Convergence criterion (MCONV)	1E-5, 1E-8
Model complexity (MC)	2, 3, and 4-class models

**Note:** <sup>a</sup>Refers to latent growth modeling (LGM; Meredith & Tisak, 1990). <sup>b</sup>Refers to latent class growth modeling (LCGM; Nagin, 1999)

The population model follows a three-class linear model following conditions outlined by Tolvanen (2008) and detailed in the Appendix to this paper. Data were generated in *R* software following the two-step procedure outlined by Hipp and Bauer (2006) and all models were fitted using *Mplus* 6.2 (Muthén & Muthén, 2010). Two hundred and fifty replications were run for each of the  $2 \times 4 \times 3 \times 5 \times 2 \times 3 = 720$  cells in Simulation 1 as well as each of the  $4 \times 3 \times 5 \times 2 = 120$  cells in Simulation 2 in a full factorial design.

Starting value quality is defined as initial parameter estimates that were thought to be in the neighborhood of the solution found via ML. Good starting values for the means of the growth parameters are defined coming from a latent class growth analysis (Jones & Nagin, 2007) and covariance parameters coming from fitting a one-class model, or LGM. This is aligned with what is believed as a reasonable approach to fitting GMMs in practice. Poor starting value quality is defined as the *Mplus* default values with no actual values given in the input file. Model complexity is measured by fitting a number of classes differing from the 3-class population model. Thus for each replicate data set 2, 3, and 4-class models were fitted.

Using multiple starting values has been recommended as a method to combat convergence to a local solution prevalent when estimating GMMs. *Mplus* allows the number of initial stage random sets of starting values to vary and 4 levels were examined which are 25, 50, 100, and 200 (the *Mplus* default value is 20 with recommendations for greater number of initial random starts). It was expected that there would be an interaction between model complexity and the necessity to increase multiple starting values. *Mplus* also allows starting values to be perturbed randomly with the magnitude of perturbation controlled by the analyst using the STSCALE command. Five perturbation levels ranging from 1 (small perturbation) to 9 (large perturbation) with the default of 5 were examined. If local solutions are present in the analysis, changing the number of final initial solutions to analyze may impact the ability of the program to conclude that a global maximization had been reached. In terms of the convergence criteria used for the EM algorithm, pilot simulations were run using convergence criteria of 1E-5 (the *Mplus* default criterion), 1E-8, and 1E-10, and results showed no significant mean outcome differences between using 1E-8 and 1E-10. So, two levels of convergence criterion were used: 1E-5 and 1E-8. Equal proportions were assumed across classes and held constant (i.e., 3-class 0.33/0.33/0.33). The sample size was fixed at  $n = 900$ , which is in the range of past GMM simulation studies (Hamilton, 2009).

## Outcomes

A factorial ANOVA was used to examine the influence of the manipulated factors and their combinations on 4 outcome measures averaged over the 250 replications for the two simulations. Outcome 1 (Dlog), the proportion of number of different loglikelihood values to the total number of loglikelihood values, is expected to be low for convergence to a proper solution. Outcome 2 (Logmatch) is the percentage of replications where the highest loglikelihood solution is also the most frequently occurring solution, and is expected to be high for good model convergence. Outcome 3 is measured as the percentage of non-converged solutions defined by negative variances (Negvariance) and/or nonconvergence (Nonconverge), and Outcome 4 (Localmax) is the percentage of local maximum likelihood solutions. Both Outcomes 3 and 4 are expected to be low for convergence to a proper solution.

## Results

To better understand which factors and/or combination of factors impacted model convergence and global optima, a factorial ANOVA was utilized where the four outcome variables were modeled as functions of the manipulated simulation conditions. Results for up to 5-way interactions for Simulation 1 and up to 3-way interactions for Simulation 2 were assessed and are reported separately for each of the outcomes. Only the effects of the manipulated factors were interpreted if they were identified to be both statistically significant ( $p$ -value  $\leq 0.05$ ) and have an effect size of  $\eta^2 \geq 0.06$  (see, e.g., Cohen, 1988, p. 283; Kohli, 2010).

Table 2 below summarizes the significant main and interaction effects for both simulation studies. Obviously, SVQ did not significantly affect the outcome variables under Simulation 1. This result was different from the findings of Jones and Nagin (2007) that using good starting values from the means of the growth parameters coming from a latent class growth analysis helps avoid the local maxima issue. Under Simulation 1, MC had the largest effect on Logmatch ( $\hat{\eta}^2 = .39$ ), Negvariance ( $\hat{\eta}^2 = .91$ ) and Localmax ( $\hat{\eta}^2 = .26$ ). For both simulation studies, Dlog, Logmatch and Localmax were all impacted by the main effects of RS and PL. FO had significant main effects on Dlog ( $\hat{\eta}^2 = .30$ ), Logmatch ( $\hat{\eta}^2 = .10$ ) and Localmax ( $\hat{\eta}^2 = .12$ ) under Simulation 1 but did not show significant effect on Logmatch in Simulation 2. Because population parameters were used as starting values to fit the generated model, it is not surprising that no cases of

nonconvergence were identified in Simulation 2. Finally, MCONV was not a significant factor for either of the two simulation studies. In terms of the interaction effects, only 2 and 3-way interaction effects were recognized for both studies. Significant 2-way interaction effect ( $PL \times MC$ ) on Dlog and Localmax and 3-way interaction effect ( $MC \times PL \times RS$ ) on Nonconverge were found in Simulation 1. For Simulation 2, significant two-way interaction effects were found for  $PL \times FO$  on Dlog, Logmatch and Localmax, for  $PL \times RS$  on Logmatch and Localmax, and for  $RS \times FO$  on Localmax. Significant 3-way interaction effect was obtained on Localmax for  $RS \times FO \times PL$ .

**Table 2.** Proportion of Variance Explained by the Outcome Variables

Reported Effects	Simulation 1					Reported Effects	Simulation 2		
	Dlog	Log-match	Neg-variance	Non-convergence	Localmax		Dlog	Log-match	Local-max
RS	8.9%	7.3%			6.6%	RS	18.9%	8.2%	14.1%
FO	29.5%	9.9%			12.3%	FO	25.2%		6.3%
PL	19.0%	11.9%			10.7%	PL	40.3%	59.7%	10.4%
MC	23.7%	39.2%	90.6%		26.5%				
PL×MC	6.2%				9.3%	PL×FO	6.7%	11.4%	10.0%
MC×PL×RS				6.4%		PL×RS		8.8%	23.1%
						RS×FO			13.8%
						FO×PL×RS			22.2%

## Results for the Main Effects

Tukey's HSD procedure was used for comparing pairs of means for the main effects for both simulation studies. Means for groups in homogeneous subsets are displayed below in Tables 3 through 6. The results presented in Table 3 show that the two simulation studies had the exact same change of directions in Dlog, Logmatch and Localmax values when the level of RS was changed. As RS increased, Dlog values decreased from .329 to .241 (for Simulation 1) and from .299 to .153 (for Simulation 2). Localmax values also decreased from .137 to .051 (for Simulation 1) and from .086 to .002 (for Simulation 2). The values of Logmatch increased from .592 to .786 for Simulation 1 and from .696 to .914 for Simulation 2. In Table 4, for both simulation studies, it was found that as FO increased, the Dlog values also increased in magnitude whereas Localmax values decreased in magnitude. Logmatch values decreased from .779 to .572 as FO

# MPLUS IN GMM ESTIMATION

increased in Simulation 2. The main effect of PL on Dlog, Logmatch, and Localmax is a little more complex. Table 5 shows that for both simulation studies, as PL increased, the Dlog values also increased. The lowest Dlog values of .194 (Simulation 1) and .110 (Simulation 2) were found at level 1 of PL. In terms of the effect of PL on Logmatch and Localmax values, Simulation 1 showed the highest value of .755 at level 2 and 3 of PL and the lowest value of .505 at level 5 of PL. Simulation 1 also had the lowest Localmax value of .052 at level 3 of PL and the highest value of .171 at level 5 of PL. For Simulation 2, increasing PL lead to decreased Logmatch values from .985 to .418 and increased Localmax values from .003 to .081. Comparing pairs of means for the main effect of MC on Dlog, Logmatch, Negvariance, and Localmax for Simulation 1 (see Table 6) suggests that as MC increased, Dlog value also increased from .209 to .370. The highest Localmax value (.177) and the lowest Logmatch value (.548) were both found at the highest level of MC. Intuitively, it seems reasonable to expect that the highest Logmatch value of .923 and the lowest Localmax value of .017 were reached for level 2 of MC (i.e., the 3-class model) because it was the model used for data generation. Results in Table 6 also indicated that Negvariance was greater for the highest MC level than for the other MC levels.

**Table 3.** Pairwise Comparisons among levels of RS for dependent variables: Dlog, Logmatch and Localmax

N	RS	Subset(Dlog)				Subset(Logmatch)				Subset(Localmax)			
		1	2	3	4	1	2	3	4	1	2	3	4
Simulation 1													
180	1				.329	.592							.137
180	2			.294			.644					.102	
180	3		.256					.713			.069		
180	4	.241							.786	.051			
Simulation 2													
30	1				.299	.696							.086
30	2			.249			.805					.008	
30	3		.195	.				.856			.004		
30	4	.153							.914	.002			

**Table 4.** Pairwise Comparisons among levels of FO for dependent variables: Dlog, Logmatch and Localmax

N	FO	Subset(Dlog)			Subset(Logmatch)			Subset(Localmax)		
		1	2	3	1	2	3	1	2	3
Simulation 1										
180	1	.188					.779			.146
180	2		.256			.700			.088	
180	3			.376	.572			.035		
Simulation 2										
40	1	.155			—	—	—			.058
40	2		.208		—	—	—		.015	
40	3			.309	—	—	—	.002		

**Note:** No subset (Logmatch) values under Simulation 2 were provided because no significant main effect on Logmatch was found.

**Table 5.** Pairwise Comparisons among levels of PL for dependent variables: Dlog, Logmatch, and Localmax

N	PL	Subset(Dlog)					Subset(Logmatch)				Subset(Localmax)				
		1	2	3	4	5	1	2	3	4	1	2	3	4	5
Simulation 1															
144	1	.194							.725						.083
144	2		.220							.755		.065			
144	3			.266							.755	.052			
144	4				.326				.679					.079	
144	5					.360	.505								.171
Simulation 2															
24	1	.110								.985	.003				-
24	2		.156								.982	.002			-
24	3			.237								.950		.006	-
24	4				.295				.753					.031	-
24	5					.322	.418								.081

**Note:** No fifth subset of Localmax was provided because only 4 subsets were identified.

**Table 6.** Pairwise Comparisons among levels of MC for dependent variables: Dlog, Logmatch, Negvariance, and Localmax for Simulation 1

N	MC	Subset (Dlog)			Subset (Logmatch)			Subset (Negvariance)		Subset (Localmax)		
		1	2	3	1	2	3	1	2	1	2	3
180	1	.209				.581		.000			.074	
180	2		.240				.923	.000		.017		
180	3			.370	.548				.029			.177

## Results for the Interaction Effects

Simple main effect pairwise comparisons were conducted to investigate the nature of significant two and three-way interaction effects for each of the simulation studies. For those significant interaction effects with a clear pattern, graphics were provided.

The only significant two-way interaction effect found in Simulation 1 was  $PL \times MC$  for dependent variables: Dlog and Localmax. Three significant two-way interaction effects in Simulation 2 were  $PL \times FO$  (for dependent variables: Dlog, Logmatch and Localmax),  $PL \times RS$  (for dependent variables: Logmatch and Localmax), and  $RS \times FO$  (for dependent variable: Localmax). Tables 7 and 8 present simple main effect results related to the  $PL \times MC$  interaction on Dlog and Localmax respectively under Simulation 1.

For Dlog, no significant mean differences were found between levels of MC at level 1 of PL. Significant mean differences were found between level 3 and level 1 and between level 3 and level 2 of MC at level 2, 3 and 4 of PL. Significant mean differences were also found for all pairs of MC levels at level 5 of PL. Further, level 3 of MC always had the highest Dlog values (.238, .296, .357, .460, and .501) across all PL levels compared with the other two MC levels, and Dlog values showed the slowest increase for level 1 of MC starting from level 3 of PL (0.207, 0.217, and 0.233). Clearly, MC is an important factor affecting Dlog measures at all PL levels.

Another important observation is that as PL increased in magnitude, Dlog increased for level 2 and 3 of MC. Dlog started to increase for level 1 of MC from level 3 of PL. Therefore, a high PL level might not be a good choice for a low Dlog solution for any of the models that were considered. In terms of the  $PL \times MC$  interaction effect on Localmax, it may be observed from the results reported



in Table 8 that the lowest Localmax mean values (.008, .004, .005, and .020) were always associated with level 2 of MC for each of the levels of PL, which seems reasonable because level 2 of MC is the 3-class model used for data generation. The lowest Localmax values were found with level 1 and 2 of MC at PL level 3 and level 3 of MC at PL level 2. It may also be noted that Localmax values started to increase markedly in magnitude at and above PL level 4 for each of the levels of MC, particularly with level 3 of MC (.186 at PL level 4 and .349 at PL level 5).

**Table 7.** Simple Main Effect Pairwise Comparisons Corresponding to PL × MC Interaction for Dlog (Simulation 1)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	MC	$\bar{X}_{PL=j, MC=k}$	PL	MC	$\bar{X}_{PL=j', MC=k'}$	
1	1	0.196	1	2	0.149	0.047
1	1	0.196	1	3	<b>0.238</b>	-0.042
1	2	0.149	1	3	0.238	-0.089
2	1	0.193	2	2	0.170	0.023
2	1	0.193	2	3	<b>0.296</b>	-0.103*
2	2	0.170	2	3	0.296	-0.126*
3	1	<b>0.207</b>	3	2	0.234	-0.027
3	1	0.207	3	3	<b>0.357</b>	-0.150*
3	2	0.234	3	3	0.357	-0.123*
4	1	<b>0.217</b>	4	2	0.303	-0.086
4	1	0.217	4	3	<b>0.460</b>	-0.243*
4	2	0.303	4	3	0.460	-0.157*
5	1	<b>0.233</b>	5	2	0.346	-0.113*
5	1	0.233	5	3	<b>0.501</b>	-0.268*
5	2	0.346	5	3	0.501	-0.155*

**Note:** The increased Dlog values for level 1 of MC at level 3, 4, and 5 level of PL and the increased Dlog values for level 3 of MC across levels of PL are in boldface.

\*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

# MPLUS IN GMM ESTIMATION

**Table 8.** Simple Main Effect Pairwise Comparisons Corresponding to PL × MC Interaction for Localmax (Simulation 1)

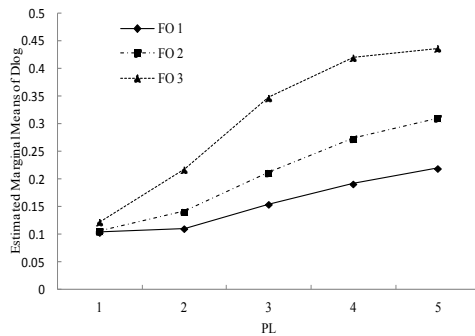
Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	MC	$\bar{X}_{PL=j, MC=k}$	PL	MC	$\bar{X}_{PL=j', MC=k'}$	
1	1	0.118	1	2	0.008	0.110*
1	1	0.118	1	3	0.123	-0.005
1	2	0.008	1	3	0.123	-0.115*
2	1	0.083	2	2	0.004	0.079
2	1	0.083	2	3	0.108	-0.025
2	2	0.004	2	3	0.108	-0.104*
3	1	0.028	3	2	0.005	0.023
3	1	0.028	3	3	0.122	-0.094*
3	2	0.005	3	3	0.122	-0.117*
4	1	<b>0.030</b>	4	2	0.020	0.010
4	1	0.030	4	3	<b>0.186</b>	-0.156*
4	2	<b>0.020</b>	4	3	0.186	-0.166*
5	1	<b>0.112</b>	5	2	0.052	0.060
5	1	0.112	5	3	<b>0.349</b>	-0.237*
5	2	<b>0.052</b>	5	3	0.349	-0.297*

Note: The increased Localmax values for 3 levels of MC at level 4 and level 5 of PL are in boldface.

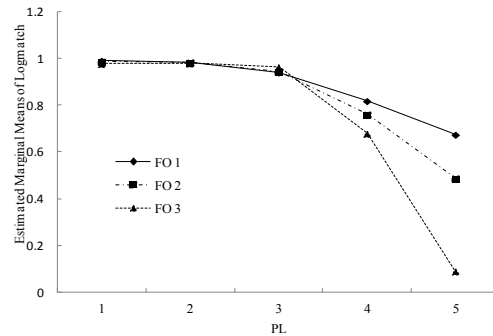
\*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

Graphic presentations of the interaction effect of PL × FO for dependent variables Dlog, Logmatch and Localmax (from Simulation 2) are provided in Figures 1 through 3. It may be observed that as PL increased, Dlog and Localmax both increased whereas Logmatch decreased for all FO levels. The highest Dlog, the lowest Logmatch and the lowest Localmax values were seen with level 3 of FO at higher PL levels. It also may be observed in Figures 2 and 3 that Logmatch and Localmax values were generally stable and similar in magnitude for pairs of FO levels at PL levels 1 and 3. However, starting from level 4 of PL, Logmatch and Localmax values both showed a sudden change, with Logmatch values dropping and Localmax rising sharply. The results reported in Tables 9 through

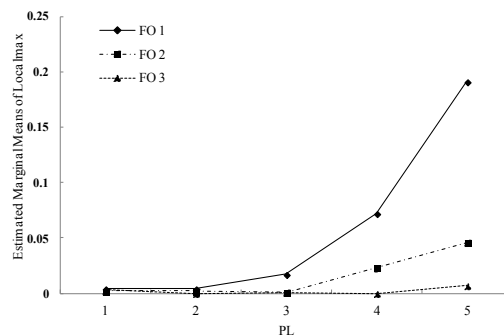
11 show the nature of the interaction between PL and FO on Dlog, Logmatch and Localmax. Significant Dlog mean differences were found between level 1 and level 3 of FO across PL levels 2-5 and there were no significant mean differences between level 1 and level 2 of FO at any PL levels (see Table 9). Tables 10 and 11 both show most significant mean differences in Logmatch and Localmax occurred at level 5 of PL between pairs of FO levels.



**Figure 1.** PL × FO For Outcome 1 (Dlog)



**Figure 2.** PL × FO Outcome 2 (Logmatch)



**Figure 3.** PL × FO Outcome 4 (Localmax)

# MPLUS IN GMM ESTIMATION

**Table 9.** Simple Main Effect Pairwise Comparisons Corresponding to PL × FO Interaction for Dlog (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	FO	$\bar{X}_{PL=j, MC=k}$	PL	FO	$\bar{X}_{PL=j', MC=k'}$	
1	1	0.103	1	2	0.105	-0.002
1	1	0.103	1	3	0.122	-0.019
1	2	0.105	1	3	0.122	-0.017
2	1	0.110	2	2	0.141	-0.031
2	1	0.110	2	3	0.217	-0.107*
2	2	0.141	2	3	0.217	-0.076
3	1	0.154	3	2	0.211	-0.057
3	1	0.154	3	3	0.347	-0.193*
3	2	0.211	3	3	0.347	-0.136*
4	1	0.191	4	2	0.273	-0.082
4	1	0.191	4	3	0.420	-0.229*
4	2	0.273	4	3	0.420	-0.147*
5	1	0.219	5	2	0.310	-0.091
5	1	0.219	5	3	0.437	-0.218*
5	2	0.310	5	3	0.437	-0.127*

**Note:** \*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

**Table 10.** Simple Main Effect Pairwise Comparisons Corresponding to PL  $\times$  FO Interaction for Logmatch (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	FO	$\bar{X}_{PL=j, MC=k}$	PL	FO	$\bar{X}_{PL=j', MC=k'}$	
1	1	0.991	1	2	0.987	0.004
1	1	0.991	1	3	0.978	0.013
1	2	0.987	1	3	0.978	0.009
2	1	0.981	2	2	0.984	-0.003
2	1	0.981	2	3	0.980	0.001
2	2	0.984	2	3	0.980	0.004
3	1	0.941	3	2	0.945	-0.004
3	1	0.941	3	3	0.964	-0.023
3	2	0.945	3	3	0.964	-0.019
4	1	0.819	4	2	0.761	0.058
4	1	0.819	4	3	0.680	0.139*
4	2	0.761	4	3	0.680	0.081
5	1	0.675	5	2	0.488	0.187*
5	1	0.675	5	3	0.091	0.584*
5	2	0.488	5	3	0.091	0.397*

**Note:** \*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

# MPLUS IN GMM ESTIMATION

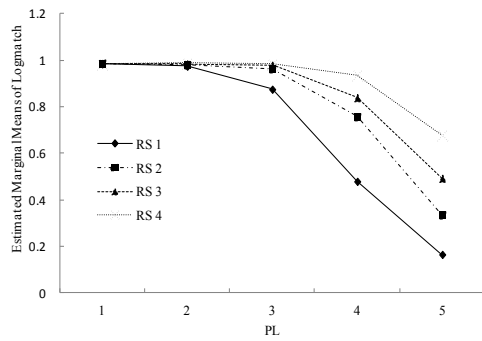
**Table 11.** Simple Main Effect Pairwise Comparisons Corresponding to PL × FO Interaction for Localmax (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	FO	$\bar{X}_{PL=j, MC=k}$	PL	FO	$\bar{X}_{PL=j', MC=k'}$	
1	1	0.004	1	2	0.002	0.002
1	1	0.004	1	3	0.003	0.001
1	2	0.002	1	3	0.003	-0.001
2	1	0.004	2	2	0.002	0.002
2	1	0.004	2	3	0.000	0.004
2	2	0.002	2	3	0.000	0.002
3	1	0.017	3	2	0.001	0.016
3	1	0.017	3	3	0.001	0.016
3	2	0.001	3	3	0.001	0.000
4	1	0.072	4	2	0.023	0.049
4	1	0.072	4	3	0.000	0.072
4	2	0.023	4	3	0.000	0.023
5	1	0.191	5	2	0.046	0.145*
5	1	0.191	5	3	0.007	0.184*
5	2	0.046	5	3	0.007	0.039

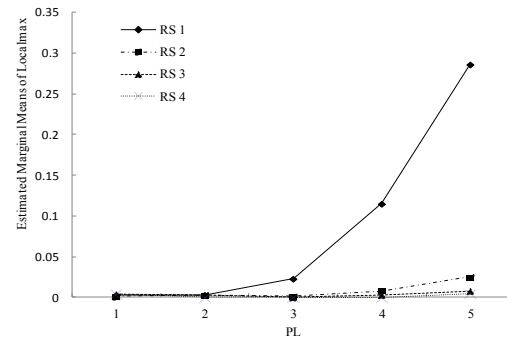
**Note:** \*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

Figures 4 and 5 graphically depict the interaction effect between PL and RS for the dependent variables: Logmatch and Localmax in Simulation 2. At level 1 and level 2 of PL, both Logmatch and Localmax values were very similar across all RS levels. From level 3 of PL, discrepancies in Logmatch and Localmax values among the RS levels started to show up and grow even larger at level 4 and level 5 of PL. It may also be noticed that Logmatch values decreased markedly at PL level 4 for all RS levels, with the sharpest decline observed at level 1 of RS. In contrast, Localmax values increased dramatically at PL level 4 for level 1 of RS. Tables 12 and 13 were provided to confirm what had been observed. For Logmatch, most significant mean differences were found between pairs of RS levels at level 4 and level 5 of PL (see Table 12). Significant Localmax mean

differences were found at levels 4 and 5 of PL between level 1 and level 2, level 1 and level 3, and level 1 and level 4 of RS (see [Table 13](#)), with no significant mean differences found for pairs of levels 2, 3, and 4 of RS.



**Figure 4.** PL × RS for Outcome 2 (Logmatch)



**Figure 5.** PL × RS for Outcome 4 (Localmax)

# MPLUS IN GMM ESTIMATION

**Table 12.** Simple Main Effect Pairwise Comparisons Corresponding to PL × RS Interaction for Logmatch (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	RS	$\bar{X}_{PL=j,RS=k}$	PL	RS	$\bar{X}_{PL=j',RS=k'}$	
1	1	0.984	1	2	0.987	-0.003
1	1	0.984	1	3	0.987	-0.003
1	1	0.984	1	4	0.984	0.000
1	2	0.987	1	3	0.987	0.000
1	2	0.987	1	4	0.984	0.003
1	3	0.987	1	4	0.984	0.003
2	1	0.974	2	2	0.980	-0.006
2	1	0.974	2	3	0.984	-0.010
2	1	0.974	2	4	0.988	-0.014
2	2	0.980	2	3	0.984	-0.004
2	2	0.980	2	4	0.988	-0.008
2	3	0.984	2	4	0.988	-0.004
3	1	<b>0.875</b>	3	2	0.963	-0.088
3	1	0.875	3	3	0.979	-0.104*
3	1	0.875	3	4	0.984	-0.109*
3	2	0.963	3	3	0.979	-0.016
3	2	0.963	3	4	0.984	-0.021
3	3	0.979	3	4	0.984	-0.005
4	1	<b>0.479</b>	4	2	0.757	-0.278*
4	1	0.479	4	3	0.839	-0.360*
4	1	0.479	4	4	0.937	-0.458*
4	2	0.757	4	3	0.839	-0.082
4	2	0.757	4	4	0.937	-0.180*
4	3	0.839	4	4	0.937	-0.098
5	1	<b>0.165</b>	5	2	0.335	-0.170*
5	1	0.165	5	3	0.493	-0.328*
5	1	0.165	5	4	0.679	-0.514*
5	2	0.335	5	3	0.493	-0.158*
5	2	0.335	5	4	0.679	-0.344*

**Note:** The decreased Logmatch values for level 1 of RS at PL levels 3, 4 and 5 are in boldface.

\*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.



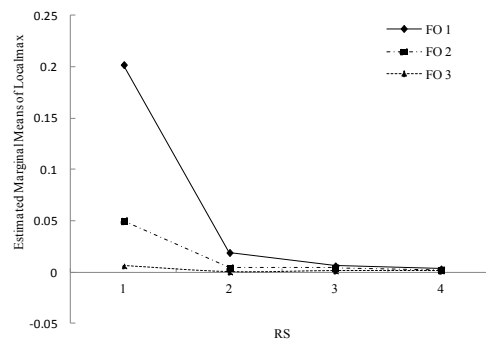
**Table 13.** Simple Main Effect Pairwise Comparisons Corresponding to PL × RS Interaction for Localmax (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	RS	$\bar{X}_{PL=j,RS=k}$	PL	RS	$\bar{X}_{PL=j',RS=k'}$	
1	1	0.003	1	2	0.001	0.002
1	1	0.003	1	3	0.004	-0.001
1	1	0.003	1	4	0.004	-0.001
1	2	0.001	1	3	0.004	-0.003
1	2	0.001	1	4	0.004	-0.003
1	3	0.004	1	4	0.004	0.000
2	1	0.003	2	2	0.003	0.000
2	1	0.003	2	3	0.003	0.000
2	1	0.003	2	4	0.000	0.003
2	2	0.003	2	3	0.003	0.000
2	2	0.003	2	4	0.000	0.003
2	3	0.003	2	4	0.000	0.003
3	1	<b>0.023</b>	3	2	0.001	0.022
3	1	0.023	3	3	0.000	0.023
3	1	0.023	3	4	0.000	0.023
3	2	0.001	3	3	0.001	0.000
3	2	0.001	3	4	0.000	0.001
3	3	0	3	4	0.000	0.000
4	1	<b>0.115</b>	4	2	0.008	0.107*
4	1	0.115	4	3	0.003	0.112*
4	1	0.115	4	4	0.000	0.115*
4	2	0.008	4	3	0.003	0.005
4	2	0.008	4	4	0.000	0.008
4	3	0.003	4	4	0.000	0.003
5	1	<b>0.286</b>	5	2	0.025	0.261*
5	1	0.286	5	3	0.009	0.277*
5	1	0.286	5	4	0.005	0.281*
5	2	0.025	5	3	0.008	0.017
5	2	0.025	5	4	0.005	0.020

**Note:** The increased Localmax values for level 1 of RS at PL level 3, level 4 and level 5 are in boldface.

\*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

Figure 6 shows how RS and FO interacted for the Dependent Variable: Localmax. Obviously, the Localmax mean values were very close in magnitude among levels of FO at level 2, 3 and 4 of RS, and as RS increased, the Localmax mean values became progressively closer in magnitude for all of the levels of FO. Localmax mean differences were clear only at level 1 of RS. Table 14 followed shows significant mean differences between level 1 and level 2 and between level 1 and level 3 of FO at RS level 1. It should also be noted that from RS level 1 to RS level 2, Localmax mean values decreased in magnitude for all FO levels (with the sharpest decrease observed for level 1 of FO), suggesting a high RS is always preferred for a low Localmax for any level of FO. It also suggests that when RS is very low at level 1, a low FO level should be considered for low percentage of Localmax solutions.



**Figure 6.** RS × FO for Outcome 4 (Localmax)

**Table 14.** Simple Main Effect Pairwise Comparisons Corresponding to RS × FO Interaction for Localmax (Simulation 2)

Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
RS	FO	$\bar{X}_{PL=j,FO=k}$	RS	FO	$\bar{X}_{PL=j',FO=k'}$	
1	1	0.202	1	2	0.050	0.152*
1	1	0.202	1	3	0.006	0.196*
1	2	0.050	1	3	0.006	0.044
2	1	0.019	2	2	0.004	0.015
2	1	0.019	2	3	0.000	0.019
2	2	0.004	2	3	0.000	0.004
3	1	0.006	3	2	0.004	0.002
3	1	0.006	3	3	0.001	0.005
3	2	0.004	3	3	0.001	0.003
4	1	0.003	4	2	0.002	0.001
4	1	0.003	4	3	0.001	0.002
4	2	0.002	4	3	0.001	0.001

**Note:** \*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of number of random starts at which the simple main effect tests were performed.

Two three-way interaction effects were found to be significant at  $\alpha = 0.05$ . One of these significant outcomes was found for the dependent variable: Nonconverge (Simulation 1) and the other for the dependent variable: Localmax (Simulation 2). Simple main effect pairwise comparisons corresponding to the two-way interaction effects from PL × RS on Nonconverge were conducted at each of the levels of MC for Simulation 1. No significant mean differences were found for the two-way interaction under either level 1 or level 2 of MC. Therefore, results for the interaction effect of PL × RS on Nonconverge under level 1 and 2 of MC are not reported. Table 15 shows only the interaction effect of PL × RS on Nonconverge under level 3 of MC. It can be seen that all significant mean outcome differences were found at level 5 of PL although there was no significant mean difference between level 3 and level 4 of RS. This findings would seems to suggest that with a complicated model (e.g., the 4-class model), if a high PL level is used, higher levels of RS (level 3 or level 4) should be considered to increase the number of converged solutions.

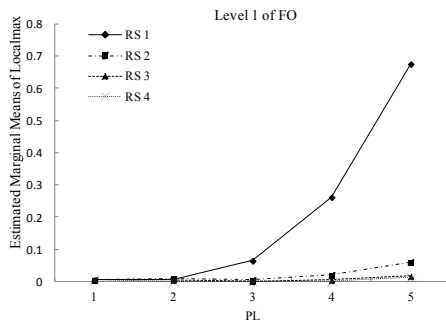
# MPLUS IN GMM ESTIMATION

**Table 15.** Simple Main Effect Pairwise Comparisons Corresponding to PL × RS Interaction for Nonconverge under Level 3 of MC

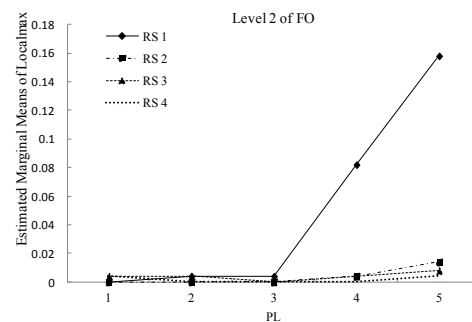
Mean 1 Factor Level			Mean 2 Factor Level			Mean Difference (Mean 1 – Mean 2)
PL	RS	$\bar{X}_{PL=j,RS=k}$	PL	RS	$\bar{X}_{PL=j',RS=k'}$	
1	1	0.000	1	2	0.000	0.000
1	1	0.000	1	3	0.000	0.000
1	1	0.000	1	4	0.001	0.000
1	2	0.000	1	3	0.000	0.000
1	2	0.000	1	4	0.001	0.000
1	3	0.000	1	4	0.001	0.000
2	1	0.000	2	2	0.001	0.000
2	1	0.000	2	3	0.001	0.000
2	1	0.000	2	4	0.000	0.000
2	2	0.001	2	3	0.001	0.000
2	2	0.001	2	4	0.000	0.001
2	3	0.001	2	4	0.000	0.001
3	1	0.000	3	2	0.000	0.000
3	1	0.000	3	3	0.000	0.000
3	1	0.000	3	4	0.000	0.000
3	2	0.000	3	3	0.000	0.000
3	2	0.000	3	4	0.000	0.000
3	3	0.000	3	4	0.000	0.000
4	1	0.000	4	2	0.000	0.000
4	1	0.000	4	3	0.000	0.000
4	1	0.000	4	4	0.000	0.000
4	2	0.000	4	3	0.000	0.000
4	2	0.000	4	4	0.000	0.000
4	3	0.000	4	4	0.000	0.000
5	1	0.001	5	2	0.002	-0.001*
5	1	0.001	5	3	0.000	0.001*
5	1	0.001	5	4	0.000	0.001*
5	2	0.002	5	3	0.000	0.002*
5	2	0.002	5	4	0.000	0.002*

**Note:** \*significant at  $\alpha = .05$  level. Family-Wise Error (FWE) was separately controlled at each level of perturbation at which simple main effect tests were performed.

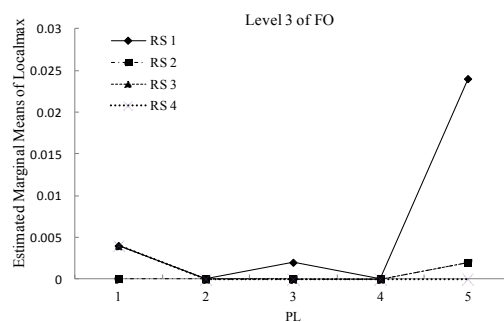
For the three-way interaction effect between  $FO \times PL \times RS$  on the dependent variable: Localmax, the interaction effect of  $PL \times RS$  was studied at each level of FO, and the results are presented only graphically in Figures 7 through 9. The three figures show clearly that for all FO levels, Localmax values were close between RS levels 2, 3 and 4 across all PL levels. At level 1 of FO, level 1 of RS diverged from the other RS levels in Localmax values at and above level 3 of PL, at level 2 of FO clear discrepancy started to occur at level 4 of PL, and at level 3 of FO a very large difference was observed between level 1 of RS and the other RS levels at the highest PL level.



**Figure 7.**  $PL \times RS$  for Localmax at level 1 of FO



**Figure 8.**  $PL \times RS$  for Localmax at level 2 of FO



**Figure 9.**  $PL \times RS$  for LogMatch at level 3 of FO

## Conclusions

Results from the factorial ANOVA analyses showed that in both simulation studies MCONV was not a significant factor affecting convergent solutions in *Mplus*, nor was SVQ studied in Simulation 1. The results related to SVQ were different from those findings obtained by Jones and Nagin (2007) who found that using informed starting values from the means of the growth parameters coming from a latent class growth analysis helps avoid the occurrence of local maxima solutions.

In terms of the main effects, RS and FO showed the same patterns of outcome means obtained under the two simulation studies. As was expected, low Dlog, high Logmatch, and low Localmax mean values were all associated with high levels of RS, indicating higher RS should be used to increase the likelihood of proper convergence of growth mixture model parameters. The story of the main effect for FO, though, was more complex and somewhat confusing. Both simulation studies showed that as levels of FO increased, the Dlog mean values became smaller (which is a desirable outcome) while the Localmax mean values increased (which is an undesirable outcome). Although Simulation 1 also showed an increased Logmatch mean value with increased levels of FO, the choice of FO still needs to be carefully considered. This is because the impacts of using various levels of FO are not consistent in terms of their impact on the four desirable properties of convergence that were considered in this study.

In terms of the main effect for PL, both simulation studies indicated that the largest Localmax value and the lowest Logmatch were found with the highest PL level considered in the study, suggesting that a high PL would not be a desirable choice for obtaining proper convergence solutions. In fact, results from Simulation 1 showed that a moderate PL level (e.g., level 3) was favored for the lowest Localmax outcome. Among all the significant factors, MC was the only factor in Simulation 1 that affected all four outcome variables and explained most of the variance in Negvariance. As was expected, when the investigated model was very complex, proper convergence solutions were negatively impacted. At its highest level, the 4-class model had the highest Dlog mean value, the lowest Logmatch mean value, the largest Negvariance mean value, and the highest Localmax mean value, all of which suggest unstable convergence solutions.

Though an assessment of main effects provides a general idea of the marginal effects of the manipulated factors, a more complete understanding of the influence of these factors on convergence to proper solutions may be obtained by assessing the presence and nature of interaction effects among these independent

variables. The only two-way interaction effect ( $PL \times MC$ ) identified in Simulation 1 showed that the highest Dlog value was with the highest level of MC across all PL levels and that Dlog values showed the slowest increase for level 1 of MC starting from level 3 of PL.

The interaction between PL and MC also affected Localmax. The highest Localmax mean values always occurred with the most complex model across levels of PL. The observation that the lowest Localmax mean values were found at (a) level 1 and 2 of MC at PL level 3 and (b) at level 3 of MC at PL level 2 suggests that a high PL was not a desirable choice for obtaining a low Localmax outcome. The assessed interaction effect between PL and FO found in Simulation 2 resulted in the following findings with respect to PL: (a) Dlog values increased for all levels of FO as PL increased, (b) Logmatch mean values started to decrease in magnitude for all levels of FO at level 4 of PL, and (c) Localmax values began to increase at or above level 4 of PL for all levels of FO. Either the increase or the decrease was wanted, which suggests that a high PL level (e.g., level 4 or 5) was not a desirable choice for obtaining effective convergence of parameter estimates.

A similar complex and confusing story with respect to the choice of a desirable level of FO occurred as what was found for the main effect of FO. As FO increased, Dlog values increased for each PL level, Lower Logmatch values were always found with higher FO for higher levels of PL, but lower Localmax values were found with higher FO for higher levels of PL. These findings indicate a conflicting situation where the choice of FO was especially challenging when a higher PL level was used.

Study of the two-way interaction effect of  $PL \times RS$  on Logmatch and Localmax showed most substantial mean differences occurred at higher levels of PL (e.g., level 4 and 5 of PL) where level 4 of RS was found having the highest Logmatch value and the lowest Localmax value, suggesting that a high RS be considered when high PL has to be used. Also, considering the decrease of Logmatch value and the increase of Localmax value for all RS levels occurred obviously at and above level 4 of PL, it was also suggested that higher PL levels not be used. The interaction effect between RS and FO for the dependent variable Localmax showed no significant mean differences between levels of FO across RS levels 2 through 4. This finding suggests that the choice of FO should not be a big concern when RS is large.

The three-way interaction effect ( $MC \times PL \times RS$ ) for the dependent variable Nonconverge found in Simulation 1 showed no significant findings at level 1 and level 2 of MC, suggesting that with a less complex model (i.e., the 2 and 3-class models), the choice of levels of PL and RS would not significantly affect

convergence solutions. At level 3 of MC, significant Nonconverge mean differences were found only at level 5 of PL, with RS level 3 and 4 (having the lowest Nonconverge value) showing no significant mean difference. This finding shows with very complex models, high RS should be used, especially when the PL level is very high. The three-way interaction effect between FO, PL and RS obtained with Localmax showed somewhat similar patterns for the two-way interactions between PL and RS for each of the three FO conditions. At middle and lower levels of PL, Localmax values were similar in magnitude for all levels of RS and remained stable. When PL level was very high, Localmax values at level 1 of RS not only increased sharply but deviated dramatically from the Localmax values at the other RS levels.

### Recommendations

Based on the findings and conclusions previously discussed, some recommendations that users of *Mplus* may wish to follow when fitting GMMs are as follows:

- Use a large number of random start values, especially when the perturbation level is high.
- Number of random starts should be at least at level 2 for lower number of local maxima when the perturbation level is high. However, the number of random starts does not greatly affect the number of non-converged solutions if moderate perturbation levels are used.
- Choice of the number of final optimizations needs to be considered carefully. Usually, a high number of final optimizations is not recommended although the choice would not be a big concern when the number of random starts is high.
- Do not use high perturbation levels (e.g., level 4 or level 5), especially with very complex growth mixture models which may show much higher rate of increase in number of local maximum solutions than less complex models. Instead, a moderate perturbation level (e.g., *Mplus* default perturbation level of 5) is recommended for obtaining better convergence solutions.
- With less complex growth mixture models, the choice of number of random starts and perturbation levels may not be so important. However, when a model is very complex (e.g., the 4-class model), a



high number of random starts should be considered for obtaining better convergence solutions, especially when perturbation level is very high.

## References

- Bauer, D. J. (2007). 2004 Cattell award address: Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42: 757-786.
- Böhning, D. (1999). *Computer-assisted analysis of mixtures and applications: Metaanalysis, disease mapping and others*. Monographs on statistics and applied probability. New York: Chapman & Hall/CRC.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39: 1-38.
- Hamilton, J. (2009). *An investigation of growth mixture models when data are collected with unequal selection probabilities: A Monte Carlo study* (Unpublished doctoral dissertation). University of Maryland, College Park.
- Harring, J. R. (2012). Finite mixtures of nonlinear mixed effects models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences*. Charlotte, NC: Information Age Publishing, Inc.
- Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, 11: 36-53.
- Hsu, J. C. (2011). *Estimation and model selection for finite mixtures of latent interaction models* (Unpublished doctoral dissertation). University of Maryland, College Park.
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42: 805-820.
- Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and a SAS procedure for estimating them. *Sociological Methods & Research*, 35: 542-571.
- Kohli, N. (2010). *Estimating unknown knots in piecewise linear-linear latent growth mixture models* (Unpublished doctoral dissertation). University of Maryland, College Park.

- Liu, J., & Harring, J. R. (2012). *A systematic investigation of within-subject and between-subject covariance structures in growth mixture models* (Unpublished doctoral dissertation). University of Maryland, College Park.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
- McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. New York: Wiley.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55: 107-122.
- Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Erlbaum.
- Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55: 463-469.
- Muthén, L., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.
- Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, 4: 139-177.
- Tolvanen, A. (2008). *Latent growth mixture modeling: A simulation study* (Unpublished doctoral dissertation). University of Jyväskylä.
- Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, 10: 635-656.

## Appendix

For the simulation, the number of occasions of measurement were fixed at  $p = 6$ . The covariance structures were constructed to be constant across groups and are defined as:

$$\text{cov}(\boldsymbol{\eta}) = \Psi = \begin{pmatrix} 1 & \\ .224 & 2 \end{pmatrix} \quad \text{cov}(\boldsymbol{\epsilon}) = \Theta = \sigma^2 \mathbf{I}, \text{ where } \sigma^2 = .75.$$

The mean vectors for the three-class model are specified as

$$\alpha_1' = (4.5, -.85)$$

$$\alpha_2' = (2.5, -.05)$$

$$\alpha_3' = (1.4, -.70)$$

Data were generated according to the two-stage approach outlined by Hipp and Bauer (2006).

## JMASM 33: A Two Dependent Samples Maximum Test Calculator: Excel

**Saverpierre Maggio**  
Wayne State University  
Detroit, MI

**Shlomo Sawilowsky**  
Wayne State University  
Detroit, MI

---

An Excel Macro was created to provide researchers with an easy to use resource in order to calculate the two dependent samples maximum test as provided in Maggio and Sawilowsky (2014), which permits conducting both the two dependent samples t-test and Wilcoxon signed-ranks test on the same data while eliminating concerns related to Type I error inflation and choice of statistical tests.

*Keywords:* Maximum test, Dependent samples t test, Wilcoxon signed-ranks test, Excel calculator, Experiment-wise Type I error inflation

---

### Introduction

Inferential errors are easy to commit, and they are compounded when conducting multiple tests (either serially or in parallel) on the same data. In the case of the two dependent samples t-test and the Wilcoxon Signed-Ranks (WSR) test, in general the former should be used if data are known or expect to be normally distributed, otherwise the latter should be used, assuming the treatment alternative is a shift in means. (Blair & Higgins, 1985; Bridge & Sawilowsky, 1999; Gerke & Randles, 2010; Wiederman & Alexandrowicz, 2011). Researchers also cannot conduct both tests on the same data without increasing the Experiment-wise Type I error rate (Sawilowsky & Fahoome, 2003).

A solution strategy is known as the maximum test, whereby the researcher puts “various score statistics together and takes the maximum of them” (Kossler, 2010, p. 2), then the maximum of the two tests are compared to a critical value obtained on a joint sampling distribution for the two tests. This strategy eliminates two concerns; (1) Type I error inflation, and (2) choice of statistic (Algina, J. et al,

---

*Dr. Maggio is an Adjunct Faculty member in the Department of Evaluation and Research, College of Education. Email him at: [bn4424@wayne.edu](mailto:bn4424@wayne.edu). Dr. Sawilowsky is Professor in the Department of Evaluation and Research, College of Education. Email him at [shlomo@wayne.edu](mailto:shlomo@wayne.edu).*

1995; Blair, 2002; Tarone, 1981; Willan, 1988; Fleming & Harrington, 1991; Lee, 1996; Ryan et al., 1999; Blair, 2002; Weichert & Hothorn, 2002; Neuhauser et al., 2004; Opdyke, 2005; Salmaso & Solari, 2005; Yang et al., 2005; Kossler, 2010; Maggio & Sawilowsky, 2014).

## Purpose

The purpose of this article is to provide researchers with an easy to use Excel macro that calculates the two dependent samples maximum test. It is based on critical values provided in Maggio and Sawilowsky (2014).

## Methodology

Download the Macro (<http://digitalcommons.wayne.edu/jmasm/vol13/iss1/32>) or contact the first author via e-mail. A screenshot of the Excel worksheet is located in Figure 1.

## Input

The process of obtaining the maximum test p-values, critical values and a determination of whether or not to “reject” or “fail to reject” a null hypothesis is as follows:

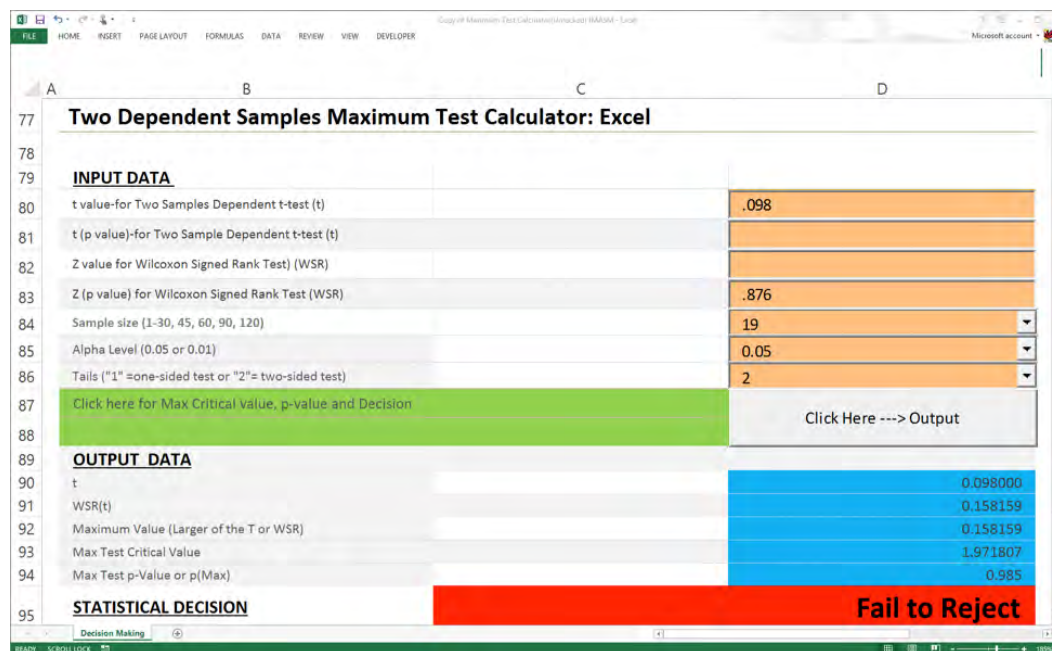
1. Obtain the t value (or p-value) for two samples dependent t test on data (e.g., via statistical software) and input that value in the appropriate cell of the worksheet. For example, the t value is placed in cell D80, or if the p value associated with the t test it is inputted in cell D81.
2. Obtain the Z (or p value) for the Wilcoxon Signed-Ranks test and place it in cell D82. (or D83)
3. Input or select the sample size in cell D84. (samples are limited to  $df = 8$  to 30, 45, 60, 90, & 120)
4. Input or select the desired alpha level for a two tailed test (0.05 or 0.01) in cell D85.

## A TWO DEPENDENT SAMPLES MAXIMUM TEST CALCULATOR: EXCEL

5. Input or select the number of tails (“2” for two-sided test; “1” for one-sided test) in cell D86.
6. Left click on the button that reads “Click here→ Output” .

### Conclusion

The macro reports the output in cells D90 through D95. If the p-value for t is inputted in cell D81 then D90 contains the corresponding t-value. The Max(t)/critical value, p(Max), and statistical decision appear in cells D93 D94, and D95, respectively.



**Figure 1.** Screenshot of the Excel Worksheet

The two dependent samples maximum test that can be used in lieu of choice between the two dependent samples t- test and Wilcoxon signed-ranks when the distribution from which samples were drawn is unknown. Both the classical parametric and non-parametric tests can be safely conducted on the same data, with the maximum of the two refereed to the new table of critical values that are

designed to maintain the Type I error rate to nominal  $\alpha$  while guaranteeing the maximum power of the two tests (Maggio & Sawilowsky, 2014).

The maximum test is easy to compute with or without an excel macro. Maggio and Sawilowsky (2014) provided the maximum test critical values for a two tailed test and a clear example to follow. Readers are encouraged to review that article.

## References

- Algina, J., Blair, R. C. & Coombs, W. T. (1995). A maximum test for scale: Type I error rate and power. *Journal of Educational and Behavioral Statistics*, 20(27): 27-39.
- Blair, R. C. (1991). New critical values for the generalized t and generalized rank-sum procedures. *Communications in Statistics*, 20, 981-994.
- Blair, C. R. (2002). Combining two nonparametric tests of location. *Journal of Modern Applied Statistical Methods*, 1(1): 13-18.
- Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of wilcoxon's sign-ranks test under various population shapes. *Psychological Bulletin*, 97(1): 119-128.
- Bridge, P. K. & Sawilowsky, S. (1999). Increasing physician's awareness of the impact of statistical tests on research outcomes: Investigating the comparative power of the wilcoxon rank-sum test and independent samples t-test to violations from normality. *Journal of Clinical Epidemiology*, 52(2): 229-236.
- Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York, NY: Wiley.
- Gerke, T. A. & Randles, H. (2010). A method for resolving ties in asymptotic relative efficiency. *Statistics and Probability Letters*, 80(13): 1065-1069.
- IMSL. (1980). *International Mathematical and Statistical Libraries*. Houston, Texas.
- Kossler, W. (2010). Max-type rank tests, u-tests, and adaptive tests for the two-sample location problem - An asymptotic power study. *Journal of Computational Statistics & Data Analysis*, 54(9): 2053-2065.
- Lee, W. J. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 52(2): 721-725.

Maggio, S. & Sawilowsky, S. (2014). A new maximum test via the dependent samples t-test and the wilcoxon sign rank test. *Applied Mathematics*, 5(10): 110-114. doi: 10.4236/am.2014.51013.

Neuhäuser, M. Büning, H., & Hothorn, L. (2004). Maximum Test versus adaptive tests for the two-sample location problem. *Journal of Applied Statistics*, 31(2), 215-227.

Opdyke, J. D. (2005). A single, powerful, nonparametric statistic for continuous- data telecommunications parity testing. *Journal of Modern Applied Statistical Methods*, 4(2), 372-393.

Ryan, L. M., Freidlin, B., Podgor, M. J., & Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics*, 55(3), 883-886.

Salmaso, L., & Solari, A. (2005). Multiple aspects testing for case-control designs. *Metrika*, 62, 331-340.

Sawilowsky, S. S., & Fahoome, G. F. (2003). Statistics through Monte Carlo Simulation with FORTRAN. Oak Park, Michigan: *Journal of Modern Applied Statistical Methods*.

Tarone, R. E. (1981). On the distribution of the maximum of the log-rank statistic and the modified Wilcoxon statistic. *Biometrics*, 37, 79-85.

Weichert, M. & Hothorn, L.A. (2002). Robust hybrid tests for the two-sample location problem. *Communications in Statistics – Simulation and Computation*, 31, 175-187.

Wiederman, W. T., & Alexandrowicz, R. W. (2011). A modified normal scores test for paired data. *European Journal of Research Methods for the Behavioral and Social Sciences*, 7(1), 25-38.

Willan, A. R. (1988). Using the maximum test statistic in the two-period crossover clinical trial. *Biometrics*, 44(1), 211-218.

Yang, S., Hsu, L., & Zhao, L. (2005). Combining asymptotically normal tests: case studies in comparison to two groups. *Journal of Statistical Planning and Inference*, 133(1), 139-158.



## Instructions for Authors

Authors wishing to submit to *JMASM* 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.

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](mailto:ea@jmasm.com)

Wayne State University Library System  
Purdy Library  
Detroit, MI 48202  
[digitalcommons@wayne.edu](mailto:digitalcommons@wayne.edu)

### Copyrights, Attribution and Usage Policies

Copyright ©2013 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 *JMASM* (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](mailto:ea@jmasm.com).

