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

JMASM Editors

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

## Recommended Citation

Editors, JMASM (2003) "Vol. 2, No. 2 (Full Issue)," Journal of Modern Applied Statistical Methods: Vol. 2 : Iss. 2 , Article 30.
DOI: 10.22237/jmasm/1067644800
Available at: http://digitalcommons.wayne.edu/jmasm/vol2/iss2/30

## announces



# The fastest, most comprehensive and robust permutation test software on the market today. 

Permutation tests increasingly are the statistical method of choice for addressing business questions and research hypotheses across a broad range of industries. Their distribution-free nature maintains test validity where many parametric tests (and even other nonparametric tests), encumbered by restrictive and often inappropriate data assumptions, fail miserably. The computational demands of permutation tests, however, have severely limited other vendors' attempts at providing useable permutation test software for anything but highly stylized situations or small datasets and few tests. Permutelt ${ }^{\text {TM }}$ addresses this unmet need by utilizing a combination of algorithms to perform non-parametric permutation tests very quickly - often more than an order of magnitude faster than widely available commercial alternatives when one sample is large and many tests and/or multiple comparisons are being performed (which is when runtimes matter most). Permutelt ${ }^{\text {TM }}$ can make the difference between making deadlines, or missing them, since data inputs often need to be revised, resent, or recleaned, and one hour of runtime quickly can become 10,20 , or 30 hours.

In addition to its speed even when one sample is large, some of the unique and powerful features of Permutelt ${ }^{\text {TM }}$ include:

- the availability to the user of a wide range of test statistics for performing permutation tests on continuous, count, \& binary data, including: pooled-variance t-test; separate-variance Behrens-Fisher t-test, scale test, and joint tests for scale and location coefficients using nonparametric combination methodology; Brownie et al. "modified" $t$-test; skew-adjusted "modified" t-test; Cochran-Armitage test; exact inference; Poisson normal-approximate test; Fisher's exact test; FreemanTukey Double Arcsine test
- extremely fast exact inference (no confidence intervals - just exact p-values) for most count data and high-frequency continuous data, often several orders of magnitude faster than the most widely available commercial alternative
- the availability to the user of a wide range of multiple testing procedures, including: Bonferroni, Sidak, Stepdown Bonferroni, Stepdown Sidak, Stepdown Bonferroni and Stepdown Sidak for discrete distributions, Hochberg Stepup, FDR, Dunnett's one-step (for MCC under ANOVA assumptions), Single-step Permutation, Stepdown Permutation, Single-step and Stepdown Permutation for discrete distributions, Permutation-style adjustment of permutation p-values
- fast, efficient, and automatic generation of all pairwise comparisons
- efficient variance-reduction under conventional Monte Carlo via self-adjusting permutation sampling when confidence intervals contain the user-specified critical value of the test
- maximum power, and the shortest confidence intervals, under conventional Monte Carlo via a new sampling optimization technique (see Opdyke, JMASM, Vol. 2, No. 1, May, 2003)
- fast permutation-style p-value adjustments for multiple comparisons (the code is designed to provide an additional speed premium for many of these resampling-based multiple testing procedures)
- simultaneous permutation testing and permutation-style p-value adjustment, although for relatively few tests at a time (this capability is not even provided as a preprogrammed option with any other software currently on the market)

For Telecommunications, Pharmaceuticals, fMRI data, Financial Services, Clinical Trials, Insurance, Bioinformatics, and just about any data rich industry where large numbers of distributional null hypotheses need to be tested on samples that are not extremely small and parametric assumptions are either uncertain or inappropriate, Permutelt ${ }^{\mathrm{TM}}$ is the optimal, and only, solution.

To learn more about how Permutelt ${ }^{\text {TM }}$ can be used for your enterprise, and to obtain a demo version, please contact its author, J.D. Opdyke, President, DataMinelt ${ }^{\text {SM }}$, at JDOpdyke@DataMinelt.com or www.DataMinelt.com.

DataMinelt ${ }^{\text {SM }}$ is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. Permutelt ${ }^{\mathrm{TM}}$ is its flagship product.

## Journal Of Modern Applied Statistical Methods

| Invited Articles |  |  |
| :---: | :---: | :---: |
| 281-292 | C. Mitchell Dayton | Model Comparisons Using Information Measures |
| 293-305 | Gregory R. Hancock | Fortune Cookies, Measurement Error, And Experimental Design |
| Regular Articles |  |  |
| 306-313 | Charles F. Bond, Jr., F. D. Richard | P* Index Of Segregation: Distribution Under Reassignment |
| 314-328 | Kimberly T. Perry | A Critical Examination Of The Use Of Preliminary Tests In Two-Sample Tests Of Location |
| 329-340 | Christopher J. Mecklin | A Comparison Of Equivalence Testing In Combination With Hypothesis Testing And Effect Sizes |
| 341-349 | Ayman Baklizi | Confidence Intervals For $\mathrm{P}(\mathrm{X}<\mathrm{Y})$ In The Exponential Case With Common Location Parameter |
| 350-358 | Vincent A. R. Camara | Approximate Bayesian Confidence Intervals For The Variance Of A Gaussian Distribution |
| 359-370 | Alfred A. Bartolucci, Shimin Zheng, Sejong Bae, Karan P. Singh | Random Regression Models Based On The Elliptically Contoured Distribution Assumptions With Applications To Longitudinal Data |
| 371-379 | Dudley L. Poston, Jr., Sherry L. McKibben | Using Zero-inflated Count Regression Models To Estimate The Fertility Of U.S. Women |
| 380-388 | Felix Famoye, Daniel E. Rothe | Variable Selection For Poisson Regression Model |
| 389-399 | Chengjie Xiong, <br> Yan Yan, Ming Ji | Test Of Homogeneity For Umbrella Alternatives In Dose-Response Relationship For Poisson Variables |
| 400-413 | Stephanie Wehry, James Algina | Type I Error Rates Of Four Methods For Analyzing Data Collected In A Groups vs Individuals Design |
| 414-424 | Terry Hyslop, Paul J. Lupinacci | A Nonparametric Fitted Test For The Behrens-Fisher Problem |


| 425-432 | David A. Walker, <br> Denise Y. Young | Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis |
| :---: | :---: | :---: |
| 433-442 | Guillermo Montes, Bohdan S. Lotyczewski | Correcting Publication Bias In Meta-Analysis: A Truncation Approach |
| 443-450 | Chin-Shang Li, Hua Liang, Ying-Hen Hsieh, Shiing-Jer Twu | Comparison Of Viral Trajectories In AIDS Studies By Using Nonparametric Mixed-Effects Models |
| 451-466 | Stephanie Wehry | Alphabet Letter Recognition And Emergent Literacy Abilities Of Rising Kindergarten Children Living In Low-Income Families |
| 467-474 | Shlomo S. Sawilowsky | Deconstructing Arguments From The Case Against Hypothesis Testing |
| Brief Reports |  |  |
| 475-477 | W. J. Hurley | A Note On MLEs For Normal Distribution Parameters Based On Disjoint Partial Sums Of A Random Sample |
| 478-480 | W. Gregory Thatcher, J. Wanzer Drane | On Treating A Survey Of Convenience Sample As A Simple Random Sample |
| Early Scholars |  |  |
| 481-496 | Katherine Fradette, J. Keselman, Lisa Lix, James Algina, Rand R. Wilcox | Conventional And Robust Paired And Independent-Samples t Tests: Type I Error And Power Rates |
| 497-511 | A. Gemperli, P. Vounatsou | Fitting Generalized Linear Mixed Models For Point-Referenced Spatial Data |
| 512-519 | Jason E. King | Bootstrapping Confidence Intervals For Robust Measures Of Association |
| JMASM Algorithms and Code |  |  |
| 520-524 | Scott J. Richter <br> Mark E. Payton | JMASM8: Using SAS To Perform Two-Way Analysis Of Variance Under Variance Heterogeneity (SAS) |
| 525-530 | David A. Walker | JMASM9: Converting Kendall’s Tau For Correlational Or Meta-Analytic Analyses (SPSS) |

Letters To The Editor
$531 \quad$ Vance W. Berger
531-532 Thomas R. Knapp
End Matter
533-534
Author
Statistical Pronouncements II

JMASM is an independent print and electronic journal (http://tbf.coe.wayne.edu/jmasm) designed to provide an outlet for the scholarly works of applied nonparametric or parametric statisticians, data analysts, researchers, classical or modern psychometricians, quantitative or qualitative evaluators, and methodologists. Work appearing in Regular Articles, Brief Reports, and Early Scholars are externally peer reviewed, with input from the Editorial Board; in Statistical Software Applications and Review and JMASM Algorithms and Code are internally reviewed by the Editorial Board.

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 computerintensive 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.

Problems may arise from applied statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation. They should relate to the social and behavioral sciences, especially education and psychology. Applications from other traditions, such as actuarial statistics, biometrics or biostatistics, chemometrics, econometrics, environmetrics, jurimetrics, quality control, and sociometrics are welcome. Applied methods from other disciplines (e.g., astronomy, business, engineering, genetics, logic, nursing, marketing, medicine, oceanography, pharmacy, physics, political science) are acceptable if the demonstration holds promise for the social and behavioral sciences.

| Editorial Assistant | Professional Staff | Production Staff | Internet Sponsor |
| :--- | :--- | :--- | :--- |
| Patric R. Spence | Bruce Fay, | Jack Sawilowsky <br> Business Manager <br> Christina Gase | Paula C. Wood, <br>  <br>  <br>  <br>  <br>  <br>  <br> Joe Musial, <br> Marketing Director |
|  | College of Education, |  |  |
|  |  | Wayne State University |  |


| Entire Reproductions and Imaging Solutions | 248.299 .8900 (Phone) | e-mail: |
| :--- | :--- | :--- |
| Internet: www.entire-repro.com | 248.299 .8916 (Fax) | sales@entire-repro.com |

# Editorial Board of Journal of Modern Applied Statistical Methods 

Subhash Chandra Bagui
Department of Mathematics \& Statistics
University of West Florida
Chris Barker
MEDTAP International
Redwood City, CA
J. Jackson Barnette

Community and Behavioral Health
University of Iowa
Vincent A. R. Camara
Department of Mathematics
University of South Florida
Ling Chen
Department of Statistics
Florida International University
Christopher W. Chiu
Test Development \& Psychometric Rsch
Law School Admission Council, PA
Jai Won Choi
National Center for Health Statistics
Hyattsville, MD
Rahul Dhanda
Forest Pharmaceuticals
New York, NY
John N. Dyer
Dept. of Information System \& Logistics
Georgia Southern University
Matthew E. Elam
Dept. of Industrial Engineering
University of Alabama
Mohammed A. El-Saidi
Accounting, Finance, Economics \&
Statistics, Ferris State University
Carol J. Etzel
University of Texas M. D.
Anderson Cancer Center
Felix Famoye
Department of Mathematics
Central Michigan University
Barbara Foster
Academic Computing Services, UT
Southwestern Medical Center, Dallas
Shiva Gautam
Department of Preventive Medicine
Vanderbilt University
Dominique Haughton
Mathematical Sciences Department
Bentley College
Scott L. Hershberger
Department of Psychology
California State University, Long Beach

Joseph Hilbe
Departments of Statistics/ Sociology
Arizona State University
Peng Huang
Dept. of Biometry \& Epidemiology
Medical University of South Carolina
Sin-Ho Jung
Dept. of Biostatistics \& Bioinformatics
Duke University
Jong-Min Kim
Statistics, Division of Science \& Math
University of Minnesota
Harry Khamis
Statistical Consulting Center
Wright State University
Kallappa M. Koti
Food and Drug Administration
Rockville, MD
Tomasz J. Kozubowski
Department of Mathematics
University of Nevada
Kwan R. Lee
GlaxoSmithKline Pharmaceuticals
Collegeville, PA
Hee-Jeong Lim
Dept. of Math \& Computer Science
Northern Kentucky University
Devan V. Mehrotra
Merck Research Laboratories
Blue Bell, PA
Prem Narain
Freelance Researcher
Farmington Hills, MI
Balgobin Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute
J. Sunil Rao

Dept. of Epidemiology \& Biostatistics
Case Western Reserve University
Brent Jay Shelton
Department of Biostatistics
University of Alabama at Birmingham
Karan P. Singh
University of North Texas Health
Science Center, Fort Worth
Jianguo (Tony) Sun
Department of Statistics
University of Missouri, Columbia
Joshua M. Tebbs
Department of Statistics
Kansas State University
Dimitrios D. Thomakos
Department of Economics
Florida International University

Justin Tobias
Department of Economics
University of California-Irvine
Jeffrey E. Vaks
Beckman Coulter
Brea, CA
Dawn M. VanLeeuwen
Agricultural \& Extension Education
New Mexico State University
David Walker
Educational Tech, Rsrch, \& Assessment
Northern Illinois University
J. J. Wang

Dept. of Advanced Educational Studies
California State University, Bakersfield
Dongfeng Wu
Dept. of Mathematics \& Statistics
Mississippi State University
Chengjie Xiong
Division of Biostatistics
Washington University in St. Louis
Andrei Yakovlev
Biostatistics and Computational Biology
University of Rochester
Heping Zhang
Dept. of Epidemiology \& Public Health
Yale University
International
Mohammed Ibrahim Ali Ageel
Department of Mathematics
King Khalid University, Saudi Arabia
Mohammad Fraiwan Al-Saleh
Department of Statistics
Yarmouk University, Irbid-Jordan
Keumhee Chough (K.C.) Carriere
Mathematical \& Statistical Sciences
University of Alberta, Canada
Debasis Kundu
Department of Mathematics
Indian Institute of Technology, India
Christos Koukouvinos
Department of Mathematics
National Technical University, Greece
Lisa M. Lix
Dept. of Community Health Sciences
University of Manitoba, Canada
Takis Papaioannou
Statistics and Insurance Science
University of Piraeus, Greece
Mohammad Z. Raqab
Department of Mathematics
University of Jordan, Jordan
Nasrollah Saebi
School of Mathematics
Kingston University, UK
Keming Yu
Statistics
University of Plymouth, UK

## INVITED ARTICLES

Model Comparisons Using Information Measures

C. Mitchell Dayton<br>University of Maryland



Methodologists have criticized the use of significance tests in the behavioral sciences but have failed to provide alternative data analysis strategies that appeal to applied researchers. For purposes of comparing alternate models for data, information-theoretic measures such as Akaike AIC have advantages in comparison with significance tests. Model-selection procedures based on a $\min (A I C)$ strategy, for example, are holistic rather than dependent upon a series of sometimes contradictory binary (accept/reject) decisions.

Key words: Akaike AIC, significance tests, information measures

## Introduction

Quantitative researchers have been trained to evaluate effects of interest utilizing the methods of statistical inference. In a single research study it is not unusual to see several dozen, or even several hundred, significance tests applied to assess, for example, multiple correlations, differences among multiple correlations and regression coefficients. However, the appropriateness of the use of significance tests in social and behavioral research settings has been
C. Mitchell Dayton is a Professor of Measurement \& Statistics at the University of Maryland. His major research interests deal with the topics of latent class analysis and simultaneous inference. He recently published a Sage book dealing with latent class scaling models, a topic on which he has published widely. His long standing interest in simultaneous inference has led to a focus on model-comparison approaches utilizing information theory and posterior Bayes factors.
debated for more than 40 years. In particular, Rozeboom (1960) summarized criticisms of significance testing that have resurfaced in various guises from time to time. Generally, these criticisms have focused on the issue of binary decision-making (e.g., accept/reject null hypotheses) as opposed to considerations related to weight of evidence (e.g., measures of strength of effect or effect sizes).

The fundamental error, as seen by Rozeboom, "...lies in mistaking the aim of a scientific investigation to be a decision, rather than a cognitive evaluation of propositions (op. cit., page 212)." Although distinctions can be drawn between significance testing in the Fisherian (1959) sense and hypothesis testing in the Neyman-Pearson (1933) sense, current teaching and practice in the behavioral sciences blur these distinctions and the terms can be considered as essentially interchangeable in practice. However, it is likely that Fisher himself would concur with many of the criticisms as suggested by the following quotes (Fisher, 1959):
...the calculation \{of significance levels\} is absurdly academic, for in fact no scientific worker has a fixed
level of significance at which from year to year, and in all circumstance, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (page 42)

On the whole the ideas (a) that a test of significance must be regarded as one of a series of similar tests applied to a succession of similar bodies of data, and (b) that the purpose of the test is to discriminate or 'decide' between two or more hypotheses, have greatly obscured their understanding, when taken not as contingency possibilities but as elements essential in their logic. (page 42)

Advocates for change have urged minimizing (or, even eliminating) the role of significance tests in behavioral research and elevating the roles of procedures such as confidence intervals, measures of effect size (e.g., Carver, 1993) or replicability measures (e.g., Thompson, 1994). Although these advocacy positions have been well articulated and widely disseminated among applied statisticians, there is scant evidence for change in practice by applied researchers in the behavioral sciences.

For example, the Fall 1995, Winter 1995 and Spring 1996 issues of the American Educational Research Journal contained 11 databased articles in the Teaching, Learning and Human Development section of the journal. The number of significance tests per article (with some allowances for counting errors) are, in rank order: 3, 29, 33, 35, 40, 48, 94, 212, 290, 335 and 448 for a total of 1567 tests or an average of 522 significance tests per issue of the journal.

Although the lowest number, 3, might lead to useful interpretations within a single research study, it is highly doubtful that 29 , much less 448 , such tests in a single study can be interpreted in manner that provides much scientific value. Indeed, the lack of popularity for alternative procedures to significance testing has, itself, a long history as evidenced by Heermann and Braskamp (1970) who wrote in the Introduction to Part 4, Testing Statistical Hypotheses, of their book of readings:
there is considerable agreement among statisticians and behavioral scientists that there has been an unfortunate emphasis on the part of the latter on hypothesis testing to the exclusion of other inferential techniques....In spite of this widely known fact, behavioral scientists continue to employ significance tests to the exclusion of other more informative techniques. (page 154)

It can be argued that a major reason for the apparent resistance to change from significance tests to other techniques is that the alternatives that have been proposed are unattractive to applied researchers. Consider the relatively simple example of multiple comparisons among a set of, say, five sample means. A typical traditional approach would be the use of Tukey tests (or one of the plethora of variations such as Games-Howell tests). In effect, 10 significance tests would be conducted and referred to the appropriate theoretical distribution (e.g., studentized range).

If a researcher were to follow Carver's (1993) advice, the Tukey tests would be replaced by "...estimates of effect size and of sampling error such as standard errors and confidence intervals 89)." However, the q statistic per se can be viewed as an effect size (i.e., difference between two means divided by the estimated standard error of a mean) and how does the researcher arrive at a unified interpretation of the 10 confidence intervals? But Carver (1993) has additional advice: "Better yet, by conducting multiple studies, replication of results can replace statistical significance testing." This is not a particularly attractive option given the obstacles that may exist to replication and the fact that the researcher really needs to interpret the present study in order to decide whether or not replication is a worth while expenditure of time and resources.

A premise of this paper is that significance tests are appropriate for only certain, highly constrained purposes but have enjoyed much wider use because of the failure of methodologists to popularize other, more appropriate statistical methods. In particular, significance tests are useful for interpreting data that arise from controlled experimental or quasi-
experimental designs in which the role of specific hypotheses is well-defined. For nonexperimental settings, researchers typically utilize significance tests for purposes of comparing alternate models for data or for interpreting effects within specific models. It is this application that is better served by procedures specifically designed for comparisons among models and is ill-served by significance tests.

An increasingly popular technique for comparing models involves informationtheoretic measures such as Akaike $(1973,1978)$ AIC or measures based on posterior Bayes factors such as Schwarz (1978) BIC. In either case, these measures may be viewed as penalized log-likelihoods and are computed separately for each model under consideration. Then, a preferred model, among those being compared, can be selected.

This permits a very wide range of applications and even avoids some technical issues in applying statistical tests for model comparisons (e.g., for comparing number of components for discrete mixture models such as latent class models). Model comparison procedures are holistic in the sense that a variety of competing models can be assessed simultaneously and a best model selected by applying a single rule. Attempting to compare models using significance tests is, by contrast, piece-meal with the final selection of a model based on results from sometimes conflicting outcomes.

Consider, for example, the procedure that is often used when fitting polynomial regression models to bivariate data. Assume there are five distinct levels of a quantitative independent variable, $X$, so that models corresponding to linear, quadratic, cubic and quartic regression can meaningfully be fit to the data. Typically, the differences in fit of increasingly more complex models are evaluated by significance tests based on differences in multiple correlations (of, equivalently, differences in explained variability).

Thus, four distinct hypotheses are tested with, say, four hierarchical F statistics each at some specified level of significance. Since four independent tests are being conducted, an initial decision is whether or not to control the Type I
error rate for the set of tests or, simply, to use a conventional .05 level for each test. This decision, it should be noted, can dramatically affect the interpretation of results. On the other hand, a holistic, model-comparison approach entails computing, say, an Akaike AIC statistic for each regression model and then selecting a "best" model corresponding to the minimum value of AIC.

Another consideration in selecting an approach to comparing models is the logic of the decision-making strategy itself. In applying significance tests, the null hypothesis corresponds to some restricted form of a model (e.g., a test for quadratic regression involves a null hypothesis stating that the regression coefficient for the quadratic term is zero and this corresponds to a simpler, linear regression model). The validity of the test depends upon assuming that the simpler model is true and that deviations from the model are due to chance. But this is a gross over-simplification of the scientific process. In a holistic, modelcomparison approach the underlying goal is to select the best approximating model from among the models under consideration. It is not necessary to assume that any given model is "true" and there is no need to posit that a true model exists among those being compared.

In this article, the rationale for information-theoretic model comparison procedures is presented and two specific areas of application are discussed - pairwise comparisons and analysis of finite mixtures.

## Information Criteria

Akaike (1973) suggested that the Kullback-Leibler (1951) information measure provides a natural criterion for ordering alternate models for data. He developed a sample-based estimate, AIC, for this information measure that he incorporated into a decision-making strategy. For any specific model, the form of AIC is $-2 L L+2 p$ where LL is the log-likelihood for the model and p is the number of independent parameters that are estimated in fitting the model to data.

For example, assuming normally distributed residuals for a homogeneous linear regression model for three independent
variables, $p$ would equal five and comprise three partial regression coefficients, the mean of the dependent variable (or the Y-intercept) and the variance of the residuals.

A summary of the technical development for the AIC measure can be found in Dayton (2003a) whereas a detailed analysis of the measure is presented by de Leeuw (1992). In general terms, Kullback-Leibler information is a measure of the discrepancy between the true distribution for a random variable (possibly vector-valued) and the distribution specified by some particular model. Although the true model is never known, Akaike managed to derive an estimate of this discrepancy by considering the distribution of a future sample conditional on knowing the maximum-likelihood estimator for parameters in the model.

Fundamentally, AIC involves the notion of cross-validation, but only in a theoretical sense. Given AIC values for two or more alternate models, the model satisfying $\min (A I C)$ is, in this information-theoretic sense, most representative of the true model and, on this basis, may be interpreted as the best approximating model among those being considered. A useful interpretation of AIC is that it estimates the loss of precision (or, increase in information) that results from substituting maximum likelihood estimates for the true parametric values in the likelihood function. Thus, among the models under consideration, it can be argued that the preferred model (i.e., $\min (\mathrm{AIC})$ model) has the smallest expected loss of precision relative to the true, but unknown, model.

It should be noted that AIC does not depend directly on sample size. Bozdogan (1987) noted that, because of this, AIC lacks certain properties of asymptotic consistency and he proposed a related measure, CAIC, by applying his own heuristic to the development of the estimate for Kullback-Leibler information. In particular, for a sample of N cases, CAIC $=-2 L L+(\ln (N)+1) p$.

This measure is very similar to the BIC measure proposed by Schwarz (1978) that takes the form $B I C=-2 L L+\ln (N) p, \quad$ although Schwarz developed his measure as an estimate for a particular posterior Bayes factor not
directly related to Kullback-Leibler information. In any case, both CAIC and BIC reflect sample size and have properties of asymptotic consistency although the importance of this property for the interpretation of data for any specific sample setting can be disputed since, unlike significance tests, the interpretation of AIC does not depend on long-range sampling notions. AIC, CAIC and BIC may each be viewed as a penalized log-likelihood (Sclove, 1987) with penalties per parameter of $2, \ln (\mathrm{~N})+1$ and $\ln (N)$, respectively. For all reasonable sample sizes, CAIC and BIC apply larger penalties than AIC and, thus, other factors being equal, they tend to select simpler models than does AIC.

Among the reasons for preferring the use of a model selection procedure such as AIC in comparison to traditional significance tests are:
(a) A single, holistic decision can be made concerning the model that is best supported by the data in contrast to what is usually a series of possibly conflicting significance test. Moreover, models can be ranked from best to worst supported by the data, thus, extending the possibilities of interpretation.
(b) Models with various parameterizations can be compared even when the models do not obey hierarchic relations.
(c) Homogeneous and heterogeneous versions of models can be compared; in particular, the homogeneity of variance (homoscedasticity) assumptions required by many significance tests can be circumvented and the selection of the most appropriate model can be based on the information criteria.
(d) Considerations related to underlying distributions for random variables can be incorporated into the decision-making process rather than being treated as an assumption whose robustness must be considered (e.g., models based on normal densities and on log-normal densities can be compared).

Various arguments have been presented against the use of information criteria such as AIC although some of these are difficult to follow. For example, McDonald and Marsh (1990) seem to argue as follows: major premise - the saturated model is always the true model; minor premise - for sufficiently large sample
size, AIC will always select the saturated model; conclusion - AIC is defective and cannot be used in practice. In a context such as pairedcomparisons among K means, a saturated model based on normal densities would comprise K unique means and variances. Thus, no other model could possibly fit the data better in an absolute sense (i.e., yield a larger loglikelihood). However, if two of the group means are truly equally and very large samples are involved, measures such as AIC will tend to select the correct model, not the saturated model.

As noted above, others are concerned with the fact that AIC does not directly depend upon sample size and, therefore, lacks properties of asymptotic consistency (Bozdogan, 1987). However, variations on AIC such as Schwarz's (1978) BIC and Bozdogan's (1987) CAIC do reflect sample size considerations. In practice, it is not necessarily the case that the property of asymptotic consistency leads to a better procedure in a true-model identification sense.

For example, in the context of comparing non-nested latent clas (mixture) models, Lin and Dayton (1997) found that AIC was superior to BIC when the "true model" was relatively complex (i.e., was based on a relatively large number of parameters). Similarly, Huang and Dayton (1995) report that, for multiple comparisons among bivariate mean vectors, AIC tended to outperform BIC and CAIC when "the null case was excluded and, in general, for heterogeneous cases." However, for multiple regression analysis, the results for AIC and BIC reported by Gagné and Dayton (2002) are more complex but consistent with the observation that AIC is more successful with more complex models.

Clearly, further research around the issue of competing information measures is needed but that does not alter the fact that this class of procedures often provides a highly desirable alternative to traditional significance testing techniques. Finally, it should be pointed out that information measures themselves depend upon certain asymptotic properties of chi-square statistics and, thus, issues of robustness must be considered. This is a researchable topic about which little is known at present. Of course, very similar distributional issues must be considered for significance tests
and, despite years of research, the best advice has always been to use large samples.

A technical point about the calculation of AIC (or CAIC or BIC) is that the loglikelihood, LL, often involves the estimation of theoretical variances. The maximum-likelihood estimate for a variance is biased since the denominator for the computation is the sample size, N , regardless of the number of parameters that are estimated in fitting the model to data. In regression analysis with p independent variables, for example, the unbiased estimate for the residual variance is computed by dividing the residual sum of squares by $\mathrm{N}-\mathrm{p}-1$ but in the context of computing AIC the divisor for the maximum likelihood estimate is N .

The computation of AIC for any specific model requires the specification of a distributional form (e.g., univariate normal, multivariate normal, multinomial, Poisson, etc.). Then, the log-likelihood, LL, for the sample is computed based on the model and the specified distributional form. In multiple regression analysis, for example, residuals may be assumed to follow a univariate normal density with variances that are homogeneous conditional on the independent variables.

However, unlike conventional significance tests, the set of alternate models being considered may include different specifications and different distributional assumptions. For example, residuals may be characterized as heterogeneous or dependent on the independent variables in various ways. On the other hand, residuals may be assumed to follow a mixture of homogeneous univariate normal densities. In any case, the $\min (A I C)$ criterion can be used to order and select among these models.

To illustrate these ideas in the context of real data, consider the plot (Figure 1) for mathematics achievement scores as a function of weekend television watching activity based on a $5 \%$ random sample of cases from the public use for the National Education Longitudinal Study (NELS). The distinct non-linear trend based on 1092 cases seems to invite a quadratic regression model (the television watching categories were coded at their upper values except that the final category was coded 6). Conventional F tests for increments to explained variability $\left(\Delta \mathrm{R}^{2}\right)$ using a
direct notation are $\mathrm{F}_{\text {linear }}=5.34, \mathrm{~F}_{\text {quad }}=41.05$ and $\mathrm{F}_{\text {cubic }}=1.30$. The linear and quadratic terms are significant at the conventional $5 \%$ level whereas the cubic term is non-significant. Thus, the three significance tests can be interpreted as supporting the selection of a quadratic model for the data. As reported in Gagné and Dayton (2002), the log-likelihood for homogeneous multiple regression models can be computed directly from the residual sum of squares $\left(\mathrm{SS}_{\mathrm{e}}\right)$ and sample size:

$$
\begin{equation*}
\mathrm{LL}=-.5 N \cdot\left[\ln (2 \pi)+\ln \left(\frac{S S_{e}}{N}\right)+1\right] . \tag{1}
\end{equation*}
$$

The AIC values for linear, quadratic and cubic models are, respectively, 8140.02, 8101.62 and 8127.30 leading to the choice of the quadratic model as the best approximating model among these three models (using BIC leads to the same preferred model). But, other models might be explored for these data. For example, using the reciprocal of weekend television watching as a predictor (actually, reciprocal of $\mathrm{X}+1$ due to the presence of 0 's), the AIC value is 8144.16 which is less preferred than any of the polynomial models. Note that from a conventional point of view, a test of significance can be run for the regression coefficient in the reciprocal regression model $(\mathrm{t}=-1.095, \mathrm{p}=.274)$ but there is no direct way of testing the difference in fit between, say, the linear model and the reciprocal model since they are not nested.


Figure 1. Mathematics achievement scores as a function of weekend television watching.

Paired Comparisons Information Criterion
Dayton (1998, 2003a) proposed a method for comparisons among means using information criteria such as Akaike's AIC. He advocated this approach rather than standard pairwisecomparison procedures such as Tukey tests in order to avoid or minimize the following problems with conventional procedures.
(a) Tukey tests (and variations) have been proposed based on some arbitrary method for controlling the family-wise type I error rate for the set of correlated pairwise contrasts. Release 11.5 of SPSS, for example, provides options for 18 different post hoc pairwise comparison procedures that are based on several different approaches to controlling type I error.
(b) Unequal sample sizes and heterogeneity of variance pose difficulties for many procedures. The classic Tukey test, for example, assumes constant sample size and homogeneous variances, an often unrealistic set of assumptions. Modifications of Tukey tests such as Games-Howell tests allow for both unequal sample sizes and heterogeneous variances but only provide approximate control of the familywise type I error rates by means of an adjustment to degrees of freedom.
(c) Intransitive decisions are routinely encountered with pairwise-comparison procedures in general and pose serious interpretive problems if some overall conclusion is desired for the set of means. For three means in rank order, an intransitive decision entails rejecting the difference between the highest and lowest mean but retaining the null hypotheses for comparisons of these means with the middle mean. It has been argued that this really doesn't pose a problem if the main concern of a study is to draw conclusions about the separate pairwise differences. However, if the focus is on individual pairwise contrasts, what rationale is there for sacrificing power and adopting a family-wise error rate rather than simply running separate $t$ tests for each pair of means?

The method based on information criteria described below and known as pairedcomparisons information-criterion, or PCIC, has been the topic of simulations by Cribbie \& Keselman (2003) who suggest that PCIC has allpairs power that is typically superior to standard pairwise comparison procedures (e.g., Tukey

HSD). The method has been extended to repeated observations as well as to data in the form of proportions.
(A) Independent Samples of Means

Consider a design comprising J independent, random groups of respondents with sample sizes, $\quad n_{j}$, sample means $\bar{Y}_{j}$ and unbiased variance estimates, $S_{j}^{2}$, with $N=\sum_{j=1}^{J} n_{j}$. In PCIC, AIC (or similar measure) is computed for each possible, different ordered subset of means. Thus, only non-overlapping subsets of means are compared rather than all possible subsets. In general, for J groups there are $2^{J-1}$ distinct patterns of subsets based on ordered means. For example, with three groups with the means ranked and labeled 1, 2, 3, the $2^{2}=4$ ordered subsets are $\{123\},\{1,23\},\{12,3\}$, and $\{1,2,3$, where a comma is used to separate subsets with unequal means. Focusing on ordered subsets of means and using a $\min (\mathrm{AIC})$ [or $\min (\mathrm{BIC})$ ] strategy avoids the intransitivity problem that may arise when using traditional pairedcomparisons techniques without sacrificing interpretability of results.

Assuming homogeneity of variance, the loglikelihood for the $\mathrm{m}^{\text {th }}$ model can be written as:

$$
\begin{align*}
L L_{m}= & -\frac{N}{2} \operatorname{Ln}(2 \pi)-\frac{N}{2} \operatorname{Ln}\left(\hat{\sigma}_{W}^{2}\right) \\
& \frac{1}{2 \hat{\sigma}_{W}^{2}} \sum_{j=1}^{J \quad n_{j}} \sum_{i=1}^{n_{i j}}\left(Y_{i j}\right)^{2} \tag{2}
\end{align*}
$$

where $\hat{\sigma}_{W}^{2}$ is computed from the ANOVA within-groups sum of squares but with denominator $N$ rather than $N-J$. Means for the $\mathrm{m}^{\text {th }}$ model are estimated assuming that the model is correct. The independent parameters estimated for a model comprise the variance and means, as necessary. If variances are assumed to be equal in the same pattern as means, the case is termed the restricted heterogeneous variance case (for other cases, see Dayton, 1998). Assuming the restricted heterogeneous variance case, an estimated variance for a subset of means can be obtained either by pooling
variance estimates as appropriate from the separate groups or by computing the (biased) sample variance from the appropriate combined group. For the latter preferred case, the sample variance for a $\{23\}$ subset of means, for example, would be

$$
\begin{equation*}
\hat{\sigma}_{23}^{2}=\frac{\sum_{i=1}^{n_{2}}\left(Y_{i 2}-\hat{\mu}_{23}\right)^{2}+\sum_{i=1}^{n_{3}}\left(Y_{i 3}-\hat{\mu}_{23}\right)^{2}}{\left(n_{2}+n_{3}\right)} . \tag{3}
\end{equation*}
$$

Assume that, for the $\mathrm{m}^{\text {th }}$ model, the pattern of sample means has been partitioned into K nonoverlapping subsets. Then,

$$
\begin{align*}
L L_{m} & =-\frac{N}{2}[\operatorname{Ln}(2 \pi)+1] \\
& -\frac{1}{2} \sum_{k=1}^{K} n_{m k} \ln \left(\hat{\sigma}_{m k}^{2}\right) \tag{4}
\end{align*}
$$

where $\hat{\sigma}_{m k}^{2}$ is the (biased) variance estimate and $n_{m k}$ is the sample size for the $\mathrm{k}^{\text {th }}$ subset.

Table 1 (following page) summarizes NELS data for standardized reading scores for five racial/ethnic as identified in the data base. Tukey tests, as well as Games-Howell tests that lack the homogeneity of variance assumption, yield a typical intransitive pattern of differences with three overlapping, non-significant ranges comprising, in rank order of means from high to low, $\{123\},\{34\}$ and $\{45\}$. The three smallest AIC values assuming homogeneity of variance and not making this assumption are shown in Table 1.

Note that $\min (A I C)$ occurs for the pattern $\{12,345\}$ assuming the restricted heterogeneous variance case although several models show quite similar AIC values. An interesting feature of model comparisons with AIC and related information measures is that, although a single preferred model is identified, a ranking of alternative models is provided. Additional illustrative analyses for both the homogeneous and heterogeneous cases are presented in Dayton $(1998,2003 a)$ as well as in connection with a Gauss program (Aptech Systems, 1997) for conducting these tests (Dayton, 2001a).

| Table 1NELS Reading Standardized Scores |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| "Race" | n | Mean | Variance | Homogeneity |  | Restricted Heterogeneity |  |
| White Non- | 798 | 52.55 | 98.21 | \{1,2,345\} | 8926.90 | \{12,345\} | 8926.62 |
| $\begin{aligned} & \text { Hispanic } \\ & \text { API } \end{aligned}$ | 75 | 50.40 | 97.66 | \{1,2,3,45\} | 8927.59 | \{1,2,345\} | 8927.37 |
| Hispanic | 140 | 47.36 | 92.13 | \{12,34,5\} | 8928.30 | \{12,3,45\} | 8927.56 |
| Black NonHispanic | 152 | 46.16 | 77.68 |  |  |  |  |
| American Indian | 44 | 46.00 | 70.39 |  |  |  |  |
|  | 1209 |  |  |  |  |  |  |

(B) Means Based on Repeated Observations

Consider a cohort of individuals that is measured on the same variable at several points in time. Assuming multivariate normality, the parameters of the distribution are means and variances for the occasions of measurement as well as covariances among occasions. As with independent observations, attention is focused on the distinct ordered subsets of means and AIC, or a related information measure, can be used to select a preferred pattern.

As for the case of independent groups, variances and covariances can be homogeneous, heterogeneous or restricted heterogeneous. However, the situation is more complex since these conditions can be applied separately to the variances, covariances or to both. In addition, various patterned covariance matrices may be considered to be appropriate (e.g., a simplex pattern with observations closer in time more highly correlated that those further apart in time). Dayton (2003a) presents more detailed information about this case along with illustrative data.
(C) Independent Samples of Proportions

Consider J groups of sizes $n_{j}$ with sample proportions, $p_{1}, p_{2}, \ldots, p_{J}$, for a dichotomous dependent variable. The theoretical model for the data is that responses represent a
series of $0 / 1$ Bernoulli trials with a true population probability, $\pi_{j}$, of a favorable outcome (e.g., 1 or positive) for the $\mathrm{j}^{\text {th }}$ group. The log-likelihood for any specific ordered outcome (e.g., 0110 for proportions based on four outcomes) in the $j^{\text {th }}$ group is $n_{j} p_{j} \ln \left(p_{j}\right)+n_{j}\left(1-p_{j}\right) \ln \left(1-p_{j}\right)$ and the loglikelihood for the total sample is found by summing across the J groups:

$$
\begin{equation*}
L L=\sum_{j=1}^{J}\left[n_{j} p_{j} \ln \left(p_{j}\right)+n_{j}\left(1-p_{j}\right) \ln \left(1-p_{j}\right)\right] . \tag{5}
\end{equation*}
$$

Note that $n_{j} p_{j}$ is the expected number of favorable outcomes and $n_{j}\left(1-p_{j}\right)$ is the expected number of unfavorable outcomes. The sample proportion, $p_{j}$, is the MLE for the corresponding population proportion. Unlike the situation for sample means, there is no need to consider homogeneous and heterogeneous cases since each Bernoulli process is based on a single parameter, $\pi_{j}$. Otherwise, model selection follows the same reasoning as for independent sample means (Dayton, 2001a). That is, there is a total of $2^{J-1}$ distinct patterns of subsets of proportions to evaluate and proportions for a model are estimated assuming that the model is correct. Illustrative analyses for this case are presented in Dayton (2001a, 2003a).

## PCIC for Distributions

Standard pairwise comparison procedures, such as Tukey HSD and its many variations, have been the subject of a good deal of research directed toward assessing their robustness with respect to distributional assumptions. Typically, non-normal distributions with varying degrees of skew and kurtosis are selected for comparison (e.g., Keselman, Lix \& Kowalchuk, 1998, report simulations with normal distributions, three-degree-of-freedom chi-square distributions and a highly non-normal distribution with skewness and kurtosis indices equal to 6.2 and 114, respectively). The issue, then, is the degree of sensitivity of the multiple comparison procedures to departures from normality. Also, a number of simulations have dealt with the relative power of pairwise comparison procedures (e.g., Ramsey, 2002).

An alternative approach is to directly model the underlying distributions for observed data and then compute appropriate likelihoods for candidate distributions of interest. Once these distributions have been selected, procedures comparable to PCIC can be implemented. In practice, identifying the set of candidate distributions is a non-trivial problem. Two classes of plausible models that have credibility in practice than can be compared are normal and log-normal densities.

The motivation for log-normal models arises from the fact that, in contrast to an additive effect, a multiple effect for an independent variable can be modeled in loglinear terms. For example, the usual additive model for a response in a one-way ANOVA design can be represented as $Y_{i j}=\mu+\tau_{j}+\varepsilon_{i j}$ where $\mu$ is a grand mean effect, $\tau_{j}$ is the effect of the $\mathrm{j}^{\text {th }}$ treatment and $\varepsilon_{i j}$ is a residual error term. Alternatively, assuming a multiplicative, rather than an additive treatment effect, yields the model: $Y_{i j}=\mu \times \tau_{j} \times \varepsilon_{i j} \quad$ or $\ln \left(Y_{i j}\right)=\mu^{*}+\tau_{j}^{*}+\varepsilon_{i j}^{*}$ where the * superscript denotes a parameter on a logarithmic scale. In practice, many positively skewed distributions of observations are reasonably well approximated by log-linear models.

Some preliminary simulation results have been carried out for two-sample and a limited number of three-sample cases to assess how well the AIC and BIC information measures distinguish between samples based on normal and log-normal distributions (Dayton, 2003b). In one series of simulations, theoretical log-normal densities with means, standard deviations of $(0, .1),(0, .5)$ and $(0,1.0)$ in log units corresponding to $(1.00, .10),(1.13, .60)$ and $(1.65,2.16)$ in raw units were considered. The first distribution is slightly skewed (index $=$ .30) and modestly kurtotic (index $=3.16$ ), the second distribution is moderately skewed (index $=1.75$ ) and somewhat peaked (index $=8.89$ ), while the third distribution is both highly skewed (index $=6.18$ ) and highly kurtotic (index $=113.94$ ). In a second series of simulations, information criteria were compared assuming only log-normal densities but the generated data were either normal or log-normal.

Typical results for two groups are, in additional to the expected sample size differences: (a) BIC selected the correct model more often than AIC in virtually all simulated cases with an average difference ranging from about $6 \%$ to $13 \%$; (b) both information criteria were much more successful in selecting models when the true distribution was log-normal as opposed to when it was normal. This latter result occurs because, as the median increases, lognormal distributions assume a nearly symmetric shape that approximates normality. Limited results for three samples suggest that, as was true for two groups, BIC tends to select the correct pattern of means more often than does AIC and both criteria were more successful for log-normal than for normal distributions. The superiority of BIC over AIC should not be generalized at this time, however, since Dayton (1998) found for cases with several groups that neither criterion was uniformly superior to the other.

Number of Components in Mixture models
An emerging area of interest in applied research is the use of finite mixture models when distributions such as normal, Poisson and binomial fail to provide satisfactory fit to data. An impetus for considering mixtures is the phenomenon of over-dispersion which is
manifested by, for example, distributions with "heavy tails." For situations of this sort it is often reasonable to assume that observations represent a mixture from two or more subpopulations rather than arising from a single population. In general, a mixture of J distributions for some dependent variable, Y, can be represented by: $g(Y \mid \beta)=\sum_{j=1}^{J} \theta_{j} \times g_{j}\left(Y \mid \beta_{j}\right) \quad$ where $\quad \theta_{j} \quad$ are mixing fractions such that $\sum_{j=1}^{J} \theta_{j}=1, g()$ is some specified probability (e.g., binomial) or density (e.g., normal) function based on a vector of parameters, $\beta_{j}$, and $\beta$ is a vector containing all relevant parameters.

For a mixture of two heterogeneous normal densities, for example, $g_{1}\left(Y \mid \mu_{1}, \sigma_{1}^{2}\right)$ and $g_{2}\left(Y \mid \mu_{2}, \sigma_{2}^{2}\right)$ would represent normal densities with unique means and variances that are mixed in proportions $\theta_{1}$ and $\theta_{2}=1-\theta_{1}$. To fit such models to data, the parameters for the separate components as well as mixing fractions must be estimated. For a mixture of two normal densities this would entail estimating five unique parameters (two means, two variances and one mixing proportion). Some relatively simple mixtures (e.g., normal densities) can be estimated using available statistical software such as Mplus (Muthén and Muthén, 1998) but specialized programs such as LEM (Vermunt, 1993) are required in more complex cases such as latent class models.

A persistent dilemma for applications of mixture models is that models with varying numbers of components cannot be compared using conventional significance tests even though these models are hierarchical. For example, the comparison of a mixture of two normal densities to a single normal density could, seemingly, be based on a difference-chisquare test since the single normal density is a restricted form of the mixture (e.g., by setting $\theta_{2}=0$ ). However, as noted by Everitt and Hand (1981) and Titterington, Smith and Makov (1985), among others, this difference-chi-square statistic fails to satisfy theoretical requirements
related to boundaries of the parameter space and is not distributed as expected (nor is its distribution known). Some insight into the problem can be seen from observing that the single restriction, $\theta_{2}=0$ is equivalent to the two restrictions $\mu_{1}=\mu_{2}$ and $\sigma_{1}^{2}=\sigma_{2}^{2}$ since, in either case, the resulting model is a single normal density. In fact, the mixture is based on five parameters whereas the single normal distribution is based on only two parameters, yet only one restriction is required to obtain the simpler from the more complex model.

Given the failure of conventional significance tests to provide a basis for assessing the number of components in a mixture, information measures such as AIC present an attractive alternative. Information criteria provide a single summary statistic for each model being compared and avoid the asymptotic distributional issues faced by difference-chisquares tests for mixture models. Some preliminary work on assessing AIC, BIC and related measures was reported by Dayton (2001b) who focused on the issue of selecting the appropriate number of mixtures in binomial models (restricted latent class models) with four and six binary variables.

Simulations were based on samples sizes ranged from 80 to 1280 , binomial probabilities for mixtures of two and three processes were selected to represent varying degrees of discriminability of the components and mixing proportions were varied from equal splits to cases where one component represented only $20 \%$ of the cases. Cases with high discriminability involved, for two components, cases with binomial probabilities and $.1, .5$ and .1, 8 where low discriminability involved cases with binomial probabilities of $.1, .2$. All of the measures studied provided reasonable correct identification rates for the high disciminability cases (e.g., $80 \%$ and above across the conditions) but very poor correct identification rates for the low disciminability cases (e.g., $10 \%$ or less across the conditions). Dayton (2001b) concludes that this area of analysis requires "...reasonably large sample sizes and the realization that poorly defined latent structures will almost certainly go undetected."

## Conclusion

Although the recommendation has been repeated often in the past, researchers should become aware of modern alternatives to the use of significance tests when comparing alternate models is the focus of analysis. Information theoretical procedures such as Akaike AIC provide a holistic approach to ordering and selecting among competing models that avoids the piece-meal and potentially inconsistent outcomes that arise from applying multiple significance tests. This paper has summarized applications of these measures to multiple comparisons including the possibility of varying distributional assumptions and to mixture models where traditional significance tests are known to be inappropriate.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csake (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics, 30, Part A 9-14.

Aptech Systems, Inc. (1997). GAUSS for Windows NT/95: Version 3.2.32, Maple Valley, WA.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52, 345-370.

Carver, R. P. (1993). The case against statistical significance testing, revisited. The Journal of Experimental Education, 61, 287292.

Cribbie, R. A. \& Keselman, H. J. (2003). A power comparison of pairwise multiple comparison procedures: A model testing approach versus stepwise procedures. British Journal of Statistical \& Mathematical Psychology, 56, 157-182.

Dayton, C.M. (1998). Information criteria for the paired-comparisons problem. American Statistician, 52, 144-151.

Dayton, C. M. (2001a). SUBSET: Best subsets using information criteria, Journal of Statistical Software, Vol 6., Issue 02, April.

Dayton, C. M. (2001b). Performance of information criteria for number of components for product-binomial processes. Paper presented at the Mixtures 2001: Recent Developments in Mixture Modeling conference at Universitat der Bundeswehr, Hamburg, Germany, July.

Dayton, C. M. (2003a). Information criteria for pairwise comparisons. Psychological Methods, 8, 61-71.

Dayton, C. M. (2003b). A modeling approach to post-hoc comparisons of means and proportions. Paper presented at Second Workshop on Research Methodology (RM2003), Vrije University, Amsterdam, The Netherlands, June.
de Leeuw, J. (1992). Introduction to Akaike (1973) Information theory and an extension of the maximum likelihood principle. In S. Kotz \& N. L. Johnson (Eds) Breakthroughs in Statistics Volume I Foundations and Basic Theory. New York: Springer-Verlag.

Everitt, B. S. \& Hand, D. J. (1981). Finite Mixture Models. New York: Chapman \& Hall, Ltd.

Fisher, R. A. (1959). Statistical Methods and Scientific Inference (2 ${ }^{\text {nd }}$ Edition). New York: Hafner.

Gagné, P. \& Dayton, C.M. (2002). Best regression model using information criteria. Journal of Modern Applied Statistical Methods, 1, 479-488.

Heermann, E. \& Braskamp, L. A. (editors) (1970). Reading in Statistics for the Behavioral Sciences. New Jersey: Prentice-Hall.

Huang, C-C \& Dayton, C.M. (1995). Detecting patterns of bivariate mean vectors using model-selection criteria. British Journal of Mathematical \& Statistical. Psychology, 48, 129-147.

Keselman, H. J., Lix, L. M., \& Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.

Kullback, S. \& Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22, 79-86.

Lin, T. S. \& Dayton, C. M. (1997). Model-selection information criteria for nonnested latent class models. Journal of Educational and Behavioral Statistics, 22, 249264.

McDonald, R. P. \& Marsh, H. W. (1990). Choosing a multivariate model: noncentrality and goodness of fit. Psychological Bulletin, 107, 247-255.

Muthén, L. K., and Muthén, B. O. (1998), Mplus User's Guide. Los Angeles: Muthén and Muthén.

Neyman, J. \& Perarson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society (A), 231, 289337.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. Psychological Methods, 7, 504-523.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. Psychometrika, 52, 333343.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62, 157176.

Titterington, D. M. Smith, A. F. M. \& Makov, U. E. (1985). Statistical Analysis of Finite Mixture Models. New York: John Wiley \& Sons.

Vermunt, J. K. (1993). Log-linear \& event history analysis with missing data using the EM algorithm. WORC Paper, Tilburg University, The Netherlands.

# Fortune Cookies, Measurement Error, And Experimental Design 

Gregory R. Hancock<br>University of Maryland



This article pertains to the theoretical and practical detriments of measurement error in traditional univariate and multivariate experimental design, and points toward modern methods that facilitate greater accuracy in effect size estimates and power in hypothesis testing.

Keywords: measurement error, latent variables, multivariate analysis, experimental design

## Introduction

Whichever leg of my post-secondary academic journey, and with whichever campus I have had the privilege of affiliating, the vast majority of my midday meals have ended with a fortune cookie. Since my college days, in fact, I estimate that I have had lunch at some inexpensive Asian restaurant near campus well over a thousand times. My graduate school office mates and the many students and faculty whom I served as teaching assistant might even remember all the little strips of paper taped to the top of my desk, filling the entire surface with fortunes by the time I finished my doctorate.

Gregory R. Hancock is Professor in the Department of Measurement, Statistics and Evaluation at the University of Maryland. His research appears in such journals as Psychometrika, Structural Equation Modeling, and Journal of Educational and Behavioral Statistics. He serves on several journal editorial boards, and regularly conducts workshops around the U.S. Email: ghancock@umd.edu.

Today, a little more reserved in my decorative zeal, though no less so in my meal predilection, I have but a single fortune tacked outside of my office door. Amidst aging cartoons and family pictures is an enlarged photocopy of the one little rectangle of wisdom I have saved over these last decades. It reads:


Although my quantitative training precludes me from seeking fortune based on the third line, not so with the first two. Their aphorism seems replete with insight and potential on many levels, personal and professional, with the latter level serving as the inspiration for this article.

Less obtusely, in so many applied statistical analyses there seems to be a schism between the variables we have and the variables we wish we had. This is apparent in statements of theory preceding and justifying those analyses and in the interpretations and purported implications that follow. Educational policy researchers, for example, might analyze measures of teacher's job satisfaction and absenteeism and then make proclamations
regarding the apparent degree of teacher burnout. Those studying child development might start by eliciting new mothers' responses to rating scale items regarding interactions with their infants, and conclude by making inferences about those mothers' emerging maternal warmth. Health care researchers might want an understanding AIDS patients' sense of hopelessness while in group therapy, and choose measures of patients' treatment compliance to help facilitate that understanding. Such is the nature of so much applied research, particularly within the social sciences - constructs of interest such as burnout, maternal warmth, or hopelessness are generally latent, so our analyses seem resigned to rely upon the fallible measured variables as surrogates.

And therein lies the schism, in the operationalization of true constructs as errorladen measured variables. At best the imperfect connection might lead us to a distorted image of the critical relations in a population; at worst we might not even have sufficient power to draw inference at all. Within the context of experimental design specifically, the primary focus of this treatise, the implication is that treatment effectiveness might be severely underestimated, or perhaps even undetected. Of course this is not unknown. In fact, nothing written here will be new knowledge. But it is important and often-overlooked knowledge, bearing clarification and amplification. It will thus be my purpose to drive home the often underestimated (if not entirely disregarded) importance of constructs and measurement error in our univariate and multivariate experimental analyses, and to point the applied researcher toward more modern strategies for dealing with measurement error in experimental design.

## Love Truth

The purpose of applied statistics seems to be to gain insight into some truth bearing practical consequence. Drawing upon a few familiar test statistics, we attempt to use observed relations among measured variables in samples to make educated guesses about unobserved relations in the populations of which each sample serves as assumed microcosm. But what, precisely, is the population relation we hope to understand in order to have practical
consequence? What is the truth into which we seek insight?

As we learn and practice the many methods huddled under the general linear model umbrella, we typically hold as our goal a correct inference about, and often estimation of, some population relation among observed variables a true correlation between $X$ and $Y\left(\rho_{X Y}\right)$, a true predictive relation of $X_{3}$ to $Y$ holding $X_{1}$ and $X_{2}$ constant ( $\beta_{3}$ ), a true standardized effect size for the mean difference between Populations 1 and 2 on $Y\left(d_{Y}\right)$, and so on. But what does any measured $X$ or $Y$ variable really represent, and what information do any relations among such variables convey?

In the physical sciences, variables such as temperature, pressure, mass, and volume, when considered in sufficient quantities, are in their measurement as they are in name. That is, there tends to be a strong correspondence between the measurement and the entity it represents. In the social sciences, some such variables exist as well - biological sex, treatment group assignment, and political party affiliation, for example. Except for data recording or entry errors, we expect each variable to represent precisely that which its name implies. Other social science variables would also seem to have such identity, being determinable largely without interference - number of therapy sessions attended, number of children's books in the home, and the like. However, a fundamental question in many disciplines, particularly those in the social sciences, is the following: What is the underlying construct that each variable has been selected to represent?

The univariate scenario
Consider a researcher who is truly interested in a construct contrived here as InHome Reading Resources. In that case, number of children's books in the home is indeed a fairly proximal operationalization of the construct of interest. As such, estimates regarding population mean differences in number of children's books in the home, or regarding the population relations this measured variable has with other such proximal operationalizations, provides direct insight into some truth for the construct of In-Home Reading Resources. On the other hand, if a researcher is interested in a construct
designated as Parental Commitment to Literacy, and has attempted to capture the spirit of this construct using the number of children's books in the home, then we expect the measured variable to be a more distal operationalization of the desired construct. As such, inference about population differences in Parental Commitment to Literacy, or of its relation to other variables (proximally or distally operationalized), is compromised by typical general linear model analyses. Thus, the truths we seek - constructs and their population relations - are often not directly accessible.

The issue at hand, of course, is one of measurement error in our variables. As an indicator of In-Home Reading Resources, number of children's books in the home has virtually no measurement error; when reflecting Parental Commitment to Literacy, however, it has considerable error. Imagine a researcher employing a control and treatment group to draw inference about the impact of a treatment designed to enhance Parental Commitment to Literacy. Figure 1 displays hypothetical population distributions for the measured variable of number of children's books in the home ( $Y$ ), as well as for the latent construct of Parental Commitment to Literacy ( $\eta$ ). Notice that while the means of the two populations are
expected to be the same for $Y$ and $\eta$, the relative magnitude of the treatment effect on the Parental Commitment to Literacy construct would be underestimated. The standardized effect size for $Y$, which is the familiar

$$
\begin{equation*}
d_{Y}=\left(\mu_{1 Y^{-}} \mu_{2 Y}\right) / \sigma_{Y} \tag{1}
\end{equation*}
$$

(Cohen, 1988), is depicted as approximately .65; meanwhile, the standardized effect size for $\eta$,

$$
\begin{equation*}
d_{\eta}=\left(\mu_{1 \eta^{-}} \mu_{2 \eta}\right) / \sigma_{\eta} \tag{2}
\end{equation*}
$$

is near .95. For this disparity to occur, the construct's standard deviation would have to be $68.4 \%$ of the size of standard deviation of $Y$, meaning its variance is roughly $46.8 \%\left(.684^{2}\right)$ that of $Y$. That is, $46.8 \%$ of the variability in $Y$ reflects $\eta$, while $53.2 \%$ is error with respect to the construct of interest. Put directly,

$$
\begin{equation*}
d_{Y}^{2}=\rho_{Y Y} d_{\eta}^{2} \tag{3}
\end{equation*}
$$

where $\rho_{Y Y}$ is the reliability of $Y$ (.468 in the above example). Thus, while the number of children's books in the home may accurately reflect In-Home Reading Resources, with regard to Parental Commitment to Literacy it could be a relative overestimate or underestimate for any given individual.

Figure 1. Univariate population difference on measured variable and underlying construct.


As mentioned previously, the implications of measurement error for inference are two-fold. First, as seen in Figure 1, we would underestimate the magnitude of the treatment effect on Parental Commitment to Literacy. That is, we would make an incorrect estimate of the truth we seek. Second, the presence of the error variance would decrease the power of a twosample test to detect the presence of that treatment effect. An understanding of this loss of power may be communicated in terms of additional subjects needed in each group to compensate for the presence of measurement error. Assuming a desired level of power (e.g., .80) and a specific standardized effect size at the construct level (e.g., $d_{\eta}=.30$ ), the number of subjects per group for a two-sample $z$-test can easily be shown to be:

$$
\begin{equation*}
n_{Y}=\left(1 / \rho_{Y Y}\right) n_{\eta} . \tag{4}
\end{equation*}
$$

For example, conducting the test using a valid measure with reliability of $\rho_{Y Y}=.50$ would require twice as many subjects as a test that could, hypothetically, be conducted directly at the construct level. This result holds for $t$-tests as well for all but the smallest sample sizes, where appreciable changes in the critical value make the relation only approximate. Further, except for very small samples, Equations 3 and 4 hold for $k$-group between-subjects analysis of variance (ANOVA) as well using the more general $k$-group effect size measures (see Cohen, 1988).

The scenario for the univariate outcome may also be depicted symbolically using a path diagram. In Figure 2 we see the measured variable $Y$ being defined by two components, the construct of interest $\eta$ and measurement error $\varepsilon$. The connection between $\eta$ and $Y$, labeled as $\lambda$ in Figure 2, symbolically reflects the (square root of the) measured variable's reliability. The stronger the relation $\lambda$, the more proximal $Y$ 's operationalization of $\eta$ and thus the less error variance it contains; conversely, the weaker $\lambda$, the more distal $Y$ 's operationalization of $\eta$ and thus the more error variance it contains. On the left we see a grouping variable representing population membership and whose influence is being assessed; this could be a single variable for $k=2$ groups, or $k-1$ group code variables for the general $k$-group case.

As depicted, population membership $X$ has a potential bearing $\gamma$ on the construct $\eta$ underlying the measured variable $Y$, while the remaining variance in $\eta$ is accounted for by other independent but latent residual influences $\zeta$. Thus, an observed population difference on the measured variable $Y$ is actually the attenuated manifestation of a population difference on the true underlying construct of interest. The weaker the connection between the $\eta$ and $Y$ (i.e., the weaker the reliability), the less well the population difference on the construct of interest is propagated to, and thus reflected in, the observed variable.

Figure 2. Path model for univariate case


This simple univariate example underscores two needs regarding truth in experimental design and analysis. First, we must seek measured operationalizations as proximal to their constructs as possible. Certainly in the social sciences perfect operationalization is generally unrealistic, particularly given the vagaries of human behavior, perception, affect, and attitude. Notwithstanding, researchers should expend considerable effort to select or construct the most valid and reliable measures feasible. Second, to the extent that measurement error remains, we must employ analytic methods that maximize the accuracy of inference and estimation, thereby portraying population truths with the greatest clarity. These analytic methods must, to every extent possible, penetrate the measurement noise to achieve the same fidelity to truth as the theoretical questions that preceded and the practical proclamations we hope to follow. One attempt to do so lies within a multivariate scenario.

The multivariate scenario
Researchers often attempt to enhance their ability to make inference about population differences by gathering several pieces of evidence to be employed within a multivariate experimental design. In multivariate analysis of variance (MANOVA) with outcome measures $Y_{1}$ through $Y_{m}$, the hope is that the signal of population differences on some combination of variables will be detected above the noise of their measurement error. This portion of the article will address MANOVA in the presence of measurement error, and highlight its somewhat misguided attempt to get closer to truth.

Consider the multivariate scenario with $k=2$ populations, often analyzed using Hotelling's $T^{2}$. An example is depicted in Figure 3 using $m=2$ outcomes for simplicity, and with extremely large population differences for clarity. As before, assume that each $Y_{i}$ measure is an operationalization of its own specific construct $\eta_{i}$, with individual standardized effect sizes of $d_{Y_{i}}$ and $d_{\eta_{i}}$ for the univariate measured and latent population mean differences, respectively. The assessment of the multivariate population difference between centroids $\mu_{Y 1}$ and $\mu_{Y 2}$ is tantamount to evaluating the univariate
mean difference on the maximally differentiating discriminant function $W=w_{1} Y_{1}+$ $w_{2} Y_{2}=\mathbf{w} \mathbf{Y}$, with weights $\mathbf{w}$ commonly (but not necessarily) chosen so the within-group variance $\sigma_{W}^{2}=\mathbf{w}^{\prime} \Sigma_{Y} \mathbf{w}$ equals 1. Observed and latent variable distributions on each $Y_{i}$ axis, as well as on the $W$ axis, are depicted in Figure 3.

Given that $W$ is a linear combination of the observed variables, the measurement error of each $Y_{i}$ is propagated to the linear composite $W$. The standardized effect size along the $W$ axis, the effect of interest in MANOVA, is $d_{W}=\left(\mu_{1 W^{-}}\right.$ $\left.\mu_{2 W}\right) / \sigma_{W}$; it appears as approximately 3 . The square of this effect size, $d_{W}^{2}$, may be computed as the squared Mahalanobis' distance

$$
\begin{equation*}
D_{W}^{2}=\left[\boldsymbol{\mu}_{1 Y}-\boldsymbol{\mu}_{2 Y}\right]^{\prime} \boldsymbol{\Sigma}_{Y_{\text {within }}}^{-1}\left[\boldsymbol{\mu}_{1 Y}-\boldsymbol{\mu}_{2 Y}\right], \tag{5}
\end{equation*}
$$

where $\boldsymbol{\Sigma}_{Y_{\text {within }}}$ is the pooled (within-group) variance-covariance matrix reflecting the observed $Y_{i}$ measures' m-dimensional dispersion and within lurks the influence of measurement error. Specifically, $\quad \boldsymbol{\Sigma}_{Y_{\text {within }}}=\boldsymbol{\Sigma}_{\eta_{\text {within }}}+\boldsymbol{\Sigma}_{\varepsilon_{\text {within }}}$, where $\boldsymbol{\Sigma}_{\eta_{\text {witinin }}}$ is the pooled (within-groups) variance-covariance matrix of the specific constructs $\eta_{i}$ and $\boldsymbol{\Sigma}_{\varepsilon_{\text {within }}}$ is a diagonal matrix of within-group error variances, assumed independent and each equal to $\sigma_{Y_{i}}^{2}\left(1-\rho_{Y_{i} Y_{i}}\right)$. Thus,

$$
\begin{equation*}
D_{W}^{2}=\left[\boldsymbol{\mu}_{1 Y}-\boldsymbol{\mu}_{2 Y}\right]^{\prime}\left[\boldsymbol{\Sigma}_{\eta_{\text {wixhin }}}+\boldsymbol{\Sigma}_{\varepsilon_{\text {within }}}\right]^{-1}\left[\boldsymbol{\mu}_{1 Y}-\boldsymbol{\mu}_{2 Y}\right] . \tag{6}
\end{equation*}
$$

As seen in Figure 3, the population mean difference on the $W$ axis mirrors the univariate case, where the standardized effect size on the measured composite $W$ underestimates the standardized effect size on the corresponding underlying construct. In this case, the construct underlying $W$, denoted here as $\eta_{W}$, is a linear combination of the $\eta_{i}$ constructs underlying the respective measured $Y_{i}$ variables: $\eta_{W}=w_{1} \eta_{1}+$ $w_{2} \eta_{2}=w^{\prime} \eta$, where $\eta$ is the vector of $\eta_{i}$ constructs. Whereas the measured standardized


Figure 3. Multivariate population difference on measured variables and underlying constructs.
effect size on $W$ was depicted as near 3, the latent standardized effect size for $\eta_{W}$, $d_{\eta_{W}}=\left(\mu_{1 \eta}-\mu_{2 \eta}\right) / \sigma_{\eta_{W}}$, is approximately 5 . The square of this effect size, $d_{\eta_{w}}^{2}$, is also the squared Mahalanobis’ distance

$$
\begin{equation*}
D_{\eta_{w}}^{2}=\left[\boldsymbol{\mu}_{1 \eta}-\boldsymbol{\mu}_{2 \eta}\right]^{\prime} \boldsymbol{\Sigma}_{\eta_{\text {vixitin }}}^{-1}\left[\boldsymbol{\mu}_{1 \eta}-\boldsymbol{\mu}_{2 \eta}\right] \tag{7}
\end{equation*}
$$

which corresponds to Equation 6 with the error variance $\boldsymbol{\Sigma}_{\varepsilon_{\text {witini }}}$ removed. In fact, the reliability of the composite $W$ could be determined as $D_{W}^{2} / D_{\eta_{W}}^{2}$, which is just a multivariate restatement and rearrangement of Equation 3.

To get a sense of the impact of measurement error on the multivariate effect size, consider a simple scenario in which the $\eta_{i}$ constructs are uncorrelated (and hence so too are the $Y_{i}$ variables). In this case the matrix $\boldsymbol{\Sigma}_{Y_{\text {vititin }}}$ is
diagonal, and thus Equation 5 may be shown to simplify to

$$
\begin{equation*}
D_{W}^{2}=\mathbf{d}_{Y}{ }^{\prime} \mathbf{d}_{Y}=\sum_{i=1}^{m} d_{Y_{i}}^{2} \tag{8}
\end{equation*}
$$

where $\mathbf{d}_{Y}$ is the vector of standardized effect sizes for $Y_{i}(i=1 \ldots m)$ as per Equation 1. The same logic would also yield a parallel result for the latent effect size:

$$
\begin{equation*}
D_{\eta_{W}}^{2}=\mathbf{d}_{\eta}^{\prime} \mathbf{d}_{\eta}=\sum_{i=1}^{m} d_{\eta_{i}}^{2} \tag{9}
\end{equation*}
$$

where $\mathbf{d}_{\eta}$ is the vector of latent standardized effect sizes for $\eta_{i}(i=1 \ldots m)$ as per Equation 2. Taking each $Y_{i}$ variable’s measurement error into account following Equation 3, Equation 8 yields

$$
\begin{equation*}
D_{W}^{2}=\sum_{i=1}^{m} d_{\eta_{i}}^{2}\left(\rho_{Y_{i} Y_{i}}\right) \tag{10}
\end{equation*}
$$

If all $Y_{i}$ variables were of the same reliability $\rho_{Y Y}$, it further follows that

$$
\begin{equation*}
D_{W}^{2}=\left(\rho_{Y Y}\right) D_{\eta_{W}}^{2} \tag{11}
\end{equation*}
$$

(again, given uncorrelated $\eta_{i}$ constructs and homogeneous reliabilities). Assuming a desired level of power (e.g., .80) and a specific effect size at the latent multivariate level (e.g., $d_{\eta_{W}}=.30$ ), the number of subjects per group for a two-sample test can be shown to be inversely proportional to measured variable reliability for all but the smallest sample sizes (in this highly restrictive example). That is,

$$
\begin{equation*}
n_{W}=\left(1 / \rho_{Y Y}\right) n_{\eta_{W}} . \tag{12}
\end{equation*}
$$

More generally, given any correlational pattern among the $\eta_{i}$ constructs (and resulting attenuated correlations among the $Y_{i}$ variables), the resulting reliability $\rho_{W W}$ of the composite $W$ would yield the corresponding relation

$$
\begin{equation*}
n_{W}=\left(1 / \rho_{W W}\right) n_{\eta_{W}} . \tag{13}
\end{equation*}
$$

Thus, the more reliable the composite $W$, the more MANOVA's power tends toward that of a theoretical test directly on the underlying construct.

In the univariate case, two implications of measurement error were highlighted: underestimating the magnitude of the treatment effect on the underlying construct of interest, and decreased power to detect the treatment effect. As illustrated, these hold as well for the multivariate case. However, while we may tend to gain power by accommodating multiple measured variables simultaneously, it is here that we must remind ourselves of our purpose, of precisely what truth we seek. That is - what, exactly, is the construct of interest in MANOVA?

Figure 4, a conceptual path diagram for the multivariate case, will help this discussion. On the left is a group code variable (e.g., dummy) representing population membership and whose influence is being assessed. As depicted, population membership has a potential
bearing on the $\eta_{i}$ constructs underlying the measured $Y_{i}$ variables. Portions of the constructs not explained by population membership are represented in the latent residual influences $\zeta_{i}$, which are likely to be correlated (shown in Figure 4 by shared two-headed arrows). Population differences on the measured variables are the observable manifestations of differences on the true underlying constructs of interest. The connection between each $\eta_{i}$ and $Y_{i}$ reflects the (square root of the) reliability of each variable; the weaker such a relation the less well the population differences on a construct are propagated to, and thus reflected in, the observed variables. As a result of each variable's imperfect operationalization of its construct, error $\varepsilon_{i}$ contributes to the variable as well. Finally, in the case of multiple outcomes, a discriminant function $W$ is represented as a composite of the measured variables. The weights determining this composite are optimal in the sense that they maximize the relation between $W$ and $X$. Note that $W$, as a weighted sum of measured variables, is also a weighted sum of constructs and errors. That is, unless all variables are perfect operationalizations of their constructs, the composite $W$ will contain measurement error which thus hampers its ability to reflect population differences propagated by $X$.


Figure 4. Path model for multivariate case, with $m$ constructs.

So if $W$ contains measurement error, with respect to what construct does that measurement error exist? The answer, as utilized previously, is the composite implicitly formed by MANOVA
of the constructs underlying the variables. But what truth does a composite of univariate constructs represent? To this critical question there seems to be three common answers, none of which is entirely satisfactory. Each will be presented in turn, along with the concerns it inspires.

Position 1: The composite is not itself intended to be a construct; rather, it is merely a vehicle for the simultaneous examination of the $m$ individual constructs of interest.

Response 1: If the separate constructs are of interest, then a MANOVA is inconsistent with that interest. Rather, a collection of individual ANOVAs, however seemingly inelegant, would address each construct directly. An omnibus MANOVA is not generally appropriate as a Type I error control mechanism since a single false univariate null hypothesis renders the multivariate null hypothesis false, and thus control over other true univariate ("partial") nulls becomes ungoverned. If one wishes to invoke an error control mechanism at the level of the constructs of interest, such as that of Bonferroni or his descendants, it may be applied across ANOVAs.

Position 2: The univariate constructs are facets of a single meaningful whole, as represented by the discriminant function and upon which knowledge of population differences is sought.

Response 2: Measured variables having a deterministic and defining bearing on a construct have been referred to as constituting an emergent variable system (e.g., Bollen \& Lennox, 1991; Cohen, Cohen, Teresi, Marchi, \& Velez, 1990). For example, one could imagine an unmeasured construct representing stress, contributed to and defined by such variables as relationship with parents, relationship with spouse, and demands of the workplace. In this case population differences in stress might indeed be of interest.

However, the formation of the discriminant function is not done in a manner reflecting any relative theoretical contributions of the three measured variables. If population differences existed only in terms of demands of the workplace, for example, then the discriminant function would be composed of only that variable. But does that then mean that
stress is only a function of demands in the workplace? Surely not. Thus, while the notion is reasonable that variables combine to define a composite with a meaningful underlying construct, those variables' combination is not informed by the theoretical soundness of the construct, but rather only by measured variable mean differences. Forming a meaningful composite and then conducting an ANOVA on the resulting scores would seem more consistent with the beliefs underlying this variable system.

Position 3: The univariate constructs are actually a single meaningful underlying construct; the discriminant function represents that construct and allows for the assessment of population differences thereon.

Response 3: Contrary to the emergent variable system described in Response 2, the variable system here is latent. That is, all measured variables are believed to be undergirded by the same construct (but perhaps varying in the quality of their reflection), and it is on this common construct that inference is desired. Still, although a single construct exists, MANOVA remains clouded in its ability to address this construct directly.

Consider Figure 5, where $X$ codes population membership and has a potential bearing $\gamma$ on the common construct $\eta$ underlying the measured $Y_{i}$ variables. Thus, population mean differences on the measured variables are the observable manifestations of a population difference on the true underlying construct of interest. Again, the connections between the $\eta$ construct and $Y_{i}$ variables ( $\lambda_{i}$ ) embody the (square root of the) reliability of each variable; the weaker such a relations the less well the group differences will be reflected in the observed variables. Finally, the discriminant function $W$ is again shown as an optimal composite of the measured variables, where every variable in the composite contributes some part $\eta$ and some part $\varepsilon_{i}$. So the discriminant function has succeeded to some extent in being a reflection of a construct of interest; however, it has still failed to eradicate error.

Further, the function has used group mean differences to guide its definition rather than proximity of construct operationalization. Thus, even if a single common construct underlies the
measured variables, measurement error within this multivariate approach will continue to compromise the accuracy of a treatment effect's assessment as well as the power to detect that effect. That is, we must continue the search for methods that attempt to pardon error.

$$
\begin{equation*}
\mathbf{Y}=\Lambda \eta+\varepsilon \tag{14}
\end{equation*}
$$

where $\mathbf{Y}$ is a subject's $m \times 1$ vector of $Y_{i}$ scores, $\Lambda$ is an $m \times 1$ vector of unstandardized factor loadings generally assumed to hold for all

Figure 5. Path model for multivariate case, with one construct.


## Pardon Error

Having cursed the varying degrees of darkness inherent in traditional univariate and multivariate experimental analyses, I now wish to light a candle - or more accurately, introduce the candle others have lit (e.g., Muthén, 1989; Sörbom, 1974). The foundation for this illumination may be seen in Figure 5, already presented. Our real goal is not to be able to detect an overall relation between the population membership $X$ and the discriminant function $W$, but rather between $X$ and the construct $\eta$. That is, we desire an estimate of the path denoted as $\gamma$, making the discriminant function $W$ irrelevant. Fortunately, under the umbrella of structural equation modeling, a clearer attempt at a solution exists.

In Figure 5 the relations between the construct and its measured operationalizations may be expressed in a system of $m$ structural equations of the form $Y_{i}=\lambda_{i} \eta+\varepsilon_{i}(i=1 \ldots m)$. These measurement equations may in turn be represented collectively as
subjects in both populations (homogeneity of measurement), and $\varepsilon$ is a subject's $m \times 1$ vector of $\varepsilon_{i}$ measured variable residuals. More interestingly, the theoretical relation of our current focus is contained in the structural equation relating population membership to the construct,

$$
\begin{equation*}
\eta=\gamma X+\zeta . \tag{15}
\end{equation*}
$$

These structural equations, along with the simplifying (but not mandatory) assumption of independence of all exogenous elements ( $X, \varepsilon$, and $\zeta$ ), have implications for the partitioned variance-covariance matrix $\Sigma$ containing the $X$ and $Y_{i}$ variables for all populations combined. Specifically, for the $Y_{i}$ variables alone, Equation 14 implies

$$
\begin{equation*}
\Sigma_{Y}=\Lambda \phi_{\eta} \Lambda^{\prime}+\Theta_{\varepsilon}, \tag{16}
\end{equation*}
$$

where $\phi_{\eta}$ is the total construct variance for both populations combined, and $\Theta_{\varepsilon}$ is the mxm
variance-covariance matrix for the $\varepsilon_{i}$ residuals. Equation 15 has implications for $\phi_{\eta}$, such that

$$
\begin{equation*}
\phi_{\eta}=\gamma^{2} \sigma_{X}^{2}+\psi, \tag{17}
\end{equation*}
$$

where $\sigma_{X}^{2}$ is the variance of $X$ and $\psi$ is the variance of the construct residual $\zeta$. That is, $\psi$ is the part of the construct variance that is not explained by population membership; as such, it is the pooled within-groups variance for the construct. Finally, the portion of the covariance matrix relating the vector $\mathbf{Y}$ of $Y_{i}$ variables to $X$, as following from Equations 14 and 15, is

$$
\begin{equation*}
\boldsymbol{\Sigma}_{X Y}=\gamma \sigma_{X}^{2} \mathbf{\Lambda}^{\prime} . \tag{18}
\end{equation*}
$$

As implied by the model in Figure 5, the full partitioned matrix for the $X$ and $Y_{i}$ variables (respectively) is:

$$
\boldsymbol{\Sigma}=\left[\begin{array}{c|c}
\sigma_{X}^{2} & \gamma \sigma_{X}^{2} \boldsymbol{\Lambda}^{\prime}  \tag{19}\\
\hline \boldsymbol{\Lambda} \gamma \sigma_{X}^{2} & \boldsymbol{\Lambda}\left[\gamma^{2} \sigma_{X}^{2}+\psi\right] \boldsymbol{\Lambda}^{\prime}+\boldsymbol{\Theta}_{\varepsilon}
\end{array}\right] .
$$

Using maximum likelihood estimation within structural equation modeling (see, e.g., Bollen, 1989), and after fixing one factor loading to a value of 1 so as to give the construct $\eta$ a unit of measurement (i.e., that of the corresponding indicator variable), population values for all parameters in Equation 19 are chosen so as to maximize the likelihood of the observations giving rise to the sample covariance matrix $\mathbf{S}$. After conducting an assessment of the data-model fit as represented by the degree of correspondence between the observed matrix $\mathbf{S}$ and the expected matrix $\hat{\boldsymbol{\Sigma}}$ (after substituting the optimum parameter values into Equation 19), satisfactory fit allows one to proceed to the question at hand. That question involves the estimation of, and statistical test of, the population mean difference(s) on the construct $\eta$.

For the two-group case, the path from the single dummy variable $X$ to the construct $\eta$ is an estimate of the population difference on the construct. This path, $\gamma$, will also have a maximum likelihood standard error as a byproduct of the estimation process, which will
allow a statistical test of the difference between the two population means on the construct $\eta$. If $X$ is coded $0 / 1$, then a statistically significant and positive estimate of $\gamma$ implies the population coded $X=1$ has a higher mean on the construct $\eta$, whereas a negative value would imply superiority of the population coded $X=0$. An interpretation of the value of $\gamma$ itself is not generally useful because it reflects the metric that $\eta$ has been assigned by fixing a variable loading to 1 . However, given that the pooled within-groups construct variance $\psi$ has been estimated as well, we may derive an estimate of the latent standardized effect size $d_{\eta}$, where

$$
\begin{equation*}
d_{\eta}=\gamma / \sqrt{\psi} . \tag{20}
\end{equation*}
$$

Thus, if a single construct underlies our measured variables, we are able to conduct a statistical test on the construct mean difference as well as estimate the standardized effect size associated with that differences in latent means.

The simple process described above, which may be conducted using any structural equation modeling software (e.g., AMOS, EQS, LISREL, Mplus), is part of a larger class of models known as multiple-indicator multiplecause (MIMIC) models suggested for assessing latent population differences (Muthén, 1989). The procedure is not without its own assumptions and restrictions, some of which may be softened in a somewhat more complicated strategy known as structured means modeling (Sörbom, 1974). Those details are left for the interested reader, and are summarized didactically elsewhere (e.g., Hancock, in press). More importantly is that these methods exist to put the construct back at center stage, in terms of hypothesis testing and effect size estimation, and as such the theoretical benefits over a MANOVA approach should be clear.

We may also take a practical approach in comparing the MIMIC and MANOVA strategies by determining the sample sizes required to detect a specific latent standardized effect size in order to achieve a desired level of statistical power. In Table 1 we see the cases of $m=2,3$, and 4 measured variables, crossed with homogeneous sets of standardized loadings of $\lambda=.4, .6$, and .8 . The standardized latent effect
sizes included were $d_{\eta}=.2$, .5, and .8. In all conditions the necessary sample size was assessed for both MANOVA and the MIMIC approach in order to achieve .80 power using the equivalent of a two-tailed test at the .05 level. For MANOVA, sample size determination in each case followed strategies for Hotelling's $T^{2}$ outlined by Cohen (1988; Section 10.3.2.1), while for the MIMIC approach the methods derived by Hancock (2001) were used.
case of homogeneous loadings $H$ mirrors the Spearman-Brown prophecy formula as

$$
\begin{equation*}
H=m \lambda^{2} /\left[1+(m-1) \lambda^{2}\right] \tag{21}
\end{equation*}
$$

(see Hancock, 2001). For example, with $m=3$ variables, $H=.276$ for $\lambda=.4$ and $H=.529$ for $\lambda=.6$; sample size thus decreases by a multiplicative factor of $.276 / .529=.521$ for both the MIMIC and MANOVA strategies. For MANOVA this

| Table 1 <br> Sample Size Required For Two-Group .05-Level Tests With Power=. 80 |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| MIMIC |  |  |  | MANOVA |  |  |
|  | $\lambda=.4$ | $\lambda=.6$ | $\lambda=.8$ | $\lambda=.4$ | $\lambda=.6$ | $\lambda=.8$ |
| $m=2 \quad d_{\eta}=.2$ | 1424 | 742 | 504 | 1748 | 912 | 619 |
| $d_{\eta}=.5$ | 229 | 120 | 81 | 281 | 148 | 101 |
| $d_{\eta}=.8$ | 90 | 47 | 32 | 111 | 59 | 41 |
| $m=3 \quad d_{\eta}=.2$ | 1080 | 626 | 467 | 1502 | 871 | 650 |
| $d_{\eta}=.5$ | 174 | 101 | 76 | 242 | 141 | 106 |
| $d_{\eta}=.8$ | 68 | 40 | 30 | 96 | 57 | 43 |
| $m=4 \quad d_{\eta}=.2$ | 909 | 568 | 449 | 1383 | 865 | 684 |
| $d_{\eta}=.5$ | 146 | 92 | 73 | 224 | 141 | 112 |
| $d_{\eta}=.8$ | 58 | 36 | 29 | 89 | 57 | 45 |

Many points are noteworthy in Table 1. As expected, for both the MIMIC and MANOVA methods the necessary sample size decreases as effect size increases (holding all else constant). Specifically, sample size decreases were approximately proportional to corresponding increases in the square of $d_{\eta}$ (e.g., from $d_{\eta}=.2$ to $d_{\eta}=.5$, sample size necessary decreases by a multiplicative factor of $.2^{2} / .5^{2}=.16$ ). Sample size also decreases for both methods as loading magnitude increases (holding all else constant). In particular, sample size decreases were approximately proportional to corresponding increases in construct reliability as measured by coefficient $H$ (also known as maximal reliability), where for the
sample size decrease is due to the increased presence of the construct in the discriminant function; for the MIMIC approach, which already operates at the construct level, this sample size decrease is due to a decrease in the standard error associated with the $\gamma$ path.

With regard to increasing the number of variables, for the MIMIC strategy sample size decreases correspondingly (holding all else constant); this is because distributional noncentrality varies directly with construct reliability as measured by $H$ (Hancock, 2001), which increases with the addition of any nonzero loading. For MANOVA, sample size decreases with additional variables for $\lambda=.4$ and .6 , but an increase in required sample size is observed for
$\lambda=.8$. This is because at some point additional variables do not contribute sufficient new information about the construct to justify the additional degree of freedom expenditure. This was seen in a supplemental analysis as well using $\lambda=.9$ (not shown in Table 1), where for $m=2,3$ and 4 the necessary sample size per group for MANOVA increased from 540 to 590 to 635 , respectively.

Overall, as expected the sample size required for MANOVA was always greater than for MIMIC. For the $m=2$ case MANOVA sample sizes were always about $23 \%$ larger than for the MIMIC approach. For $m=3$ that number increased to around $39 \%$, while for $m=4$ required sample sizes for MANOVA were approximately $52 \%$ larger than for the MIMIC strategy. Thus, not only has the MIMIC approach's estimation and inference operated directly at the level of the construct of interest, it has done so with the same power for a considerable savings in sample size (or with greater power for the same sample size expenditure). And interestingly, at no point did we need to estimate variables' reliability; this information was implicit within the MIMIC process in the estimation of the $\lambda_{i}$ loadings.

Extensions to this latent approach exist both internally and externally, where the former refers to methods for answering the same questions under less restrictive assumptions and the latter refers to methods for addressing more complex questions. With regard to internal extensions, the primary assumption implicit in MIMIC modeling is that, because the data from the groups are combined and only one model results, the same measurement model holds across populations. This includes loadings, construct variance, and error variances. In effect, all sources of covariation among observed variables are assumed to be equal in all populations, making the assumption of identical measurement models tantamount to an assumption of equal variance/covariance matrices (as is actually assumed in MANOVA as well). As alluded to previously, a more flexible approach to assessing latent means exists in structured means modeling (Sörbom, 1974), where only the corresponding loadings are commonly constrained across populations in the complete covariance model. Further,
additional flexibility may exist to allow for some loading differences across populations under particular configurations of partial measurement invariance (Byrne, Shavelson, \& Muthén, 1989).

Externally, the methods of assessing latent means may be extended greatly. Within the MIMIC framework, the creative use of group code predictors of the latent construct of interest (e.g., dummy variables) can fairly easily facilitate inferences that parallel those of more complex one-way and factorial ANOVA designs. Also, covariates may be introduced along with the group code variables. In fact, like all other variables covariates have underlying constructs; as such, given multiple indicator variables a latent covariate construct may be incorporated into the model along with the group code variables. The disattenuation of measurement error in the covariate provides greater accuracy in the assessment and testing of the covariate's predictive role in the design, as well as of population mean differences on the outcome construct after exacting such latent control.

## Seeking Your Fortune

Inspired jointly by ancient wisdom and modern analytical methods, this article has attempted to return our focus to the constructs that underlie our experimental research endeavors. Certainly those constructs must be grounded in observable measures, but the proximity of those measures’ operationalization of the construct(s) should be acknowledged and even accommodated. I have attempted to highlight the theoretical and practical costs of imperfect operationalization within traditional experimental analyses, and pointed toward reasonably accessible strategies that circumvent our measures' necessary imperfections.

But there is no free lunch, so to speak. Although the latent variable approaches to experimental design can pardon error and thus attempt to correct for unreliability, researchers are not thereby absolved of expending considerable effort in choosing or constructing quality measures. Poor reliability in measures yields less stability in the constructs and in estimates of their relations with other variables (e.g., group code variables), as well as larger standard errors for the statistical assessment of
estimated relations. Thus, the methods described briefly herein serve to complement sound principles of instrument selection and construction.

These methods also signal the potential to reframe other aspects of the multivariate general linear model as well. Although this article has focused on experimental design, the canonical correlation model suffers from some of the same problems as MANOVA. Specifically, while $X$ and $Y$ variables are generally chosen by researchers with some constructs in mind, $X$ and $Y$ composites are formed whose primary allegiance is to the maximization of $X Y$ relations. If one used variables to define constructs in separate $X$ and $Y$ measurement models, the relations between constructs would be directly couched in theory, disattenuated of measurement error, and detectable with considerably more power than within the canonical framework. Expositions similar to those provided here for experimental design could be crafted, and would be equally compelling.

In sum, although constructs and their relations are the beloved truths that motivate most applied statistics, so many of our analytical efforts are hindered in their inferential estimation and hypothesis testing by our measures' inability to reflect those constructs satisfactorily. The current article has illustrated the detriments of failing to pardon error from our experimental inference, and has directed the applied researcher toward more modern methods that can assist researchers in getting closer to the truths they seek. It is my hope that they will pursue these and related methods as they seek their research fortunes. In the mean time, I believe I have a lunch appointment....

## References

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley \& Sons.

Bollen, K. A., \& Lennox, R. (1991).Conventional wisdom on measurement: A structural equation perspective. Psychological Bulletin, 110, 305-314.

Byrne, B. M., Shavelson, R. J., \& Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456-466.

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

Cohen, P., Cohen, J., Teresi, M., Marchi, M., \& Velez, C.N. (1990). Problems in the measurement of latent variables in structural equations causal models. Applied Psychological Measurement, 14, 183-196.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. Psychometrika, 66, 373-388.

Hancock, G. R. (in press). Experimental, quasi-experimental, and nonexperimental design and analysis with latent variables. In D. Kaplan (Ed.), Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks, CA: SAGE Publications.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. Psychometrika, 54, 557-585.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

# Regular Articles P* Index of Segregation: Distribution Under Reassignment 

Charles F. Bond<br>Texas Christian University

F. D. Richard<br>University of North Florida

Students of intergroup relations have measured segregation with a $P^{*}$ index. In this article, we describe the distribution of this index under a stochastic model. We derive exact, closed-form expressions for the mean, variance, and skewness of $P^{*}$ under random segregation. These yield equivalent expressions for a second segregation index: $\eta^{2}$. Our analytic results reveal some of the distributional properties of these indices, inform new standardizations of the indices, and enable small-sample significance testing. Two illustrative examples are presented.

Key words: Segregation index, randomization methods, quadratic assignment, Eta-squared

Introduction
Bell (1954) developed a way to measure the amount of contact between two groups. This widely-used measure has gone by several names. It has been called the exposure index (James, 1986) and the interaction index (Massey \& Denton, 1986). We, however, follow Lieberson (1980) in referring to a measure of sort devised by Bell as a $P^{*}$ index. It is intended to measure the probability that individuals from two different groups will have contact with one another.

The $P^{*}$ index has been used in studies of residential segregation - when data are available on the number of members of a minority group ( $j$ ) and a majority group ( $k$ ) who live in a particular spatially-defined unit (on the same city block, for example, or in the same census tract). It requires data on the number of minority and majority residents in a number of such units. Then $P^{*}$ is the probability that a randomly selected member of group $j$ lives in the same unit as a member of group $k$. The index is defined as

Address correspondence to Charles F. Bond, Jr., Box 298920, Department of Psychology, Texas Christian University, Fort Worth, TX, 76129, or e-mail: c.bond@tcu.edu.

$$
\begin{equation*}
{ }_{j} P_{k}^{*}=\frac{1}{N_{\bullet j}} \sum_{i=1}^{u} \frac{N_{i j} N_{i k}}{N_{i \bullet}} \tag{1}
\end{equation*}
$$

where $N_{0_{j}}$ is the total number of members of group $j, u$ is the number of units; and $N_{i j}, N_{i k}$, and $N_{i}$ 。 are the number of members of group j in unit $i$, the number of members of group $k$ in unit $i$, and the total number of people in unit $i$, respectively.
$P^{*}$ plays a role in the study of segregation. It has been used to document school, as well as residential segregation (Coleman, Kelly, \& Moore, 1975; Krivo \& Kaufman, 1999). It complements alternative indices by tapping a distinct dimension of segregation (Massey, White, \& Phua, 1996; Stearns \& Logan, 1986). Despite recurrent criticism (Taeuber \& Taeuber, 1965), the $P^{*}$ index of segregation has found application in a variety of contexts for nearly 50 years.

Researchers who measure segregation with the $P^{*}$ index have an obligation to interpret their results. $P^{*}$ is a probability. It varies between $\underline{0}$ to $\underline{1}$. However, the probability of a member of one group being exposed to a member of another group could be misleading, depending (as it does) on relative group size. To facilitate interpretation, researchers often compare an observed value of ${ }_{j} P_{k}^{*}$ with the value that would have been observed if there had been no segregation - that is, if the proportion of
members of group $j$ and group $k$ within each unit equaled the overall proportion of members of those groups across all units. Then ${ }_{j} P_{k}^{*}$ would equal $N_{\circ k} N_{0 .}{ }^{-1}$, and the probability of a member of group $j$ being in the same unit as a member of group $k$ would be the proportion of members of group $k$ across all units (Lieberson, 1980).

No doubt, students of segregation enhance understanding by providing comparison values for their measures. We wonder, however, if it is best to compare $P^{*}$ to a value which assumes that there is no segregation. After all, even if the members of two groups made residential choices entirely at random, some degree of segregation could be expected by chance (cf., Winship, 1977). In some contexts it would be informative to compare $P^{*}$ to the value it would attain under a random degree of segregation. Unfortunately, this comparison has not been possible to date because the value of $P^{*}$ that would be produced by random segregation has not been known.

In the current article, we describe the distribution of $P^{*}$ under random segregation. We develop an analytic method for determining whether the amount of intergroup contact in a particular setting differs from the amount that would be expected by chance. Exact, closedform expressions for the expected value, variance, and skewness of $P^{*}$ under random segregation are presented. These imply equivalent expressions for a second segregation index: Bell's eta-square. Our analytic results reveal some of the distributional properties of these segregation indices, inform new standardizations of the indices, and enable small-sample significance testing. For statistical characterizations of $P^{*}$, see Zoloth (1974). For distributional analyses of the widely-used index of dissimilarity, see the papers by Winship (1977) and Inman and Bradley (1991).

## Formulation of the Problem

Our goal is to determine the distribution of the statistic in equation 1) under a stochastic model. We begin by assuming that the total number of individuals in each of $u$ units is fixed - as is the total number of members of group $j$ and group $k$. Our model is that each individual is randomly assigned to a unit. We seek to
determine the distribution of the $P^{*}$ index under all possible assignments of individuals to units - assuming that each assignment that preserves the marginal totals is as likely as every other such assignment.

If all possible assignments of individuals to units could be made, then the distribution of $P^{*}$ could be constructed empirically. Ordinarily, the number of assignments will be prohibitive, however, and other methods will be required. Monte Carlo techniques could be used (cf. Taeuber \& Taeuber, 1965); but these are computationally intensive and yield no exact distributional information. Here we derive the exact mean, variance, and skewness of $P^{*}$ under all possible assignments of individuals to units.

Our derivation treats the distribution of $P^{*}$ as a quadratic assignment problem (Hubert, 1987). We begin by representing $P^{*}$ in a form that is amenable to quadratic assignment methods, so that we can draw on existing analytic results.

Denoting the total number of individuals in the analysis as $N_{\ldots}, P^{*}$ is represented in two $N_{\text {。 }} \times N_{\text {. }}$ matrices. Each row of each matrix will denote a particular individual, as will the corresponding column of the matrix. Hence, each entry in each matrix will denote a pair of individuals, matrix element $s, t$ denoting the dyad that consists of individual $s$ and individual $t$. This representation is familiar to students of social networks (Wasserman \& Faust, 1994). We define the two $N_{\text {. }} \times N_{\text {. }}$ matrices: Q (which we call the cross-group membership matrix) and R (the unit co-occupancy matrix). Both matrices are symmetric.

The cross-group membership matrix Q identifies dyads in which the intergroup contact of interest could, in principle, occur. If the researcher wishes to measure the likelihood that a member of group $j$ will have contact with a member of group $k$, the entry in the sth row and $t$ th column of this matrix is set to $\left(2 N_{\cdot j}\right)^{-1}$ whenever one of the two individuals in the dyad ( $s$ or $t$ ) belongs to group $j$ and the other individual belongs to group $k$. All other entries of the Q matrix are set to 0 .

The unit co-occupancy matrix R identifies individuals who are in the same unit. The entry in the sth row and $t$ th column of the R
matrix is set to $N_{i \bullet}^{-1}$ if the two individuals ( $s$ and $t$ ) are in unit $i$. All elements along the diagonal of R are set to 0 , as are any offdiagonal elements that denote two individuals who are in different units.

Denoting element $s, t$ of matrix Q by $q_{s t}$ and the corresponding element of matrix R as $r_{s}$ - algebra reveals that

$$
\begin{equation*}
{ }_{j} P_{k}^{*}=\sum_{s=1}^{N_{0} . .} \sum_{t=1}^{N_{. .}} q_{s t} r_{s t} \tag{2}
\end{equation*}
$$

Thus, the $P^{*}$ index can be expressed as the sum of the products of corresponding elements of two matrices. Such statistics can be analyzed with quadratic assignment methods (Hubert, 1987).

Our goal is describe the distribution of $P^{*}$ under all possible assignments of individuals to units. In our formulation, individuals are implicitly assigned to units by the R matrix. We could change the assignment of individuals to units by reordering the rows and corresponding columns of R.

Having expressed $P^{*}$ as the sum of the products of corresponding elements of two matrices, we can draw on formulas that have been derived for the mean, variance, and skewness of such statistics under all possible reorderings of the rows and columns of one of those matrices (Hubert, 1987) These provide the desired distributional information.

## Analytic Results

Our quadratic assignment formulation yields the following results. The mean of ${ }_{j} P_{k}{ }^{*}$ under all possible assignments of individuals to units is

$$
\begin{equation*}
E\left({ }_{j} P_{k}^{*}\right)=\left[N_{\bullet k} N_{\bullet \bullet}^{-1}\right]\left[\left(N_{\bullet \bullet}-u\right)\left(N_{\bullet .}-1\right)^{-1}\right] \tag{3}
\end{equation*}
$$

all symbols having been defined above.
The variance of ${ }_{j} P_{k}{ }^{*}$ under all possible assignments of individuals to units has a more complicated mathematical expression. In fact,

$$
\begin{equation*}
\operatorname{Var}\left({ }_{j} P_{k}^{*}\right)=a(b+c+d)-\left[E\left({ }_{j} P_{k}^{*}\right)\right]^{2} \tag{4}
\end{equation*}
$$

where

$$
\begin{gathered}
a=N_{\bullet k}\left[N_{\bullet j} N_{\bullet \bullet}\left(N_{\bullet \bullet}-1\right)\left(N_{\bullet \bullet}-2\right)\right]^{-1} \\
b=\left(N_{\bullet \bullet}-2\right) \sum_{i=1}^{u}\left[\left(N_{i \bullet}-1\right) N_{i \bullet}^{-1}\right] \\
c=\left(N_{\bullet j}+N_{\bullet k}-2\right) \sum_{i=1}^{u}\left[\left(N_{i \bullet}-1\right)\left(N_{i \bullet}-2\right) N_{i \bullet}^{-1}\right]
\end{gathered}
$$

and

$$
\begin{aligned}
d= & {\left[\left(N_{\bullet}-j-1\right)\left(N_{\bullet}-k\right)\left(N_{\bullet} \bullet-3\right)^{-1}\right] } \\
& \left\{\left(N_{\bullet}-u\right)^{2}-2 \sum_{i=1}^{u}\left[\left(N_{i} \bullet-1\right)\left(2 N_{i} \bullet-3\right) N_{i}^{-1}\right]\right\} .
\end{aligned}
$$

We have derived the coefficient of skewness of ${ }_{j} P_{k}{ }^{*}$ under all possible assignments of individuals to units. Appendix A presents an analytic expression for this statistic, which we symbolize $\gamma_{1}\left({ }_{j} P_{k}{ }^{*}\right)$.

Because these analytic expressions are intricate, it may be helpful to begin by noting some quantities they omit. Neither the mean, the variance, nor the skewness of ${ }_{j} P_{k}{ }^{*}$ are affected by the number of members of group $j$ or $k$ in any particular unit. These expressions reflect only marginal totals - the size of the two groups, and the size of the $u$ units. Values that would appear as entries in a unit x group contingency table do not enter into the equations because these are moments of a distribution of the possible values of ${ }_{j} P_{k}^{*}$ over all possible entries that would preserve the marginal totals.

Equation (3) yields insight into the impact of random segregation on $P^{*}$. In the absence of any segregation, the probability of a member of group $j$ being in the same unit with a member of group $k$ equals $N_{\bullet k} N_{\bullet_{0}}^{-1}$, as earlier researchers noted. This probability is lower under random segregation. Relative to the probability of intergroup exposure under no segregation, the random expectation for ${ }_{j} P_{k}{ }^{*}$ is lower by a factor of $\left(N_{. .}-u\right)\left(N_{. .}-1\right)^{-1}$, as equation 3) indicates. This difference might be negligible if the units under analysis were sufficiently large; it could be appreciable if the units were sufficiently small.

## Other $P^{*}$ Indices

The probability of a member of one group interacting with a member of a second group will not, in general, equal the probability of a member of the second group interacting with a member of the first (Lieberson, 1980). However, these probabilities have a simple relationship to one another.

$$
\begin{equation*}
{ }_{k} P_{j}^{*}=N_{\bullet j} N_{\bullet k}-1{ }_{j} P_{k}^{*} \tag{5}
\end{equation*}
$$

This implies that

$$
\begin{align*}
& \mathrm{E}\left({ }_{k} P_{j}^{*}\right)=N_{\circ j} N_{\bullet k}{ }^{-1} \mathrm{E}\left({ }_{j} P_{k}^{*}\right)  \tag{6}\\
& \operatorname{Var}\left({ }_{k} P_{j}^{*}\right)=N_{\circ j}^{2} N_{\bullet k}{ }^{-2} \operatorname{Var}\left({ }_{j} P_{k}^{*}\right) \\
& \alpha_{1}\left({ }_{j} P_{k}^{*}\right)=\alpha_{1}\left({ }_{k} P_{j}^{*}\right)
\end{align*}
$$

and

These equations permit a comparison of the distributions of complementary exposure indices. In skewness, the distributions of ${ }_{j} P_{k}{ }^{*}$ and ${ }_{k} P_{j}^{*}$ are identical. In expectation and variance, these two distributions are identical if groups $j$ and $k$ are the same size. If group $j$ is smaller than group $k$, then ${ }_{j} P_{k}{ }^{*}$ will have a higher expectation and greater variability than ${ }_{k} P_{j}^{*}$. If group $j$ is larger than group k , then ${ }_{j} P_{k}{ }^{*}$ will have a lower expectation and less variability than ${ }_{k} P_{j}{ }^{*}$. The greater the difference in the size of two groups, the greater will be the difference in expectation and variability of the two exposure indices involving those groups.

Often students of segregation wish to measure the likelihood that a member of a group will be in the same unit as other members of that group. They have done so with a isolation index (Bell, 1954).

$$
\begin{equation*}
{ }_{j} P_{j}^{*}=\frac{1}{N_{\bullet j}} \sum_{i=1}^{u} \frac{N_{i j} N_{i j}}{N_{i \bullet}} \tag{7}
\end{equation*}
$$

The methods above must be adapted to describe the distribution of ${ }_{j} P_{j}^{*}$. One begins by applying the formulas above to an index for the exposure of individuals who are members of group $j$ to individuals who are not members of group $j$. Having obtained results for the exposure index jPnot-j ${ }^{*}$, results for the corresponding isolation index follow when one recognizes that

$$
\begin{equation*}
{ }_{j} P_{j}^{*}=1-{ }_{j} P_{\text {not-j }}{ }^{*} \tag{8}
\end{equation*}
$$

Then it should be apparent that

$$
\begin{align*}
& \mathrm{E}\left(P_{j}^{*}\right)=1-\mathrm{E}\left(\mathrm{j}_{\mathrm{not-j}}{ }^{*}\right)  \tag{9}\\
& \operatorname{Var}\left({ }_{j} P_{j}^{*}\right)=\operatorname{Var}\left({ }_{j} P_{\text {not-j-j}}^{*}\right) \\
& \gamma_{1}\left({ }_{j} P_{j}^{*}\right)=-\gamma_{1}\left({ }_{j} P_{\text {not. }}{ }^{*}\right)
\end{align*}
$$

And $\quad \gamma_{1}\left(j P_{j}^{*}\right)=-\gamma_{1}\left(j P_{\text {not-j }}{ }^{*}\right)$

Eta-square
Bell (1954) also proposed a revised index of isolation

$$
\begin{equation*}
I=\left({ }_{j} P_{j}^{*}-\frac{N_{\bullet j}}{N_{\bullet \bullet}}\right)\left(1-\frac{N_{\bullet j}}{N_{\bullet \bullet}}\right)^{-1} \tag{10}
\end{equation*}
$$

noting that the value of this statistic would invariably lie between 0 and 1.

As Duncan and Duncan (1955) observed, Bell's revised index of isolation is identical to $\eta^{2}$ for predicting a dicotomous group membership variable ( $1=$ member of group $j$, $0=$ not a member of group $j$ ) from unit. $\eta^{2}$, the correlation ratio, equals the percentage of variance in group membership accounted for by unit, a familiar metric for describing strength of association.

Having derived the mean, variance, and skewness of the distribution of ${ }_{j} P_{j}{ }^{*}$ under random segregation, we can use equation 10) to obtain equivalent expressions for Bell's $\eta^{2}$ measure

$$
\begin{align*}
& E\left(\eta^{2}\right)=(u-1)\left(N_{\bullet \bullet}-1\right)^{-1}  \tag{11}\\
& \operatorname{Var}\left(\eta^{2}\right)=\operatorname{Var}\left({ }_{j} P_{n o t-j}^{*}\right) N_{\bullet \bullet}^{2}\left(N_{\bullet \bullet}-N_{\bullet j}\right)^{-2} \\
& \gamma_{1}\left(\eta^{2}\right)=-\gamma_{1}\left({ }_{j} P_{n o t-j}^{*}\right)
\end{align*}
$$

where $\operatorname{Var}\left({ }_{j} P_{n o t-j}^{*}\right)$ can be obtained from equation 4) above and $\quad \gamma_{1}\left({ }_{j} P_{n o t-j}^{*}\right)$ can be obtained from Appendix A.

Standardization and Significance Testing
Often, researchers want to compare the levels of intergroup contact in different locales. Locales may differ from one another in a number of ways - in-group composition, for example, and in the size of units. If some researchers want their comparisons of intergroup
contact to reflect differences in-group composition across locales (Massey, White, \& Phua, 1996), others would prefer to make these comparisons in a standardized metric.

Many scholars have treated Bell's $\eta^{2}$ index as a standardized measure of intergroup contact. In this role, $\eta^{2}$ has some limitations. Neither the expectation nor the variance of $\eta^{2}$ are fixed under the assumption of random segregation, as the equations in 11) reveal. Given completely random segregation in two locales, the expected value of $\eta^{2}$ in the two locales would not in general be equal. Ordinarily, one of the locales would have larger units than the other, hence a lower expected $\eta^{2}$.

For standardized comparisons of intergroup contact, we propose the following measure

$$
\begin{equation*}
\mathrm{Z}=\left[{ }_{j} P_{k}^{*}-\mathrm{E}\left({ }_{j} P_{k}^{*}\right)\right]\left[\operatorname{Var}\left({ }_{j} P_{k}^{*}\right)\right]^{-5} \tag{12}
\end{equation*}
$$

or the analogous Z-statistic for $\eta^{2}$. Under random segregation, these statistics would have an expected value of $\underline{0}$ and a variance of $\underline{1}$ in any locale - regardless of group composition or unit size.

Researchers may wish to determine whether an observed level of intergroup contact differs to a statistically significant degree from the level that would be produced by random segregation. Although in principle an exact test might be constructed with the multiple hypergeometric distribution (Agresti, 1990), we propose a less cumbersome alternative. We suggest that segregation researchers refer the Zstatistic of equation 12) to some reference value. Liberal reference values could be taken from the standard normal distribution, conservative reference values from Chebyschev's inequality. These would imply that intergroup contact departs from the level expected under random segregation at an alpha-level of .05 if the absolute value of Z exceeds 1.96 (by the normal criterion) or 4.47 (by the Chebyschev criterion). Intermediate reference values could be obtained by incorporating the skewness of the segregation measure into a Type III Fisher's distribution. See Hubert (1987) for details.

For samples of the size analyzed in
many studies of residential segregation,
significance testing may be moot. In such large samples, every departure from expectation may be highly significant (Taeuber \& Taeuber, 1965), and associations between group membership and unit occupancy may be amenable to traditional chi-square tests. Our standardization methods would nonetheless be of value.

The inferential test we are proposing should be more useful for small data sets, where the statistical significance of intergroup contact is not a foregone conclusion, and chi-square approximations would be suspect. Such data sets may be uniquely suited to a $P^{*}$ analysis - the members of a unit being most likely to have contact with one another when the units are small.

## Examples

For illustrative purposes, we will analyze intergroup contact at a mid-sized American University. We will consider two examples - an example of contact between minority and non-minority faculty members, and an example of contact between minority and non-minority students.

Table 1 presents data on the number of minority and non-minority faculty members serving in eight different units of this University, as published by the University's Office of Institutional Research. These units are housed in different buildings. Faculty tend to interact within these units of the University, not across units.

To assess intergroup contact in this setting, we begin by computing the probability of a minority faculty member serving in the same unit of the University as a non-minority faculty member. Results show that ${ }_{m} P_{\text {non-m }}{ }^{*}=$ .8724 , a sizeable probability. Of course, one needs to consider that the overall proportion of non-minority faculty members is .8868 . It is noteworthy that the observed probability of a minority faculty member serving in the same unit as a non-minority is slightly lower than the proportion of non-minorities as a whole. Does this imply that minority faculty members tend to be isolated from non-minorities? Is this tendency greater than would be expected if these faculty members were distributed across the eight units at random?

Table 1. Faculty Members at an American University, By Educational Unit and Minority Status

| Educational Unit: | Minority | Non-Minority |
| :--- | ---: | :--- |
| Humanities |  |  |
| Social Science | 11 | 48 |
| Natural Science | 5 | 47 |
| Fine Arts | 7 | 76 |
| Nursing | 6 | 42 |
| Business | 1 | 21 |
| Education | 7 | 42 |
| Divinity | 3 | 21 |

To answer these questions, we used the present analytic methods. Application of equation 3 ) above shows that the observed level of minority exposure to non-minorities ( ${ }_{m} P_{\text {non-m }}{ }^{*}$ $=.8724$ ) is slightly greater than what would be produced by random segregation: $\mathrm{E}\left({ }_{m} P_{\text {non-m }}{ }^{*}\right)=$ .8700. There is little dispersion in the values of ${ }_{m} P_{\text {non }-m}{ }^{*}$ across all possible assignments of these faculty members to the 8 units; the square root of equation 4) yields S.D. $\left({ }_{m} P_{\text {non-m }}{ }^{*}\right)=.0089$. Applying the equations in the Appendix, we find that the distribution of ${ }_{m} P_{\text {non-m }}$ is negatively skewed: $\gamma_{1}\left({ }_{m} P_{\text {non }-m}{ }^{*}\right)=-1.25$. Plugging into equation 12), a standardized measure of minority faculty exposure to non-minorities is $\mathrm{Z}=+.27$. By any significance testing criterion, this level of the intergroup contact could have been produced by chance.

Although the isolation of minority faculty members could be expressed in terms of a complementary $P^{*}$ index ( ${ }_{m} P_{m}{ }^{*}=.1132$ with Z $=-.27$ ), we will express it in terms of Bell's $\eta^{2}$.

The observed value of $\eta^{2}=.0162$ - a value that is close to what would be expected under random segregation: $\mathrm{E}\left(\eta^{2}\right)=.0189$. For an analogue to the Z -statistic in equation 12 ), we could divide the difference between observed and expected values of $\eta^{2}$ by .01004 (the standard deviation of $\eta^{2}$ ) and find that in this standardized metric $\mathrm{Z}=-.27$ - the same value that was found for the $P^{*}$ isolation index. These values will always be the same.

Even if minority faculty are integrated at this institution, students may be segregated. We checked for segregation among some undergraduates who were enrolled in a Psychology course. Weekly, students choose to attend any one of the six laboratory sessions that are taught in conjunction with the course. Table 2 depicts the number of minority and nonminority students who attended different laboratory sessions one week during the Spring semester of 2000. Each student's minority status was reported by a laboratory supervisor who was unaware of the purpose of the report.

Table 2. Students Enrolled in a Psychology Class, By Laboratory and Minority Status

| Laboratory Attended: | Minority | Non-Minority |
| :---: | :---: | ---: |
| Monday 1:00 PM |  |  |
| Monday 3:00 PM | 1 |  |
| Wednesday 1:00 PM | 0 | 7 |
| Wednesday 3:00 PM | 0 | 6 |
| Friday 1:00 PM | 2 | 2 |
| Friday 3:00 PM | 2 | 4 |
|  | 2 | 4 |

## Conclusion

Do students avoid intergroup contact by choosing to attend laboratories with peers of their own ethnicity? To address this question, we computed the probability of a minority student attending the same laboratory session as another minority student. Computations showed that the isolation index ${ }_{m} P_{m}{ }^{*}=.494-$ far greater than total proportion of minority students in this sample (.233), and somewhat greater than the isolation index that random laboratory choices would have produced: $\mathrm{E}\left({ }_{m} P_{m}{ }^{*}\right)=.366$.

In this sample, random laboratory choices produce sufficient variability in values of the isolation index [S.D. $\left({ }_{m} P_{m}{ }^{*}\right)=.073$ ] that the observed degree of minority isolation would not (by a two-tailed test) differ significantly from its expected value ( $\mathrm{Z}=+1.75$ ). Bell's $\eta^{2}$ index (.340) also exceeds its expected value (.172) by an amount that yields the same value of $\mathrm{Z}(+1.75$, with S.D. $=.095)$. Of course, these small-scale examples are only illustrative. Larger data sets might yield different conclusions.

We hope that these analytic techniques will be useful to students of segregation. They require no assumption about the sampling of observations, or the form of any population distribution. They reflect randomizations of the data at hand (Edington, 1995).

## References

Agresti, A. (1990). Categorical data analysis. New York: John Wiley.

Bell, W. (1954). A probability model for the measurement of ecological segregation. Social Forces, 32, 357-364.

Coleman, J. S., Kelly, S. D., \& Moore, J.A. (1975). Trends in school segregation, 196873. Washington, DC: Urban Institute.

Duncan, O. D., \& Duncan, B. (1955). A methodological analysis of segregation indices. American Sociological Review, 20, 210-217.

Edington, E. S. (1995). Randomization tests (Third edition). New York: Marcel Dekker.

Hubert, L. J. (1987). Assignment methods in combinatorial data analysis. New York: Marcel Dekker.

Inman, H. F., \& Bradley, E. L., Jr. (1991). Approximations to the mean and variance of the index of dissimilarity in $2 \times \mathrm{C}$ tables under a random allocation model. Sociological Methods \& Research, 20, 242-255.

James, F. J. (1986). A new generalized "exposure-based" segregation index. Sociological Methods \& Research, 14, 301-316.

Krivo, L. J., \& Kaufman, R. L. (1999). How low can it go? Declining black-white segregation in a multiethnic context. Demography, 36, 93-109.

Lieberson, S. (1980). A piece of the pie: Blacks and white immigrants since 1880. Berkeley and Los Angeles: University of California Press.

Massey, D. S., \& Denton, N. A. (1988). The dimensions of residential segregation. Social Forces, 67, 281-315.

Massey, D. S., White, M. J., \& Phua, V. C. (1996). The dimensions of segregation revisited. Sociological Methods \& Research, 25, 172-206.

Sterns, L. B., \& Logan, J. R. (1986). Measuring trends in desegregation: Three dimensions, three measures. Urban Affairs Quarterly, 22, 124-150.

Taeuber, K. E., \& Taeuber, A. F. (1965). Negroes in Cities: Residential segregation and neighborhood change. Chicago: Aldine.

Wasserman, S., \& Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge University Press.

Winship, C. (1977). A revaluation of indexes of residential segregation. Social Forces, 55, 1058-1066.

Zoloth, B. S. (1974). An investigation of alternative measures of school segregation. Madison: University of Wisconsin Institute for Research on Poverty.

## Appendix

The coefficient of skewness of $P^{*}$ is defined as $\gamma_{1}\left({ }_{j} P_{k}{ }^{*}\right)=E\left[{ }_{j} P_{k}^{*}-E\left({ }_{j} P_{k}^{*}\right)\right]^{3}\left[\operatorname{Var}\left({ }_{j} P_{k}^{*}\right)\right]^{-3 / 2}$. Below we present an expression for $E\left[\left({ }_{j} P_{k}^{*}\right)^{3}\right]$. Skewness can be computed from

$$
\begin{aligned}
& \gamma_{1}\left(P_{k}{ }^{*}\right)=\left\{E\left[\left({ }_{j} P_{k}^{*}\right)^{3}\right]-\left[E\left({ }_{j} P_{k}^{*}\right)\right]^{3}-3 E\left({ }_{j} P_{k}^{*}\right) \operatorname{Var}\left({ }_{j} P_{k}^{*}\right)\right\}\left[\operatorname{Var}\left({ }_{j} P_{k}^{*}\right)\right]^{-3 / 2} \\
& E\left[\left({ }_{j} P_{k}^{*}\right)^{3}\right]=N_{\bullet k}\left[N_{. .}\left(N_{. .}-1\right)\right]^{-1} N_{\bullet j}^{-2}[A+B+C+D+F+G] \text {, where } \\
& A=\sum_{i=1}^{u}\left[\left(N_{i \bullet}-1\right) N_{i \bullet}^{-2}\right], \quad B=3\left(N_{\bullet j}+N_{\bullet k}-2\right)\left(N_{\bullet \bullet}-2\right)^{-1}\left\{u+\sum_{i=1}^{u}\left[\left(2-3 N_{i \bullet}\right) N_{i \bullet}^{-2}\right]\right\} \text {, } \\
& C=3\left(N_{\bullet j}-1\right)\left(N_{\bullet k}-1\right)\left[\left(N_{\bullet \bullet}-2\right)\left(N_{\bullet \bullet}-3\right)\right]^{-1}\left[u\left(u-N_{\bullet \bullet}\right) \sum_{i=1}^{u} N_{i \bullet}^{-1}+2\left(N_{\bullet \bullet}-8 u+16 \sum_{i=1}^{u} N_{i \bullet}^{-1}-9 \sum_{i=1}^{u} N_{i \bullet}^{-2}\right)\right] \text {, } \\
& D=\left[\left(N_{\bullet j}-1\right)\left(N_{\bullet j}-2\right)+\left(N_{\bullet k}-1\right)\left(N_{\bullet k}-2\right)\right]\left[\left(N_{\bullet \bullet}-2\right)\left(N_{\bullet \bullet}-3\right)\right]^{-1} \\
& \left(N_{. \bullet}-6 u+11 \sum_{i=1}^{u} N_{i \bullet}^{-1}-6 \sum_{i=1}^{u} N_{i \bullet}^{-2}\right) \text {, } \\
& F=3\left(N_{\bullet j}-1\right)\left(N_{\bullet k}-1\right)\left(N_{\bullet j}+N_{\bullet k}-4\right)\left[\left(N_{\bullet \bullet}-2\right)\left(N_{\bullet \bullet}-3\right)\left(N_{\bullet \bullet}-4\right)\right]^{-1} \sum_{i=1}^{u}\left(N_{i \bullet}-1\right)\left(N_{i \bullet}-2\right)\left[N_{i \bullet}\left(N_{\bullet \bullet}-u\right)-6\left(N_{i \bullet}-2\right)\right] N_{i \bullet}^{-2} \\
& G=\left(N_{\bullet j}-1\right)\left(N_{\bullet_{j}}-2\right)\left(N_{\bullet k}-1\right)\left(N_{\bullet k}-2\right)\left[\left(N_{. \bullet}-2\right)\left(N_{. \bullet}-3\right)\left(N_{. .}-4\right)\left(N_{\bullet \bullet-5}\right)\right]^{-1} \\
& {\left[\left(N_{. .}-u\right)^{3}-6\left(N_{. .}-u\right)\left(2 N_{. .}-5 u+3 \sum_{i=1}^{u} N_{i \bullet}^{-1}\right)+8\left(5 N_{. .}-22 u+32 \sum_{i=1}^{u} N_{i \bullet}^{-1}-15 \sum_{i=1}^{u} N_{i \bullet}^{-2}\right)\right] \text {. }}
\end{aligned}
$$

# A Critical Examination Of The Use Of Preliminary Tests In Two-Sample Tests Of Location 

Kimberly T. Perry<br>Pfizer Inc.<br>Kalamazoo, Michigan

This paper explores the appropriateness of testing the equality of two means using either a $t$ test, the Welch test, or the Wilcoxon-Mann-Whitney test for two independent samples based on the results of using two classes of preliminary tests (i.e., tests for population variance equality and symmetry in underlying distributions).

Key words: $t$ test, Welch, Wilcoxon-Mann-Whitney, Levene, preliminary test for variance, triples test, test of symmetry, test selection

Introduction
In practice, the two-sample $t$ test is widely used to test the equality of two means. However, it is well known that the assumptions of independence (which will not be discussed in this paper), variance homogeneity and normality must be met for the two-sample $t$ test to perform well. Results from Zimmerman and Williams (1989), Gans (1981), Murphy (1976), and Snedecor \& Cochran (1967) have demonstrated that the Welch test or the Wilcoxon-MannWhitney (WMW) test is more robust in certain cases of variance heterogeneity or nonnormality.

Based on the above results for testing the equality of means, we conclude the following:

1. The $t$ test is robust when the distributions are symmetric and the variances are equivalent.
2. The Welch test is robust when the distributions are symmetric and the variances are unequal.

Kimberly T. Perry is a Senior Research Advisor, Pfizer Inc., Kalamazoo, Michigan. Her areas of interest are innovated clinical study designs, multiple endpoint analysis, and interim analysis. Email: Kimberly.t.perry@pfizer.com. Michael Stoline is acknowledged for his support.
3. The Wilcoxon-Mann-Whitney test is robust when the distributions are asymmetric and the variances are equivalent.
4. None of the above three methods are robust when the distributions are asymmetric and the variances are unequal.

Therefore it would be useful to use the results from two classes of preliminary test to determine which of the three tests, the $t$ test, the Welch test, or the Wilcoxon-Mann-Whitney test, should be used to test the hypothesis $H_{0}: \mu_{1}=$ $\mu_{2}$. One class of preliminary tests determines whether the population variances differ, and the other class ascertains if the underlying distributions are symmetric or skewed.

## Tests of Variances Used as Preliminary Tests

The goal of the preliminary test for variance heterogeneity is to indicate when to avoid using mean tests that are sensitive to variance heterogeneity.

Many methods for testing variance homogeneity have been developed and compared. Brown and Forsythe (1974), Conover, M.E. Johnson, and M.M. Johnson (1981), Loh (1987), and O'Brien (1979) have conducted simulations to examine the robustness of many popular methods for testing variance homogeneity. The $L_{50}$, the Levene test using the median, was found to be robust for the nonnormal cases and was one of the procedures
recommended by Conover et al. (1981) as well as the other authors cited above. Based on the above cited literature, the Levene test using the median might be a robust preliminary test procedure.

Furthermore, Olejnik (1987) conducted a study where the Levene test using the median was compared to the O'Brien procedure (1979) as a preliminary test procedure preceding the means test. His results showed the Levene test and the O'Brien procedure used as preliminary tests of variance homogeneity were only slightly more robust than using the $t$ test alone. It is noted that Olejnik (1987) used significance levels of $5 \%$ and $10 \%$ for testing variance homogeneity in the preliminary test procedure.

It is of interest to examine the performance of the $\mathrm{L}_{50}$ test as a preliminary test procedure with a higher significance level. A higher significance level would aid in controlling the Type II error. For this simulation the Levene test at a significance level of $25 \%$ was arbitrary selected.

Test of Symmetry Used as Preliminary Tests
Randles, Fligner, Policello, and Wolfe (1980) compared three procedures for testing whether a univariate population is symmetric about some unspecified value compared to a large class of asymmetric distribution alternatives. These are the Triples test, Gupta's skewness test (Gupta, 1967) and Gupta's nonparametric procedure (Gupta, 1967). Their results show that the Triples test is superior to either competitor for testing the hypothesis of symmetry while possessing good power for detecting asymmetric alternative distributions (Randles et al., 1980).

In addition, Cabilio \& Masaro (1996) and Perry and Stoline (2002) compared the Triples test to other tests of symmetry and the Triples test continued to perform well both on robustness and power. Based on the above studies, the Triples test is selected as a possible preliminary test of symmetry/skewness prior to the testing of means equality in a test selection procedure. A significance level of $5 \%$ for testing of symmetry was arbitrary chosen for this simulation.

Test Selection Procedure
The test selection procedure, hereafter denoted as the TS procedure, will select either a $t$ test, the Welch test, or the Wilcoxon-MannWhitney test based on the results of the two preliminary tests. One class of preliminary tests determines whether the population variances differ, and the other class ascertains if the underlying distributions are symmetric or skewed. The "recommended" $\mathrm{L}_{50}$ test (hereafter denoted Levene test) will be assessed as preliminary test for variance homogeneity, whereas, the Triples test will be assessed as a preliminary test of symmetry/skewness. Based on the results of the two preliminary tests, the TS procedure is constructed in the following way:

1. The $t$ test is used to test the equality of means if symmetry is accepted and variance homogeneity is accepted.
2. The Welch test is used to test the equality of means if symmetry is accepted and variance homogeneity is rejected.
3. The Wilcoxon-Mann-Whitney test is used to test the equality of means if symmetry is rejected and variance homogeneity is accepted.
4. The Welch test is used to test the equality of means if symmetry is rejected and variance homogeneity is rejected.

It is noted that robust methods exist for testing $H_{0}: \mu_{1}=\mu_{2}$ for cases \#1-3 above, but no robust method exists for case \#4.

## Methodology

This section contains the details describing the two-sample methodology used to test the equality of means and variance homogeneity under selected distributions.

Let $x_{11}, \ldots, x_{1 n_{1}}$ be a random sample with sample size of $n_{1}$ from a distribution denoted $f_{1}\left(x ; \mu_{1}, \sigma_{1}\right)$; and $x_{21}, \ldots, x_{2 \mathrm{n}_{2}}$ be a random sample with sample size of $n_{2}$ from a distribution denoted $f_{2}\left(x ; \mu_{2}, \sigma_{2}\right)$. It is assumed that $\mathrm{E}\left(x_{i j}\right)=\mu_{i}$ and $\operatorname{Var}\left(x_{i j}\right)=\sigma_{i}^{2}$ for each $i=1,2$ and $\mathrm{j}=1, \ldots, n_{i}$. The two samples are assumed to be independent. Let the sample mean and
sample variance for $x_{i 1}, \ldots, x_{\mathrm{in}_{\mathrm{i}}}$ be denoted as $\mathrm{x}_{\mathrm{i}}$ and $s_{\mathrm{i}}^{2}$ for $i=1,2$, respectively.

Testing the Equality of Means
The $t$ test, the Welch test, and the Wilcoxon-Mann-Whitney test procedures of $H_{0}$ : $\mu_{1}=\mu_{2}$ vs. $H_{1}: \mu_{1} \neq \mu_{2}$, are now described.

The $t$ test is the given as

$$
\begin{gather*}
t=\frac{\left(\bar{X}_{1}-\bar{X}_{2}\right)}{\sqrt{s^{2}\left(1 / n_{1}+1 / n_{2}\right)}}  \tag{1a}\\
\text { where } \mathrm{s}^{2}=\frac{\left(n_{1}-1\right) s_{1}^{2}+\left(n_{2}-1\right) s_{2}^{2}}{\left(n_{1}+n_{2}-2\right)} \tag{1b}
\end{gather*}
$$

is the pooled estimate of $\sigma^{2}$, assuming $\sigma_{1}^{2}=\sigma_{2}^{2}=$ $\sigma^{2}$.

The Welch test statistic is

$$
\begin{equation*}
t_{w}=\frac{\left(\bar{X}_{1}-\bar{X}_{2}\right)}{\sqrt{\left(s_{1}^{2} / n_{1}\right)+\left(s_{2}^{2} / n_{2}\right)}} \tag{2a}
\end{equation*}
$$

which uses Satterthwaite's (1946) approximation for the degrees of freedom:

$$
\begin{equation*}
d f=\frac{\left(s_{1}^{2} / n_{1}+s_{2}^{2} / n_{2}\right)^{2}}{\left(s_{1}^{2} / n_{1}\right)^{2} /\left(n_{1}-1\right)+\left(s_{2}^{2} / n_{2}\right)^{2} /\left(n_{2}-1\right)} . \tag{2b}
\end{equation*}
$$

The Wilcoxon-Mann-Whitney statistic is

$$
\begin{equation*}
z=\frac{S-n_{1}\left(n_{1}+1\right) / 2-n_{1} n_{2} / 2}{\sqrt{n_{1} n_{2}\left(n_{1}+n_{2}+1\right) / 12}} \tag{3}
\end{equation*}
$$

where $S$ is the sum of the ranks assigned to the sample observations from group 1 , and $z$ is an approximate normal deviate.

The $\alpha$-level tests of $H_{0}: \mu_{1}=\mu_{2}$ vs. $H_{1}$ : $\mu_{1} \neq \mu_{2}$ are $|t|>t_{\alpha / 2, n_{1}+n_{2}-2,\left|t_{w}\right|>t_{\alpha / 2, d f} \text {, and }|z|}$ $>z_{\alpha / 2}$ for the $t$ test, the Welch test, and the Wilcoxon-Mann-Whitney test, respectively, where $z_{\alpha}$ is the upper $\alpha$-point of the standard unit normal distribution and $t_{\alpha, r}$ is the upper $\alpha$-point of a $t$ distribution with $r$ degrees of freedom.

Testing the Equality of Variances
The Levene test of $H_{0}: \sigma_{1}^{2}=\sigma_{2}^{2}$ vs. $\mathrm{H}_{1}$ : $\sigma_{1}^{2} \neq \sigma_{2}^{2}$ is now described, assuming the sampling conditions described above hold.

The Levene $\alpha$-level test is

$$
\begin{equation*}
L=\frac{\sum n_{i}\left(z_{i .}-z_{. .}\right)^{2}}{\sum \sum\left(z_{i j}-z_{i .}\right)^{2} /\left(n_{1}+n_{2}-2\right)}>F_{\alpha, 1, n_{1}+n_{2}-2} \tag{2.5}
\end{equation*}
$$

which is the one-way analysis of variance $F$-test computed on the $\mathrm{z}_{\mathrm{ij}}$ values, where $\mathrm{z}_{\mathrm{ij}}=\mid \mathrm{x}_{\mathrm{ij}}{ }^{-}$ median of group i|.

Testing of Symmetry
The Triples test, as described in a paper by Randles, Fligner, Policello, and Wolfe (1980), is a test to determine if a distribution is symmetric. The procedure used to obtain the test statistic is outlined in Perry and Stoline (2002) and is not repeated here.

Selected Configurations of Distributions, Sample Sizes and Variance Ratios Used in the Simulation

Type I error rates for testing the homogeneity of means were simulated under a variety of conditions using four probability distributions. Each of these four distributions is classified into one of two groups: (1) symmetric and (2) asymmetric.

The Results section examines the use of the TS procedure using two classes of preliminary tests (i.e., testing for variance homogeneity and testing for symmetry) preceding the test of equality of means, $H_{0}: \mu_{1}=$ $\mu_{2}$ for the two symmetric distributions: (1) normal and (2) double exponential. In addition, the Results section examines the TS procedure for the two asymmetric distributions: (1) lognormal and (2) gamma.

To evaluate the performance of the preliminary test of variance homogeneity, the following standard deviation ratios $R=\sigma_{1} / \sigma_{2}$ are used: $0.25,0.50,1.0,2.0$, and 4.0 . Clearly the standard deviations are equal when $R=1$. Sample size configurations $\left(n_{1}: n_{2}\right)$ used in the simulations are: $(10: 10),(10: 20),(10: 40)$, and
(20:20). This allows for both direct and indirect pairings to be examined.

Direct pairing occurs when either $R=$ 0.25 and 0.50 holds with any of the imbalanced samples (10:20) and (10:40). Direct pairing occurs when the group with the smaller $\sigma$ is associated with the group with the smaller sample size.

Indirect pairing occurs when either $R=$ 2.0 and 4.0 holds with any of the imbalanced sample sizes (10:20) and (10:40). Indirect pairing occurs when the group with the smaller $\sigma$ is associated with the group with the larger sample size.

Generation of Random Realizations
This section contains an outline of how the random realizations are generated for each specified distribution. As before, let $x_{11}, \ldots, x_{1 n_{1}}$ be a random sample of size $n_{1}$ from the distribution $f_{1}\left(x ; \mu_{1}, \sigma_{1}\right)$; and $x_{21}, \ldots, x_{2 \mathrm{n}_{2}}$ be a random sample of size $n_{2}$ from the distribution
$f_{2}\left(x ; \mu_{2}, \sigma_{2}\right)$, where it is assumed that the two samples are independent.

The random realizations from the standardized distribution $f_{2}\left(x ; \mu_{2}, \sigma_{2}\right)$ are generated for each of the selected distributions. For the first sample, $f_{1}\left(x ; \mu_{1}, \sigma_{1}\right)$, the random realizations are generated in the same fashion, but shape parameters and scale parameters are adjusted to yield the desired standard deviation ratio $R=\sigma_{1} / \sigma_{2}$. Details on each of the four selected distributions are outlined in Perry and Stoline (2002). The IMSL random number generator RNSET, which initializes the seed, is used in all of the simulations.

Testing the Equality of Means Using the TS Procedure

The TS procedure has been described in the Introduction section. Figure 1 is a diagram of how the TS procedure is constructed.

Figure 1. Components of the TS procedure

|  |  | $\mathrm{H}_{0}:$ Symmetry |  |
| :--- | :--- | :--- | :--- |
|  |  | $\frac{\text { Accepted }}{\downarrow}$ | $\frac{\text { Rejected }}{\downarrow}$ |
| $\mathrm{H}_{0}:$ Variance Homogeneity |  |  |  |
| $\left(\mathrm{H}_{0}: \sigma_{1}=\sigma_{2)}\right.$ | Accepted $\rightarrow$ | $t$ test | WMW test |
|  | Rejected $\rightarrow$ | Welch test | Welch test |

Notes: WMW = Wilcoxon-Mann-Whitney.

Asymmetry is concluded if at least one of the samples is declared skewed. Another alternative would be that skewness is declared significant only if both samples are skewed. It was arbitrary chosen for this simulation to use the former approach with asymmetry being concluded if at least one of the samples is declared skewed.

## Results

In this section, the performance of the TS procedure is evaluated. The "TS procedure" denotes the results of the test selection procedure using the $5 \%$ Triples test for testing symmetry and the $25 \%$ Levene test for testing variance homogeneity.

Symmetric Distributions
For each of the two symmetric distributions (i.e., normal and double exponential) as defined in Perry and Stoline (2002), the simulations are conducted for the four selected sample size combinations ( $n_{1}: n_{2}$ )= (10:10), (10:20), (10:40), and (20:20). For each of the four sample size combinations, the simulated null rejection rate is generated for the specified ratio $R=\sigma_{1} / \sigma_{2}$. These are: (1) $R=$ 0.25 , (2) $R=0.50$, (3) $R=1$ (equal variance), (4) $R=2.0$, and (5) $R=4.0$.

The results of the simulations for the two symmetric distributions are combined in Table 1. The proportions of rejections are expressed as a percent for the $t$ test, the Welch test, the Wilcoxon-Mann-Whitney test, and the TS procedure. These proportions are tabulated for each $R$ grouping combined over all (8) combinations of sample size pairs (4) and distributions (2) for the five categories listed below:

1. $\mathrm{x} \leq 2.5$ (extremely conservative)
2. $2.5<x \leq 4.0$ (conservative)
3. $4.0<\mathrm{x} \leq 6.0$ (robust)
4. $6.0<\mathrm{x} \leq 10.0$ (liberal)
5. $x>10.0$ (extremely liberal)

The value x represents the percentage of rejections for testing $\mathrm{H}_{0}: \mu_{1}=\mu_{2}$ based on 10,000
simulations for each sample size. Each entry in the following tables denotes the frequency at which $\mathrm{a}<\mathrm{x} \leq \mathrm{b}$ occurs. The outcome of the "test" is defined to be robust if the simulated null rejection rate is $>4.0$ and $\leq 6.0$.

Equal Variance Cases ( $\mathrm{R}=1$ )
Table 1 shows, as anticipated, that the $t$ test is robust for the equal variance cases. However, the other procedures are also robust. None of the procedures examined show simulated rejection rates $\leq 4.0 \%$ or $>6 \%$.

Unequal Variance Cases
Table 1 shows the $t$ test is extremely conservative in $50 \%$ of the simulations for the R $=0.25$ and 0.50 cases. The WMW test is liberal for the $\mathrm{R}=0.50$ cases and can be extremely conservative for both the $\mathrm{R}=0.25$ and the $\mathrm{R}=$ 0.50 cases. The Welch test and the TS procedure are robust for both the $\mathrm{R}=0.25$ and $\mathrm{R}=0.50$ cases.

For the $\mathrm{R}=2.0$ cases the $t$ test is extremely liberal. The WMW test tends to be liberal and can be extremely liberal. The TS procedure is reasonably robust. The Welsh test is robust.

For the $\mathrm{R}=4.0$ cases, the $t$ test and the WMW test are extremely liberal in $50 \%$ of the simulations. The Welsh test and the TS procedure are reasonably robust.

Table 1. Summary Of Symmetric Distributions Using TS Procedure: Frequency (\%) Of Simulated Null Rejection Rate (\%) With Nominal 5\% Level.

| R | Test | Extremely Conservative $\leq 2.5$ | Conservative $2.5<x \leq 4$ | Robust $4<x \leq 6$ | Liberal $6<x \leq 10$ | Extremely Liberal $x>10$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1.00 | $\sigma_{1}=\sigma_{2}$ |  |  |  |  |  |
|  | t | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | W | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | WMW | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | TS | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| $\sigma_{1} \neq \sigma_{2}$ |  |  |  |  |  |  |
| 0.50 | t | 50.0 | 0.0 | 50.0 | 0.0 | 0.0 |
|  | W | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | WMW | 25.0 | 25.0 | 0.0 | 50.0 | 0.0 |
|  | TS | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| 0.25 | t | 50.0 | 0.0 | 50.0 | 0.0 | 0.0 |
|  | W | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | WMW | 25.0 | 25.0 | 50.0 | 0.0 | 0.0 |
|  | TS | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| 2.0 | t | 0.0 | 0.0 | 50.0 | 12.5 | 37.5 |
|  | W | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
|  | WMW | 0.0 | 0.0 | 50.0 | 37.5 | 12.5 |
|  | TS | 0.0 | 0.0 | 75.0 | 25.0 | 0.0 |
| 4.0 | t | 0.0 | 0.0 | 37.5 | 12.5 | 50.0 |
|  | W | 0.0 | 12.5 | 87.5 | 0.0 | 0.0 |
|  | WMW | 0.0 | 0.0 | 0.0 | 50.0 | 50.0 |
|  | TS | 0.0 | 12.5 | 87.5 | 0.0 | 0.0 |

Notes: Table is based on the two symmetric distributions (normal and double exponential) and four sample sizes. $\mathrm{W}=$ Welch, WMW $=$ Wilcoxon-Mann-Whitney.

Based on the above simulation results, the Welch test and the TS procedure are reasonably robust for testing the Ho: $\mu_{1}=\mu_{2}$ for the symmetric cases examined.

Results For Asymmetric Distributions
To evaluate the overall performance of the procedures for varying degrees of variance heterogeneity, the results of the simulation for the two asymmetric distributions as defined in Perry and Stoline (2002) are combined in Table 2 using the same format as previously defined for the symmetric distributions.

For the gamma $(2,1)$ distribution the coefficient of skewness ranged from 0.4 when R $=0.25$ to approximately 5.7 when $\mathrm{R}=4.0$. For
the lognormal $(0,0.40)$ distribution, the coefficient of skewness ranged from 0.3 when $R$ $=0.25$ to approximately 9.6 when $\mathrm{R}=4.0$. For each value of $R$ within the gamma and lognormal case, a skewness ratio has been calculated. The skewness ratio is the skewness of distribution \#1 divided by the skewness of distribution \#2 within each gamma and lognormal case. The skewness ratios are displayed in Table 2.

Equal Variance Cases ( $\mathrm{R}=1$ )
A summary of the simulated null rejection rates for the two asymmetric distributions for the equal variance cases are presented in Table 2. The WMW test and $t$ test
are robust for the $R=1$ cases. The Welsh test is robust for approximately $88 \%$ of the $R=1$ cases. The TS procedure tends to be liberal for approximately $38 \%$ of these cases. None of the procedures are extremely liberal, extremely conservative, or conservative.

## Unequal Variance Cases

Table 2 shows the Welch test is robust in approximately $75 \%$ of the $R=0.50$ cases. The Welch test can be liberal for some $R=0.50$ cases. The $t$ test is conservative or extremely conservative for approximately $50 \%$ of the $\mathrm{R}=$ 0.50 cases. Furthermore, the $t$ test is liberal in approximately $38 \%$ of the simulations for the R $=0.50$ cases. The WMW test and the TS procedure are liberal or extremely liberal in at approximately $63 \%$ and $50 \%$, respectively, for the $\mathrm{R}=0.50$ cases.

For the $R=0.25$ cases, none of the test procedures are robust. The Welch test and the TS procedure tend to be liberal. The $t$ test is liberal $(50 \%)$ as well as extremely conservative ( $50 \%$ ). The WMW test is liberal or extremely liberal in approximately $88 \%$ of the simulations for the $R=0.25$ cases.

Table 2 shows all procedures tend to be liberal or extremely liberal for the $R=2.00$ cases. Furthermore, all procedures are extremely liberal for $100 \%$ of the $R=4$ cases.

In summary for the $R=1$ cases, the $t$ test, the Welsh test, and the WMW test are robust in at least $87 \%$ of the simulations. The TS procedure is robust in approximately $63 \%$ of the simulations for the $\mathrm{R}=1$ cases. For the $R=0.50$ cases, the Welch test is robust for approximately $75 \%$ of the simulated cases. For the $R=0.25,2.0$ and 4.0 cases, all procedures tend to be liberal. The degree of liberal bias increases as the degree of variance heterogeneity increases.

Frequency (\%) Each Means Test Is Used
In addition to the simulated null rejection rates, the TS procedure can report the frequency (\%) at which each of the test procedures is used for a given sample size and $R$ value. Results for the imbalanced case $n_{1}=10$ and $n_{2}=20$, and the balanced case $n_{1}=n_{2}=20$ are summarized for the two symmetric distribution cases combined and the two asymmetric distribution cases combined.

Tables 3 and 4 summarize the frequency (\%) at which each of the test procedures is used for the two symmetric distributions cases combined, and the two asymmetric cases combined, respectively. The format for Tables 3 and 4 is as follows. For each $R$ value, the frequency at which the $t$ test, the Welch-S test, the WMW test, and the Welch-AS test was selected by the TS procedure is reported. In these tables, the $t$ test, Welch-S, WMW, and Welch-AS denote the following:
$t$ test: The $t$ test was used because the TS procedure concluded $\sigma_{1}=\sigma_{2}$ and symmetry was accepted.

Welch-S: The Welch test was used because the TS procedure concluded $\sigma_{1} \neq \sigma_{2}$ and symmetry was accepted.

WMW: The WMW test was used because the TS procedure concluded $\sigma_{1}=\sigma_{2}$ and symmetry was rejected.

Welch-AS: The Welch test was used because the TS procedure concluded $\sigma_{1} \neq \sigma_{2}$ and symmetry was rejected.

Table 2. Summary Of Asymmetric Distributions Using TS Procedure: Frequency (\%) Of Simulated Null Rejection Rate With Nominal 5\% Level.


Notes: Table is based on the two asymmetric distributions [lognormal $(0,0.40) \& G(2,1)]$ and four sample sizes. The skewness ratio is the skewness for distribution \#1/distribution \#2 for each gamma and lognormal case, respectively, at each R value. $\mathrm{W}=\mathrm{Welch}$, WMW = Wilcoxon-Mann-Whitney.

## Symmetric Cases

Table 3 contains the frequency (\%) at which each of the test procedures is used in the two symmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Includes the Imbalanced and Balanced Cases)

For the $R=1.00$ case with equal sample sizes, the $t$ test is known to be robust for the symmetric distributions. Results in Table 3 show that the TS procedure correctly selected the $t$ test
for approximately $69 \%$ of the simulations. The Welch-S test was incorrectly selected for approximately $22 \%$ of the simulations when using the TS procedure. The WMW test was incorrectly selected for only $7 \%$ of the simulations when using the TS procedure.

For the $R=1.00$ case with unequal sample sizes, Table 3 shows that the TS procedure selected the $t$ test for $70 \%$ of the simulations. The TS procedure incorrectly selected the Welch-S test for nearly $23 \%$ of the simulations. However, the WMW test was incorrectly selected for less
than $6 \%$ of the simulations when using the TS procedure.

Unequal Variances (Includes the Imbalanced and Balanced Cases)

For the $R=0.50$ and 2.0 cases with equal sample sizes, Table 3 shows the TS procedure correctly selected the Welch-S test for approximately $81 \%$ of the simulations. The TS procedure incorrectly selected the $t$ test in
approximately $10 \%$ of the simulations and incorrectly concluded asymmetry in approximately $9 \%$ of the simulations.

For the $R=0.50$ and 2.0 cases with unequal sample sizes, Table 3 shows the TS procedure correctly selected the Welch-S test for about $70 \%-73 \%$ of the simulations. The TS procedure incorrectly selected the $t$ test for about $20 \%-23 \%$ of the simulations.

Table 3. Frequency (\%) At Which Each Means Test Is Used In The TS Procedure For The Symmetric Distributions.

| $\mathrm{n}_{1}, \mathrm{n}_{2}$ | $\sigma_{1}, \sigma_{2}$ | $R$ | $t$ test | Welch-S | WMW | Welch-AS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 68.91 | 21.96 | 7.10 | 2.04 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.09 | 90.78 | $<0.01$ | 9.13 |
|  |  | 0.50 | 10.44 | 80.43 | 1.30 | 7.84 |
|  |  | 2.00 | 10.33 | 80.54 | 1.28 | 7.86 |
| 10,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 70.30 | 22.69 | 5.64 | 1.38 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.58 | 92.41 | 0.05 | 7.00 |
|  |  | 0.50 | 20.02 | 72.97 | 2.06 | 4.96 |
|  |  | 2.00 | 22.76 | 70.23 | 1.92 | 5.10 |
|  |  | 4.00 | 0.97 | 92.02 | 0.15 | 6.87 |

For the $R=0.25$, and 4.0 symmetric cases, the Welch test is known to be robust. Table 3 shows the TS procedure correctly used the Welch-S test for about $90 \%-92 \%$ of the simulations regardless of the sample size configurations. The Welch-AS test was incorrectly used for about 7-9\% of the simulations for each of these same cases.

In summary, for the combined symmetric cases, the TS procedure correctly selected the $t$ test for approximately $70 \%$ of the $R=1$ cases regardless of the sample size configuration. For the $R=0.50$ and 2.0 cases, the TS procedure correctly selected the Welch-S test for approximately $81 \%$ of the simulations with equal sample sizes and about $70 \%-73 \%$ of the simula-
tions with unequal sample sizes. For the $R=0.25$ and 4.0 cases, regardless of sample size configuration, the TS procedure correctly used the Welch-S test for about $90 \%-92 \%$ of the simulations. It is noted for the $R \neq 1$ cases, the TS procedure incorrectly concluded asymmetry for about $7 \%-9 \%$ of the simulations.

Asymmetric Cases
Table 4 contains the frequency (\%) at which each of the test procedures is used in the two asymmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Includes the Imbalanced and Balanced Cases)

For the $R=1$ case with equal sample sizes, the WMW test is known to be robust for the asymmetric distributions. Results in Table 4 shows the TS procedure correctly selected the WMW test for approximately $42 \%$ of the simulations. The TS procedure incorrectly selected the Welch-AS test in approximately $12 \%$ of the simulations with homogeneous variances. The $t$ test was incorrectly selected by the TS procedure in approximately $33 \%$ of the simulations.

For the $R=1$ cases with unequal sample sizes, Table 4 shows the TS procedure correctly selected the WMW test for approximately $31 \%$ of the simulations. As also seen for the balanced sample size cases, the TS procedure incorrectly selected the Welch-AS test in approximately $8 \%$ of these cases. In addition, the $t$ test was incorrectly selected by the TS procedure in approximately $45 \%$ of the simulations.

Unequal Variances (Imbalanced and Balanced Cases)

For the equal sample size cases, Table 4 shows the TS procedure incorrectly selected the Welsh-S in approximately $50 \%$ of the $\mathrm{R}=0.50$ cases and approximately $10 \%$ of the $\mathrm{R}=2.0$ cases. Furthermore, the TS procedure incorrectly selected the WMW test in approximately $6 \%$ of the $R=0.50$ cases and approximately $36 \%$ in the $R=2.0$ cases. The Welch-AS test was correctly selected for approximately $35 \%$ and $47 \%$ of the $R$ $=0.50$ and 2.0 cases, respectively, when using the TS procedure.

For the $R=0.50$ and 2.0 cases with imbalanced sample sizes, results in Table 4 shows the same trends as was seen for the equal sample size cases. The TS procedure incorrectly used the WMW test for approximately $10 \%$ of the $R=0.50$ and approximately $28 \%$ in the $R=2.0$ cases; and correctly selected the Welch-AS test for about 25$26 \%$ of the $R=0.50$ and 2.0 cases.

Results in Table 4 shows for the balanced case that the TS procedure correctly selected the Welch-AS test for approximately $37 \%$ of the $R=$ 0.25 cases. The WMW test was incorrectly used for about $2 \%$ of the $R=0.25$ cases.

Results in Table 4 for the unequal sample size case show that the TS procedure correctly used the Welch-AS test for approximately $35 \%$ of the $R=0.25$ cases, whereas the WMW test and the Welch-S test were each incorrectly selected for about $20 \%$ and $65 \%$, respectively, of the $R=$ 0.25 cases.

The TS procedure incorrectly used the WMW test for approximately $43 \%$ of the $R=4.0$ cases and the Welch-AS test was correctly used for about $52 \%$ of the $R=4.0$ equal sample size. For the $\mathrm{R}=4.0$ unequal sample size cases, the TS procedure incorrectly used the WMW test for approximately $32 \%$ of the simulations and the Welsh-AS test was correctly used for approximately $43 \%$ of the simulations.

In summary, for the $R=1$ cases regardless of the sample size configuration, the TS procedure used the WMW test correctly for about $31 \%-42 \%$ of the simulations. For the $R=0.50$ cases, the WMW test was incorrectly selected for about $6 \%-10 \%$ of the simulations when using the TS procedure. The TS procedure generally correctly used the Welch-AS test for about $35 \%$ $37 \%$ of the 0.25 cases. For the R= 2.0 cases, the TS procedure selected the Welsh-AS test correctly for about $25 \%-47 \%$ of the simulations and the WMW test incorrectly for about $28 \%-36 \%$ of the simulations. The TS procedure selected the Welch-AS test correctly for about $43 \%-52 \%$ of the simulations and the WMW test incorrectly each for about $32 \%-43 \%$ of the simulations for the $\mathrm{R}=4.0$ cases.

Summary of the TS Procedure Using an Alpha Level of $5 \%$ of the Triple's Test

For the cases where variance homogeneity and symmetry each are unknown to the practicing statistician, an overall test using the TS procedure yielded improved results with respect to robustness over using the $t$ test or the Wilcoxon-Mann-Whitney test alone, except for the asymmetry unequal variance cases, where no method maintained the stated Type I error rate. The Welch test is recommended as a robust test for testing $H_{0}: \mu_{1}=\mu_{2}$ for the symmetric cases examined. The TS procedure is also reasonably robust.

Table 4. Frequency (\%) At Which Each Means Test Is Used In The TS Procedure For The Asymmetric Distributions.

| $\mathrm{n}_{1}, \mathrm{n}_{2}$ | $\sigma_{1}, \sigma_{2}$ | $R$ | $t$ test | Welch-S | WMW | Welch-AS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 32.98 | 12.43 | 42.22 | 12.38 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.02 | 63.31 | 0.02 | 36.66 |
|  |  | 0.50 | 9.47 | 49.95 | 5.90 | 34.69 |
|  |  | 2.00 | 6.88 | 9.81 | 36.21 | 47.11 |
|  |  | 4.00 | 1.79 | 2.63 | 43.26 | 52.33 |
| 10,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 45.02 | 16.59 | 30.50 | 7.90 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.43 | 64.58 | 0.20 | 34.79 |
|  |  | 0.50 | 17.73 | 46.61 | 9.79 | 25.88 |
|  |  | 2.00 | 21.73 | 24.83 | 28.32 | 25.13 |
|  |  | 4.00 | 9.74 | 15.74 | 31.55 | 42.98 |

The performance of the TS procedure was also evaluated by the frequency at which the TS procedure selected the most appropriate test of means. For the symmetric equal variance cases, the TS procedure correctly selected the $t$ test for approximately $70 \%$ of the simulated. For the symmetric cases with unequal variances $(R=$ $0.25,0.50,2.0$, and 4.0 ), the frequency at which the Welch test was correctly selected was about $70 \%-92 \%$ for the TS procedure. Asymmetry was incorrectly concluded for about $7 \%-9 \%$ of the simulated symmetric cases when using the TS procedure.

The TS procedure correctly concluded asymmetry for about $35 \%-96 \%$ of the simulated cases for the families of asymmetric distributions examined. For the asymmetric equal variance cases, the TS procedure correctly selected the Wilcoxon-Mann-Whitney test for about $31 \%-42 \%$ of the simulations. For the asymmetric cases with unequal variances, the TS procedure correctly concluded asymmetry and variance heterogeneity for about $25 \%-52 \%$ of the simulations.

Results showed that the TS procedure concluded symmetry too often (for $45 \%-62 \%$ of the asymmetric cases with equal variances).

Since the TS procedure examined in this simulation study concluded symmetry too often, it would be of interest to examine the performance
of an TS procedure using the Triples test for testing of symmetry at a higher significance level such as $\alpha=0.25$.

Further Investigation of the TS Procedure Using an Alpha Level of $25 \%$ for the Test of Symmetry

As the results above showed that the TS procedure was concluding symmetry too often, the simulations were repeated using the TS procedure with the alpha level set at $25 \%$ for the Triples test. To compare the TS procedure using the Triples test at alpha level $25 \%$ versus $5 \%$, only the results of the frequency (\%) at which each means test is used are displayed.

Tables 5 and 6 summarize the frequency (\%) at which each of the test procedures is used for the two symmetric distributions cases combined, and the two asymmetric cases combined, respectively. The format for Tables 5 and 6 is the same as described above in section "Frequency (\%) at Which Each Mean Test is Used."

Frequency (\%) Each Means Test is Used For Symmetric Cases

Table 5 contains the frequency (\%) at which each of the test procedures is used in the two symmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Imbalanced and Balanced Cases)

For the $R=1.00$ case with equal sample sizes, the $t$ test is known to be robust for the symmetric distributions. Results in Table 5 show that the TS procedure correctly selected the $t$ test for approximately $46 \%$ of the simulations. The Welch-S test was incorrectly selected for approximately $14 \%$ of the simulations when using the TS procedure. The WMW test was incorrectly selected for only $32 \%$ of the simulations when using the TS procedure.

For the $R=1.00$ case with unequal sample sizes, Table 5 shows that the TS procedure selected the $t$ test for approximately $48 \%$ of the simulations. The TS procedure incorrectly selected the Welch-S test for approximately $15 \%$ of the simulations. However, the WMW test was incorrectly selected for about $30 \%$ of the simulations when using the TS procedure.

Unequal Variances (Imbalanced and Balanced Cases)

For the $R=0.50$ and 2.0 cases with equal sample sizes, Table 5 shows the TS procedure correctly selected the Welch-S test for approximately $58 \%$ of the simulations. The TS procedure incorrectly selected the $t$ test in approximately $2 \%$ of the simulations and incorrectly concluded asymmetry in approximately $40 \%$ of the simulations.

For the $R=0.50$ and 2.0 cases with unequal sample sizes, Table 5 shows the TS procedure correctly selected the Welch-S test for about $54 \%-57 \%$ of the simulations. The TS procedure incorrectly selected the $t$ test for about $6 \%-9 \%$ of the simulations.

For the $R=0.25$, and 4.0 symmetric cases, the Welch test is known to be robust. Table 5 shows the TS procedure correctly used the Welch-S test for about $60 \%-63 \%$ of the simulations regardless of the sample size configurations. The Welch-AS test was incorrectly used for about $37 \%-40 \%$ of the simulations for each of these same cases.

Table 5. Frequency (\%) At Which Each Means Test Is Used In The TS Procedure For The Symmetric Distributions.

| $\mathrm{n}_{1}, \mathrm{n}_{2}$ | $\sigma_{1}, \sigma_{2}$ | $R$ | $t$ test | Welch-S | WMW | Welch-AS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 46.38 | 13.77 | 32.42 | 7.44 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.00 | 60.15 | 0.01 | 39.85 |
|  |  | 0.50 | 2.38 | 57.77 | 1.78 | 38.08 |
|  |  | 2.00 | 2.46 | 57.69 | 1.67 | 38.19 |
|  |  | 4.00 | 0.00 | 60.15 | 0.00 | 39.81 |
| 10,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 47.89 | 15.14 | 29.97 | 7.01 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.02 | 63.00 | 0.01 | 36.98 |
|  |  | 0.50 | 6.24 | 56.78 | 4.31 | 32.68 |
|  |  | 2.00 | 9.25 | 53.77 | 6.22 | 30.76 |
|  |  | 4.00 | 0.11 | 62.92 | 0.07 | 36.91 |

In summary, for the combined symmetric cases, the TS procedure correctly selected the $t$ test for approximately $47 \%$ of the $R=1$ cases regardless of the sample size configuration. For the $R=0.50$ and 2.0 cases, the TS procedure correctly selected the Welch-S test for approximately $58 \%$ of the simulations with equal sample sizes and about $54 \%-57 \%$ of the simulations with unequal sample sizes. For the $R$ $=0.25$ and 4.0 cases, regardless of sample size configuration, the TS procedure correctly used the Welch-S test for about $60 \%-63 \%$ of the simulations. It is noted for the $R \neq 1$ cases, the TS procedure incorrectly concluded asymmetry for about $37 \%-40 \%$ of the simulations.

Frequency (\%) Each Means Test is Used For Asymmetric Cases

Table 6 contains the frequency (\%) at which each of the test procedures is used in the two asymmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Imbalanced and Balanced Cases)

For the $R=1$ case with equal sample sizes, the WMW test is known to be robust for the asymmetric distributions. Results in Table 6 show
the TS procedure correctly selected the WMW test for approximately $67 \%$ of the simulations. The TS procedure incorrectly selected the WelchAS test in approximately $22 \%$ of the simulations with homogeneous variances. The $t$ test was incorrectly selected by the TS procedure in approximately $8 \%$ of the simulations.

For the $R=1$ cases with unequal sample sizes, Table 6 shows the TS procedure correctly selected the WMW test for approximately $60 \%$ of the simulations. As also seen for the balanced sample size cases, the TS procedure incorrectly selected the Welch-AS test in approximately $19 \%$ of these cases. In addition, the $t$ test was incorrectly selected by the TS procedure in approximately $15 \%$ of the simulations.

Unequal Variances (Imbalanced and Balanced Cases)

For the equal sample size cases, Table 6 shows the TS procedure incorrectly selected the Welsh-S in approximately $25 \%$ of the $\mathrm{R}=0.25$ cases. Furthermore, the TS procedure incorrectly selected the WMW test in approximately $12 \%$ of the $R=0.50$ cases and approximately $43 \%$ in the $R=2.0$ cases. The Welch-AS test was correctly selected for approximately $67 \%$ and $55 \%$ of the $R$ $=0.50$ and 2.0 cases, respectively, when using the TS procedure.

Table 6. Frequency (\%) For Means Test In The TS Procedure For The Asymmetric Distributions.

| $\mathrm{n}_{1}, \mathrm{n}_{2}$ | $\sigma_{1}, \sigma_{2}$ | $R$ | $t$ test | Welch-S | WMW | Welch-AS |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 20,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 8.05 | 3.48 | 66.95 | 21.53 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.01 | 24.85 | 0.03 | 75.12 |
|  |  | 0.50 | 3.43 | 17.52 | 11.85 | 67.17 |
|  |  | 2.00 | 1.16 | 1.49 | 42.72 | 54.65 |
|  |  | 4.00 | 0.16 | 0.22 | 45.08 | 54.55 |
| 10,20 | $\sigma_{1}=\sigma_{2}$ | 1.00 | 14.78 | 6.34 | 60.04 | 18.85 |
|  | $\sigma_{1} \neq \sigma_{2}$ | 0.25 | 0.22 | 27.49 | 0.40 | 71.90 |
|  |  | 0.50 | 7.15 | 18.80 | 20.83 | 53.23 |
|  |  | 2.00 | 5.36 | 6.27 | 44.27 | 44.11 |
|  |  | 4.00 | 1.88 | 3.05 | 39.18 | 55.90 |

For the $R=0.50$ and 2.0 cases with imbalanced sample sizes, results in Table 6 shows the same trends as was seen for the equal sample size cases. The TS procedure incorrectly used the WMW test for approximately $21 \%$ of the $R=0.50$ and approximately $44 \%$ in the $R=2.0$ cases; and correctly selected the Welch-AS test for about $44 \%-53 \%$ of the $R=0.50$ and 2.0 cases.

Results in Table 6 shows for the balanced case that the TS procedure correctly selected the Welch-AS test for approximately $75 \%$ of the $R=$ 0.25 cases. The Welch-S test was incorrectly used for about $25 \%$ of the $R=0.25$ cases.

Results in Table 6 for the unequal sample size case show that the TS procedure correctly used the Welch-AS test for approximately $72 \%$ of the $R=0.25$ cases, whereas the Welch-S test was incorrectly selected for about $27 \%$ of the $R=0.25$ cases.

The TS procedure incorrectly used the WMW test for approximately $45 \%$ of the $R=4.0$ cases and the Welch-AS test was correctly used for about $55 \%$ of the $R=4.0$ equal sample size. For the $\mathrm{R}=4.0$ unequal sample size cases, the TS procedure incorrectly used the WMW test for approximately $39 \%$ of the simulations and the Welsh-AS test was correctly used for approximately $56 \%$ of the simulations.

In summary, for the $R=1$ cases regardless of the sample size configuration, the TS procedure used the WMW test correctly for about $60 \%-67 \%$ of the simulations. For the $R=0.50$ cases, the WMW test was incorrectly selected for about $12 \%-21 \%$ of the simulations when using the TS procedure. The TS procedure generally correctly used the Welch-AS test for about 72\%$75 \%$ of the 0.25 cases. For the $\mathrm{R}=2.0$ cases, the TS procedure selected the Welsh-AS test correctly for about $44 \%-55 \%$ of the simulations and the WMW test incorrectly for about $43 \%-44 \%$ of the simulations. The TS procedure selected the Welch-AS test correctly for about $55 \%-56 \%$ of the simulations and the WMW test incorrectly each for about $39 \%-45 \%$ of the simulations for the $\mathrm{R}=4.0$ cases.

## Conclusion

For the TS procedure using the Triples test with an alpha level of $5 \%$, results showed that the TS
procedure concluded symmetry too often (for $45 \%-62 \%$ of the asymmetric cases with equal variances).

For the TS procedure using the Triples test at an alpha level of $25 \%$, results showed that the TS procedure concluded asymmetry for the symmetric distributions in $37 \%-40 \%$ of the $\mathrm{R} \neq 1$ cases.

Recommendations for alternative approaches in the future, would be to examine the performance of an TS procedure which concludes asymmetry at an alpha level between $5 \%$ and $25 \%$ (i.e., $15 \%$ ) or concludes asymmetry only if both samples were judged to be nonsymmetric at $\alpha=$ 0.25 . In addition, there was a trend, especially in the asymmetric distributions, of concluding variance homogeneity too often for the $\mathrm{R} \neq 1$ cases. Therefore, it would be recommended to increase alpha level for testing of variance homogeneity to a higher alpha level beyond $\alpha=$ 0.25 .

## References

Brown, M. B. \& Forsythe, A. B. (1974, June). Robust tests for the equality of variances. Journal of the American Statistical Association, 69 (346), 364-367.

Brown, M. B. \& Forsythe, A.B. (1974b). The small behavior of some statistics which test the equality of several means. Technometrics, 16, 129-132.

Cabilio, P. \& Masaro, J. (1996). A simple test of symmetry about an unknown median. The Canadian Journal of Statistics, 24(3), 349-361.

Conover, W. J., Johnson, M. E., \& Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23 (4), 351-361.

Gans, D. J. (1981). Use of a preliminary test in comparing two sample means. Communication in Statistics B, Simulation \& Computation, B10 (2), 163-174.

Gupta, M. K. (1967). An asymptotically nonparametric test of symmetry. Annals of Mathematical Statistics, 38, 849-866.

IMSL. (1989, January). Math/Library User's Manual (Version 1.1). Houston, Texas: Author.

IMSL. (1989, December). Math/Library User's Manual (Version 1.1). Houston, Texas: Author.

Loh, W. (1987, May 21). Some modifications of Levene's test of variance homogeneity. Journal of Statistical Computation \& Simulation, 28, 213-226.

Murphy, B. P. (1976). Comparison of some two sample means tests by simulation. Communication in Statistics B, Simulation and Computation, B5 (1), 23-32.

O’Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. Journal of the American Statistical Association, 74, 877-880.

Olejnik, S. (1987). Conditional ANOVA for mean differences when population variances are unknown. Journal of Experimental Education, 55, 141-148.

Perry, K. T. \& Stoline, M. R. (2002). A comparison of the D'Agostino $S_{U}$ test to the Triples test for testing of symmetry versus asymmetry as a preliminary test to testing the equality of means. Journal of Modern Applied Statistical Methods, 1 (2), 316-325.

Randles, R. H., Fligner, M. A., Policello II, G.E., \& Wolfe, D.A. (1980, March). An asymptotically distribution-free test for symmetry versus asymmetry. Journal of the American Statistical Association, 75 (369), 168-172.

Snedecor, G. W. \& Cochran, W. G. (1967). Statistical methods. Ames, Iowa: The Iowa State University Press.

Zimmerman, D. W. \& Williams, R. H. (1989). Power comparisons of the student $t$-test and two approximations when variances and sample sizes are unequal. Journal of Indian Society Agricultural Statistics, 41 (2), 206-217.

# A Comparison Of Equivalence Testing In Combination With Hypothesis Testing And Effect Sizes 

Christopher J. Mecklin<br>Department of Mathematics and Statistics<br>Murray State University

Equivalence testing, an alternative to testing for statistical significance, is little used in educational research. Equivalence testing is useful in situations where the researcher wishes to show that two means are not significantly different. A simulation study assessed the relationships between effect size, sample size, statistical significance, and statistical equivalence.

Key words: Equivalence testing, statistical significance, effect size

## Introduction

The use of statistical inference, particularly via null hypothesis significance testing, is an extremely common but contentious practice in educational research. Both the pros and the cons of hypothesis testing have been argued in the literature for several decades. A recent monograph edited by Harlow, Mulaik, and Steiger (1997) was devoted to these arguments. Some classic references criticizing standard hypothesis testing include Boring (1919), Berkson (1938, 1942), Rozeboom (1960), Meehl (1967, 1978), and Carver (1978). More recently, some support the continued usage of significance testing (Abelson, 1997; Hagan, 1997, 1998; Harris, 1997; McLean \& Ernest, 1998), while others desire a greater reliance on alternatives such as confidence intervals or effect sizes (Cohen, 1992, 1994; Knapp, 1998, 2002; Meehl, 1997; Serlin, 2002; Thompson, 1998, 2001; Vacha-Haase, 2001), and still others advocate an outright ban on significance testing (Carver, 1993; Falk, 1998; Hunter, 1997; Nix \& Barnette, 1998; Schmidt \& Hunter, 1997).

Christopher Mecklin is an Assistant Professor of Mathematics \& Statistics. His research interests include goodness-of-fit, educational statistics and statistical ecology. He enjoys working with faculty and students from various disciplines. Email:christopher.mecklin@murraystate.edu.

The references included here are by no means close to an exhaustive list. This debate is not limited to educational research and the social sciences; for instance, it is also being argued in ecology (McBride, 1999; Anderson, Burnham, \& Thompson, 2000). Many in the statistical community outside of the niche of educational and psychological research, though, are either unaware of this debate or feel that it is trivial (Krantz, 1999).

The objective of this paper is not to continue this heated argument, but rather to borrow the method of equivalence testing from biostatistics, as suggested by Bartko (1991), and using it in conjunction with standard hypothesis testing in educational research. Lehmann (1959) anticipated the need for interval testing in his classic volume on the theory of hypothesis testing. Many of the currently employed methods of equivalence testing were developed in the 1970's and 1980's to address biostatistical and pharmaceutical problems (Westlake, 1976, 1979; Schuirmann, 1981, 1987; Anderson \& Hauck, 1983; Patel \& Gupta, 1984). Rogers, Howard, and Vessey (1993) introduced the use of equivalence testing methods to the social sciences. Serlin (1993) essentially suggested equivalence testing when he suggested the use of range, rather than point, null hypotheses.

## Methodology

Standard null hypothesis significance testing dates back to the pioneering theoretical work of

Fisher, Neyman, and Pearson. Hypothesis testing can be found in almost every textbook of statistical methods and thus will not be further elaborated on here. Equivalence testing, on the other hand, is a newer technique and one that is unfamiliar to most researchers in education and the social sciences.

Equivalence testing was developed in biostatistics to address the situation where the goal is not to show that the mean of one group is greater than the mean of another group (i.e. the superiority of one treatment to another), but rather to establish that two methods are equal to one another. A common application of this idea in biostatistics is to show that a less expensive "generic" medication is as effective as the more expensive "brand-name" medication. In equivalence testing, the null hypothesis is that the two groups are not equivalent to one another, and hence rejection of the null indicates that the two groups are equivalent. This differs from standard significance testing where the null hypothesis states that the group means are equal and rejection of the null indicates that the two groups are statistically different. A common methodological mistake in research is to conclude that the null hypothesis is true (i.e. two groups have equal means) based on the failure to reject it. This action fails to recognize that the failure to reject the null is often merely a Type II error, especially when the sample sizes are small and the power of the test is low.

An explanation of the theory of equivalence testing can be found in Berger and Hsu (1996); Blair and Cole (2002) give a less technical explanation. Here, we will merely review the most commonly implemented method used for establishing the equivalence of two population means for an additive model, where the difference of means is considered. The multiplicative model, which looks at the ratio of means, will not be considered further in this paper. The commonly used procedure in biostatistics for this problem is to use the "two one-sided tests" procedure, or TOST (Westlake, 1976, 1979; Schuirmann, 1981, 1987). With the TOST, the researcher will consider two groups equivalent if he can show that they differ by less than some constant $\tau$, the equivalence bound, in both directions. The constant $\tau$ is often chosen to be a percentage (such as $10 \%$ or $20 \%$ ) of the
mean of the control group, although $\tau$ can also be chosen to be a constant that is the smallest absolute difference between two means that is large enough to be practically important.

The null hypothesis (i.e. the means are different) for the TOST is $H_{0}:\left|\mu_{1}-\mu_{2}\right| \geq \tau$. The alternative hypothesis (i.e. the means are equivalent) is $H_{1}:\left|\mu_{1}-\mu_{2}\right|<\tau$.

The first one-sided test seeks to reject the null hypothesis that the difference between two means is less than or equal to ${ }^{-\tau}$; similarly, the second one-sided test seeks to reject the null hypothesis that the difference in the means is greater than or equal to ${ }^{\tau}$. If the one-sided test with the larger p -value leads to rejection, then the two groups are considered to be equivalent.

For the first one-sided test, we compute the test statistic

$$
t_{1}=\frac{x_{1}-x_{2}+\tau x_{2}}{s_{p} \sqrt{1 / n_{1}+1 / n_{2}}}
$$

where $S_{p}$ is the pooled standard deviation of the two samples and compute the p -value as

$$
p_{1}=P\left(t_{v}>t_{1}\right)
$$

where $t_{v}$ is a random variable from the t distribution with $v=n_{1}+n_{2}-2$ degrees of freedom.

The second one-sided test is similar to the first. The test statistic is

$$
t_{2}=\frac{x_{1}-x_{2}-\tau x_{2}}{s_{p} \sqrt{1 / n_{1}+1 / n_{2}}}
$$

and the $p$-value is

$$
p_{2}=P\left(t_{v}<t_{2}\right) .
$$

If we let $p=\max \left(p_{1}, p_{2}\right)$, then the null hypothesis of nonequivalence is rejected if $p<\alpha$.

The choice of $\tau$ is a difficult choice that is up to the researcher. This choice is analogous to the selection of an appropriate alpha level in
standard significance testing, an appropriate level of confidence in interval estimation, or a sufficiently large effect size, and should be made carefully. Knowledge of the situation at hand should be used to specify the maximum difference between population means that would be considered clinically trivial. Researchers in biostatistics typically have the choice made for them by government regulation.

As in standard hypothesis testing, an equivalency confidence interval can also be constructed. If the entire confidence interval is within $(-\tau, \tau)$, then equivalence between the groups is indicated. If the entire confidence interval is within either $(-\tau, 0)$ or $(0, \tau)$ (i.e.
zero is not in the interval), then we would reject the null hypotheses of both a significance and an equivalence test. In that case, we could make the somewhat discomforting conclusion that the difference of means was both statistically significant and equivalent.

It is important to note that the equivalency confidence interval is expressed at the $100(1-2 \alpha) \%$ level of confidence. Rogers et al. (1993) noted that if one performs both a standard significance test and an equivalence test on the same data set, making either a "reject" or "fail to reject" decision, that there are four possibilities. These four conditions are given in Table 1.

Table 1. Possible Combinations of Significance and Equivalence Testing

| Significance Test | Equivalence Test | Term |
| :--- | :--- | :--- |
| Fail to reject | Reject | Equivalent |
| Reject | Reject | Equivalent and Different |
| Reject | Fail to reject | Different |
| Fail to reject | Fail to reject | Equivocal |

The second condition "equivalent and different", a simultaneous rejection of both inferential procedures, could happen in a situation where large samples provide "too much power", resulting in a trivial difference in means being statistically significant. The equivalence test (and the effect size) should detect the small magnitude of these mean differences. The fourth condition indicates that there is insufficient evidence to conclude that the groups are either equivalent or different. This would most likely occur when the samples are very small and/or the group variances are very large.

The effect size for the difference of means is the standardized difference between the groups (Fan, 2001). We will use the parameter

$$
\delta=\frac{\mu_{1}-\mu_{2}}{\sigma}
$$

to represent the effect size of the population, where $\mu_{1}$ and $\mu_{2}$ are the population means and $\sigma^{2}$ is the common variance.

Of course, $\delta$ is typically unknown and
needs to be estimated. Cohen's d (1988) is a statistic often used for this purpose. The effect size (ES) is found with

$$
d=\frac{x_{1}-x_{2}}{S_{\text {pooled }}}
$$

where

$$
s_{\text {pooled }}=\frac{\left(n_{1}-1\right) s_{1}^{2}+\left(n_{2}-1\right) s_{2}^{2}}{n_{1}+n_{2}-2}
$$

is the pooled standard deviation of the two samples. We stress that Cohen's $d$ is a sample statistic and has a sampling distribution like other estimates.

Cohen (1988) gave some suggestions for interpreting d . An effect size of $\mathrm{d}=0.2$ is deemed "small", $\mathrm{d}=0.5$ is "medium", and $\mathrm{d}=0.8$ is "large". It is becoming, rather regrettably in our opinion, common for researchers to rigidly apply Cohen's suggestions. Absolute reliance on Cohen's rule of thumb is as misguided as blind adherence to a particular level of significance
(e.g. $\alpha=0.05$ ). As Thompson (2001) said, "we would merely be being stupid in another metric."

Results
Rogers et al. (1993) provided empirical examples of the application of equivalence testing on data from the psychological literature. We will do the same with an example from the educational research literature. This will demonstrate that there often exist situations where a statistically significant difference between groups coincides with the groups being statistically equivalent. This is the "equivalent and different" condition that is typically associated with a small to moderate effect size, as opposed to the strong effect sizes that typically occur with the "different" condition and the weak effect sizes that occur with the "equivalent" condition.

Benson (1989), in a study concerning statistical test anxiety, presented means and variances for a sample of 94 males and 123 females on seven variables. Using standard hypothesis testing methods (i.e. t-tests), significant group differences were found for: prior math courses, math self-concept, self-
efficacy, and statistical test anxiety. However, after calculating Cohen's d as an effect size (ES) measure and the use of the TOST equivalence test, we see that only prior math courses and statistical test anxiety are "different" between males and females. Not surprisingly, the two largest effect sizes are found for these two variables. Table 2 shows results of both traditional significance and equivalence tests for the Benson data.

Statistical significance was defined as a rejection of $H_{0}$ with $\alpha=0.05$ and equivalence was defined as a rejection of $H_{0}$ with $\alpha=0.10$. The reason for the two different significance levels is because while a traditional significance test at level $\alpha$ corresponds to a $100(1-\alpha) \%$ confidence interval, an equivalence test at level $\alpha$ corresponds to a $100(1-2 \alpha) \%$ equivalence interval. We selected $\tau=0.2$ (i.e. $20 \%$ of the mean of the female group). This choice was arbitrary and by no means should be taken as a choice recommended for all equivalence problems. The results could differ with different choices for ${ }^{\tau}$.

Table 2. Comparing Significance and Equivalence Testing for the Benson Data

| Descriptive Statistics |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\begin{gathered} \hline \text { Males } \\ (\mathrm{N}=94) \end{gathered}$ |  | $\begin{aligned} & \text { Females } \\ & (\mathrm{N}=123) \end{aligned}$ |  |  |  |  |  |
| Variable | M | SD | M | SD | Effect Size | Sig. pvalue | Equiv. pvalue | Category |
| GPA | 3.05 | 0.44 | 3.16 | 0.47 | -0.24 | 0.040 | $<0.001$ | Equiv. \& Diff. |
| Prior Math Courses | 3.45 | 2.14 | 2.20 | 2.01 | 0.60 | $<0.001$ | 0.998 | Different |
| Math SelfConcept | 25.77 | 5.96 | 23.20 | 7.05 | 0.39 | 0.002 | 0.012 | Equiv. \& Diff. |
| Self-efficacy | 12.68 | 1.77 | 11.62 | 2.30 | 0.51 | $<0.001$ | <0.001 | Equiv. \& Diff. |
| General Test Anxiety | 36.38 | 0.49 | 40.62 | 12.25 | -0.37 | 0.004 | 0.007 | Equiv. \& Diff. |
| Achievement | 32.56 | 5.68 | 32.26 | 7.55 | 0.04 | 0.374 | $<0.001$ | Equivalent |
| Statistical <br> Test Anxiety | 32.65 | 12.57 | 41.84 | 14.83 | -0.66 | <0.001 | 0.663 | Different |

For a test of statistical significance, power is the probability of rejecting the null hypothesis that the population means are equal when they are in fact not equal. The power of an equivalence test is the probability of rejecting that the means are different by at least some equivalence bound $\tau$ when the means are in fact equivalent (i.e. differ by less than ${ }^{\tau}$ ).

Of interest to us is the probability of rejecting both the null hypotheses (of nonsignificance and non-equivalence) simultaneously. We designed a small simulation study to assess the power of simultaneously concluding that two means are both statistically different and equivalent.

As is always the case with Monte Carlo studies, the choices of simulation parameters are difficult to make and are somewhat arbitrary. We endeavored to simulate situations that were likely to be encountered in actual quantitative data analysis. We also made some simplifying assumptions to keep the number of simulations and associated tables and figures to a reasonable level.

We assumed that both of our populations were always normally distributed with a common variance $\sigma^{2}=1$. Six different sample sizes per group ( $\mathrm{n}=10,20,50,100,200,500$ ) were chosen; only equally sized groups were used in this study. Six different values for the effect size parameter ( $\delta=0,0.1,0.2,0.3,0.4,0.5$ ) were used, reflecting situations from no effect (i.e. equivalent population means) to a "medium" effect size (i.e. population means that differ by
one half of a standard deviation). Three different equivalence bounds ( $\tau=0.1,0.2,0.4$ ) were used, defining the minimum difference between means that is practically important (i.e. nonequivalent) to be $10 \%, 20 \%$ or $40 \%$ of $\mu_{1}$.

Hence, we have a fully crossed design with 6 X 6 X $3=108$ cells. Within each cell (i.e. combination of sample size, effect size, and equivalence bound), 10000 simulations were run. The R statistical computing environment was used to conduct the simulations. Each simulation consisted of generating $n$ random normal variates with mean $0+\delta$ and variance 1 and a second, independent set of $n$ random normal variates with mean 0 and variance 1 . The independent samples t -test and the TOST with equivalence bound $\tau$ was conducted for each simulation, and the number of rejections of each test, along with the number of simultaneous rejections of both procedures and the number of failures to reject either procedure, were noted.

Tables 3 through 8 show the number of rejections of the null hypotheses of the equivalence test, both tests, the significance test, and neither test. Columns involving the equivalence test are in italics; columns involving the significance test are in boldface. Note that the power of the equivalence test for each situation can be found by dividing the sum of the italicized columns by 10000. Similarly, the power of the significance test is obtained by dividing the sum of the columns in boldface by 10000.

Table 3. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0$

| Equivalence Bound $\tau$ | $\underset{\mathrm{N}}{\text { Sample Size }}$ | Number of Rejections (10000 Simulations) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Equivalent | Both | Different | Neither |
| 0.1 | 10 | 0 | 0 | 506 | 9494 |
|  | 20 | 0 | 0 | 500 | 9500 |
|  | 50 | 0 | 0 | 476 | 9524 |
|  | 100 | 0 | 0 | 535 | 9465 |
|  | 200 | 0 | 0 | 504 | 9496 |
|  | 500 | 2337 | 0 | 511 | 7152 |
| 0.2 | 10 | 0 | 0 | 496 | 9504 |
|  | 20 | 0 | 0 | 507 | 9493 |
|  | 50 | 0 | 0 | 485 | 9515 |
|  | 100 | 1063 | 0 | 546 | 8391 |
|  | 200 | 5121 | 0 | 514 | 4365 |
|  | 500 | 9386 | 3 | 490 | 121 |
| 0.4 | 10 | 10 | 0 | 486 | 9504 |
|  | 20 | 370 | 0 | 469 | 9161 |
|  | 50 | 5279 | 0 | 481 | 4240 |
|  | 100 | 8757 | 0 | 457 | 786 |
|  | 200 | 9493 | 444 | 63 | 0 |
|  | 500 | 9483 | 517 | 0 | 0 |

Table 4. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0.1$

| Equivalence Bound $\tau$ | Number of Rejections (10000 Simulations) |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | N | Equivalent | Both | Different | Neither |
| 0.1 | 10 | 0 | 0 | 535 | 9494 |
|  | 20 | 0 | 0 | 606 | 9500 |
|  | 50 | 0 | 0 | 817 | 9524 |
|  | 100 | 0 | 0 | 1118 | 9465 |
|  | 200 | 0 | 0 | 1652 | 9496 |
|  | 500 | 709 | 0 | 3366 | 7152 |
| 0.2 | 10 | 0 | 0 | 521 | 9504 |
|  | 20 | 0 | 0 | 605 | 9493 |
|  | 50 | 1 | 0 | 786 | 9515 |
|  | 100 | 793 | 0 | 1090 | 8391 |
|  | 200 | 3452 | 0 | 1687 | 4365 |
|  | 500 | 6192 | 15 | 3486 | 121 |
| 0.4 | 10 | 11 | 0 | 565 | 9424 |
|  | 20 | 347 | 0 | 622 | 9031 |
|  | 50 | 4759 | 0 | 772 | 4469 |
|  | 100 | 7902 | 0 | 1044 | 1054 |
|  | 200 | 8361 | 1196 | 443 | 0 |
|  | 500 | 6521 | 3475 | 4 | 0 |

Table 5. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0.2$

| Equivalence Bound $\tau$ | $\underset{\mathrm{N}}{\text { Sample Size }}$ | Number of Rejections (10000 Simulations) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Equivalent | Both | Different | Neither |
| 0.1 | 10 | 0 | 0 | 727 | 9273 |
|  | 20 | 0 | 0 | 962 | 9038 |
|  | 50 | 0 | 0 | 1727 | 8273 |
|  | 100 | 0 | 0 | 2865 | 7135 |
|  | 200 | 0 | 0 | 5193 | 4807 |
|  | 500 | 16 | 0 | 8880 | 1104 |
| 0.2 | 10 | 0 | 0 | 699 | 9301 |
|  | 20 | 0 | 0 | 950 | 9050 |
|  | 50 | 0 | 0 | 1678 | 8322 |
|  | 100 | 408 | 0 | 2908 | 6684 |
|  | 200 | 951 | 0 | 5207 | 3842 |
|  | 500 | 915 | 7 | 8924 | 154 |
| 0.4 | 10 | 8 | 0 | 734 | 9258 |
|  | 20 | 296 | 0 | 967 | 8737 |
|  | 50 | 3397 | 0 | 1677 | 4926 |
|  | 100 | 5485 | 0 | 2890 | 1625 |
|  | 200 | 4886 | 2800 | 2314 | 0 |
|  | 500 | 1167 | 8534 | 299 | 0 |

Table 6. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0.3$

| Equivalence Bound $\tau$ | Sample Size | Number of Rejections (10000 Simulations) |  |  |  |
| :---: | :---: | ---: | ---: | ---: | ---: |
|  | N | Equivalent | Both | Different | Neither |
|  | 10 | 0 | $\mathbf{0}$ | $\mathbf{9 4 7}$ | 9053 |
|  | 20 | 0 | $\mathbf{0}$ | $\mathbf{1 5 4 0}$ | 8460 |
|  | 50 | 0 | $\mathbf{0}$ | $\mathbf{3 1 4 4}$ | 6856 |
|  | 100 | 0 | $\mathbf{0}$ | $\mathbf{5 5 9 4}$ | 4406 |
|  | 200 | 0 | $\mathbf{0}$ | $\mathbf{8 4 8 2}$ | 1518 |
|  | 500 | 0 | $\mathbf{0}$ | $\mathbf{9 9 7 3}$ | 27 |
| 0.2 | 10 | 0 | $\mathbf{0}$ | $\mathbf{9 8 5}$ | 9015 |
|  | 20 | 0 | $\mathbf{0}$ | $\mathbf{1 5 0 1}$ | 8499 |
|  | 0 | 0 | $\mathbf{0}$ | $\mathbf{3 2 0 3}$ | 6797 |
|  | 104 | $\mathbf{0}$ | $\mathbf{5 6 8 1}$ | 4215 |  |
|  | 95 | $\mathbf{0}$ | $\mathbf{8 5 2 4}$ | 1381 |  |
|  | 100 | 19 | $\mathbf{1}$ | $\mathbf{9 9 7 3}$ | $\mathbf{7}$ |
|  | 200 | 11 | $\mathbf{0}$ | $\mathbf{9 9 1}$ | 8998 |
|  | 500 | 225 | $\mathbf{0}$ | $\mathbf{1 5 6 3}$ | 8212 |
|  | 10 | 2061 | $\mathbf{0}$ | $\mathbf{3 1 3 3}$ | 4806 |
|  | 20 | 2796 | $\mathbf{0}$ | $\mathbf{5 6 0 2}$ | 1602 |
|  | 1516 | $\mathbf{2 3 7 4}$ | $\mathbf{6 1 1 0}$ | 2167 |  |
|  | 23 | $\mathbf{6 1 1 5}$ | $\mathbf{3 8 6 2}$ | 0 |  |
|  | 100 |  |  |  |  |

Table 7. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0.4$

| Equivalence Bound $\tau$ | Sample Size | Number of Rejections (10000 Simulations) |  |  |  |
| :---: | :---: | ---: | ---: | ---: | ---: |
|  | N | Equivalent | Both | Different | Neither |
| 0.1 | 10 | 0 | $\mathbf{0}$ | $\mathbf{1 3 3 5}$ | 8665 |
|  | 20 | 0 | $\mathbf{0}$ | $\mathbf{2 3 3 3}$ | 7667 |
|  | 50 | 0 | $\mathbf{0}$ | $\mathbf{5 0 1 5}$ | 4985 |
|  | 100 | 0 | $\mathbf{0}$ | $\mathbf{8 0 6 9}$ | 1931 |
|  | 200 | 0 | $\mathbf{0}$ | $\mathbf{9 7 6 9}$ | 231 |
|  | 500 | 0 | $\mathbf{0}$ | $\mathbf{1 0 0 0 0}$ | 0 |
| 0.2 | 10 | 0 | $\mathbf{0}$ | $\mathbf{1 3 4 4}$ | 8656 |
|  | 20 | 0 | $\mathbf{0}$ | $\mathbf{2 3 4 1}$ | 7659 |
|  | 50 | $\mathbf{0}$ | $\mathbf{5 0 7 7}$ | 4923 |  |
|  | 23 | $\mathbf{0}$ | $\mathbf{8 1 1 0}$ | 1867 |  |
|  | 100 | 1 | $\mathbf{0}$ | $\mathbf{9 7 8 4}$ | 215 |
|  | 200 | 0 | $\mathbf{0}$ | $\mathbf{1 0 0 0 0}$ | 0 |
|  | 500 | 9 | $\mathbf{0}$ | $\mathbf{1 4 0 2}$ | 8589 |
|  | 10 | 164 | $\mathbf{0}$ | $\mathbf{2 3 4 6}$ | 7490 |
|  | 20 | 933 | $\mathbf{0}$ | $\mathbf{5 0 9 9}$ | 3968 |
|  | 532 | $\mathbf{0}$ | $\mathbf{8 0 7 5}$ | 993 |  |
|  | 232 | $\mathbf{8 0 6}$ | $\mathbf{8 9 6 2}$ | 0 |  |
|  | 0 | $\mathbf{1 0 2 5}$ | $\mathbf{8 9 7 5}$ | 0 |  |
|  | 100 |  |  |  |  |

Table 8. Simulated Power of the Tests of Statistical Equivalence and Significance, Effect Size $\delta=0.5$

| Equivalence Bound $\tau$ | $\underset{\mathrm{N}}{\text { Sample Size }}$ | Number of Rejections (10000 Simulations) |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Equivalent | Both | Different | Neither |
| 0.1 | 10 | 0 | 0 | 1897 | 8103 |
|  | 20 | 0 | 0 | 3383 | 6617 |
|  | 50 | 0 | 0 | 6981 | 3019 |
|  | 100 | 0 | 0 | 9428 | 572 |
|  | 200 | 0 | 0 | 9985 | 15 |
|  | 500 | 0 | 0 | 10000 | 0 |
| 0.2 | 10 | 0 | 0 | 1804 | 8196 |
|  | 20 | 0 | 0 | 3437 | 6563 |
|  | 50 | 0 | 0 | 6905 | 3095 |
|  | 100 | 1 | 0 | 9429 | 570 |
|  | 200 | 0 | 0 | 9987 | 13 |
|  | 500 | 0 | 0 | 10000 | 0 |
| 0.4 | 10 | 7 | 0 | 1866 | 8127 |
|  | 20 | 117 | 0 | 3425 | 6458 |
|  | 50 | 370 | 0 | 6936 | 2692 |
|  | 100 | 236 | 0 | 9378 | 386 |
|  | 200 | 13 | 108 | 9879 | 0 |
|  | 500 | 0 | 28 | 9972 | 0 |

## Conclusion

The data originally collected and analyzed with traditional significance tests by Benson (1989) showed a statistically significant difference between the means of male and female statistics students on six variables (GPA, number of prior math courses, math self-concept, self-efficacy, general test anxiety, and statistical test anxiety) and failed to find a significance for only one variable (achievement). We computed Cohen's $d$ as an effect size. Not surprisingly, the smallest absolute effect size of 0.04 was found for the non-significant variable, while the absolute effect sizes of the six significant variables ranged from 0.24 to 0.66 .

We then re-analyzed Benson's data using the TOST procedure for testing for statistical equivalence. This analysis showed that only two variables, number of prior math courses and statistical test anxiety, were "different" (i.e. significant and not equivalent). Not coincidentally, these were the two variables with the strongest absolute effect sizes of 0.60 and 0.66 . The non-significant variable (achievement) was found to be statistically equivalent, and the absolute effect size was virtually zero. Four of the variables (GPA, math self-concept, self-efficacy, and general test anxiety) yielded conflicting results of "equivalent and different" since they rejected the null hypotheses of both the statistical and equivalence tests. It is likely that the difference in the means of these four variables, while statistically significant, is trivial. The absolute effect sizes of these four variables ranged from 0.24 to 0.51 . This encompasses a range of effect sizes that is often classified as "small" to "medium" (Cohen, 1988), notwithstanding Lenth’s (2001) warnings against using "canned" effect sizes.

We noticed that whenever the effect size $\delta$ is less than the equivalence bound $\tau$, then the power of the equivalence test was approaching unity as $n$ increased. This convergence was slow when $\delta$ was nearly equal to $\tau$. Essentially, if the effect size parameter is less than the minimum difference that the researcher considers to be practically important (i.e. the minimum difference between means
large enough to matter), we will reject the null of the TOST and conclude equivalence with power increasing to unity with larger sample sizes.

If $\delta>\tau$, the power of the significance test approaches unity and the power of the equivalence test approaches zero as the sample size increases. This is the situation where the effect size parameter exceeds the specified maximum for practical importance; we will reject the t-test and conclude statistical significance with power increasing to unity as the sample size increases.

When $\delta=\tau$, then the power of the equivalence test will approach twice the nominal alpha level. This occurs because the effect size parameter happens to coincide with the specified equivalence bound. Rejecting the TOST (i.e. concluding equivalence) is a type I error, made with probability $2 \alpha$. The probability is twice the nominal $\alpha$ since an equivalence test at level $\alpha$ corresponds to a $100(1-2 \alpha) \%$ equivalence interval.

When $0<\delta<\tau$, then the power of both the significance and equivalence tests approaches unity (often slowly) as $n$ increases. This is the situation where the null hypothesis of a significance test is false (i.e. the difference of means is not equal to zero), but the true difference is too small to be considered practically significant, where $\tau$ is the minimum difference between means that is considered important.

It appears to be somewhat common with real data to have situations where the tests of statistical significance and equivalence are simultaneously rejected for reasonable choices of significance level $\alpha$ and equivalence bound $\tau$. Our re-analysis of the Benson (1989) data yielded 4 simultaneous rejections out of 7 variables.

The simulated power of simultaneous rejection showed that the probability of simultaneous rejection was low when the assumptions of the inferential tests (i.e. normality, equal variances, equal sample sizes between groups) were true except when both $n$ and $\tau$ were large. It is possible that "simultaneous rejection" will be more likely with real data than (at least our) simulated data
because real data will surely violate the normality and homoscedasticity assumptions. We speculate that simultaneous rejection will be more common, and thus potentially more problematic for the researcher using equivalence testing in conjunction with standard hypothesis testing, when the data is non-normal and heteroscedastic.

Sawilowsky and Yoon (2002) demonstrated that large effect sizes could be found in situations where the results of a hypothesis test are 'not significant' (i.e. p>.05). Similarly, we found the magnitude of effect sizes obtained from the statistical re-analysis of typical educational research data to be troubling. Benson's data was of a decent size (groups of 94 and 123 subjects), but an effect size as large as 0.51 yielded both statistical significance (rejecting that the male mean was equal to the female mean) and equivalence (rejecting that the absolute difference of the male and female means were within a constant $\tau$ ). We make the conjecture that the effect size conventions of Cohen (i.e. 0.2 is small, 0.5 is medium, 0.8 is large) might not be large enough. It is even possible that making any recommendation about the desired magnitude of an effect size independent of the sample sizes and variability of the populations might be futile (Lenth, 2001).

It would be desirable to extend the simulation study to consider several scenarios ignored here. In particular, more attention needs to be given to situations where one or more of the following conditions are true:

1. The populations are non-normal
2. The variances are not equal
3. The sample sizes of the groups are not equal.

It would also be desirable to analytically determine the power function for simultaneous rejection of the significance and equivalence tests, if possible. We will continue to strive for a greater understanding of the link between the effect size and the results of the significance and equivalence tests. It appears that sole reliance on any standard methodology, be it hypothesis testing, confidence intervals, effect sizes, or equivalence testing is ill advised.

## References

Abelson, R. (1997). A retrospective on the signficance test ban of 1999 (if there were no significance tests, they would have to be invented). In L. Harlow, S. Mulaik, \& J. Steiger, What if there were no significance tests? (p. 117-141). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, D., Burnham, K., \& Thompson, W. (2000). Null hypothesis testing: Problems, prevalance, and an alternative. Journal of Wildlife Management, 64(4), 912923.

Anderson, S., \& Hauck, W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics: Theory and Methods, 12, 2663-2692.

Bartko, J. (1991). Proving the null hypothesis. American Psychologist, 46(10), 801803.

Benson, J. (1989). Structural components of statistical test anxiety in adults: An exploratory model. Journal of Experimental Education, 57(3), 247-261.

Berger, R., \& Hsu, J. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, 11(4), 283-319.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. Journal of the American Statistical Association, 37, 325-335.

Blair, R. C., \& Cole, S. R. (2002). Twosided equivalence testing of the difference between two means. Journal of Modern Applied Statistical Methods, 1(1), 139-142.

Boring, E. (1919). Mathematical vs. scientific importance. Psychological Bulletin, 16, 335-338.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

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

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

Cohen, J. (1994). The Earth is round ( $p<.05$ ). American Psychologist, 49(12), 9971003.

Falk, R. (1998). In criticism of the null hypothesis statistical test. American Psychologist, 53, 798-799.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. Journal of Educational Research, 94(5), 275-282.

Hagan, R. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52(1), 15-24.

Hagan, R. (1998). A further look at wrong reasons to abandon statistical testing. American Psychologist, 53(7), 801-803.

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

Harris, R. (1997). Reforming significane testing via three-valued logic. In L. Harlow, S. Mulaik, \& J. Steiger (Eds.), What if there were no significance tests? (p.145-174). Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. (1997). Needed: A ban on the significance test. Psychological Science, 8(1), 37.

Knapp, T. R. (1998). Comments on the statistical significance testing articles. Research in the Schools, 5(2), 39-41.

Knapp, T. R. (2002). Some reflections on significance testing. Journal of Modern Applied Statistical Methods, 1(2), 240-242.

Krantz, D. (1999). The null hypothesis testing controversy in psychology. Journal of the American Statistical Association, 94, 13721381.

Lehmann, E. (1959). Testing statistical hypotheses (1st ed.). New York: Wiley.

Lenth, R. (2001). Some practical guidelines for effective sample size determination. The American Statistician, 55(3), 187-193.

McBride, G. (1999). Equivalence tests can enhance environmental science and management. Australian and New Zealand Journal of Statistics, 41(1), 19-29.

McLean, J., \& Ernest, J. (1998). The role of statistical significane testing in educational research. Research in the Schools, 5(2), 15-22.

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806834.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. Harlow, S. Mulaik, \& J. Steiger (Eds.), What if there were no significance tests? (p. 393-426). Mahwah, NJ: Lawrence Erlbaum Associates.

Nix, T., \& Barnette, J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. Research in the Schools, 5(2), 3-14.

Patel, H., \& Gupta, G. (1984). A problem of equivalence in clinical trials. Biometrical Journal, 26, 471-474.

Rogers, J., Howard, K., \& Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. Psychological Bulletin, 113(3), 553-565.

Rozeboom, W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Sawilowsky, S. S., \& Yoon, J. (2002). The trouble with trivials (p>.05). Journal of Modern Applied Statistical Methods, 1(1), 143144.

Schmidt, F., \& Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik, \& J. Steiger (Eds.), What if there were no significance tests? (р. 37-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Schuirmann, D. (1981). On hypothesis testing to determine if the mean of the normal distribution is contained in a known interval. Biometrics, 37, 617.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmokinetics and Biopharmaceutics, 15, 657680.

Serlin, R. (1993). Confidence intervals and the scientific method: A case for Holm on the range. Journal of Experimental Education, 61, 350-360.

Serlin, R. (2002). Constructive criticism. Journal of Modern Applied Statistical Methods, 1(2), 202-227.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. Research in the Schools, 5(2), 33-38.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. Journal of Experimental Education, 70(1), 80-93.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. Educational and Psychological Measurement, 61, 219-224.

Westlake, W. (1976). Symmetric confidence intervals for bioequivalence trials. Biometrics, 32, 741-744.

Westlake, W. (1979). Statistical aspects of comparative bioequivalence trials. Biometrics, 35, 273-280.

# Confidence Intervals For $\mathrm{P}(\mathrm{X}<\mathrm{Y})$ In The Exponential Case With Common Location Parameter 

Ayman Baklizi<br>Department of Statistics<br>Yarmouk University<br>Irbid - Jordan

The problem considered is interval estimation of the stress - strength reliability $R=P(X<Y)$ where $X$ and Y have independent exponential distributions with parameters $\theta$ and $\lambda$ respectively and a common location parameter $\mu$. Several types of asymptotic, approximate and bootstrap intervals are investigated. Performances are investigated using simulation techniques and compared in terms of attainment of the nominal confidence level, symmetry of lower and upper error rates, and expected length. Recommendations concerning their usage are given.

Key words: Bootstrap, exponential distribution, interval estimation, stress-strength model

## Introduction

The problem of making inference about $\mathrm{R}=$ $\mathrm{P}(\mathrm{X}<\mathrm{Y})$ has received a considerable attention in literature. This problem arises naturally in the context of mechanical reliability of a system with strength X and stress Y . The system fails any time its strength is exceeded by the stress applied to it. Another interpretation of R is that it measures the effect of the treatment when X is the response for a control group and Y refers to the treatment group. Beg (1980) obtained the (MVUE) of R when X and Y are independent exponential random variables with unequal scale and unequal location parameters.

Gupta and Gupta (1988) obtained the maximum likelihood estimator (MLE), the MVUE, and a Bayes estimator of R in case of different location parameters and a common scale parameter. Various other versions of this problem have been discussed in literature, see Johnson et al. (1994).

Ayman Baklizi is an Assistant Professor of Applied Statistics. His research interests are in accelerated life tests and censored data. Email: baklizi1@hotmail.com.

The problem of developing confidence intervals for the stress - strength probability has received relatively little attention; Halperin (1987) and Hamdy (1995) developed distribution free confidence intervals, while Bai and Hong (1992) discussed point and interval estimation of in the case of two independent exponentials with common location parameter, they derived two types of approximate intervals but did not study their finite sample properties and did not give an idea about how do they compare with each other.

In this article, for the same problem considered by Bai and Hong (1992), we shall investigate and compare the performance of the two intervals of Bai and Hong together with some other types of confidence intervals like intervals based on the transformed maximum likelihood estimator, the likelihood ratio statistic and intervals based on the bootstrap (Efron \& Tibshirani, 1993). The model and maximum likelihood estimation of its parameters will be presented in section 2. The "non-bootstrap" confidence intervals will be presented in section 3, while bootstrap intervals are discussed in section 4. A Monte Carlo study designed to investigate and compare the intervals is described in section 5. Results and conclusions are given in the final section.

The Model and Maximum Likelihood Estimation

In this study, X and Y are independently exponentially distributed random variables with scale parameters $\theta$ and $\lambda$ respectively and a common location parameter $\mu$, that is

$$
\begin{gathered}
f_{X}(x, \theta, \mu)=\theta e^{-\theta(x-\mu)}, x \geq \mu \\
f_{Y}(y, \lambda, \mu)=\lambda e^{-\lambda(y-\mu)}, y \geq \mu
\end{gathered}
$$

Let $X_{1}, \ldots, X_{n_{1}}$ be a random sample for X and $Y_{1}, \ldots, Y_{n_{2}}$ be a random sample for $Y$. The parameter $R$ we want to estimate is $R=p(X<Y)=\frac{\theta}{\theta+\lambda}$. The likelihood function can be written as
$L(\theta, \lambda, \mu)=\theta^{n_{1}} \lambda^{n_{2}} \exp \left(-\theta \sum_{i=1}^{n_{1}}\left(x_{i}-\mu\right)-\lambda \sum_{i=1}^{n_{i}}\left(y_{i}-\mu\right)\right) I(z \geq \mu)$
where $z=\min \left(x_{1}, \ldots, x_{n_{1}}, y_{1}, \ldots, y_{n_{2}}\right)$ and $I($. indicates the usual indicator function.

The maximum likelihood estimators of $\theta, \lambda$, and $\mu$ are given by (Ghosh \& Razmpour,
1984) $\hat{\mu}=z, \hat{\theta}=\frac{n_{1}}{T_{1}}$, and $\hat{\lambda}=\frac{n_{2}}{T_{2}}$, where $T_{1}=\sum_{i=1}^{n_{1}}\left(x_{i}-z\right)$ and $T_{2}=\sum_{i=1}^{n_{2}}\left(y_{i}-z\right)$. The maximum likelihood estimator of R is therefore $\hat{R}=\frac{n_{1} T_{2}}{n_{2} T_{1}+n_{1} T_{2}}$. Now we will describe the various intervals under study.

## Confidence Intervals for $R$

Exact confidence intervals that are convenient to use for $R$ are not available and hence approximate methods that exist in a simple closed form are needed. In this section and the following section we shall develop various types of intervals for the stress strength reliability $(R)$.

Intervals Based on the Asymptotic Normality of the MLE (AN Intervals)

Bai and Hong (1992) showed that if $n=n_{1}+n_{2} \rightarrow \infty$ such that $\frac{n_{1}}{n} \rightarrow \gamma, 0<\gamma<1$. Then $\quad \sqrt{n}(\hat{R}-R) \rightarrow N\left(0, \sigma^{2}\right) \quad$ where $\sigma^{2}=\frac{R^{2}(1-R)^{2}}{\gamma(1-\gamma)}$. This fact can be used to construct approximate confidence intervals for R . The intervals are of the form

$$
\left(\hat{R} \pm z_{1-\alpha / 2} \frac{\hat{R}(1-\hat{R})}{\sqrt{n\left(n_{1} / n\right)\left(1-n_{1} / n\right)}}\right)
$$

where $Z_{1-\alpha / 2}$ is the $1-\alpha / 2$-quantile of the standard normal distribution.

Intervals Based on the Asymptotic Normality of the Transformed MLE (TRAN Intervals)

When the maximum likelihood estimator of the parameter of interest has its range in only a part of the real line, a monotone transformation of this parameter with continuous derivatives and range in the entire real line will generally be better approximated by an asymptotic normal distribution as suggested by many authors including Lawless (1982) and Nelson (1982). Let $K(R)$ be a monotone function of $R$ and let $K^{\prime}(R)$ be the first derivative, then by applying the delta method (Serfling, 1980) we get

$$
\sqrt{n}(K(\hat{R})-K(R)) \rightarrow N\left(0, K^{\prime}(R)^{2} V(\hat{R})\right) .
$$

Using this, a $1-\alpha$ confidence interval for $R$ may be obtained as

$$
\binom{K^{-1}\left(K(\hat{R})-z_{1-\alpha / 2} g K(\hat{R})^{2} \hat{V}(\hat{R})\right)}{K^{-1}\left(K(\hat{R})+z K^{\prime}(\hat{R})^{2} \hat{V}(\hat{R})\right)} .
$$

An appropriate transform is the $\tan ^{-1}$ (Jeng \& Meeker, 2003). Using this transform a $1-\alpha$ confidence interval for $R$ is given by

$$
\left\{\tan \binom{\tan ^{-1}(\hat{R}) \pm z_{1-\alpha / 2}}{\frac{R(1-\hat{R})}{\left(1+\hat{R}^{2}\right)\left(n\left(n_{1} / n\right)\left(1-n_{1} / n\right)\right)^{1 / 2}}}\right\}
$$

Bai and Hong's Intervals (BH intervals)
Ghosh and Razmpour (1984) showed that $\left(T_{1}, T_{2}, Z\right)$ is a complete sufficient for $(\theta, \lambda, \mu)$ and that the joint probability density function of $\left(T_{1}, T_{2}\right)$ which is independent of $Z$ is

$$
\begin{aligned}
& g\left(t_{1}, t_{2}\right)= \\
& \left(\frac{\theta^{n_{1}} \lambda^{n_{2}}}{n_{1} \theta+n_{2} \lambda}\right)\left(\frac{n_{2} t_{1}^{n_{1}-1} t_{2}^{n_{2}-2}}{\Gamma\left(n_{1}\right) \Gamma\left(n_{2}-1\right)}+\frac{n_{1} t_{1}^{n_{1}-2} t_{2}^{n_{2}-1}}{\Gamma\left(n_{1}-1\right) \Gamma\left(n_{2}\right)}\right) \\
& \exp \left(-\theta t_{1}-\lambda t_{2}\right) \\
& t_{1}, t_{2}>0
\end{aligned}
$$

Using standard transformation techniques, it can be shown that the probability density function of the random variable $U=\frac{\theta T_{1}}{\theta T_{1}+\lambda T_{2}}$ is given by (Bai and Hong, 1992)

$$
\begin{aligned}
& g\left(u, \pi, n_{1}, n_{2}\right)=\pi b\left(u, n_{1}-1, n_{2}\right) \\
& +(1-\pi) b\left(u, n_{1}, n_{2}-1\right), \quad 0 \leq u \leq 1
\end{aligned}
$$

where $\pi=\frac{n_{1} \theta}{n_{1} \theta+n_{2} \lambda}$ and
$b(u, r, s)=\frac{\Gamma(r+s)}{\Gamma(r) \Gamma(s)} u^{r-1}(1-u)^{s-1}, \quad 0 \leq u \leq 1$.
is the beta probability density function with parameters r and s. Bai and Hong (1992) showed that an approximate $1-\alpha$ interval for R is of the form

$$
\left(\frac{k_{\alpha / 2} t_{2}}{\left(1-k_{\alpha / 2}\right) t_{1}+k_{\alpha / 2} t_{2}}, \frac{k_{1-\alpha / 2} t_{2}}{\left(1-k_{1-\alpha / 2}\right) t_{1}+k_{1-\alpha / 2} t_{2}}\right)
$$

where $t_{1}$ and $t_{2}$ are the observed values of $T_{1}$ and $T_{2}$ respectively, and $k_{\alpha}$ is such that $G\left(k_{\alpha}, \hat{\pi}, n_{1}, n_{2}\right)=\alpha$. Here $\hat{\pi}$ is an estimator of $\pi$ obtained by substituting the maximum likelihood estimators of $\theta$ and $\lambda$ in the formula of $\pi$, and $G$ is the distribution function of mixed beta random variable U .

Intervals Based on the Likelihood Ratio Statistic (LR Intervals)

The likelihood function of $(\theta, \lambda, \mu)$ is given by

$$
\begin{aligned}
& L(\theta, \lambda, \mu)=\theta^{n_{1}} \lambda^{n_{2}} \exp \\
& \left(-\theta \sum_{i=1}^{n_{1}}\left(x_{i}-\mu\right)-\lambda \sum_{i=1}^{n_{1}}\left(y_{i}-\mu\right)\right) I(z \geq \mu)
\end{aligned}
$$

The likelihood ratio statistic for testing $H_{0}: R=R_{0}$ is defined as (Barndorff-Nielsen and Cox, 1994) $W=2(l(\Omega)-l(\varpi))$, where $l(\Omega)$ is the log-likelihood function evaluated at the values of the unrestricted maximum likelihood estimator of $(\theta, \lambda, \mu)$. While $l(\varpi)$ is the log-likelihood function evaluated at the values of the restricted maximum likelihood estimator under the null hypothesis. Recall that the unrestricted maximum likelihood estimators are $\hat{\mu}=z, \quad \hat{\theta}=\frac{n_{1}}{T_{1}}, \quad$ and $\quad \hat{\lambda}=\frac{n_{2}}{T_{2}}$, where $T_{1}=\sum_{i=1}^{n_{1}}\left(x_{i}-z\right)$ and $T_{2}=\sum_{i=1}^{n_{2}}\left(y_{i}-z\right)$. Under the null hypothesis $H_{0}: R=R_{0}$ we find readily that $\lambda=\frac{1-R_{0}}{R_{0}} \theta$ and thus the maximum likelihood estimator of $\theta$ is $\tilde{\theta}=\frac{n_{1}+n_{2}}{T_{1}+\frac{1-R_{0}}{R_{0}} T_{2}}$. Substituting in the formula
of the likelihood ratio statistic and simplifying we get

$$
W\left(R_{0}\right)=2\left[\begin{array}{l}
n_{1} \ln \left(\frac{n_{1}}{T_{1}}\right)+n_{2} \ln \left(\frac{n_{2}}{T_{2}}\right)- \\
\left(\begin{array}{l}
n_{2} \ln \left(\frac{1-R_{0}}{R_{0}}\right)+\left(n_{1}+n_{2}\right) \\
\ln \left(\frac{n_{1}+n_{2}}{T_{1}+\frac{1-R_{0}}{R_{0}} T_{2}}\right)
\end{array}\right] . . . ~ . ~ . ~ . ~
\end{array}\right] .
$$

The distribution of $W\left(R_{0}\right)$ is $\chi_{1}^{2}$ (BarndorffNielsen and Cox, 1994). The bounds of likelihood ratio confidence intervals with $(1-\alpha)$ nominal coverage probability are the two roots of $W\left(R_{0}\right)=\chi_{\alpha, 1}^{2}$, where $\chi_{\alpha, 1}^{2}$ is the upper $\alpha$ quantile of the chi square distribution with one degree of freedom.

## Parametric Bootstrap Intervals

The following methods of deriving confidence intervals are based on the Bootstrap approach (Efron \& Tibshirani, 1993). They are computer intensive methods based on resampling with replacement from the original data and then using these Bootstrap samples to study the behaviour of estimators and tests. When the parametric form of the distribution from which the data are generated is known except for some unknown parameters, we generate from this distribution after its parameters are replaced by their estimates. The advantage of bootstrap methods is their wide applicability and remarkable accuracy, especially in situations where the traditional methods do not work. There are several Bootstrap based intervals discussed in the literature (Efron and Tibshirani, 1993), the most common ones are the bootstrap -t interval, the percentile interval and the bias corrected and accelerated ( $B C_{a}$ ) interval.

The Bootstrap - t Interval Based on the MLE (BTST Intervals)

Let $\hat{R}$ be the maximum likelihood estimator of $R$ and let $\hat{R}^{*}$ be the maximum likelihood estimator calculated from the bootstrap sample. Let $z_{\alpha}^{*}$ be the $\alpha$ quantile of the bootstrap distribution of

$$
Z^{*}=\frac{\left(\hat{R}^{*}-\hat{R}\right)}{\left(\hat{V}\left(\hat{R}^{*}\right)\right)^{\frac{1}{2}}}
$$

where $\hat{V}\left(\hat{R}^{*}\right)$ is estimated variance of $\hat{R}$ calculated from the bootstrap sample. The bootstrap-t interval is given by $\left(\hat{R}-\mathrm{z}_{\alpha / 2}^{*} \hat{V}(\hat{R}), \hat{R}+\mathrm{z}_{1-\alpha / 2}^{*} \hat{V}(\hat{R})\right)$ where $\mathrm{z}_{\alpha}^{*}$ is determined by simulation.

The Bootstrap - t Interval Based on the Transformed MLE (TRBTST Intervals)

Let $\hat{R}$ be the maximum likelihood estimator of $R$ and let $\hat{R}^{*}$ be the maximum likelihood estimator calculated from the bootstrap sample. Let $z_{\alpha}^{*}$ be the $\alpha$ quantile of the bootstrap distribution of

$$
Q^{*}=\frac{\left(\tan ^{-1}\left(\hat{R}^{*}\right)-\tan (\hat{R})\right)}{\left(1+\hat{R}^{* 2}\right)^{-1}\left(\hat{V}\left(\hat{R}^{*}\right)\right)^{\frac{1}{2}}},
$$

where $\hat{V}\left(\hat{R}^{*}\right)$ is estimated variance of $\hat{R}$ calculated from the bootstrap sample. The bootstrap-t interval is given by

$$
\binom{\hat{R}-q_{\alpha / 2}^{*} V(\hat{R})^{\frac{1}{2}}\left(1+\hat{R}^{2}\right)^{-1},}{\hat{R}+q_{1-\alpha / 2}^{*} V(\hat{R})^{\frac{1}{2}}\left(1+\hat{R}^{2}\right)^{-1}}
$$

where $q_{\alpha / 2}^{*}$ and $q_{1-\alpha / 2}^{*}$ are the quantiles of the bootstrap distribution of $Q^{*}$ determined by simulation.

The Percentile Interval (PRC Interval)
Here we simulate the bootstrap distribution of $\hat{R}^{*}$ by resampling repeatedly from the parametric model of the original data and calculating $\hat{R}_{i}^{*}, i=1, \ldots, B$ where $B$ is the number of bootstrap samples. Let $\hat{H}$ be the cumulative distribution function of $\hat{R}^{*}$, then the $1-\alpha$ interval is given by $\left(\hat{H}^{-1}\left(\frac{\alpha}{2}\right), \hat{H}^{-1}\left(1-\frac{\alpha}{2}\right)\right)$.

The Bias Corrected and Accelerated Interval (BCa Interval)

The bias corrected and accelerated interval is calculated also using the percentiles of the bootstrap distribution of $\hat{R}^{*}$, but not necessarily identical with the percentile interval described in the previous subsection. The percentiles depend on two numbers $\hat{a}$ and $\hat{z}_{0}$ called the acceleration and the bias correction. The $1-\alpha$ interval is given by $\left(\hat{G}^{-1}\left(\alpha_{1}\right), \hat{G}^{-1}\left(\alpha_{2}\right)\right)$ where

$$
\begin{aligned}
& \alpha_{1}=\Phi\left(\hat{z}_{0}+\frac{\hat{z}_{0}+z_{\alpha / 2}}{1-\hat{a}\left(\hat{z}_{0}+z_{\alpha / 2}\right)}\right), \\
& \alpha_{2}=\Phi\left(\hat{z}_{0}+\frac{\hat{z}_{0}+z_{1-\alpha / 2}}{1-\hat{a}\left(\hat{z}_{0}+z_{1-\alpha / 2}\right)}\right),
\end{aligned}
$$

$\Phi($.$) is the standard normal cumulative$ distribution function, $z_{\alpha}$ is the $\alpha$ quantile of the standard normal distribution. The values of $\hat{a}$ and $\hat{z}_{0}$ are calculated as follows;

$$
\hat{a}=\frac{\sum_{i=1}^{n}(\hat{R}(.)-\hat{R}(i))^{3}}{6\left\{\sum_{i=1}^{n}(\hat{R}(.)-\hat{R}(i))^{2}\right\}^{3 / 2}}
$$

where $\hat{R}(i)$ is the maximum likelihood estimator of R using the original data excluding the i-th observation and

$$
\hat{R}(.)=\frac{\sum_{i=1}^{n} \hat{R}(i)}{n}
$$

The value of $\hat{z}_{0}$ is given by

$$
\hat{z}_{0}=\Phi^{-1}\left(\frac{\#\left\{\hat{R}^{*}<\hat{R}\right\}}{B}\right)
$$

Small Sample Performance of the Intervals
For the confidence intervals with nominal confidence coefficient $(1-\alpha)$, we use the criterion of attainment of lower and upper error probabilities which are both equal to $\frac{\alpha}{2}$. Attainment of lower and upper nominal error probabilities is important because otherwise we will use an interval with unknown error probabilities and our conclusions therefore are imprecise and can be misleading. Attainment of nominal error probabilities (assumed equal) means that if the interval fails to contain the true value of the parameter, it is equally likely to be above as to be below the true value. Users of two sided confidence intervals expect the lower and upper error probabilities to be symmetric because they are using symmetric percentiles of the approximating distributions to form their confidence intervals. However, symmetry of error probabilities may not occur due to the skewness of the actual sampling distribution Jennings (1987).

Another criterion for comparing confidence intervals is their expected lengths, obviously the shortest confidence interval among intervals having the same confidence level is the best. We have simulated the expected lengths of the three considered intervals.

A simulation study is conducted to investigate the performance of the intervals. The indices of our simulations are:
$\left(n_{1}, n_{2}\right)=(10,10),(20,20),(30,30),(40,40),(10,40),(40,10),(20,40),(40,20)$ $R$ : The true value of $\mathrm{R}=\mathrm{p}(\mathrm{X}<\mathrm{Y})$ and is taken to be $0.5,0.7,0.9,0.95$.

For each combination of $n_{1}, n_{2}$ and R , 2000 samples were generated for X taking $\theta=1$, $\mu=0$, and 2000 samples for Y with $\lambda=\frac{1}{R}-1, \mu=0$. The intervals are calculated, we used B = 1000 for bootstrap calculations.

The following quantities are simulated for each interval using the results of the 2000 samples; the expected width of the interval (W): The average of the widths of the 2000 intervals. Lower error rates (L): The fraction of intervals that fall entirely above the true parameter. Upper error rates (U): The fraction of intervals that fall entirely below the true parameter. Total error rates ( T ): The fraction of intervals that did not contain the true parameter value.

Table 1: Simulated error rates and expected lengths of the intervals

| $\left(n_{1}, n_{2}\right)$ | $R$ |  | AN | TRAN | BH | LR | BTST | TRBTST | PRC | BCa |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $(10,10)$ | 0.50 | L | 0.0425 | 0.0500 | 0.0245 | 0.0275 | 0.0070 | 0.0110 | 0.0310 | 0.0160 |
|  |  | U | 0.0455 | 0.0255 | 0.0305 | 0.0340 | 0.0110 | 0.0030 | 0.0285 | 0.0235 |
|  |  | T | 0.0880 | 0.0755 | 0.0550 | 0.0615 | 0.0180 | 0.0140 | 0.0595 | 0.0395 |
|  |  | W | 0.4160 | 0.4160 | 0.4230 | 0.3870 | 0.5240 | 0.4940 | 0.4210 | 0.4270 |
|  | 0.70 | L | 0.0175 | 0.0395 | 0.0245 | 0.0210 | 0.0355 | 0.0095 | 0.0115 | 0.0235 |
|  |  | U | 0.0475 | 0.0550 | 0.0425 | 0.0200 | 0.0650 | 0.0095 | 0.0100 | 0.0200 |
|  |  | T | 0.0650 | 0.0945 | 0.0670 | 0.0410 | 0.1010 | 0.0190 | 0.0215 | 0.0435 |
|  |  | W | 0.3610 | 0.3570 | 0.3230 | 0.4270 | 0.6000 | 0.4500 | 0.4470 | 0.3760 |
|  | 0.90 | L | 0.0010 | 0.0095 | 0.0075 | 0.0180 | 0.0035 | 0.0080 | 0.0550 | 0.0255 |
|  |  | U | 0.1080 | 0.0975 | 0.0565 | 0.0425 | 0.0195 | 0.0145 | 0.0095 | 0.0240 |
|  |  | T | 0.1090 | 0.1070 | 0.0640 | 0.0605 | 0.0230 | 0.0225 | 0.0645 | 0.0495 |
|  |  | W | 0.1550 | 0.1560 | 0.1630 | 0.1060 | 0.2030 | 0.2170 | 0.1640 | 0.1890 |
|  | 0.95 | L | 0.0000 | 0.0030 | 0.0120 | 0.0175 | 0.0145 | 0.0095 | 0.0655 | 0.0230 |
|  |  | U | 0.1370 | 0.1110 | 0.0655 | 0.0480 | 0.0270 | 0.0230 | 0.0150 | 0.0185 |
|  |  | T | 0.1370 | 0.1140 | 0.0775 | 0.0655 | 0.0415 | 0.0325 | 0.0805 | 0.0415 |
|  |  | W | 0.0813 | 0.0825 | 0.0863 | 0.0772 | 0.1080 | 0.1160 | 0.0872 | 0.1080 |
| (20, 20) | 0.50 | L | 0.0390 | 0.0500 | 0.0250 | 0.0290 | 0.0140 | 0.0145 | 0.0340 | 0.0290 |
|  |  | U | 0.0450 | 0.0295 | 0.0305 | 0.0340 | 0.0175 | 0.0215 | 0.0325 | 0.0195 |
|  |  | T | 0.0840 | 0.0795 | 0.0555 | 0.0630 | 0.0315 | 0.0360 | 0.0665 | 0.0485 |
|  |  | W | 0.3018 | 0.3010 | 0.3028 | 0.3355 | 0.3354 | 0.3250 | 0.3028 | 0.3050 |
|  | 0.70 | L | 0.0175 | 0.0275 | 0.0200 | 0.0225 | 0.0145 | 0.0125 | 0.0385 | 0.0170 |
|  |  | U | 0.0605 | 0.0420 | 0.0430 | 0.0365 | 0.0205 | 0.0160 | 0.0195 | 0.0235 |
|  |  | T | 0.0780 | 0.0695 | 0.0630 | 0.0590 | 0.0350 | 0.0285 | 0.0580 | 0.0405 |
|  |  | W | 0.2546 | 0.2560 | 0.2594 | 0.2305 | 0.2835 | 0.2830 | 0.2570 | 0.2630 |
|  | 0.90 | L | 0.0030 | 0.0115 | 0.0110 | 0.0155 | 0.0130 | 0.0170 | 0.0485 | 0.0195 |
|  |  | U | 0.0800 | 0.0605 | 0.0455 | 0.0430 | 0.0325 | 0.0160 | 0.0135 | 0.0275 |
|  |  | T | 0.0830 | 0.0720 | 0.0565 | 0.0585 | 0.0455 | 0.0330 | 0.0620 | 0.0470 |
|  |  | W | 0.1103 | 0.1110 | 0.1135 | 0.0845 | 0.1249 | 0.1300 | 0.1134 | 0.1230 |
|  | 0.95 | L | 0.0035 | 0.0060 | 0.0125 | 0.0190 | 0.0200 | 0.0235 | 0.0490 | 0.0270 |
|  |  | U | 0.0850 | 0.0855 | 0.0470 | 0.0380 | 0.0225 | 0.0240 | 0.0125 | 0.0260 |
|  |  | T | 0.0885 | 0.0915 | 0.0595 | 0.0570 | 0.0425 | 0.0475 | 0.0615 | 0.0530 |
|  |  | W | 0.0585 | 0.0586 | 0.0610 | 0.0558 | 0.0665 | 0.0682 | 0.0604 | 0.0662 |
| $(30,30)$ | 0.50 | L | 0.0305 | 0.0455 | 0.0255 | 0.0265 | 0.0175 | 0.0205 | 0.0290 | 0.0220 |


| $\left(n_{1}, n_{2}\right)$ | $R$ |  | AN | TRAN | BH | LR | BTST | TRBTST | PRC | BCa |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | U | 0.0310 | 0.0265 | 0.0265 | 0.0270 | 0.0190 | 0.0210 | 0.0280 | 0.0255 |
|  |  | T | 0.0615 | 0.0720 | 0.0520 | 0.0535 | 0.0365 | 0.0415 | 0.0570 | 0.0475 |
|  |  | W | 0.2488 | 0.2480 | 0.2461 | 0.3249 | 0.2663 | 0.2600 | 0.2492 | 0.2500 |
|  | 0.70 | L | 0.0205 | 0.0320 | 0.0225 | 0.0225 | 0.0180 | 0.0210 | 0.0400 | 0.0230 |
|  |  | U | 0.0565 | 0.0435 | 0.0345 | 0.0355 | 0.0240 | 0.0230 | 0.0230 | 0.0255 |
|  |  | T | 0.0770 | 0.0755 | 0.0570 | 0.0580 | 0.0420 | 0.0440 | 0.0630 | 0.0485 |
|  |  | W | 0.2097 | 0.2100 | 0.2129 | 0.2060 | 0.2249 | 0.2240 | 0.2110 | 0.2140 |
|  | 0.90 | L | 0.0035 | 0.0100 | 0.0090 | 0.0155 | 0.0155 | 0.0180 | 0.0365 | 0.0320 |
|  |  | U | 0.0600 | 0.0610 | 0.0395 | 0.0305 | 0.0205 | 0.0220 | 0.0125 | 0.0255 |
|  |  | T | 0.0635 | 0.0710 | 0.0485 | 0.0460 | 0.0360 | 0.0400 | 0.0490 | 0.0575 |
|  |  | W | 0.0903 | 0.0907 | 0.0922 | 0.0762 | 0.0977 | 0.0999 | 0.0919 | 0.0968 |
|  | 0.95 | L | 0.0030 | 0.0080 | 0.0145 | 0.0210 | 0.0225 | 0.0225 | 0.0425 | 0.0235 |
|  |  | U | 0.0700 | 0.0645 | 0.0470 | 0.0320 | 0.0270 | 0.0250 | 0.0175 | 0.0275 |
|  |  | T | 0.0730 | 0.0725 | 0.0615 | 0.0530 | 0.0495 | 0.0475 | 0.0600 | 0.0510 |
|  |  | W | 0.0479 | 0.0480 | 0.0480 | 0.0449 | 0.0520 | 0.0529 | 0.0489 | 0.0526 |
| $(40,40)$ | 0.50 | L | 0.0300 | 0.0380 | 0.0295 | 0.0260 | 0.0180 | 0.0210 | 0.0255 | 0.0240 |
|  |  | U | 0.0335 | 0.0185 | 0.0335 | 0.0290 | 0.0230 | 0.0155 | 0.0280 | 0.0205 |
|  |  | T | 0.0635 | 0.0565 | 0.0630 | 0.0550 | 0.0410 | 0.0365 | 0.0535 | 0.0445 |
|  |  | W | 0.2163 | 0.2160 | 0.2164 | 0.2989 | 0.2271 | 0.2240 | 0.2162 | 0.2170 |
|  | 0.70 | L | 0.0170 | 0.0320 | 0.0165 | 0.0255 | 0.0210 | 0.0270 | 0.0350 | 0.0220 |
|  |  | U | 0.0470 | 0.0280 | 0.0345 | 0.0295 | 0.0235 | 0.0170 | 0.0245 | 0.0260 |
|  |  | T | 0.0640 | 0.0600 | 0.0510 | 0.0550 | 0.0445 | 0.0440 | 0.0595 | 0.0480 |
|  |  | W | 0.1811 | 0.1830 | 0.1819 | 0.2003 | 0.1906 | 0.1920 | 0.1819 | 0.1850 |
|  | 0.90 | L | 0.0090 | 0.0165 | 0.0170 | 0.0210 | 0.0220 | 0.0225 | 0.0395 | 0.0225 |
|  |  | U | 0.0605 | 0.0470 | 0.0405 | 0.0350 | 0.0230 | 0.0235 | 0.0190 | 0.0270 |
|  |  | T | 0.0695 | 0.0635 | 0.0575 | 0.0560 | 0.0450 | 0.0460 | 0.0585 | 0.0495 |
|  |  | W | 0.0782 | 0.0791 | 0.0796 | 0.0687 | 0.0829 | 0.0848 | 0.0792 | 0.0824 |
|  | 0.95 | L | 0.0050 | 0.0035 | 0.0145 | 0.0215 | 0.0265 | 0.0160 | 0.0325 | 0.0265 |
|  |  | U | 0.0560 | 0.0575 | 0.0300 | 0.0390 | 0.0190 | 0.0245 | 0.0165 | 0.0255 |
|  |  | T | 0.0610 | 0.0610 | 0.0445 | 0.0605 | 0.0455 | 0.0405 | 0.0490 | 0.0520 |
|  |  | W | 0.0415 | 0.0414 | 0.0425 | 0.0400 | 0.0441 | 0.0443 | 0.0421 | 0.0442 |
| $(10,40)$ | 0.50 | L | 0.0270 | 0.0345 | 0.0415 | 0.0250 | 0.0175 | 0.0205 | 0.0530 | 0.0485 |
|  |  | U | 0.0490 | 0.0335 | 0.0165 | 0.0295 | 0.0100 | 0.0105 | 0.0080 | 0.0055 |
|  |  | T | 0.0760 | 0.0680 | 0.0580 | 0.0545 | 0.0275 | 0.0310 | 0.0610 | 0.0540 |
|  |  | W | 0.3359 | 0.3350 | 0.3345 | 0.3369 | 0.3828 | 0.3700 | 0.3351 | 0.3370 |
|  | 0.70 | L | 0.0105 | 0.0215 | 0.0375 | 0.0205 | 0.0115 | 0.0185 | 0.0790 | 0.0435 |
|  |  | U | 0.0855 | 0.0585 | 0.0230 | 0.0395 | 0.0155 | 0.0080 | 0.0065 | 0.0085 |
|  |  | T | 0.0960 | 0.0800 | 0.0605 | 0.0600 | 0.0270 | 0.0265 | 0.0855 | 0.0520 |
|  |  | W | 0.2790 | 0.2820 | 0.3033 | 0.2630 | 0.3358 | 0.3400 | 0.2770 | 0.2810 |
|  | 0.90 | L | 0.0020 | 0.0055 | 0.0220 | 0.0175 | 0.0145 | 0.0130 | 0.1055 | 0.0625 |
|  |  | U | 0.1185 | 0.0945 | 0.0520 | 0.0550 | 0.0195 | 0.0160 | 0.0025 | 0.0045 |
|  |  | T | 0.1205 | 0.1000 | 0.0740 | 0.0725 | 0.0340 | 0.0290 | 0.1080 | 0.0670 |
|  |  | W | 0.1190 | 0.1190 | 0.1440 | 0.0913 | 0.1557 | 0.1640 | 0.1181 | 0.1250 |
|  | 0.95 | L | 0.0010 | 0.0015 | 0.0125 | 0.0175 | 0.0190 | 0.0230 | 0.1120 | 0.0700 |
|  |  | U | 0.1265 | 0.1330 | 0.0475 | 0.0535 | 0.0170 | 0.0225 | 0.0015 | 0.0065 |
|  |  | T | 0.1275 | 0.1340 | 0.0600 | 0.0710 | 0.0360 | 0.0455 | 0.1135 | 0.0765 |
|  |  | W | 0.0625 | 0.0615 | 0.0720 | 0.0631 | 0.0843 | 0.0864 | 0.0614 | 0.0686 |


| $(20,40)$ | 0.50 | L | 0.0320 | 0.0260 | 0.0320 | 0.0280 | 0.0210 | 0.0110 | 0.0370 | 0.0370 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | U | 0.0395 | 0.0290 | 0.0215 | 0.0300 | 0.0170 | 0.0205 | 0.0175 | 0.0115 |
|  |  | T | 0.0715 | 0.0550 | 0.0535 | 0.0580 | 0.0380 | 0.0315 | 0.0545 | 0.0485 |
|  |  | W | 0.2632 | 0.2630 | 0.2666 | 0.3312 | 0.2838 | 0.2780 | 0.2631 | 0.2650 |
|  | 0.70 | L | 0.0175 | 0.0285 | 0.0260 | 0.0240 | 0.0185 | 0.0255 | 0.0470 | 0.0310 |
|  |  | U | 0.0620 | 0.0415 | 0.0235 | 0.0325 | 0.0180 | 0.0160 | 0.0155 | 0.0125 |
|  |  | T | 0.0795 | 0.0700 | 0.0495 | 0.0565 | 0.0365 | 0.0415 | 0.0625 | 0.0435 |
|  |  | W | 0.2214 | 0.2220 | 0.2304 | 0.2086 | 0.2434 | 0.2430 | 0.2216 | 0.2240 |
|  | 0.90 | L | 0.0025 | 0.0055 | 0.0195 | 0.0170 | 0.0180 | 0.0110 | 0.0625 | 0.0340 |
|  |  | U | 0.0830 | 0.0840 | 0.0390 | 0.0360 | 0.0230 | 0.0215 | 0.0075 | 0.0180 |
|  |  | T | 0.0855 | 0.0895 | 0.0585 | 0.0530 | 0.0410 | 0.0325 | 0.0700 | 0.0520 |
|  |  | W | 0.0950 | 0.0942 | 0.0987 | 0.0797 | 0.1077 | 0.1090 | 0.0953 | 0.1010 |
|  | 0.95 | L | 0.0040 | 0.0025 | 0.0135 | 0.0185 | 0.0160 | 0.0190 | 0.0715 | 0.0410 |
|  |  | U | 0.0940 | 0.0825 | 0.0420 | 0.0410 | 0.0240 | 0.0245 | 0.0090 | 0.0180 |
|  |  | T | 0.0980 | 0.0850 | 0.0555 | 0.0595 | 0.0400 | 0.0435 | 0.0805 | 0.0590 |
|  |  | W | 0.0498 | 0.0494 | 0.0524 | 0.0462 | 0.0571 | 0.0576 | 0.0499 | 0.0540 |
| $(40,20)$ | 0.50 | L | 0.0430 | 0.0500 | 0.0230 | 0.0325 | 0.0170 | 0.0160 | 0.0220 | 0.0210 |
|  |  | U | 0.0315 | 0.0170 | 0.0315 | 0.0310 | 0.0230 | 0.0200 | 0.0470 | 0.0230 |
|  |  | T | 0.0745 | 0.0670 | 0.0545 | 0.0635 | 0.0400 | 0.0360 | 0.0690 | 0.0440 |
|  |  | W | 0.2631 | 0.2630 | 0.2666 | 0.3289 | 0.2839 | 0.2770 | 0.2631 | 0.2640 |
|  | 0.70 | L | 0.0205 | 0.0360 | 0.0145 | 0.0245 | 0.0135 | 0.0170 | 0.0235 | 0.0195 |
|  |  | U | 0.0465 | 0.0340 | 0.0475 | 0.0265 | 0.0225 | 0.0235 | 0.0305 | 0.0260 |
|  |  | T | 0.0670 | 0.0700 | 0.0620 | 0.0510 | 0.0360 | 0.0405 | 0.0540 | 0.0455 |
|  |  | W | 0.2227 | 0.2240 | 0.2171 | 0.2084 | 0.2373 | 0.2370 | 0.2258 | 0.2260 |
|  | 0.90 | L | 0.0070 | 0.0135 | 0.0140 | 0.0230 | 0.0190 | 0.0175 | 0.0275 | 0.0180 |
|  |  | U | 0.0550 | 0.0500 | 0.0470 | 0.0305 | 0.0240 | 0.0235 | 0.0290 | 0.0330 |
|  |  | T | 0.0620 | 0.0635 | 0.0610 | 0.0535 | 0.0430 | 0.0410 | 0.0565 | 0.0510 |
|  |  | W | 0.0973 | 0.0982 | 0.0944 | 0.0795 | 0.1031 | 0.1060 | 0.1015 | 0.1040 |
|  | 0.95 | L | 0.0045 | 0.0080 | 0.0125 | 0.0265 | 0.0145 | 0.0180 | 0.0195 | 0.0235 |
|  |  | U | 0.0550 | 0.0510 | 0.0400 | 0.0300 | 0.0205 | 0.0220 | 0.0230 | 0.0250 |
|  |  | T | 0.0595 | 0.0590 | 0.0525 | 0.0565 | 0.0350 | 0.0400 | 0.0425 | 0.0485 |
|  |  | W | 0.0518 | 0.0521 | 0.0506 | 0.0467 | 0.0548 | 0.0560 | 0.0545 | 0.0558 |
| $(40,10)$ | 0.50 | L | 0.0525 | 0.0605 | 0.0185 | 0.0325 | 0.0090 | 0.0145 | 0.0120 | 0.0160 |
|  |  | U | 0.0260 | 0.0140 | 0.0400 | 0.0255 | 0.0210 | 0.0175 | 0.0625 | 0.0370 |
|  |  | T | 0.0785 | 0.0745 | 0.0585 | 0.0580 | 0.0300 | 0.0320 | 0.0745 | 0.0530 |
|  |  | W | 0.3354 | 0.3350 | 0.3345 | 0.3402 | 0.3839 | 0.3670 | 0.3349 | 0.3330 |
|  | 0.70 | L | 0.0370 | 0.0425 | 0.0175 | 0.0335 | 0.0120 | 0.0120 | 0.0080 | 0.0160 |
|  |  | U | 0.0380 | 0.0295 | 0.0425 | 0.0245 | 0.0240 | 0.0235 | 0.0610 | 0.0340 |
|  |  | T | 0.0750 | 0.0720 | 0.0600 | 0.0580 | 0.0360 | 0.0355 | 0.0690 | 0.0500 |
|  |  | W | 0.2900 | 0.2890 | 0.2774 | 0.2740 | 0.3194 | 0.3130 | 0.2993 | 0.2890 |
|  | 0.90 | L | 0.0070 | 0.0275 | 0.0105 | 0.0240 | 0.0100 | 0.0185 | 0.0110 | 0.0205 |
|  |  | U | 0.0565 | 0.0520 | 0.0505 | 0.0285 | 0.0260 | 0.0255 | 0.0460 | 0.0325 |
|  |  | T | 0.0635 | 0.0795 | 0.0610 | 0.0525 | 0.0360 | 0.0440 | 0.0570 | 0.0530 |
|  |  | W | 0.1293 | 0.1320 | 0.1277 | 0.0901 | 0.1377 | 0.1440 | 0.1430 | 0.1370 |
|  | 0.95 | L | 0.0055 | 0.0140 | 0.0065 | 0.0240 | 0.0170 | 0.0185 | 0.0125 | 0.0190 |
|  |  | U | 0.0600 | 0.0585 | 0.0495 | 0.0325 | 0.0275 | 0.0285 | 0.0435 | 0.0345 |
|  |  | T | 0.0655 | 0.0725 | 0.0560 | 0.0565 | 0.0445 | 0.0470 | 0.0560 | 0.0535 |
|  |  | W | 0.0703 | 0.0706 | 0.0698 | 0.0672 | 0.0745 | 0.0768 | 0.0798 | 0.0755 |

Conclusion
Our simulations indicate that the performance of intervals based on asymptotic normality (AN intervals) are not satisfactory even for relatively large samples, they are quite anti-conservative in the sense that their coverage probabilities are often higher than the nominal confidence level. Also they are quite asymmetric, especially for values of $R$ far from 0.5 . The performance of the intervals based on the transformed maximum likelihood estimator (TRAN intervals) is about similar to that of AN intervals, but their anticonservativeness and asymmetry being slightly less severe than AN intervals. Concerning Bai and Hong (BH) intervals, they often attain the nominal sizes but are asymmetric for values of $R$ away from 0.5. On the other hand, the Likelihood ratio (LR) intervals attain the nominal size and are almost symmetric even for small sample sizes.

For the Bootstrap intervals, it appears that the bootstrap - t intervals (BTST) and (TRBTST) are symmetric but tend to be conservative for small sample sizes, while the percentile interval (PRC) attains the nominal level but tends to be asymmetric for values of $R$ far from 0.5 . The bias corrected and accelerated interval appear to be the best interval based on the bootstrap principle, they attain the nominal level and are symmetric in almost all situations considered.

With regard to interval widths, our simulation results suggest that all intervals have about equal performance. No intervals appear to be uniformly shorter or longer than the others.

Overall, the ( $\mathbf{B C a}$ ) interval appears to have the best performance according to the criteria of attainment of coverage probability, symmetry and expected length followed by the (LR) intervals. Although the other intervals (especially AN intervals) are anti-conservative and sometimes extremely asymmetric, which limit their usefulness, especially when lower or upper confidence bounds are desired.

## References

Bai, D. S. (1992) Estimation of $\mathrm{p}(\mathrm{X}<\mathrm{Y})$ in the exponential case with common location parameter. Communications in Statistics, 21(1): 269-282.

Barndorff-Nielsen, O., \& Cox, D. R. (1994). Inference and asymptotics. NY: Chapman and Hall.

Beg, M. A. (1980). Estimation of $\mathrm{p}(\mathrm{Y}<\mathrm{X})$ for truncation parameter distributions. Communications in Statistics, 9, $327-345$.

Efron, B., \& Tibshirani, R. (1993). An introduction to the bootstrap. New York: Chapman and Hall.

Ghosh, M., \& Razmpour, A. (1984) Estimation of the common location parameter of several exponentials. Sankhya, A46:383-394.

Gupta R. D., \& Gupta R. C. (1988). Estimation of $p\left(Y_{p}>\operatorname{Max}\left(Y_{1}, Y_{2}, \ldots, Y_{p-1}\right)\right)$ in the exponential case. Communications in Statistics, A17, 911-924.

Jennings, D. (1987). How do we judge confidence intervals adequacy? The American Statistician, 41(4), 335-337.

Johnson, N., Kotz, S., \& Balakrishnan, N. (1994). Continuous univariate distributions. (Vol. 1). New York: Wiley.

# Approximate Bayesian Confidence Intervals <br> For The Variance Of A Gaussian Distribution 

Vincent A. R. Camara<br>University of South Florida

The aim of the present study is to obtain and compare confidence intervals for the variance of a Gaussian distribution. Considering respectively the square error and the Higgins-Tsokos loss functions, approximate Bayesian confidence intervals for the variance of a normal population are derived. Using normal data and SAS software, the obtained approximate Bayesian confidence intervals will then be compared to the ones obtained with the well known classical method. The Bayesian approach relies only on the observations. It is shown that the proposed approximate Bayesian approach relies only on the observations. The classical method, that uses the Chi-square statistic, does not always yield the best confidence intervals.

Key words: Estimation, loss functions, statistical analysis

Introduction
There is a significant amount of research in Bayesian analysis and modeling, which has been published the last twenty-five years; see references. A Bayesian analysis implies the exploitation of a suitable prior information and the choice of a loss function in association with Bayes' Theorem. It rests on the notion that a parameter within a model is not merely an unknown quantity but rather behaves as a random variable, which follows some distribution. In the area of life testing, it is indeed realistic to assume that a life parameter is stochastically dynamic. This assertion is supported by the fact that the complexity of electronic and structural systems is likely to cause undetected component interactions resulting in an unpredictable fluctuation of the life parameter. Recently, Drake (1966) gave an excellent account for the use of Bayesian statistics in reliability problems.

Vincent A. R. Camara earned a Ph.D. in Mathematics/Statistics. His research interests are in the theory and applications of Bayesian and empirical Bayes analyses with emphasis on the computational aspect of modeling. He is featured in the 2003 edition of Marquis Who's Who in America. (E-mail: gvcamara@ij.net)

As he pointed out, "He (Bayesian) realizes ... that his selection of a prior (distribution) to express his present state of knowledge will necessarily be somewhat arbitrary. But he greatly appreciates this opportunity to make his entire assumptive structure clear to the world..."

In the present study, we shall consider a classical and useful underlying model. That is, we shall consider the normal underlying model characterized by

$$
\begin{align*}
& f(x)=\frac{1}{\sqrt{2 \pi} \sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}} ;  \tag{1}\\
& -\infty \prec x \prec \infty,-\infty \prec \mu \prec \infty, \sigma \succ 0 .
\end{align*}
$$

As we well know, once the underlying model is found to be normally or approximately normally distributed, the classical approach uses the Chi-square statistic and considers the following confidence interval for the population variance $\sigma^{2}$ :

$$
\begin{equation*}
\left[\frac{(n-1) s^{2}}{\chi_{n-1, \alpha / 2}^{2}}, \frac{(n-1) s^{2}}{\chi_{n-1,1-\alpha / 2}^{2}}\right] . \tag{2}
\end{equation*}
$$

For the above model (1), approximate Bayesian confidence bounds for the parameter
$\sigma^{2}$ will be derived to challenge the classical approach (2). In the study, we shall denote the inverse of the population variance $\sigma^{2}$ by $\theta$ and its corresponding estimate by $\hat{\theta}$.

Although there is no specific analytical procedure that allows us to identify the appropriate loss function to be used, the most commonly used is the square error loss function. One of the reasons for selecting this loss function is because of its analytical tractability in Bayesian analysis. As it will be shown, selecting the square error loss does not always lead to the best approximate Bayesian confidence intervals. However, the obtained approximate Bayesian confidence intervals corresponding to the square error and the Higgins-Tsokos loss functions will be respectively used to challenge the classical method (2). The loss functions that will be used are given below, along with a statement of their key characteristics.

Square error loss function
The popular square error loss function places a small weight on estimates near the true value and proportionately more weight on extreme deviation from the true value of the parameter. Its popularity is due to its analytical tractability in Bayesian modeling. The square error loss is defined as follows:

$$
\begin{equation*}
L_{S E}(\hat{\theta}, \theta)=(\hat{\wedge}-\theta)^{2} \tag{3}
\end{equation*}
$$

Higgins-Tsokos loss function
The Higgins-Tsokos loss function places a heavy penalty on extreme over- or underestimation. That is, it places an exponential weight on extreme errors. The Higgins-Tsokos loss function is defined as follows:

$$
L_{H T}(\hat{\theta}, \theta)=\frac{f_{1} e^{f_{2}(\hat{\theta}-\theta)}+f_{2} e^{-f_{1}(\hat{\theta}-\theta)}}{f_{1}+f_{2}}-1,
$$

$$
\begin{equation*}
f_{1}, f_{2} \succ 0 . \tag{4}
\end{equation*}
$$

We shall assume that $\theta$ behaves as a random variable and is being characterized by the Pareto probability density function given by

$$
\begin{equation*}
f_{1}(\theta)=\frac{a}{b}\left(\frac{b}{\theta}\right)^{a+1} ; \theta \geq b \succ 0, a \succ 0 \tag{5}
\end{equation*}
$$

where $\theta=1 / \sigma^{2}$.
The Pareto prior has been selected because of its mathematical tractability. Using observations from normal distributions, we will approximate the Pareto prior (5) in such a way that good approximate Bayesian estimates of $\theta$ are obtained.

Preliminaries
Let $x_{1}, x_{2}, \ldots \ldots, x_{n}$ denote the observations of a given system that are being characterized by the normal distribution. Replacing $1 / \sigma^{2}$ by $\theta$, we obtain the following characterization of the normal underlying model defined in (1).

$$
\begin{align*}
& f(x)=\frac{1}{\sqrt{2 \pi}} \theta^{\frac{1}{2}} e^{-\theta \frac{(x-\mu)^{2}}{2}} ;  \tag{6}\\
& -\infty \prec x \prec \infty, \theta \succ 0
\end{align*}
$$

This leads to the following posterior distribution:

$$
\begin{equation*}
h(\theta \backslash x) \frac{\theta^{\frac{n}{2}-a-1} e^{-\theta \sum_{1}^{n} \frac{\left(x_{i}-\mu\right)^{2}}{2}}}{\int_{b}^{\infty} \theta^{\frac{n}{2}-a-1} e^{-\theta \sum_{1}^{n} \frac{\left(x_{i}-\mu\right)^{2}}{2}} d \theta}, \theta \succ b . \tag{7}
\end{equation*}
$$

## Methodology

Approximate Bayesian confidence bounds of $\sigma^{2}$ when the population mean $\mu$ is known.

With respectively the following approximate priors for the square error and the Higgins-Tsokos loss functions, good
approximate Bayesian estimates of $\theta$ are obtained.

Approximate prior for the square error loss:

$$
\begin{align*}
\bar{g}(\theta) & =\frac{a 1}{b 1}\left(\frac{b 1}{\theta}\right)^{a 1+1} ; \\
a 1 & =\frac{n}{2}-1, b 1=\frac{n-2}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}} \tag{8}
\end{align*}
$$

Approximate prior for the Higgins-Tsokos loss:

$$
\begin{align*}
& \bar{g}_{1}(\theta)=\frac{a o}{b o}\left(\frac{b o}{\theta}\right)^{a o+1} ;  \tag{9}\\
& ; a_{0}=\frac{n}{2}-1, b o=\frac{n-1}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}-F(n, \mu)
\end{align*}
$$

where
$F(n, \mu)=\frac{1}{f_{1}+f_{2}} \operatorname{Ln}\left(\frac{\sum_{i=1}^{n} \frac{\left(x_{i}-\mu\right)^{2}}{2}+f_{2}}{\sum_{i=1}^{n} \frac{\left(x_{i}-\mu\right)^{2}}{2}-f_{1}}\right)$
with

$$
f_{1} \prec \sum_{i=1}^{n} \frac{\left(x_{i}-\mu\right)^{2}}{2} .
$$

It's easily shown that the approximate Bayesian estimate of the parameter $\theta$, subject to the square error loss, is the same as the Bayesian estimate of $\theta$ under the HigginsTsokos loss. They are equal to

$$
\frac{n-1}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}
$$

Using respectively the approximate posterior distributions that correspond to (8) and (9), along with the equalities $P(\theta \succ L \mid x)=1-\alpha / 2$ and $P(\theta \succ U \mid x)=\alpha / 2$, we respectively obtain the following lower and upper confidence bounds for $\theta$ :

Approximate Bayesian confidence bounds of $\theta$ corresponding to the square error loss function when $\mu$ is known:

$$
\begin{gather*}
L_{\theta(S E)}=\frac{n-2-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}} \\
U_{\theta(S E)}=\frac{n-2-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}} \tag{10}
\end{gather*}
$$

Approximate Bayesian confidence bounds of $\theta$ corresponding to the Higgins-Tsokos loss function when $\mu$ is known:

$$
\begin{gather*}
L_{\theta(H T)}=\frac{n-1-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}-F(n, \mu) \\
U_{\theta(H T)}=\frac{n-1-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}-F(n, \mu) . \tag{11}
\end{gather*}
$$

Thus when the population mean is known, (10) and (11) respectively yield the following $100(1-\alpha) \%$ approximate Bayesian confidence bounds for the normal population variance $\sigma^{2}$ :

Confidence bounds corresponding to the square error loss:

$$
\begin{gather*}
L_{\sigma^{2}(S E)}=\frac{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}{n-2-2 \ln (\alpha / 2)} \\
U_{\sigma^{2}(S E)}=\frac{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}{n-2-2 \ln (1-\alpha / 2)}, \tag{12}
\end{gather*}
$$

Confidence bounds corresponding to the Higgins-Tsokos loss:

$$
\begin{align*}
& L_{\sigma^{2}(H T)}=\frac{1}{\frac{n-1-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}-F(n, \mu)} \\
& U_{\sigma^{2}(H T)}=\frac{1}{\frac{n-1-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\mu\right)^{2}}-F(n, \mu)} . \tag{13}
\end{align*}
$$

Approximate Bayesian confidence bounds of $\sigma^{2}$ when the population mean $\mu$ is unknown.

In the case where the population mean $\mu$ is unknown, it is estimated by the sample mean $\bar{x}$ and we obtain the following:
Approximate Bayesian confidence bounds of $\theta$ corresponding to the square error loss function when $\mu$ is unknown:

$$
\begin{gather*}
L_{\theta(S E)}=\frac{n-2-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}} \\
U_{\theta(S E)}=\frac{n-2-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}} \tag{14}
\end{gather*}
$$

Approximate Bayesian confidence bounds of $\theta$ corresponding to the Higgins-Tsokos loss function when $\mu$ is unknown:

$$
\begin{gather*}
L_{\theta(H T)}=\frac{n-1-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}-F(n, \bar{x})  \tag{15}\\
U_{\theta(H T)}=\frac{n-1-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}-F(n, \bar{x}) .
\end{gather*}
$$

Thus when $\mu$ is unknown (14) and (15) respectively yield the following $100(1-\alpha) \%$ approximate Bayesian confidence bounds for the normal population variance $\sigma^{2}$ :

Confidence bounds corresponding to the square error loss:

$$
\begin{gather*}
L_{\sigma^{2}(S E)}=\frac{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}{n-2-2 \ln (\alpha / 2)} \\
U_{\sigma^{2}(S E)}=\frac{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}{n-2-2 \ln (1-\alpha / 2)}, \tag{16}
\end{gather*}
$$

Confidence bounds corresponding to the Higgins-Tsokos loss:

$$
\begin{align*}
& L_{\sigma^{2}(H T)}=\frac{1}{\frac{n-1-2 \operatorname{Ln}(\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}-F(n, \bar{x})} \\
& U_{\sigma^{2}(H T)}=\frac{1}{\frac{n-1-2 \operatorname{Ln}(1-\alpha / 2)}{\sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}-F(n, \bar{x})} .
\end{align*}
$$

Results
In order to compare the proposed approximate Bayesian approach to the classical method, samples that have been obtained from normally distributed populations (Examples 1, 2, 3, .4, 7) as well as approximately normal populations (Examples 5, 6) will be considered. SAS software is used to obtain the normal population parameters $\mu$ and $\sigma$
corresponding to each of the examples. The proposed approximate Bayesian estimates of the variance (16) (17) will be used. For the Higgins-Tsokos loss function, we will consider $f_{1}=1, f_{2}=1$. The lengths of the classical and approximate Bayesian confidence intervals are respectively denoted by $l_{C}, l_{S E}$ and $l_{H T}$.

Example 1. (Data obtained from Prem S. Mann, Introductory Statistics, Third edition, page 504, 1998).
24, 28, 22, 25, 24, 22, 29, 26, 25, 28, 19, 29.
Normal population distribution
obtained with SAS:
$N(\mu=25.083, \sigma=3.1176)$ Population and
sample variances: $\sigma^{2}=9.71943$,
$s^{2}=9.719696$.

Table 1. Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the first data set.

| $\begin{aligned} & \hline \text { C L. } \\ & \% . \end{aligned}$ | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :---: | :---: | :---: | :---: |
| 80 | $\begin{aligned} & 6.18- \\ & 19.16 \end{aligned}$ | 7.32-10.47 | 8.08-10.23 |
| 90 | $\begin{gathered} \hline 5.43- \\ 23.36 \end{gathered}$ | 6.68-10.58 | 7.32-10.47 |
| 95 | $\begin{aligned} & 4.87- \\ & 28.01 \end{aligned}$ | $6.15-10.63$ | $6.68-10.58$ |
| 99 | $\begin{aligned} & 3.99- \\ & 41.07 \end{aligned}$ | $5.19-10.68$ | 5.56-10.67 |


| Confidence <br> level | $\left(l_{C}\right) \div\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.1193 | 6.0455 |
| $90 \%$ | 4.6021 | 5.6927 |
| $95 \%$ | 5.1589 | 5.9373 |
| $99 \%$ | 6.7538 | 7.2636 |

Example 2. Data obtained from Prem S. Mann, Introductory Statistics, Third edition, page 504, 1998.
$13,11,9,12,8,10,5,10,9,12,13$.
Normal population distribution obtained with SAS: $N(\mu=10.182, \sigma=2.4008)$. Population and sample variances: $\quad \sigma^{2}=5.76384$, $s^{2}=5.763636$.

Table 2: Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the second data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | :--- | :--- |
| 80 | $3.60-$ | $4.23-6.25$ | $4.71-6.10$ |
| 11.84 |  |  |  |


| Confidence <br> level | $\left(l_{C}\right) \div$ <br> $\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.0777 | 5.9530 |
| $90 \%$ | 4.6157 | 5.6804 |
| $95 \%$ | 5.2426 | 6.0051 |
| $99 \%$ | 7.0734 | 7.5801 |

Example 3. Data obtained from Prem S. Mann, Introductory Statistics, Third edition, page 504, 1998.
$16,14,11,19,14,17,13,16,17,18,19,12$.

Normal population distribution obtained with SAS: $N(\mu=15.5, \sigma=2.6799)$. Population and sample variances: $\sigma^{2}=7.18186$, $s^{2}=7.181818$.

Table 3. Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the third data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | :--- | :--- |
| 80 | $4.57-$ | $5.40-7.73$ | $5.97-7.56$ |
| 14.16 |  |  |  |
| 90 | $4.01-$ | $4.94-7.81$ | $5.40-7.73$ |
| 17.26 |  | $4.94-7.81$ |  |
| 95 | $3.60-$ | $4.54-7.86$ |  |
| 99 | $2.95-$ | $3.83-7.89$ | $4.11-7.88$ |
|  | 30.34 |  |  |


| Confidence <br> level | $\left(l_{C}\right) \div\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.1194 | 6.0456 |
| $90 \%$ | 4.6022 | 5.6926 |
| $95 \%$ | 5.1592 | 5.9375 |
| $99 \%$ | 6.7539 | 7.2636 |

Example 4. Data obtained from Prem S. Mann, Introductory Statistics, Third edition, page 504, 1998. 27, 31, 25, 33, 21, 35, 30, 26, 25,31.33.30, 28.

Normal population distribution obtained with SAS: $N(\mu=28.846, \sigma=3.9549)$. Population and sample variances: $\quad \sigma^{2}=15.64123$, $s^{2}=15.641025$.

Table 4. Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the fourth data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | :---: | :---: |
| 80 | 10.11 | $12.02-$ | $13.20-$ |
| - | 16.74 | 16.39 |  |
| 29.77 |  | 16.90 | 16.74 |
| 90 | $8.92-$ | $11.04-$ | $12.02-$ |
| 95 | $8.04-$ | $10.21-$ | $11.04-$ |
|  | 42.61 | 16.98 | 16.90 |
| 99 | $6.63-$ | $8.69-$ | $9.28-$ |
|  | 61.05 | 17.04 | 17.03 |


| Confidence level | $\left(l_{C}\right) \div\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.1688 | 6.1471 |
| $90 \%$ | 4.6063 | 5.7243 |
| $95 \%$ | 5.1059 | 5.9013 |
| $99 \%$ | 6.5129 | 7.0273 |

Example 5. Data obtained from James T. McClave/Terry Sincich A first course in Statistics, page 301, Sixth edition, 1997

$$
52,33,42,44,41,50,44,51,45,38 \text {, }
$$ $37,40,44,50,43$.

Normal population distribution
with $N(\mu=43.6, \sigma=5.4746)$. Population and sample variances: $\quad \sigma^{2}=29.97124$, $s^{2}=29.971428$.

Table 5. Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the fifth data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | :---: | :---: |
| 80 | $19.92-$ | $23.83-31.76$ | $25.87-$ |
| 53.86 |  | 31.20 |  |
| 90 | 17.71 | $22.09-32.02$ | $23.83-$ |
|  | - |  | 31.76 |
| 95 | $16.06-$ | $20.59-32.15$ | $22.09-$ |
|  | 74.54 |  | 32.02 |
| 99 | $13.39-$ | $17.78-32.25$ | $18.89-$ |
|  | 102.96 |  | 32.22 |


| Confidence <br> level | $\left(l_{C}\right) \div\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.2814 | 6.3629 |
| $90 \%$ | 4.6465 | 5.8198 |
| $95 \%$ | 5.0583 | 5.8889 |
| $99 \%$ | 6.1902 | 6.7170 |

Example 6. Data obtained from James T. McClave/Terry Sincich A first course in Statistics, page 301, Sixth edition, 1997.
$52,43,47,56,62,53,61,50,56,52$, $53,60,50,48,60,55$.

Normal population distribution obtained with SAS: $N(\mu=53.625, \sigma=5.4145)$. Population and sample variances: $\quad \sigma^{2}=29.31681$, $s^{2}=29.316666$.

Table 6. Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the sixth data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes <br> . <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | :---: | :---: |
| 80 | $19.71-$ | $23.63-$ |  |
| 51.45 | 30.94 | $25.53-30.44$ |  |
| 90 | $17.59-$ | $21.99-$ | $23.63-30.94$ |
| 60.56 | 31.18 |  |  |
| 95 | $15.99-$ | $20.57-$ | $21.99-31.18$ |
| 99 | $13.40-$ | $17.87-$ | $18.94-31.36$ |
|  | 95.57 | 31.38 |  |


| Confidence <br> level | $\left(l_{C}\right) \div\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 4.3422 | 6.4743 |
| $90 \%$ | 4.6781 | 5.8754 |
| $95 \%$ | 5.0551 | 5.9636 |
| $99 \%$ | 6.0822 | 6.6163 |

Example 7. The following observations have been obtained from the collection of SAS data sets.

50, 65, 100, 45, 111, 32, 45, 28, 60, 66, 114, 134, 150, 120, 77, 108, 112, 113,80,77, 69, 91, 116, 122, 37, 51, 53, 131, 49, 69, 66, 46, 131, 103, 84, 78.

Normal population distribution obtained with SAS: $N(\mu=82.861, \sigma=33.226)$. Population and sample variances: $\quad \sigma^{2}=1103.96716$, $s^{2}=1103.951587$.

Table 7: Classical and approximate Bayesian confidence intervals of $\sigma^{2}$ corresponding to the seventh data set.

| C L. <br> $\%$. | Classical <br> Bounds | Approx.Bayes. <br> Bounds (SE) | Approx.Bayes. <br> Bounds (HT) |
| :--- | :--- | ---: | :---: |
| 80 | $839.4-$ <br> 1556.4 | $1000.8-$ | $1038.1-$ |
| 1129.4 | 1121.6 |  |  |
| 90 | $776.4-$ | $966.1-$ | $1000.8-$ |
| 1717.1 | 1133.0 | 1129.4 |  |
| 95 | $726.8-$ | $933.7-$ | $966.1-$ |
|  | 1874.5 | 1134.7 | 1133.0 |
| 99 | $641.6-$ | $866.3-$ | $894.1-$ |
|  | 2240.2 | 1136.0 | 1135.7 |


| Confidence <br> level | $\left(l_{C}\right) \div$ <br> $\left(l_{S E}\right)$ | $\left(l_{C}\right) \div$ <br> $\left(l_{H T}\right)$ |
| :---: | :---: | :---: |
| $80 \%$ | 5.5772 | 8.5808 |
| $90 \%$ | 5.6388 | 7.3176 |
| $95 \%$ | 5.7119 | 6.8792 |
| $99 \%$ | 5.9277 | 6.6181 |

All seven Tables show that the proposed approximate Bayesian confidence intervals contain the population variance $\sigma^{2}$. Also, the lengths of the obtained classical confidence intervals are more than four times greater than the ones corresponding to the proposed approach.

## Conclusion

In the present study, approximate Bayesian confidence intervals for the variance of a normal population under two different loss functions have been derived. The loss functions that are employed are the square error and the Higgins-Tsokos loss functions. Based on the above numerical results we can conclude the following:

The classical method used to construct confidence intervals for the variance of a normal population does not always yield the
best coverage accuracy. In fact, each of the obtained approximate Bayesian confidence intervals contains the population variance and is strictly included in the corresponding confidence interval obtained with the classical method.

Contrary to the classical method that uses the Chi-square statistic, the proposed approach relies only on the observations.

With the proposed approach, approximate Bayesian confidence intervals for a normal population variance are easily computed for any level of significance.

The approximate Bayesian approach under to the popular square error loss function does not always yield the best approximate Bayesian results. In fact, the Higgins-Tsokos loss function performs better in the above examples.

Bayesian analysis contributes to reinforcing well-known statistical theories such as the estimation theory.

## References

Bhattacharya, S. K. (1967). Bayesian approach to life testing and reliability estimation. Journal American Statistical Association, 62, 48-62.

Bernard, H. (1976). A survey of statistical methods in system reliability using Bernoulli sampling of components. Proceedings of the conference on the theory and applications of Reliability with emphasis on Bayesian and Nonparametric Methods. NY: Academic.

Britney, R. R., \& Winkler, R. L. (1968). Bayesian III point estimation under various loss functions. American Statistical Association, 356-364.

Camara, V. A. R., \& Tsokos, C. P. (1996). Effect of Loss Functions on Bayesian Reliability Analysis. Proceedings of International Conference on Nonlinear Problems in Aviation and Aerospace, 75-90.

Camara, V. A. R., \& Tsokos, C. P. (1998). Bayesian reliability modeling with applications. UMI Publishing Company.

Camara, V. A. R., \& Tsokos, C. P. (1999). Bayesian estimate of a parameter and choice of the loss function. Nonlinear Studies Journal, 6, 55-64.

Camara, V. A. R., \& Tsokos, C. P. (1999). The effect of loss functions on empirical bayes reliability analysis. Journal of Engineering Problems, Boston University, http://bujep.bu.edu.

Camara, V. A. R. (2002). Approximate Bayesian confidence intervals for the variance of a Gaussian distribution. 2002 Proceedings of the American Statistical Association, Statistical Computing Section New York, NY: American Statistical Association.

Canfield, R. V. (1970). A Bayesian approach to reliability estimation using a loss function. IEEE Trans. Reliability R-19 (1), 1316.

Vincent A. R. Camara \& Tsokos, C. P. (1999, November). Sensitivity behavior of Bayesian reliability analysis for different loss functions. International Journal of Applied Mathematics.

Drake, A. W. (1966). Bayesian statistics for the reliability eng. proc. Annual Symposium on Reliability, 315-320.

Higgins, J. J., \& Tsokos, C. P. (1976). Comparison of Bayes estimates of failure intensity for fitted priors of life data. Proceedings of the Conference on the Theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods. NY: Academic.

Higgins, J. J. \& Tsokos, C. P. (1976). On the behavior of some quantities used in Bayesian reliability demonstration tests. IEEE Trans. Reliability R-25, (4), 261-264.

Higgins, J. J., \& Tsokos, C. P. (1980). A study of the effect of the loss function on bayes estimates of failure intensity, MTBF, and reliability. Applied Mathematics and Computation, 6, 145-166.

McClave, J. T., \& Sincich, T. A. (1997). First course in statistics ( $6^{\text {th }}$ ed.). Upper Saddle River, NJ: Prentice Hall.

Mann, P. S. (1998), Introductory Statistics, ( ${ }^{\text {rd }}$ ed). NY: Wiley.

Schafer, et al. (1970). Bayesian reliability demonstration, Phase I Data for the a prior distribution. Rome Air Development Center, Griffis AFBNY RADC-TR-69-389.

Schafer, et al. (1971). Bayesian reliability, Phase II- Development of a priori distribution. Rome Air Development Center, Griffis AFR, NY RADC-YR-71-209.

Schafer, et al. (1973). Bayesian reliability demonstration Phase III Development of test plans, Rome Air development Center, Griffs AFB, NY RADC-TR-73-39.

Shafer, R. E., \& Feduccia, A. J. (1972). Prior distribution fitted to observed reliability data. IEEE Trans. Reliability $R$-21, (3), 148-154

Tsokos, C. P. \& Shimi, I. (Eds). (1976). Proceedings of the Conference on the theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Methods, Methods, I, II. NY: Academic.

Winkler, R. L. (1972). Introduction to Bayesian inference and decision-making. NY: Holt, Rinehart and Winston.

# Random Regression Models Based On The Elliptically Contoured Distribution Assumptions With Applications To Longitudinal Data 

Alfred A. Bartolucci, Shimin Zheng<br>Department of Biostatistics<br>University of Alabama at Birmingham and<br>Nanjing Audit University, P. R. China

Sejong Bae, Karan P. Singh<br>Department of Biostatistics<br>Health Science Center at Fort Worth<br>University of North Texas

We generalize Lyles et al.'s (2000) random regression models for longitudinal data, accounting for both undetectable values and informative drop-outs in the distribution assumptions. Our models are constructed on the generalized multivariate theory which is based on the Elliptically Contoured Distribution (ECD). The estimation of the fixed parameters in the random regression models are invariant under the normal or the ECD assumptions. For the Human Immunodeficiency Virus Epidemiology Research Study data, ECD models fit the data better than classical normal models according to the Akaike (1974) Information Criterion. We also note that both univariate distributions of the random intercept and random slope and their joint distribution are non-normal short-tailed ECDs, and that the error term is distributed as a non-normal long-tailed ECD if we don't use the low undetectable limit or half of it to replace the undetectable values. Instead, we use the ECD cumulative distribution function to calculate the contribution to the likelihood due to the undetectable values.

Key words: Generalized multivariate analysis, power exponential distributions, Gamma distributions, maximum likelihood functions, censoring, informative drop-outs, empirical Bayes

## Introduction

In clinical studies of human immunodeficiency virus (HIV) infection the number of copies of HIV ribonucleic acid (RNA) per milliliter of plasma is often used to measure the progression of the disease. When the number of copies per milliliter is below or equal to 500 , the observation is considered as undetectable, missing, or left-censored, since the copy numbers below 500 are not quantifiable.

Correspondence regarding this article should be emailed to Alfred A. Bartolucci: albartol@uab.edu. The authors acknowledge assistance from Robert H. Lyles with SAS and S-Plus programming; and the HERS Study Group for providing the Human Immunodeficiency Virus (HIV) Epidemiology Research Study data.

On the other hand, illness or death caused by an early drop-out is known as an informative drop-out. If either a left-censored or an informative drop-out is present, as Lyles et al. (2000) pointed out, random effects linear models (Laird \& Ware, 1982) and generalized estimating equations (GEE) (Liang \& Zeger, 1986) produce biased estimates of key parameters, such as the population average HIV RNA slope and intercept. Louis (1982) used asymptotic approximation methods to deal with the problem of left-censored and informative drop-out data. Both Hughes (1999) and Schluchter (1992) implemented Maximum Likelihood (ML) estimation via Expectation and Maximization (EM) algorithm to handle the problem of left-censored and informative dropout data. Lyles et al. (2000) combined the approaches of Hughes (1999) and Schluchter (1992) into a single likelihood integrating subject-specific random slopes and intercepts which took both informative drop-out and undetectable data into account. Then, they maximized the likelihood function with respect to fixed effects and other variables. Our
approach follows Lyles et al. (2000) and we extend their normal distribution assumptions to the ECD assumptions since when the number of undetectable observations exceeds a certain number, or when the random intercept and random slope have a bell shaped and long-tailed or short-tailed distribution the ECD distribution improves the fit of the data over the normal distribution.

We used the data from the study of Lyles et al. in this paper. From April 1993 to June 1998 there were 528 HIV-infected women ( $16-55$ years old) in the HIV Epidemiology Research Study (HERS) and 1,864 RNA measurements were collected. Overall, there were 25 (4.7\%) drop-out events which resulted in 77 informative drop-out observations, according to Lyles et al.'s (2000) definition.

We used $\delta$ as an indicator which was set to 1 if an observation was an informative dropout and to 0 otherwise. For these 25 individuals the time on study was set as the minimum of the time from the base-line to death or the time from the base-line to 3 months beyond the last visit. For other non-informative drop-out women the censored time was set equal to the time from the base-line to the last visit date. Overall, 745 ( $40 \%$ ) out of 1,864 HIV RNA observations were undetectable or left-censored (below 500 copies per milliliter).

Power Exponential Distributions and Models
The power exponential distributions can be used to model both light and heavy tailed, symmetric and unimodal continuous data sets. Gomez et al. (1998) generalized the Univariate Power Exponential (UPE) distribution, which was established by Subbotin (1923), to the Multivariate Power Exponential (MPE) distribution. Both Johnson (1979) and Gomez et al. (1998) discussed the relationship between the UPE distribution and a Gamma distribution. Gomez et al. (1998) studied the properties of MPE intensively, including the stochastic representation, the moments, the characteristic function and the marginal and conditional distributions and asymmetry and kurtosis coefficients. Obviously, the family of MPE distribution is a subset of the class of ECDs. Gomez et al. (1998) defined the MPE distribution as follows:

$$
\begin{align*}
f(y ; \mu, \Sigma, \beta)= & \frac{n \Gamma\left(\frac{n}{2}\right)}{\pi^{\frac{n}{2}} \sqrt{|\Sigma| \Gamma\left(1+\frac{n}{2 \beta}\right) 2^{\left(1+\frac{n}{2 \beta}\right)}}}  \tag{1}\\
& \exp \left(-\frac{1}{2}\left[(y-\mu)^{\prime} \Sigma^{-1}(y-\mu)\right]^{\beta}\right),
\end{align*}
$$

where $-\infty<\mu<\infty, \Sigma>0,0<\beta<\infty$. If y is distributed as an MPE distribution with parameters $\mu, \Sigma$ and $\beta$, we write $\mathrm{y} \sim \operatorname{MPE}(\mu, \Sigma$, $\beta$ ) and we write $\mathrm{y} \sim \operatorname{UPE}(\mu, \sigma, \beta)$ if $n=1$. The parameter $\beta$ is called the shape parameter.

We use the following linear randomeffects regression model (LRRM):

$$
\begin{equation*}
y_{i j}=\alpha+a_{i}+\left(\beta+b_{i}\right) t_{i j}+e_{i j} . \tag{2}
\end{equation*}
$$

We take the response $\mathrm{y}_{i j}$ to be the base 10 logarithm of HIV RNA measured at the $j t h$ time point $t_{i j}\left(j=1,2, \ldots, n_{i}\right)$ for the $i$ th woman ( $i=$ $1, \ldots, 528,1 \leq n_{i} \leq 5$ for our data set). We assume that the error terms $e_{i j}$ are distributed as UPE $\left(\mu, \sigma^{2}, v_{1}\right)$, the random intercept deviations $a_{i}$ are distributed as UPE $\left(\mu, \sigma_{1}^{2}, v_{2}\right)$ and the random slope deviations $b_{i}$ are distributed as UPE ( $\mu, \sigma_{2}{ }^{2}, v_{2}$ ) with $\operatorname{cov}\left(a_{i}, b_{i}\right)=c \sigma_{12}$ where $c$ is the correction coefficient and $v_{2}$ is a shape parameter. The joint distribution of $a_{i}$ and $b_{i}$ is $\operatorname{MPE}_{2}\left(0, \Sigma_{2}, v_{2}\right)$, where $\Sigma_{2}=\left(\sigma_{i j}\right)$. Based on the trivariate normal distribution model (Schluchter, 1992) we assume the 3 -dimensional random vector $\left(a_{i}, b_{i}, T_{i}^{0}\right)^{\prime}$ distributed as trivariate power exponential, i.e.

$$
\left(\begin{array}{c}
a_{i} \\
b_{i} \\
T_{i}^{0}
\end{array}\right) \sim M P E_{3}\left(\mu, \Sigma_{3} v_{2}\right),
$$

where

$$
\mu=\left(\begin{array}{c}
0 \\
0 \\
\mu_{t}
\end{array}\right), \quad \Sigma_{3}=\left(\begin{array}{ccc}
\sigma_{1}^{2} & \sigma_{12} & \sigma_{a t} \\
\sigma_{12} & \sigma_{2}^{2} & \sigma_{b t} \\
\sigma_{a t} & \sigma_{b t} & \sigma_{t}^{2}
\end{array}\right), \quad r k\left(\Sigma_{3}\right)=3 .
$$

The joint pdf of $\left(a_{i}, b_{i}, T_{i}\right)^{\prime}$ is given as

$$
\begin{aligned}
& f\left(a_{i}, b_{i}, T_{i}^{0}\right)=\frac{3 \Gamma\left(\frac{3}{2}\right)}{\pi^{\frac{3}{2}} \sqrt{\left|\Sigma_{3}\right|} \Gamma\left(1+\frac{3}{2 v_{2}}\right) 2^{\left(1+\frac{3}{2 v_{2}}\right)}} \\
& \exp \left(-\frac{1}{2}\left[\left(a_{i}, b_{i}, T_{i}^{0}-\mu_{t}\right) \Sigma_{3}^{-1}\left(a_{i}, b_{i}, T_{i}^{0}-\mu_{t}\right)^{\prime}\right]^{v_{2}}\right)
\end{aligned}
$$

where $T_{i}^{0}$ is the natural logarithm of the "survival" time for subject $i$.

Maximum Likelihood Functions
In this section we utilize general integrated likelihood expressions given by Lyles et al. (2000), in order to facilitate estimation and inference for the ECD case.
(a) The Maximum Likelihood (ML) function without accounting for undetects and informative drop-outs: By the conditional probability formulae the ML function without accounting for undetects and informative dropouts is given by

$$
\begin{align*}
& L(\theta, Y, T)= \\
& \prod_{i=1}^{k}\left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_{i}} f\left(Y_{i j} \mid a_{i}, b_{i}\right) f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right) d a_{i} d b_{i}\right] \tag{3}
\end{align*}
$$

where $\theta=\left(\alpha, \beta, \sigma_{1}{ }^{2}, \sigma_{2}{ }^{2}, \sigma_{12}, \sigma^{2}\right)^{\prime}, Y$ is a vector consisting of $Y_{\mathrm{ij}}, T$ is a vector consisting of $t_{i j}$ and

$$
\begin{gathered}
f\left(Y_{i j} \mid a_{i}, b_{i}\right)=\frac{1}{\sigma \Gamma\left(1+\frac{1}{2} v_{1}\right) 2^{\left(1+\frac{1}{2} v_{1}\right)}} \\
\exp \left(-\frac{1}{2} \Sigma\left[\frac{Y_{i j}-\left[\alpha+a_{i}+\left(\beta+b_{i}\right) t_{i j}\right]}{\sigma}\right]^{v_{1}}\right), \\
f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right)=f\left(a_{i}, b_{i}\right)=\frac{1}{\pi \sqrt{\left|\Sigma_{2}\right|}\left(1+\frac{1}{v_{2}}\right) 2^{\left(1+\frac{1}{v_{2}}\right)}} \\
\exp \left(-\frac{1}{2}\left[\left(a_{i}, b_{i}\right) \Sigma_{2}^{-1}\left(a_{i}, b_{i}\right)^{\prime}\right]^{v_{2}}\right) .
\end{gathered}
$$

(b) The ML function accounting for undetectable values only:
We use $d$ to denote the operable limit of detection. We assume that the first $n_{i 1}$ measurements are detectable values and there are $n_{i}-n_{i 1}$ undetectable values for subject $i$. We
use the probability distribution function (pdf) to calculate the contribution to the likelihood due to the observed values for subject $i$. On the other hand we use the cumulative distribution function (cdf) to calculate the contribution to the likelihood due to the undetectable values. Therefore, the complete-data likelihood function is given by

$$
\begin{aligned}
& L(\theta, Y)= \\
& \prod_{i=1}^{k}\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{n_{n}} f\left(Y_{i j} \mid a_{i}, b_{i}\right) \prod_{j=n_{j i}+1}^{n_{i}} F_{Y}\left(d \mid a_{i}, b_{i}\right) f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right) d a_{i} d b_{i}\right],
\end{aligned}
$$

where $f\left(\mathrm{Y}_{i j} \mid a_{i}, b_{i}\right)$ and $f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right)$ are given in (3) and

$$
\begin{align*}
F_{Y}\left(d \mid a_{i}, b_{i}\right)=\int_{-\infty}^{b} & \frac{1}{\sigma \Gamma\left(1+\frac{1}{2} v_{1}\right) 2^{\left(1+\frac{1}{2} v_{1}\right)}} \exp \left(-\frac{1}{2}\left(\frac{y}{\sigma}\right)^{v_{1}}\right) d y \\
& \frac{1}{2}\left[P\left\{u \leq \frac{1}{2}\left(\frac{b}{\sigma}\right)^{2 v_{1}}\right\}+1\right], \text { if } b \geq 0 \\
& \frac{1}{2}-\frac{1}{2} P\left\{u \leq \frac{1}{2}\left(-\frac{b}{\sigma}\right)^{2 v_{1}}\right\}, \tag{4}
\end{align*} \quad \text { if } b<0
$$

where $b=y_{i}-\left[\alpha+a_{i}+\left(\beta+b_{i}\right) t_{i}\right], y_{i}$ is the censored value for subject $i$ and $u \sim \Gamma\left(1, \frac{1}{2 \nu_{1}}\right)$.
(c) The ML function accounting for informative drop-outs only:
We use $T_{i}^{0}$ to denote the natural logarithm of the "survival" time for subject $i$ and $c_{i}$ to denote the natural logarithm of the time from the base-line to the study end. Let $T_{i}=\min \left(T_{i}{ }^{0}, c_{i}\right)$.
i) If subject $i$ did not drop out early we have $\delta_{i}=0$ and use $1-F_{T}\left(c_{i} \mid a_{i}, b_{i}\right)$ to compute the contribution to the likelihood due to the right censored values, where $F$ is the cdf of $T$ given $a_{i}$ and $b_{i}$. That is

$$
\begin{align*}
& F_{T}\left(c_{i} \mid a_{i}, b_{i}\right)=\frac{\frac{3}{2} \Gamma\left(\frac{3}{2}\right) \sqrt{\left|\Sigma_{2}\right| \Gamma\left(1+\frac{1}{\nu_{2}}\right)}}{\sqrt{\pi} \sqrt{\left|\Sigma_{3}\right| \Gamma\left(1+\frac{3}{2 v_{2}}\right)} 2^{\left(\frac{1}{2 v_{2}}\right)}}  \tag{5}\\
& \int_{-\infty}^{c_{i}} \exp \left\{\begin{array}{l}
-\frac{1}{2}\left[\left(a_{i}, b_{i}, z-\mu_{t}\right) \sum_{3}^{-1}\left(a_{i}, b_{i}, z-\mu_{t}\right)\right]^{\prime} \\
+\frac{1}{2}\left[\left(a_{i}, b_{i}\right)\left(\begin{array}{ll}
\sigma_{1}^{2} & \sigma_{12} \\
\sigma_{12} & \sigma_{2}^{2}
\end{array}\right)^{-1}\binom{a_{i}}{b_{i}}\right]^{1 / 2}
\end{array}\right\} d z,
\end{align*}
$$

where $\Sigma_{2}, \Sigma_{3}$ and $v_{2}$ were defined in LRRM.
(ii) If subject $i$ dropped out early we have $\delta_{i}=1$ and $T_{i}=T_{i}^{0}$ and use the pdf $f\left(T_{i}^{0} \mid a_{i}\right.$ ,$b_{i}$ ) to compute the contribution to the likelihood due to the informative drop-out values. Therefore, the likelihood function accounting for informative drop-outs and the right censored data is given by

$$
\begin{align*}
& L(\theta, Y, T)= \\
& \prod_{i=1}^{n}\left[\begin{array}{l}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f\left(Y_{i} \mid a_{i}, b_{i}\right) f\left(T_{i}^{0} \mid a_{i}, b_{i}\right)^{\delta_{i}} \\
{\left[1-F_{T}\left(c_{i} \mid a_{i}, b_{i}\right)\right]^{1-\delta_{i}} f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right) d a_{i} d b_{i}}
\end{array}\right], \tag{6}
\end{align*}
$$

where $\theta=\left(\alpha, \beta, \sigma_{1}{ }^{2}, \sigma_{2}{ }^{2}, \sigma_{12}, \sigma^{2}, \mu_{t}, \sigma_{a t}, \sigma_{b t}, \sigma_{\mathrm{t}}^{2}\right)^{\prime}$. Thus, the complete ML function is given by

$$
\begin{align*}
& L(\theta, Y, T)= \\
& \prod_{i=1}^{528}\left[\begin{array}{l}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{n_{n}} f\left(Y_{i j} \mid a_{i}, b_{i}\right) \prod_{j=n_{i}+1}^{n_{i}} F_{Y}\left(d \mid a_{i}, b_{i}\right) \\
f\left(T_{i}^{0} \mid a_{i}, b_{i}\right)^{\delta_{i}}\left[1-F_{T}\left(c_{i} \mid a_{i}, b_{i}\right)\right]^{1-\delta_{i}} \\
f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right) d a_{i} d b_{i}
\end{array}\right] . \tag{7}
\end{align*}
$$

Computing Empirical Bayes Estimates of Random Intercepts \& Random Slopes

In this section we discuss the calculation of empirical Bayes estimates of random intercepts and random slopes in the presence of drop-outs and undetectable values based on the ECD assumptions. Specifically, we calculate the estimate of the random intercept $a_{i}$ and random slope $b_{i}$ by substituting the ML estimators of $\theta$ based on the ML function (7) developed in the last section into the analytic expressions for the posterior means given the observed data $\left(Y_{i}, T_{i}\right)$.

Specifically, the empirical Bayes estimates of the random intercept $a_{i}$ and slope $b_{i}$ for subject $i$ are given, respectively, by
$\hat{a}_{i}=E\left(a_{i} \mid Y_{i}, T_{i}\right)=f^{*}\left(Y_{i}, T_{i}, \theta\right)^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_{a}\left(a_{i}, b_{i}\right) d a_{i} d b_{i}$,
where

$$
\begin{aligned}
& g_{a}\left(a_{i}, b_{i}\right)= \\
& a_{i} f^{*}\left(Y_{i} \mid a_{i}, b_{i}\right) f\left(T_{i} \mid a_{i}, b_{i}\right) f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right), \\
& \hat{b}_{i}=E\left(b_{i} \mid Y_{i}, T_{i}\right)= \\
& f^{*}\left(Y_{i}, T_{i}, \theta\right)^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_{b}\left(a_{i}, b_{i}\right) d a_{i} d b_{i},
\end{aligned}
$$

where

$$
\begin{aligned}
& g_{b}\left(a_{i}, b_{i}\right)= \\
& b_{i} f^{*}\left(Y_{i} \mid a_{i}, b_{i}\right) f\left(T_{i} \mid a_{i}, b_{i}\right) f\left(a_{i} \mid b_{i}\right) f\left(b_{i}\right) .
\end{aligned}
$$

The above empirical Bayes estimates were given by Lyles et al., (2000). Note that $f^{*}\left(\mathrm{Y}_{i} \mid a_{i}, b_{i}\right)$ is different from $f\left(\mathrm{Y}_{i} \mid a_{i}, b_{i}\right)$, the one with asterisk indicates that the data vector $\mathrm{Y}_{i}$ may include one or more undetectable values.

## Computation

The software package we have used to obtain the ML estimates of variance components and fixed effects corresponding to models discussed in this chapter is SAS PROC IML. The ML function is constructed within PROC IML first. The initial parameter estimates are obtained from Lyles et al. (2000). The ML function is maximized through the NLPQN routine in IML with respect to the parameters stated in this paper. The double integration was computed by quadrature for each subject. The Hessian matrix (the dispersion matrix of the estimated parameters) was found through the NLPFDD routine in IML. There are no built-in generic non-normal ECD functions in SAS. We used the theorems of relationship between a UPE and a Gamma distribution developed in another paper to compute the probability of UPE distribution below or above a certain point. However, this method can not be used to deal with MPE distribution or the conditional and marginal MPE distributions since there is no
existing useful relationship between an MPE and a Gamma distribution and the conditional or the marginal distributions of an MPE are not necessarily MPEs, which can be much more complicated ECD distributions. We used approximation methods to integrate such integrands. The Simpson's rule has been adopted which requires much less computing time and can reach highly accurate results. S-Plus and SAS PROC IML were used to obtain the empirical Bayes estimates of the random intercept and random slope for each subject and the critical values of UPE distribution and the Simpson's rule has also been used for nonnormal situations.

## Results

We used the Akaike (1974) Information Criterion (AIC) which was used by Lindsey (1999) and among others to compare the classical multivariate normal model and the multivariate power exponential model. In version 8 of SAS/STAT software AIC is defined as 'smaller-is-better'. Specifically, AIC=2l+2d, where $l$ denotes the maximum value of the log likelihood, $d$ denotes the dimension of the model, i.e., the number of parameters estimated in the ML function. Six models were considered:

Model 1 (M1): In this model we assumed the normal distributions. There were six parameters $\left(\alpha, \beta, \sigma_{1}^{2}, \sigma_{2}^{2}, \sigma_{12}, \sigma^{2}\right)$ estimated in the ML function accounting for undetectable values which were constructed as in equation (2) of Lyles et al. (2000, p.488).

Model 2 (M2): As in model M1, the normal distributions were assumed. There were ten parameters $\left(\alpha, \beta, \sigma_{1}{ }^{2}, \sigma_{2}{ }^{2}, \sigma_{12}, \sigma^{2}, \mu_{t}, \sigma_{a t,} \sigma_{b t}\right.$, $\sigma_{\mathrm{t}}^{2}$ ) estimated in the ML function accounting for both undetects and informative drop-outs which were constructed as in equation (5) of Lyles et al. (2000, p.489).

Model 3 (M3): ECDs were assumed in this model. This model accounts for undetectable values only. Furthermore we assumed that two shape parameters were equal, i.e., $v_{1}=v_{2}$. There were seven parameters ( $\alpha, \beta$, $\sigma_{1}{ }^{2}, \sigma_{2}{ }^{2}, \sigma_{12}, \sigma^{2}, v_{1}$ ) estimated using the ML function.

Model 4 (M4): This model is the same as M3 except that we don't assume $v_{1}=v_{2}$.

Model 5 (M5): ECDs were assumed in this model. Undetectable, informative drop-out and right censored values were considered at the same time in this model. Also, we assume $v_{1}=$ $\nu_{2}$.

Model 6 (M6): This model is the same as M5 except that we don't assume $v_{1}=v_{2}$.

Next, we summarize what we have found from the HERS data analysis.
(1). ECDs fit the data much better than the classical normal distributions.
Among models M1, M3 and M4 we account for undetectable values only. Model M1 is based on the normal distribution assumptions while model M3 and M4 are based on ECD assumptions. The value of AIC changes from 3932.216 to 3928.101 when the model, M3, is used whereas the value reduces to 3908.833 using the model, M4. Overall, model M4 is the best according to the AIC standard if we consider undetectable values only in our analysis.

Among models M2, M5 and M6 we treat undetects, informative drop-outs and right censored observations simultaneously. Model M2 is based on the normal distribution assumptions, but model M5 and M6 are based on the ECD assumptions. Model M5 reduces AIC from 4083.556 of M2 to 4079.746 (see Table 2). Overall, model M6 (4064.791) is the best by AIC standard if we consider all possible situations.
(2). The dispersion matrix of an MPE random vector is proportional to $\sum$ as defined in section 2. Hence, multiplying the ML estimate $\hat{\Sigma}$ by a coefficient we transformed $\hat{\Sigma}$ to the estimated dispersion matrix whose elements are listed in Table 1. As expected, variance and covariance estimates are very close under the six different models. This proportional relationship provides us a short cut to gain the ML estimates. That is, we can get the ML estimate of the dispersion matrix under the normal distribution assumption first and then utilize this estimated dispersion matrix to estimate the shape parameters. This method is very useful and effective, especially when we have a large number of parameters to estimate or when we deal with a very large data set where computing CPU time and memory space are prohibiting. The estimates of the fixed intercept and the fixed slope for all subjects are almost exactly the same
under the six different models. This is because that $\hat{\alpha}$ and $\hat{\beta}$ only involve the data set which is given and the dispersion matrices of random effects and error terms which are invariant under the normal distribution assumptions and the ECD assumptions as we discussed.
(3). The estimates of the shape parameters in Table 2 strongly suggest that we should make the power exponential distribution assumptions instead of classical normal distribution assumptions since our simulations revealed that less than 0.94 or greater than 1.15 shape parameters indicate the distribution departs significantly from the normal distribution at $\alpha=0.05$ level. The shape parameter estimates $\hat{v}=0.6574$ (S.E. $=0.116$ ) under model M3 and $\hat{v}=0.6997$ (S.E. $=0.099$ ) under model M5 indicate that $40 \%$ undetects contribute to a long tailed non-normal distribution. In model M4 and M6 we don't assume $v_{1}=v_{2}$. The estimate of the second shape parameter is $\hat{v}_{2}=1.8089$ (S.E. $=0.490$ ) in model M4 and $\hat{v}_{2}=1.3706$ (S.E. $=0.215$ ) in model M6. The shape parameter estimate $\hat{v}_{2}$ in both models M4 and model M6 are much larger than 1 which shows that both univariate distributions of the random intercept and the random slope and their joint distribution are non-normal. They are thintailed $E C D s$, concentrated around 0 means.

## Possible Extensions

First, power exponential distributions are just a member of larger ECD family. To extend the power exponential distribution assumptions for the models we have discussed is a challenging task and of great interest in both theory and practice. Second, we used approximation methods to compute probability distribution function values at a certain given point and the probability on some interval or within a certain given high dimension rectangle for the non-normal power exponential distributions. The CPU time and memory space required for this kind of task are prohibitive. This highly intensive computing problem will be eased if we could find an exact or asymptotic relationship between distributions (such as nonnormal MPEs and marginal or conditional distributions of a non-normal MPE). Third, we
have used simulation methods to assess different distributions, like normal or non-normal characteristics as per the shape parameter. If we could construct a statistic related to the shape parameter and get an explicit, exact or asymptotic distribution of the statistic we could do a formal accurate hypothesis testing about the shape parameter of the distribution. This is another challenging task for future research. All source code provided in this paper is in SAS (Appendix).

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second International Symposium on Inference Theory.

Fang, K. T., \& Zhang, Y. T. (1993). Generalized multivariate analysis. Science Press.

Gomez, E., Gomez-Villegas, M. A., \& Marin, J. M. (1998). A multivariate generalization of the power exponential family of distributions. Communications in Statistics, A27, 589-600.

Harville, D. A. (1977). Maximum likelihood approached to variance component estimation and to related problems. Journal of the American Statistical Association, 722 (358), 320 - 338.

Harville, D. (1976). Extension of the GaussMarkov theorem to include the estimation of random effects. The Annals of Statistics, 4(2), 384-395.

Hughes, J. P. (1999). Mixed effects models with censored data with application to HIV RNA levels. Biometrics, 55, 625-629.

Laird, N. and Ware, J. (1982). RandomEffects models for longitudinal data. Biometrics 38, 963-974.

Lindsey, J. K. (1999). Multivariate elliptically contoured distributions for repeated measurements. Biometrics, 55, 1277-1280.

Lyles, R. H., Lyles, C. M., \& Taylor, D. J. (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative dropouts. Applied Statistics, 49 (4), 485-497.

Lyles, R. H., Lyles, C. M. \& Taylor, D. J. (2000). SAS programs simulation data sets. http://www.blackwellpublishing.com/rss/Volum es/Cv49p4.htm.

Schluchter, M. D. (1992). Methods for the analysis of informatively censored longitudinal data. Statistics in Medicine, Vol. 11, 1861-1870.

Smith, D. K., Warren, D. L., Vlahov, D., Schuman, P, Stein, M. D., Greenberg, B.L., \& Holmberg, S. D. (1997). Design and baseline participant characteristics of the human immunodeficiency virus epidemiology research (HER) study: A prospective cohort study of human immunodeficiency virus infection in US women. American Journal of Epidemiology, 146, 459-469.

Table 1. Results from HERS data: ML Estimates.

|  | $\alpha$ | $\beta$ | $\sigma_{1}{ }^{2}$ | $\sigma_{2}{ }^{2}$ | $\sigma_{12}$ | $\sigma^{2}$ | $\mu_{\mathrm{t}}$ | $\sigma_{a t}$ | $\sigma_{b t}$ | $\sigma_{t}{ }^{2}$ |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| M1 | 2.89 | 0.058 | 0.721 | 0.037 | 0.061 | 0.383 | - | - | - | - |
|  | $(0.033)$ | $(0.016)$ | $(0.088)$ | $(0.008)$ | $(0.022)$ | $(0.023)$ |  |  |  |  |
| M2 | 2.88 | 0.062 | 0.718 | 0.039 | 0.060 | 0.382 | 2.32 | 0.165 | 0.035 | 0.298 |
|  | $(0.050)$ | $(0.016)$ | $(0.088)$ | $(0.008)$ | $(0.022)$ | $(0.023)$ | $(0.158)$ | $(0.062)$ | $(0.022)$ | $(0.096)$ |
| M3 | 2.91 | 0.058 | 0.747 | 0.040 | 0.050 | 0.387 | - | - | - | - |
|  | $(0.057)$ | $(0.017)$ | $(0.149)$ | $(0.009)$ | $(0.015)$ | $(0.076)$ |  |  |  |  |
| M4 | 2.89 | 0.050 | 0.695 | 0.044 | 0.054 | 0.410 | - | - | - | - |
|  | $(0.002)$ | $(0.001)$ | $(0.417)$ | $(0.028)$ | $(0.052)$ | $(0.023)$ |  |  |  |  |
| M5 | 2.90 | 0.062 | 0.833 | 0.047 | 0.059 | 0.383 | 2.258 | 0.173 | 0.042 | 0.269 |
|  | $(0.053)$ | $(0.017)$ | $(0.137)$ | $(0.008)$ | $(0.015)$ | $(0.068)$ | $(0.144)$ | $(0.036)$ | $(0.010)$ | $(0.060)$ |
| M6 | 2.90 | 0.062 | 0.833 | 0.047 | 0.059 | 0.383 | 2.258 | 0.173 | 0.042 | 0.269 |
|  | $(0.053)$ | $(0.017)$ | $(0.137)$ | $(0.008)$ | $(0.015)$ | $(0.068)$ | $(0.144)$ | $(0.036)$ | $(0.010)$ | $(0.060)$ |

Note. Numbers in parentheses are Standard Errors of the corresponding estimates.

Table 2. Results from HERS data: Shape parameter estimates and AIC.

|  | $v_{1}$ | $v_{2}$ | $d$ | $-2 \log$-likelihood | AIC |
| :--- | :---: | :---: | :--- | :---: | :---: |
| M1 | - | - | 6 | 3920.216 | 3932.216 |
| M2 | - | - | 10 | 4063.556 | 4083.556 |
| M3 | $0.6574(0.116)$ | - | 7 | 3914.101 | 3928.101 |
| M4 | $0.4694(0.060)$ | $1.8090(0.490)$ | 8 | 3892.833 | 3908.833 |
| M5 | $0.6997(0.099)$ | - | 11 | 4057.746 | 4079.746 |
| M6 | $0.5173(0.055)$ | $1.3706(0.215)$ | 12 | 4040.791 | 4064.791 |

Note. Numbers in parentheses are Standard Errors of the corresponding estimates.

## Appendix

SAS Program for Taking Left-censored into Account
Acknowledgments: The following program was created originally by Dr. Robert H. Lyles. We have changed his distribution assumptions normal to ECD and added five nonlinear constraints. We really appreciate Dr. Lyles's providing this program.
Description: Calculation of the ML estimates of the fixed effects and the variance matrix.
We assume the underlying distributions are ECDs. Also, we take the undetectable observations into account under the model described by the likelihood equation (4) in this paper.
data test;
infile '/herscens1.dat';
input obsn id time nondet response fail survtyrs logsurvt;
*Compute ML estimates via PROC MIXED on complete data (which would not be available in practice). That is, using the actual values for the response and all 1864 measurements;
proc mixed data=test method=ml;
class id;
model response=time / s ddfm=bw;
random intercept time /type=un subject=id;
title2 "ml estimates for full data set (unavailable in
practice)"; run;
data test2;
set test;
if nondet=1 then do;
observed=0;
end;
else if nondet=0 then do; observed=1;
end;
label response="Base 10 log HIVRNA value"
time ="time of measurement"
id ="subject id"
observed="indicator for whether value was observed"
fail="indicator for whether subject dropped out"
survtyrs="Years to dropout"
logsurvt="Natural log of dropout time";

* Compute ML estimates ignoring left censoring and drop-outs using PROC MIXED with random intercept and slope. These naive estimates will be used as starting values for the six parameters of the mixed effects model;
proc mixed data=test2 method=ml;
class id;

```
    model response=time / s ddfm=bw;
    random intercept time /type=un subject=id;
    title2 "ml estimates ignoring left censoring and dropouts";
run;
***Create dataset to be read into IML for maximizing likelihood in
Eqn. 2, accounting for left censoring: *;
data test; set test;
    if nondet=1 then do;
        observed=0;
    end;
    else if nondet=0 then do;
        observed=1;
    end; run;
proc iml worksize=999216000 symsize=999999900;
*********************************************************************
* define IML function which will be used to maximize the likelihood
***********************************************************************
start likeli1(parms);
* lower and upper boundaries and stepsize for numerical integration;
        nsteps=31;
a_l =-5;
a_u = 5;
step_a=(a_u-a_l)/(nsteps-1);
b_l = -1.5;
b_u = 1.5;
step_b=(b_u-b_l)/(nsteps-1);
pi=2*arsin(1);
* variables corresponding to input parameters from vector 'parms';
        sigsq1 =parms[1]; * random intercept effect variance;
sig12 =parms[2]; * covariance between random intercept and slope;
sigsq2 =parms[3]; * random slope effect variance;
sigsq =parms[4]; * within subject variance;
alpha =parms[5]; * fixed effect intercept;
beta =parms[6]; * fixed effect slope;
v =parms[7];
* determine number of subjects in dataset;
use test;
        read all var {id} into subjects;
        close test;
* compute number of subjects and create vector for each subjects
    contribution to the likelihood;
subjects=ncol(unique(subjects));
```

```
terms=j(subjects,1,.);
* get vector of indicators for observed vs. censored responses for
subject i;
do i=1 to subjects;
    use test;
    read all var {observed} into d_i where (id=i);
    close test;
* number of observations, number of observed values, and number of
censored
    values, respectively, for subject i;
    n_i=nrow(d_i);
    o_i=sum(d_i);
    c_i=n_i-o_i;
* create vectors of censored values and the associated time of
measurement;
    if c_i>0 then do;
        use test;
        read all var {response} into cens_i where (id=i & observed=0);
        read all var {time} into c_time_i where (id=i & observed=0);
        close test;
        end;
* create vectors of observed values and the associated time of
measurement;
    if o_i>0 then do;
        use test;
        read all var {response} into y_i where (id=i & observed=1);
        read all var {time} into time_i where (id=i & observed=1);
        close test;
        end;
    * set initial value for likelihood contribution by subject i to zero;
    func_i=0;
* define quadrature points for numerical integration;
    do a_i=a_l to a_u by step_a;
    do b_i=b_l to b_u by step_b;
        * contribution to likelihood due to observed values for subject
i;
        if o_i=0 then func_i1=1;
            else do;
            t_i1=(y_i-alpha-beta*time_i);
            t_i2=(a_i+b_i*time_i);
            func_i1=(1/(sqrt(sigsq)*gamma(1+0.5/v)*(2##(1+0.5/v)))**o_i)*
                        exp(-0.5*sum(((t_i1-t_i2)##2/sigsq)##v));
            end;
```

* contribution to likelihood due to censored values for subject i;
func_i2=1;
if c_i>0 then do $j=1$ to c_i;
b=cens_i[j,1]-alpha-a_i-beta*c_time_i[j,1]-
b_i*c_time_i[j,1];
if $b>=0$ then
temp_i2=0.5*(1+probgam(0.5*(b/sqrt(sigsq))**(2*v),(1/(2*v))));
else temp_i2=0.5*(1-probgam(0.5* (-
b/sqrt(sigsq))**(2*v),(1/(2*v))));
func_i2=func_i2*temp_i2;
end;
* compute correlation coefficient between intercept and slope; r=sig12/sqrt(sigsq1*sigsq2);
* compute joint distribution of intercept and slope;
w=(sigsq1||sig12)//(sig12||sigsq2); u=det(w); $y=i n v(w)$;
x_i=(a_i||b_i);
func_i3=(2/(pi*sqrt(u)*gamma(1+1/v)*(2\#\#(1+1/v))))* exp(-0.5*(x_i*y*x_i`)\#\#v);
* compute contribution of subject 'i' to objective function;
func_i=func_i+(func_i1*func_i2*func_i3*step_a*step_b); end; end;
* add subject i's contribution to vector of likelihood terms;
terms[i,1]=func_i;
end;
* compute -2 log likelihood;
loglik2=-2*sum(log(terms));
return(loglik2);
finish likeli1;

The following is the main body of the program (which calls the minimization function, computes the Hessian, etc.)

;
initial estimates from preliminary analysis; parms $=\{.24-.012 .015$. 2013.21 .0391 .0$\}$;

```
* options vector for minimization function;
* matrix of lower (row 1) and upper (row 2) bound contraints on
parameters
    (sigsq1 > 0, sig12 <> 0, sigsq2 > 0, sigsq > 0, alpha <> 0, beta <>
0);
/* con={1E-5 . 1E-5 1E-5 . .,
    . . . . . .}; */
* The following are five non-linear restrictions;
start c_h(parms);
    c=j(5,1,0.);
    c[1]=parms[1];
    c[2]=parms[3];
    c[3]=parms[4];
    c[4]=parms[1]-(parms[2]##2/parms[3]);
    c[5]=parms[7];
        return(c);
finish c_h;
* call function minimizer in IML;
optn=j(1,11,.); optn[1]=0; optn[2]=3; optn[10]=5; optn[11]=0;
call nlpqn(rc, xres, "likeli1", parms, optn) nlc="c_h";
* create vector of mle's computed using function minimizer;
parms=xres`;
* compute numerical value of Hessian (and covariance matrix) using
    mle's calculated above;
call NLPFDD(crit, grad, hess, "likeli1", parms);
cov_mat=2*inv(hess);
se_vec =sqrt(vecdiag(cov_mat));
print cov_mat se_vec;
*******************************************************;
The following program is used to transform MLE of ECD Sigma matrix int the variance matrix;
proc iml;
sig1={ 0.221828 -0.014873 0.011839};
sig = 0.141499; a = 2.906699; b = 0.057614;
beta=0.657391;
c1=2**(1/beta)*gamma(2/beta)/(2*gamma(1/beta));
c2=2**(1/beta)*gamma(1.5/beta)/(gamma(0.5/beta));
sig11=c1*sig1; sig0=c2*sig;
sigma=sig11||sig0||a||b;
print sigma; /* with ECD */
sigmaOld={0.720710 -0.060955 0.037333 0.382976 2.886360 0.058335};
print sigmaOld; /* without ECD */
```


# Using Zero-inflated Count Regression Models <br> To Estimate The Fertility Of U. S. Women 

Dudley L. Poston, Jr. Sherry L. McKibben<br>Department of Sociology<br>Texas A\&M University

In the modeling of count variables there is sometimes a preponderance of zero counts. This article concerns the estimation of Poisson regression models (PRM) and negative binomial regression models (NBRM) to predict the average number of children ever born (CEB) to women in the U.S. The PRM and NBRM will often under-predict zeros because they do not consider zero counts of women who are not trying to have children. The fertility of U.S. white and Mexican-origin women show that zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models perform better in many respects than the Poisson and negative binomial models. Zero-inflated Poisson and negative binomial regression models are statistically appropriate for the modeling of fertility in low fertility populations, especially when there is a preponderance of women in the society with no children.

Key words: Poisson regression, negative binomial regression, demography, fertility, zero counts

## Introduction

When analyzing variation in the number of children that women have born to them, demographers frequently use Poisson and negative binomial regression models rather than ordinary least squares models. Poisson and negative binomial regression models are statistically more appropriate for predicting a woman's children ever born (CEB), particularly in societies where mean fertility is low (Poston, 2002). Most women in such populations have children at the lower parities, including zero parity, and few have children at the higher parities. The CEB variable, which by definition

Dudley L. Poston, Jr. is Professor of Sociology, and the George T. and Gladys H. Abell Professor of Liberal Arts. He is co-editing (with Michael Micklin) the Handbook of Population (Klewer Plenum, 2004). E-mail him at: dudleyposton@yahoo.com. Sherry L. McKibben is a Lecturer in the Department of Sociology. Email: sherrymc. sherrymckibben@yahoo.com.
is a count variable, i.e., a nonnegative integer, is hence heavily skewed with a long right tail.

The statistical modeling of these kinds of CEB data is best based on approaches other than the ordinary least squares (OLS) linear regression model because using it to predict a count outcome, such as CEB, will often "result in inefficient, inconsistent, and biased estimates" (Long, 1997, p. 217) of the regression parameters. Poisson regression models (PRM) and negative binomial regression models (NBRM) have been shown to be statistically more appropriate (Poston, 2002).

However, sometimes there are so many zeros in the count dependent variable that both the PRM and the NBRM under-predict the number of observed zeros; the resulting regression models, therefore, often do not fit the data. Zero-inflated count regression models were introduced by Lambert (1992) and Greene (1994) for those situations when the PRM and the NBRM failed to account for the excess zeros and resulted in poor fit. This paper examines the use and application of zero-inflated count regression models to predict the number of children ever born to U.S. women.

## Methodology

The most basic approach for predicting a count variable, such as CEB, is the Poisson regression model (PRM). In the PRM, the dependent variable, namely, the number of events, i.e., in the case of this paper, the number of children ever born (CEB), is a nonnegative integer and has a Poisson distribution with a conditional mean that depends on the characteristics (the independent variables) of the women (Long, 1997; Long \& Freese, 2001). The PRM incorporates observed heterogeneity according to the following structural equation:

$$
\mu_{i}=\exp \left(a+X_{1 i} b_{1}+X_{2 i} b_{2}+\ldots+X_{k i} b_{k}\right)
$$

where $\mu_{I}$ is the expected number of children ever born for the $\boldsymbol{i}^{\text {th }}$ woman; $\mathbf{X}_{1 i}, \mathbf{X}_{2 i} \ldots \mathbf{X}_{\mathbf{k} i}$ are her characteristics; and a, $\boldsymbol{b}_{1}, \boldsymbol{b}_{\mathbf{2}} \ldots \boldsymbol{b}_{\mathrm{k}}$ are the Poisson regression coefficients.

The PRM is appropriate when the mean and the variance of the count distribution are similar, and is less applicable when the variance of the distribution exceeds the mean, that is, when there is over-dispersion in the count data. If there is significant over-dispersion in the distribution of the count, the estimates from the PRM will be consistent, but inefficient. "The standard errors in the Poisson regression model will be biased downward. Resulting in spuriously large z -values and spuriously small p values" (Long \& Freese, 2001; Cameron \& Trivedi, 1986), which could lead the investigator to make incorrect statistical inferences about the significance of the independent variables.

This is addressed by adding to the PRM "a parameter that allows the conditional variance of (the count outcome) to exceed the conditional mean" (Long, 1997, 230). This extension of the Poisson regression model is the negative binomial regression model (NBRM). The NBRM adds to the Poisson regression model the error term $\boldsymbol{\varepsilon}$ according to the following structural equation:

$$
\mu_{i}=\exp \left(a+X_{1 i} b_{1}+X_{2 i} b_{2}+\ldots+X_{k i} b_{k}+\varepsilon_{i}\right)
$$

However, sometimes there are many more zeros in the count dependent variable than
are predicted by the PRM or NBRM, resulting in an overall poor fit of the model to the data. Zeroinflated models respond to this problem of excess zeros "by changing the mean structure to allow zeros to be generated by two distinct processes" (Long \& Freese, 2001, p. 250).

Consider a few examples of excess zeros. Suppose one wishes to survey visitors to a national park to predict the number of fish they caught. Suppose that some of the visitors did not fish, but data were not available on who fished and who did not fish. The data gathered hence have a preponderance of zeros, some of which apply to persons who fished and caught no fish, and others to persons who did not fish (Stata, 2001; Cameron \& Trivedi, 1998).

Or consider the problem of predicting the number of publications written by scientists. Some scientists will never publish either because they have chosen not to do so, or, perhaps, because they are not permitted to do so. But assume that there are no data telling which scientists have a zero probability of ever publishing. As with the example of the number of fish caught, there will be a preponderance of zeros among scientists with regards to the number of articles published. Some of the zeros will apply to scientists who tried to publish but were not successful and others to scientists who did not try to publish (Long \& Freese, 2001; Long, 1990).

Finally, consider the example to be addressed in this paper, namely, the number of children born to women. Some women will choose not to have children and are referred to as voluntarily childless women. Other women will try to have children but will not be successful in their attempts and are referred to as involuntarily childless women (Poston, 1976; Poston \& Kramer, 1983). But, assume that it is not directly known to which group each woman belongs. Thus among women of the childbearing ages of $15-49$, there will be many zeros on the CEB dependent variable; some of the zeros will apply to women who tried to produce children but were not successful, and others to women who voluntarily opted against having children.

Long and Freese (2001) stated that in zero-inflated models it is assumed that "there are two latent (i.e., unobserved) groups. An individual in the Always-0 Group (Group A) has
an outcome of 0 with a probability of 1 , while an individual in the Not Always-0 Group (Group $\sim$ A) might have a zero count, but there is a nonzero probability that she has a positive count" (p. 251).

In all cases, the investigator does not know into which of the two groups the respondents fall. If it was known into which group each subject was placed, one could subtract the persons belonging to the Always-0 Group from the total sample, and estimate Poisson or negative binomial regression models. But typically one does not have this kind of information, thus requiring the introduction of zero-inflated regression.

The estimation of zero-inflated regression models involves three steps: 1) predicting membership in the two latent groups, Group A and Group $\sim \mathrm{A}$; 2) estimating the number of counts for persons in Group $\sim A$; and 3) computing "the observed probabilities as a mixture of the probabilities for the two groups" (Long \& Freese, 2001, p. 251).

To analyze the fertility of U.S. women, one would follow these steps (for detail, see Long \& Freese, 2001. p. 251-252; Cameron \& Trivedi, 1998, p. 125-127, 211-215).

In Step 1, use a logistic regression model to predict the woman's membership in Group A (never have children) or Group $\sim \mathrm{A}$ (may or may not have children). The independent variables used in the logistic equation may be "referred to as inflation variables since they serve to inflate the number of 0s" (Long \& Freese, 2001, p. 251).

In Step 2, for women in Group ~A (may or may not have children), depending on whether or not there is over-dispersion in the CEB dependent variable, use either a Poisson regression model or a negative binomial regression model to predict the probabilities of counts 0 to $y$ (where $y$ is the maximum number of children born to a woman). The independent variables used in Step 2 may or may not be the same as those used in Step 1. In the examples shown below, the same independent variables are used in both steps. Using the same variables in both steps is not required. Different variables could be used in each step.

In Step 3, the results from the preceding steps are used to determine the overall
probability of 0 's, which is "a combination of the probabilities of 0 's from each group, weighted by the probability of an individual (woman) being in the group" (Long, 1997, p. 242-243). The probabilities of counts other than 0 are adjusted in a similar way.

## Results

Data are available for 1995 for U.S. (nonHispanic) white and Mexican-origin women, gathered in Cycle 5 of the National Survey of Family Growth (National Center for Health Statistics, 1995). The data are based on personal interviews conducted in the homes of a national sample of 10,847 females between the ages of 14 and 44 in the civilian, non-institutionalized population in the United States. Table 1 reports the descriptive data on children born (CEB) for U.S. white and Mexican-origin women in 1995.

White women have a mean CEB of 1.2 with a variance of 1.6. Mean CEB for Mexicanorigin women is 1.9 with a variance of 2.8 . For both white and Mexican-origin women, the variance of CEB is greater than the mean of CEB. There are several ways for determining if there is over-dispersion in the CEB data (see Poston, 2002). It turns out that there is not a significant amount of over-dispersion in the CEB data for whites, justifying the use of a Poisson regression model. There is a significant amount of over-dispersion in the CEB data for Mexican-origin women, so that a negative binomial regression model will be appropriate.
Poisson Regression versus Zero-inflated Poisson Regression

A Poisson regression model is thus estimated for the white women that predicts their CEB with socioeconomic and location characteristics that have been shown in the demographic literature to be associated with fertility. The independent variables pertain to education, rural residence, poverty status, age, regional location, and religion. Some are measured as dummy variables and others as interval.

Table 1. Data for Children Ever Born: U.S. White and Mexican-Origin Women, Ages 15-49.

| Group | Mean | Standard Dev. Variance | No. of Cases |  |
| :--- | :--- | :--- | :--- | :--- |
| White | 1.2471 | 1.2839 | 1.6486 | 6,456 |
| Mexican | 1.8864 | 1.6592 | 2.7531 | 924 |

Source of Data: National Center for Health Statistics (1995).

They are the following: $X_{1}$ is the woman's education measured in years of school completed; $\mathrm{X}_{2}$ is a dummy variable indicating whether the woman lives in a rural area; $\mathrm{X}_{3}$ is a dummy variable indicating whether the woman is classified as being in poverty (poverty status is based on whether the woman's family income is below the national poverty threshold, adjusted for family size).

Continuing, $\mathrm{X}_{4}$ is the woman's age measured in years; $\mathrm{X}_{5}$ to $\mathrm{X}_{7}$ are three dummy variables representing the woman's region of residence, namely, $\mathrm{X}_{5}$ residence in the Midwest, $X_{6}$ residence in the South, and $X_{7}$ residence in the West; residence in the Northeast is the reference category; and $\mathrm{X}_{8}$ to $\mathrm{X}_{10}$ are three dummy variables reflecting the woman's religion, as follows: $\mathrm{X}_{8}$ indicates if the woman's religion is Protestant, $X_{9}$ if she is Catholic, and $\mathrm{X}_{10}$ if she has no religion, or religion is not specified; Jewish religion is the reference category. The first panel of Table 2 reports the results of the Poisson regression equation predicting CEB for U.S. white women in 1995.

According to the Poisson coefficients shown in the first panel of Table 2, four of the ten independent variables are significantly related with the CEB of white women. The higher the woman's education, the fewer her CEB; the older her age, the higher her CEB. If she is a rural resident or in poverty, she will have more children than urban residents or women not living in poverty. The geographic location and religion variables are not statistically significant.

Using the above Poisson regression results, the predicted probabilities of each white woman may be calculated for each count of CEB from 0 to 10 . The mean of the predicted probabilities at each count may then be determined, using this formula (Long \& Freese, 2001):

$$
\overline{\operatorname{Pr}}(y=m)=\frac{1}{N} \sum_{i=1}^{N} \hat{\operatorname{Pr}}\left(y_{i}=m \mid x_{i}\right)
$$

where $\mathrm{y}=\mathrm{m}=$ the count of children ever born, and $\mathrm{x}_{\mathrm{i}}$ are the above ten independent variables.

Table 2. Poisson Regression Model, and Zero-inflated Poisson Regression Model, U.S. White (non-Hispanic) Women, 1995.

| Independent Variable | Poisson Model |  | Zero-inflated Poisson Model |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | b | z | Logit |  | Poisson |  |
|  |  |  | b | z | b | z |
|  | Panel 1 |  | Panel 2 |  | Panel 3 |  |
| $\mathrm{X}_{1}$ Education | -. 070 | -15.45 | . 905 | 12.83 | -. 058 | 11.84 |
| $\mathrm{X}_{2}$ Rural Residence | . 111 | 3.87 | -. 336 | -1.55 | . 097 | 3.35 |
| $\mathrm{X}_{3}$ Poverty Status | . 377 | 10.48 | -. 482 | -2.06 | . 336 | 8.96 |
| $\mathrm{X}_{4}$ Age | . 076 | 48.41 | -. 781 | -14.79 | . 034 | 16.01 |
| $\mathrm{X}_{5}$ Midwest | . 045 | 1.37 | -. 714 | -2.63 | . 007 | . 22 |
| $\mathrm{X}_{6}$ South | -. 023 | -. 07 | -. 289 | -1.06 | -. 048 | -1.39 |
| $\mathrm{X}_{7}$ West | . 019 | . 53 | -. 187 | -0.63 | . 011 | . 31 |
| $\mathrm{X}_{8}$ Protestant | . 074 | . 80 | -2.483 | -3.02 | -. 030 | -. 31 |
| $\mathrm{X}_{9}$ Catholic | . 052 | . 56 | -1.871 | -2.28 | -. 035 | -. 37 |
| $\mathrm{X}_{10}$ No Religion | -. 110 | -1.14 | -2.020 | -2.15 | -. 223 | -2.23 |
| Constant | -1.504 | -11.98 | 8.780 | 8.12 | . 096 | . 14 |
| Likelihood Ratio $\chi^{2}$ | 2857.59, $\mathrm{P}=0.000$ |  | $458.66, \mathrm{P}=0.000$ |  |  |  |

Vuong Test of Zip vs. Poisson: 20.16 P = 0.000

Figure 1 is a plot of the mean Poisson predicted probabilities at each count of CEB (the green $x$ symbols), and they may be compared with the observed empirical distribution of CEB (the blue circles). Just over 40 percent (proportion of .4046) of U.S. white women have no children ever born, but the Poisson regression results predict a mean probability at zero count of .361 , which is an under-prediction of the observed CEB. The Poisson regression results over-predict the observed CEB data at count one, under-predict at count two, and are more consistent with the observed CEB data at the third and higher counts.

But, a central issue for this paper is the under-prediction by the Poisson regression model of the observed zero counts of CEB for white women. In such a situation, it would be
appropriate to estimate a zero-inflated Poisson regression model. The $2^{\text {nd }}$ and $3^{\text {rd }}$ panels of Table 2 present the results of such a model. Recall from the previous section that the first two steps in estimating a zero-inflated model involve 1) using a logistic regression model to predict the woman's group membership in Group A (never have children) or Group ~A (may or may not have children), and 2) for women in Group ~A (may or may not have children), using a Poisson regression model to predict her number of children ever born. Thus there are two panels of zero-inflated Poisson results reported in Table 2. Panel 2, titled "Logit" are the logit coefficients obtained in Step 1, and Panel 3, titled "Poisson" are the Poisson coefficients obtained in Step 2.


Figure 1. Distributions of CEB, PRM, \& ZIP, U.S. White Women

The coefficients in the "Logit" panel (Panel 2) of Table 2 are the logit coefficients predicting a woman's membership in Group A (never having children). The higher her education the greater the likelihood of her not having children. If she is in poverty, she is likely to not have children. The older her age, the less likely she will not have children. If she lives in the Midwest, she will be less likely than women living in the Northeast to not have children. And if she is a Catholic, or a Protestant, or has no religion, she will be less likely than Jewish women to have no children. The rural, South, and West variables are not significant.

For the purpose of this paper, the more relevant coefficients are shown in the "Poisson" panel (Panel 3) of Table 2; these are the zeroinflated Poisson coefficients predicting the woman's CEB. The higher her education, the less number of children shill will have. If she is a rural resident, or in poverty, she will have more children. The older her age, the more the children. If she has no religion, she will have fewer children than Jewish women. The other variables are not significant.

A relevant comparison is between the zero-inflated Poisson coefficients (Panel 3) and
the Poisson coefficients (Panel 1). Note first that the Poisson coefficients (Panel 1) for most of the independent variables are slightly larger than those for the zero-inflated Poisson coefficients (Panel 3). However, the z-scores for many of the Poisson coefficients are quite a bit larger than the z -scores for the corresponding zero-inflated Poisson coefficients. Thus although the two sets of Poisson coefficients are not too different in magnitude, the standard errors for the zeroinflated coefficients will tend to be larger than they are for the Poisson coefficients.

Regarding issues of interpretation and statistical inference, the results of the two Poisson models allow the investigator to conclude that the effects on a woman's CEB of her education, rural residence, poverty status and age are all statistically significant. However, the zero-inflated Poisson results, but not the basic Poisson results, also allow the investigator to conclude that the "no religion" variable has a statistically significant negative effect on CEB. Women who report no religion have fewer children than women in the reference (Jewish religion) category. This inference would not have been made using the results of the Poisson model (Panel 1).

Is the zero-inflated Poisson regression model (ZIP) statistically preferred over the basic Poisson model regression model (PRM)? There is a formal test statistic, the Vuong test (Vuong, 1989) that determines statistically whether the zero-inflated model is a significant improvement over the Poisson model (for details, see Long, 1997, p. 248; Long \& Freese, 2001, p. 261-262). The Vuong statistic is asymptotically normal; if its value is $>1.96$, the ZIP model is preferred over the PRM. If Vuong $<1.96$, the PRM is preferred. The Vuong test statistic is shown at the base of Table 2, Vuong $=20.16$. This is clear evidence that the zero-inflated Poisson regression results are preferred over the Poisson regression results.

Another way to judge whether ZIP is preferred over PRM is to ascertain if the results from a ZIP regression improve the prediction of the mean probability at count zero. Recall from the discussion of Figure 1 (above) that the Poisson regression results under-predicted the mean probability at count zero. Figure 1 also contains mean predicted probabilities at each count that are based on the results of the zeroinflated Poisson model.

The ZIP predictions are shown in Figure 1 as maroon diamonds. The ZIP model predicts a probability at count zero of .3988 , which is very close to the observed proportion of CEB at count zero of .4046. The expected probabilities from the ZIP model at counts 1, 2 and 3 are also closer to the corresponding observed CEB counts than are those predicted by the PRM. The ZIP results seem to do a much better job predicting the observed CEB at counts $0,1,2$, and 3 than do the results from the PRM.

Negative Binomial Regression versus Zeroinflated Negative Binomial Regression

In the above example predicting CEB for U.S. white women, it was first determined that there was not a significant amount of overdispersion in CEB, thus justifying modeling CEB with the PRM. But recall that for the U.S. Mexican-origin women, the variance of CEB was significantly greater than the mean of CEB. In such a case, the PRM is not appropriate. Instead a negative binomial regression model (NBRM) is preferred.

The first panel of Table 3 reports the results of a negative binomial regression (NBRM). The same independent variables are used in this regression, as were used in the regressions shown in Table 2, except that age is excluded and "no religion" is used as the reference religion category. Excluding age resulted in a better fit of the negative binomial model with the data. The "no religion" variable is removed from the equation and used as the reference category because the Jewish variable was removed altogether from the equation; only 2 of the 924 Mexican-origin women were Jewish, so there was insufficient variation in this variable. Thus, regression results are shown in Table 3 for eight independent variables.

The negative binomial regression coefficients in the first panel of Table 3 indicate that only two of the eight independent variables are significantly related with the CEB of Mexican-origin women. The higher the woman's education, the lower her fertility; and if she is living in poverty, she will have more children than women not in poverty. The other independent variables are not statistically significant.

There is a large number of zeros for the CEB of Mexican-origin women. Almost 27 percent of them have zero children. Although this is not quite as high as the level of zero parity among white women ( 40.4 percent of the white women have no children), a zero-inflated negative binomial regression model (ZINB) was estimated to see if model fit would be improved over that of the NBRM. Its results may be compared with those of the NBRM shown in the first panel of table 3.

Recall that zero-inflated models produce two sets of coefficients (see the discussion above). Thus, the coefficients in the "Logit" panel (Panel 2) of Table 3 are the logit coefficients predicting a Mexican-origin woman's membership in Group A (never having children). The higher her education the greater her likelihood of not having children. And if she is Catholic, she is less likely than women with no religion to have no children. The other independent variables are not significant.

Table 3. Negative Binomial Regression Model, and Zero-inflated Negative Binomial Regression Model, U.S. Mexican-origin Women, 1995


Vuong Test of Zero-inflated Negative Binomial versus Negative Binomial 70.67, $\mathrm{P}=0.000$

A comparison may be made between the zero-inflated negative binomial coefficients (Panel 3) and the negative binomial coefficients (Panel 1). Note first that the NBRM coefficients (Panel 1) for four of the independent variables are slightly larger than those for the ZINB coefficients (Panel 3). And the z-scores for four of the NBRM coefficients are larger than those for the corresponding ZINB coefficients. The results of the two models in Panels 1 and 3 allow the investigator to conclude that the effects on a woman's CEB of her education and poverty status are statistically significant. However, the zero-inflated negative binomial regression results, but not the negative binomial results, also allow the investigator to conclude that Mexican-origin women living in the West have more children than those living in the Northeast. The NBRM results did not allow this inference
to be made.
One may compare the ZINB regression results with the NBRM results to determine if one is statistically preferred over the other. The Vuong test statistic provided at the base of Table 3 has a value of 70.67. Clearly the zero-inflated negative binomial regression results are preferred over the basic negative binomial regression results.

## Conclusion

This article considered a situation that frequently occurs when modeling count variables, namely, that there is a preponderance of zero counts. The application addressed in this paper involved the estimation of Poisson regression models (PRM) and negative binomial regression models
(NBRM) to predict the average number of children ever born (CEB) to women in the U.S. This is a count variable, and in a low fertility society such as the U.S., it is skewed with a long right tail.

It was noted in this article that many U.S. women have no children, resulting in a very large percentage of zero counts. But two groups of women have no children; one group will have zero CEB because they have chosen to never have children; another group will have no children even though they are trying to do so. PRM and NBRM are best suited to predict CEB counts among women who are having, or trying to have, children. Thus these models end up under-predicting zero counts because strictly speaking they are not able to consider the zero counts of women who are not trying to have children. Zero-inflated Poisson (ZIP) and zeroinflated negative binomial (ZINB) models have been proposed to handle such situations.

Analyses conducted in this paper of the fertility of U.S. white and Mexican-origin women in 1995 demonstrated that the zeroinflated models performed better in many respects than the straightforward Poisson and negative binomial models. Not only were the coefficients in the ZIP and ZINB models different from those in the PRM and NBRM, it was also shown that errors of statistical inference, in terms of failing to include significant effects, would have been made had the investigator only relied on the results of the PRM and NBRM.

It would appear that the use of zeroinflated Poisson and negative binomial regression models are statistically appropriate for the modeling of fertility in low fertility populations. This is especially the case when there is a preponderance of women in the society with no children.

## References

Cameron, A. C., \& Trivedi, P. K. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. Journal of Applied Econometrics, 1, 29-53.

Cameron, A. C., \& Trivedi, P. K. (1998). Regression analysis of count data. Cambridge, U.K.: Cambridge University Press.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Stern School of Business, Department of Economics, Working Paper Number 94-10.

Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics, 34, 1-14.

Long, J. S. (1990). The origins of sex differences in science. Social Forces, 68, 12971315.

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks, California: Sage Publications.

Long, J. S., \& Freese, J. (2001). Regression models for categorical dependent variables using Stata. College Station, Texas: Stata Press.

National Center for Health Statistics (NCHS). (1995). National Survey of Family Growth, Cycle V. Hyattsville, Maryland: Department of Health and Human Services.

Poston, D. L., Jr. (1976). Characteristics of voluntarily and involuntarily childless wives. Social Biology, 23, 198-209

Poston, D. L., Jr. (2002). The statistical modeling of the fertility of Chinese women. Journal of Applied Statistical Methods, 1(2), 387-396.

Poston, D. L., Jr., \& Kramer, K. B. (1983). Voluntary and involuntary childlessness in the United States, 1955-1973. Social Biology, 30, 290-306.

StataCorp. (2001). Stata statistical software: Release 7.0. Vol. IV, College Station, Texas: Stata Corporation, 481.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57: 307-333

# Variable Selection for Poisson Regression Model 

Felix Famoye<br>Department of Mathematics<br>Central Michigan University

Daniel E. Rothe<br>Alpena Community College

Poisson regression is useful in modeling count data. In a study with many independent variables, it is desirable to reduce the number of variables while maintaining a model that is useful for prediction. This article presents a variable selection technique for Poisson regression models. The data used is log-linear, but the methods could be adapted to other relationships. The model parameters are estimated by the method of maximum likelihood. The use of measures of goodness-of-fit to select appropriate variables is discussed. A forward selection algorithm is presented and illustrated on a numerical data set. This algorithm performs as well if not better than the method of transformation proposed by Nordberg (1982).

Key words: Transformation, goodness-of-fit, forward selection, R-square

## Introduction

Regression models using count data have a wide range of applications in engineering, medicine, and social sciences. Other forms of regression such as logistic regression are well established in various social science and medical fields. For example, in epidemiology, researchers study the relationship between the chance of occurrence of a disease and various suspected risk factors. However, when the outcomes are counts, Signorini (1991) and others point out that Poisson regression gives adequate results.

The social sciences often perform studies that involve count data. Sociology, psychology, demography, and economics all perform studies using the type of data that can make use of the Poisson regression model. Sociology applications involve situations where researchers wish to predict an individual's

Felix Famoye, Department of Mathematics, Central Michigan University, Mt. Pleasant, MI, 48859. E-mail: felix.famoye@cmich.edu. Daniel E. Rothe, Alpena Community College, 666 Johnson Street, Alpena, MI, 49707. Email: rothed@alpenacc.edu. The first author acknowledges the support received from Central Michigan University FRCE Committee under Grant \#48136.
behavior based on a particular group of observed characteristics and experiences. D'Unger et al. (1998) examined categories of criminal careers using Poisson latent class regression models. They assert that Poisson regression models are appropriate for modeling delinquent behavior and criminal careers.

Gourieroux et al. (1984) and Cameron and Trivedi (1986) described the use of Poisson regression in economics applications such as the daily number of oil tankers' arrivals in a port, the number of accidents at work by factory, the number of purchases per period, the number of spells of unemployment, the number of strikes in a month, or the number of patents applied for and received by firms. Gourieroux et al. (1984) concluded that the use of Poisson regression model is justified in a situation where the dependent variable consists of counts of the occurrence of an event during a fixed time period.

Christiansen and Morris (1997) listed applications of Poisson regression in a variety of fields. Poisson regression has been used in literary analysis of Shakespeare's works and the Federalist Papers, Efron and Thisted (1976). Home run data has been analyzed using these types of methods, Albert (1992). Poisson regression and count data in general are very important in a wide range of fields and thus deserve special attention. Often these models
involve many independent variables. Hence there is a need to consider variable selection for the Poisson regression model.

Variable selection techniques are well known for linear regression. See for example Efroymson (1960). Beale (1970) summarizes the various familiar methods: forward, backward, stepwise, and several other methods. Krall et al. (1975) discussed a forward selection technique for exponential survival data. They used the likelihood ratio as the criterion for adding significant variables. Greenberg et al. (1974) discussed a backward selection and use a log likelihood ratio step-down procedure for elimination of variables. For other nonlinear regressions and Poisson regression in particular, little is available in the literature.

Nordberg (1982) considered a certain data transformation in order to change the variable selection problem for a general linear model including the Poisson regression model into a variable selection problem in an ordinary unweighted linear regression model. Thus, ordinary linear regression variable selection software can be used.

In this article, we provide the Poisson regression model and describe some goodness-of-fit statistics. These statistics will be used as selection criteria for the variable selection method. A variable selection algorithm is described. We present the results of a simulation study to compare the variable selection algorithm with the method suggested by Nordberg (1982). The algorithm is illustrated with a numerical example and it is compared with the method suggested by Nordberg. Finally, we give some concluding remarks.

Poisson Regression Model and Goodness-of-fit Measures

The Poisson regression model assumes the response variable $y_{i}$, which is a count, has a Poisson distribution given by
$P\left(y_{i} ; \mu_{i}\right)=\frac{\mu^{y_{i}} e^{-\mu_{i}}}{y_{i}!}, y_{i}=0,1,2, \ldots$
$x_{i j}\left(j=0,1, \ldots, k\right.$ and $\left.x_{i 0}=1\right)$ are independent variables, and $\beta_{j}(j=0,1,2, \ldots, k)$ are regression parameters. The mean and variance of $y_{i}$ are equal and this is given by

$$
\begin{align*}
& \mathrm{E}\left(y_{i} \mid x_{i j}\right)=\mathrm{V}\left(y_{i} \mid x_{i j}\right)=\mu_{i},  \tag{2}\\
& i=1,2, \ldots, n, \text { and } j=0,1, \ldots, k .
\end{align*}
$$

Throughout this article, a log linear relationship $\mu_{i}=\exp \left(\sum_{j=0}^{k} \beta_{j} x_{i j}\right) \quad$ will be considered. However, the results can be modified to accommodate other types of relationships. Frome et al. (1973) described the use of the maximum likelihood (ML) method to estimate the unknown parameters for the Poisson regression model.

Several measures of goodness-of-fit for the Poisson regression model have been proposed in the literature. The Akaike information criterion (AIC) is a commonly used measure (Akaike, 1973). It is defined as

$$
\begin{equation*}
A I C=-\log L+(k+1) \tag{3}
\end{equation*}
$$

where $k+1$ is the number of estimated parameters and $L$ is the likelihood function. The smaller the value of the AIC statistic, the better the fit of the model. The log likelihood could be used as a measure of goodness-of-fit. However, the AIC criterion also includes $k$ as an adjustment for the number of independent variables, so that a model with many variables included is not necessarily better using this statistic.

Merkle and Zimmermann
(1992)
suggested some measures similar to the $R^{2}$ statistic for linear regression. They define

$$
\begin{equation*}
R_{D}^{2}=\frac{l\left(\hat{\mu}_{i}\right)-l(\bar{y})}{l\left(y_{i}\right)-l(\bar{y})} \tag{4}
\end{equation*}
$$

where

$$
\begin{aligned}
& l\left(\hat{\mu}_{i}\right)=\sum_{i=1}^{n}\left(y_{i} \log \hat{\mu}_{i}-\hat{\mu}_{i}-\log y!\right), \\
& l(\bar{y})=\sum_{i=1}^{n}\left(y_{i} \log \bar{y}-\bar{y}-\log y!\right)
\end{aligned}
$$

and

$$
l\left(y_{i}\right)=\sum_{i=1}^{n}\left(y_{i} \log y_{i}-y_{i}-\log y!\right) .
$$

The quantity $R_{D}^{2}$ measures the goodness-of-fit by relating the explained increase in the log-likelihood to the maximum increase possible. The interpretation is that higher $R_{D}^{2}$ indicates a better fit from the model. The numerator of $R_{D}^{2}$ is the deviance statistic. Cameron and Windmeijer (1996) analyzed Rsquared measures for count data. They establish five criteria for judging various $R^{2}$ measures. Among all $R^{2}$ measures considered, only the $R_{D}^{2}$ defined by Merkle and Zimmermann (1992) satisfies all the five criteria.

## Selection Criteria Statistics

Variable selection procedures need criteria for adding significant variables. We propose two selection criteria statistics (SCS). The first SCS is the Akaike information criterion (AIC) defined earlier. The smaller the value of the AIC statistic, the better the fit of the model.

The second SCS is a modification of the $R_{D}^{2}$ suggested by Cameron and Windmeijer (1996) by taking the number of parameters into account. We define $R_{a d j}^{2}$ as

$$
\begin{equation*}
R_{a d j}^{2}=\frac{\sum_{i=1}^{n}\left[y_{i} \log \left(\hat{\mu}_{i} / \bar{y}\right)-\left(\hat{\mu}_{i}-\bar{y}\right)\right]}{\sum_{i=1}^{n} y_{i} \log \left(y_{i} / \bar{y}\right)} \cdot \frac{(n-1)}{(n-k-1)} \tag{5}
\end{equation*}
$$

where $n$ is the sample size and $k$ is the number of independent variables. Either of the selection criteria statistics in (3) and (5) can be used to
determine which variable to add in the selection procedure. These variable selection criteria measures are adjusted to include the number of parameters. In this way, an additional variable being added to the model may not necessarily result in an improvement to the measure.

Selection Algorithm
The transformation suggested by Nordberg (1982) for log-linear Poisson regression model takes the form
$u_{i j}=x_{i j} \sqrt{\hat{\mu}_{i}}$, where $j=0,1,2, \ldots k$, and $i=1$, $2, \ldots n$

$$
\begin{equation*}
z_{i}=\left(\frac{y_{i}-\hat{\mu}_{i}}{\sqrt{\hat{\mu}_{i}}}\right)+\sum_{j=0}^{k} \hat{\beta}_{j} u_{i j} \tag{6}
\end{equation*}
$$

where $\hat{\mu}_{i}$ 's are the estimates of the predicted values from the full Poisson regression model. The variable selection procedure is as follows. Compute the ML estimate of $\beta$ in the full Poisson regression model. Transform the data using (6) and (7). Perform variable selection on the linear model with $z_{i}$ as the dependent variable and $u_{i j}$ as the independent variables. Identify the subset of the $u_{i j}$ variables that is selected and choose the corresponding $x_{i j}$ variables. This gives the Poisson regression submodel. Compute the maximum likelihood estimate for the Poisson regression on the chosen $x_{i j}$ variables. This gives the final result of variable selection through transformation.

Nordberg (1982) indicated that the success of this technique depends on the accuracy of the approximation of the loglikelihood function given by

$$
\begin{equation*}
\log L(\underline{\beta}) \approx \log L(\underline{\hat{\beta}})-(Q(\underline{\beta})-Q(\underline{\hat{\beta}})) / 2, \tag{8}
\end{equation*}
$$

where $Q(\beta)$ is given by

$$
Q(\beta)=\sum_{i=1}^{n}\left(z_{i}-\sum_{j=0}^{k} \beta_{j} u_{i j}\right)^{2} .
$$

The error in (8) is given by

$$
\begin{equation*}
\mathrm{E}=\frac{1}{6} \sum_{i=1}^{n} \frac{1}{\sqrt{\hat{\mu}_{i}}}\left(\sum_{j=0}^{k}\left(\underline{\beta}_{j}-\underline{\hat{\beta}}_{j}\right) u_{i j}\right)^{3} . \tag{9}
\end{equation*}
$$

Nordberg (1982) concludes that the approximation is adequate even when $30 \%$ of the $\hat{\mu}_{i}$ are less than or equal to 4 . However it is not clear what would happen to a case with say $70 \%$ of the $\hat{\mu}_{i}$ are less than or equal to 4 . We note here that Nordberg did not run simulations on such cases.

## Forward Selection Algorithm

The forward selection program begins by finding all possible regression models with one variable. The one with the best selection criteria statistic is chosen as the best one variable model. Once the best one variable model has been chosen, all models with the first variable and one additional variable are calculated and the one with the best selection criteria statistic is chosen. In this way, a two variable model is chosen. The process continues to add variables until the asymptotically normal Wald type " $t$ "-value for an added variable is not significant. The process then stops and returns the previous acceptable model.

The selection criteria statistics (SCS) and a test of significance of each variable are used to determine which variable to enter. The following is the algorithm:
[Initialize: $k=$ number of independent variables, $\alpha=$ significance level]

1. $\quad v \leftarrow 1$
2. Fit $k$ Poisson regression models with the intercept and $\nu$ independent variable
3. Select the model with the optimal SCS. Let $x_{i}$ be the independent variable chosen and $\beta_{i}$ be its parameter.
4. If the asymptotically normal Wald type " $t$ "-value associated with $\beta_{i}$ is significant at level $\alpha$,

- Retain Poisson regression model with independent variable $x_{i}$ and go to 5 . else
- Return "No variables are significant" and Stop.

5. Do while $(k \geq 2)$

- $v \leftarrow v+1$
- $k \leftarrow k-1$
- Fit $k$ Poisson regression models each with the intercept and $v$ independent variables. [The model includes all previously selected $x_{i}$ 's and one new $\left.x_{j}, j=1,2,3, \ldots k\right]$
- Select the model with the optimal SCS. Let $x_{\text {new }}$ be the independent variable added and $\beta_{\text {new }}$ be its parameter.
- If the asymptotically normal Wald type " $t$ "-value for $\beta_{\text {new }}$ is not significant at level $\alpha$,
o $\quad v \leftarrow v-1$
o go to 6 , else
0 add $x_{\text {new }}$ to the Poisson regression model
o Continue

6. The forward selection selects $\nu$ independent variables. Deliver the parameter estimates, $t$ values, and goodness-of-fit statistics for the selected model.

## Simulation Study

In order to compare the proposed method with the method proposed by Nordberg (1982), we conduct a simulation study. The Poisson regression model in (1) is generated and both methods were used for variable selection.

We generated a set of $x$-data consisting of $n(n=100,250,500$, and 1000) observations on eight explanatory variables $x_{i j}, i=1,2, \ldots, n$ and $j=0,1,2, \ldots, 7$, where $x_{i 0}=1$ (a constant term). The variables $x_{i l}, x_{i 2}, \ldots, x_{i 7}$ were generated as uncorrelated standard normal variates. All simulations were done using computer programs written in Fortran codes and the Institute of Mathematical Statistics Library (IMSL) is used.

The parameter vector $\underline{\beta}=\left(\beta_{0}, \beta_{1}, \beta_{2}, \ldots\right.$, $\beta_{7}$ ) used in the simulation study is chosen in such a way that $\beta_{5}=\beta_{6}=\beta_{7}=0$, while $\beta_{0}, \beta_{1}, \beta_{2}$, $\beta_{3}$, and $\beta_{4}$ are non-zero. For all simulations, we
chose $\beta_{1}=\beta_{2}=\beta_{3}=\beta_{4}=0.2$ and six different values of $\beta_{0}$. We consider the six values $\beta_{0}=-$ $1.0,-0.5,1.5,1.7,2.0$, and 3.0. These values were chosen so that certain percentages of fitted values $\hat{\mu}_{i}$ will be less than or equal to 4.0 . When $\beta_{0}=-1.0$ or -0.5 , all fitted values $\hat{\mu}_{i}$ from the Poisson regression model are less than or equal to 4.0 . For $\beta_{0}=1.5$, about $40 \%$ of the fitted values $\hat{\mu}_{i}$ are less than or equal to 4.0. When $\beta_{0}$ $=1.7$, about $20 \%$ of the fitted values $\hat{\mu}_{i}$ are less than or equal to 4.0 , and for $\beta_{0}=3.0$, almost all fitted values $\hat{\mu}_{i}$ exceed 4.0.

Using the $\beta$-vector and $x_{i 0}, x_{i l}, x_{i 2}, \ldots, x_{i 7}$ as explanatory variables, the observations $y_{i}, i=$ $1,2, \ldots, n$, were generated from the Poisson regression model in (1). Thus, the $y$-variates are Poisson distributed with mean

$$
\mu_{i}=\exp \left(\sum_{j=0}^{7} \beta_{j} x_{i j}\right) .
$$

The Nordberg method is used to perform variable selection on each set of data generated. The forward selection algorithm developed in this article is also used for variable selection. The result from using AIC selection criterion is presented in this article. The result from using $R_{a d j}^{2}$ selection criterion is the same as that of using AIC, and hence the result is not given.

Each simulation was repeated 1000 times by generating new $y$-variates keeping the $x$-data and the $\beta_{1}, \beta_{2}, \ldots, \beta_{7}$ constant. Since the parameters $\beta_{5}=\beta_{6}=\beta_{7}=0$, we expect $x_{5}, x_{6}$, and
$x_{7}$ not to enter into the selected model. Whenever any or all of these three variables enter a selected model, it is considered an error. The error rate from the 1000 simulations was recorded in Table 1 for both selection methods. In each simulation, the percentage of fitted values $\hat{\mu}_{i}$ less than or equal to 4 is recorded. These percentage values are averaged over the 1000 simulations and the results are presented in Table 1.

From Table 1, we notice some differences between the error rates from the forward selection method and the transformation method proposed by Nordberg. In general, the error rates from the forward selection method are smaller than the error rates from the Nordberg method. The error rates are much larger when the sample size is small, say $n=100$ or $n=250$. As the sample size increases to $n=$ 500 or $n=1000$, the two methods are closer in performance. However, the forward selection method seems to have a slight advantage over the Nordberg method. When the percentage of the fitted values $\hat{\mu}_{i}$ less than or equal to 4.0 is high, the error rates from the Nordberg method seem to be high, especially when the sample size $n$ is small.

From the simulation study, the difference between the two selection methods is not only due to whether the percentage of fitted values $\hat{\mu}_{i}$ less than or equal to 4.0 is high, it also depends on the sample size $n$. For small sample size, the Nordberg method tends to select variables $x_{5}, x_{6}$, and/or $x_{7}$ more often than the forward selection algorithm presented earlier. As the sample size increases to 1000 , the Nordberg method tends to perform as well as the forward selection algorithm.

Table 1. Error Rates For Nordberg And Forward Selection Algorithms.

| $N$ | $\beta_{0}$ | Nordberg <br> Method | Forward <br> selection | Percentage of <br> $\hat{\mu}_{i} \leq 4.0$ |
| :---: | ---: | :---: | :---: | :---: |
|  | -1.0 | 0.188 | 0.166 | 100.0 |
| 100 | -0.5 | 0.189 | 0.170 | 100.0 |
|  | 1.5 | 0.164 | 0.140 | 42.3 |
|  | 1.7 | 0.159 | 0.129 | 24.1 |
|  | 2.0 | 0.145 | 0.129 | 6.9 |
|  | 3.0 | 0.147 | 0.130 | 0.0 |
| 250 | -1.0 | 0.191 | 0.168 | 100.0 |
|  | -0.5 | 0.171 | 0.155 | 100.0 |
|  | 1.5 | 0.155 | 0.136 | 43.0 |
|  | 1.7 | 0.155 | 0.142 | 24.0 |
|  | 2.0 | 0.155 | 0.143 | 7.8 |
|  | 3.0 | 0.152 | 0.142 | 0.4 |
|  | -1.0 | 0.159 | 0.151 | 100.0 |
|  | -0.5 | 0.147 | 0.139 | 100.0 |
|  | 1.5 | 0.133 | 0.136 | 41.6 |
|  | 1.7 | 0.139 | 0.139 | 23.4 |
|  | 2.0 | 0.149 | 0.147 | 6.9 |
|  | 3.0 | 0.143 | 0.138 | 0.2 |
| 1000 | -1.0 | 0.144 | 0.144 | 100.0 |
|  | -0.5 | 0.154 | 0.146 | 100.0 |
|  | 1.5 | 0.162 | 0.160 | 38.9 |
|  | 1.7 | 0.153 | 0.147 | 20.5 |
|  | 2.0 | 0.159 | 0.156 | 5.7 |
|  | 3.0 | 0.148 | 0.145 | 0.1 |

Numerical Example
We applied the forward selection algorithm and the transformation method suggested by Nordberg (1982) to several data sets. The forward selection algorithm was implemented using AIC and $\mathrm{R}_{\text {adj }}^{2}$ as selection criteria statistics. When the percentage of the $\hat{\mu}_{i}$ less than or equal to 4 satisfied the cases considered by Nordberg (1982), both methods yielded the same sub-model. However, when the data has a much larger percentage of $\hat{\mu}_{i}$ less than or equal to 4 , we tend to obtain different results. We now present the results of a data set.

Wang and Famoye (1997) modeled fertility data using Poisson and generalized Poisson regression models. The data was from the Michigan Panel Study of Income Dynamics (PSID), a large national longitudinal data set. The particular portion of the data used in this paper was from 1989 and consisted of data from

2936 married women who were not head of households and with nonnegative total family income. The dependent variable was the number of children. Of the families, 1029 (35.05\%) had no children under age 17 . The response variable had a mean of 1.29 and a variance of 1.50 . The predicted values under full Poisson regression model were small with $54.26 \%$ less than or equal to 1 . Thus the data set was much more extreme than any of the cases considered by Nordberg (1982).

The Poisson regression model was fitted to the data using 12 covariates. The results are presented in Table 2. The forward selection algorithm was run on the data and the variables chosen are $x_{9}, x_{1}, x_{4}, x_{5}, x_{2}$, and $x_{10}$. The variables chosen are exactly the same variables that are significant in the full model. The transformation method proposed by Nordberg (1982) was applied to the data. The variables selected were

Table 2. Poisson Regression Model.

|  | Full Model |  | Forward Selection Sub-model |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Parameter | Estimate $\pm$ s.e. | t-value | Estimate $\pm$ s.e. | t-value | Step added |
| Intercept | $2.0686 \pm 0.1511$ | $13.69^{*}$ | $2.1226 \pm .0744$ | $28.5^{*}$ | -- |
| $x_{1}$ | $-0.2657 \pm 0.0356$ | $-7.46^{*}$ | $-0.2674 \pm .0351$ | $-7.61^{*}$ | 2 |
| $x_{2}$ | $-0.0193 \pm 0.0041$ | $-4.71^{*}$ | $0.0196 \pm .0041$ | $-4.82^{*}$ | 5 |
| $x_{3}$ | $-0.1226 \pm 0.0651$ | -1.88 | -- | -- | -- |
| $x_{4}$ | $-0.2811 \pm 0.0379$ | $-7.42^{*}$ | $-0.2629 \pm .0368$ | $-7.1^{*}$ | 3 |
| $x_{5}$ | $0.3057 \pm 0.0575$ | $5.32^{*}$ | $0.3002 \pm .0567$ | $5.29^{*}$ | 4 |
| $x_{6}$ | $-0.0050 \pm 0.0087$ | -0.57 | -- | -- | -- |
| $x_{7}$ | $0.0035 \pm 0.0071$ | 0.49 | -- | -- | -- |
| $x_{8}$ | $-0.0143 \pm 0.0187$ | -0.76 | -- | -- | -- |
| $x_{9}$ | $-0.0211 \pm 0.0038$ | $-5.55^{*}$ | $-0.0217 \pm .0038$ | $-5.76^{*}$ | 1 |
| $x_{10}$ | $-0.0147 \pm 0.0066$ | $-2.23^{*}$ | $-0.0132 \pm .0059$ | $-2.25^{*}$ | 6 |
| $x_{11}$ | $0.0118 \pm 0.0078$ | 1.51 | -- | -- | -- |
| $x_{12}$ | $-0.0545 \pm 0.0340$ | -1.60 | -- | -- | -- |

*Significant at 5\% level.
$x_{6}, x_{7}, x_{8}, x_{I I}, x_{9}$, and $x_{10}$. These are not the same variables chosen by the forward selection procedure. Only two of the variables are chosen by both methods. The results from the transformation method are shown in Table 3. The parameters corresponding to $x_{6}, x_{7}, x_{8}$, and $x_{11}$ are not significant in the full Poisson regression model (see Table 2), causing concerns about the accuracy of the transformation method.

Goodness-of-fit statistics for the models are provided in Table 4. The goodness-of-fit statistics for the forward selection sub-model are close to those for the full model even though the number of independent variables is now six. This is not the case for the transformation submodel. All these results are in support of the simulation study reported earlier.

Table 3. Nordberg's Transformation Method.

|  | Transformation sub-model |  |  |
| :---: | ---: | ---: | :---: |
| Parameter | Estimate $\pm$ s.e. | t-value | Step <br> added |
| Intercept | $1.8062 \pm .1484$ | $12.17^{*}$ | -- |
| $x_{6}$ | $-0.0140 \pm .0086$ | -1.63 | 1 |
| $x_{7}$ | $0.0075 \pm .0070$ | -1.07 | 2 |
| $x_{8}$ | $-0.0063 \pm .0186$ | -0.34 | 3 |
| $x_{9}$ | $-0.0376 \pm .0017$ | $-22.12^{*}$ | 5 |
| $x_{10}$ | $-0.0027 \pm .0008$ | $-3.04^{*}$ | 6 |
| $x_{11}$ | $0.0124 \pm .0077$ | 1.61 | 4 |

* Significant at 5\% level.

Table 4. Goodness-of-fit For The Poisson Model.

| Statistic | Full <br> Model | Forward <br> Selection | Nordberg |
| :--- | :---: | :---: | :---: |
| Deviance | 3277.84 | 3286.14 | 3430.81 |
| d.f. | 2923.0 | 2929.0 | 2929.0 |
| Log- <br> likelihood | -2410.09 | -2414.24 | -2486.58 |
| AIC | 2423.09 | 2421.24 | 2493.58 |
| $\mathrm{R}_{\text {adj }}^{2}$ | 0.2014 | 0.1994 | 0.1639 |

Conclusion
The size of the predicted values affected the usefulness of the transformation method. In the data set, the predicted values are relatively small ( $54.3 \%$ less than or equal to 1). Since the approximation error E in (9) for the transformation method involves division by the square root of the predicted value, one should be concerned when many predicted values are small. Dividing by small values may cause this error term to become large and make the approximation inaccurate. Although many other data sets analyzed indicate that the transformation method can be useful when the
predicted values are large, it may run into problems when predicted values are small. Real world data may not necessarily have large predicted values. It would be ideal to have an algorithm that is not dependent on the size of the predicted values. The forward selection method presented performed well regardless of the size of the predicted values.

The forward selection algorithm may take much more computer time than the transformation method proposed by Nordberg (1982). In these days of better computer technology, more computer time should not be a reason for using a method that may not always produce an adequate result. From our simulation study, the forward selection algorithm performs as well if not better than the transformation method.

In this article, a forward selection algorithm was developed. Similar methods could be developed using backward or stepwise selection for the class of generalized linear models. In addition, other selection criteria statistics could be used. Count data occur very frequently in real world applications. The size of the predicted values cannot be controlled within a particular study. Thus a selection method that can deal with any size of predicted values is desirable.

## References

Akaike, H. (1973). Information theory and an extension of the likelihood principle, in Proceedings of the Second International Symposium of Information Theory, Eds. Petrov, B.N., \& Csaki, F. Budapest: Akademiai Kiado.

Albert, J. (1992), A Bayesian analysis of a Poisson random-effects model for Home Run hitters, The American Statistician, 46, 246-253.

Beale, E. M. L. (1970). Note on procedures for variable selection in multiple regression, Technometrics, 12, 909-914.

Cameron, A. C., \& Trivedi, P. (1986). Econometric models based on count data: comparisons and applications of some estimators and tests, Journal of Applied Econometrics, 1, 29-53.

Cameron, A. C., \& Windmeijer, F. A. G. (1996). R-squared measures for count data regression models with applications to healthcare utilization, Journal of Business and Economic Statistics, 14, 209-220.

Christiansen, C. L., \& Morris, C. N. (1997). Hierarchical Poisson regression modeling, Journal of the American Statistical Association, 92, 618-632.

D’Unger, D. V., Land, K. C., McCall, P. L., \& Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses, American Journal of Sociology, 103, 1593-1630.

Efron, B., \& Thisted, R. (1976). Estimating the number of unseen species: how many words did Shakespeare know? Biometrika, 63, 435-447.

Efroymson, M. A. (1960). Multiple regression analysis, Mathematical Methods for Digital Computers. Eds. A. Ralston \& H.S.Wilf, New York: John Wiley \& Sons, Inc.

Frome, E. L., Kutner, M. H., \& Beauchamp, J. J. (1973). Regression analysis of Poisson-distributed data, Journal of the American Statistical Association, 68, 935-940.

Gourieroux, C., Monfort, A., \& Trognon, A. (1984). Pseudo maximum likelihood methods: applications to Poisson models, Econometrica, 52, 701-720.

Greenberg, R. A., Bayard, S., \& Byar, D. (1974). Selecting concomitant variables using a likelihood ratio step-down procedure and a method of testing goodness-of-fit in an exponential survival model, Biometrics, 30, 601608.

Institute for Social Research, The University of Michigan (1992). A Panel Study of Income Dynamics. Ann Arbor, MI.

Krall, J. M., Uthoff, V. A., \& Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival, Biometrics, 31, 49-57

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, The Canadian Journal of Statistics, 15, 209-225.

Merkle, L., \& Zimmermann, K. F. (1992). The demographics of labor turnover: a comparison of ordinal probit and censored count data models, Recherches Economiques de Louvain, 58, 283-307.

Nordberg, L. (1982). On variable selection in generalized linear and related regression models, Communications in Statistics Theory and Method, 11, 2427-2449.

Signorini, D. F. (1991). Sample size for Poisson regression, Biometrika, 78, 446-450.

Wang, W., \& Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression, Journal of Population Economics, 10, 273-283.

# Test Of Homogeneity For Umbrella Alternatives In Dose-Response Relationship For Poisson Variables 

Chengjie Xiong<br>Division of Biostatistics<br>Washington University in St. Louis

Yan Yan<br>Division of Biostatistics<br>Department of Surgery

Graduate School of Public Health
San Diego State University
Washington University in St. Louis


#### Abstract

This article concerns the testing and estimation of a dose-response effect in medical studies. We study the statistical test of homogeneity against umbrella alternatives in a sequence of Poisson distributions associated with an ordered dose variable. We propose a test similar to Cochran-Armitage's trend test and study the asymptotic null distribution and the power of the test. We also propose an estimator to the vertex point when the umbrella pattern is confirmed and study the performance of the estimator. A real data set pertaining to the number of visible revertant colonies associated with different doses of test agents in an in vitro mutagenicity assay is used to demonstrate the test and estimation process.


Key words: $C(\alpha)$ statistic, maximum likelihood estimate, monotone trend test, Poisson distribution, vertex point

## Introduction

Medical studies often evaluate treatment effects at several doses of a test drug. One usually assumes a priori, based either on past experience with the test drug or on theoretical considerations, that if there is an effect on a parameter of interest, the response is likely monotonic with dose, i.e., the effect of the drug is expected to increase or decrease monotonically with increasing dose levels. Comparing several doses with a placebo in a clinical dose study is then typically performed by one-sided many-to-one comparisons or trend tests assuming an order restriction. Monotonicity of dose-response relationship, however, is far from universal.

Instances may be found where a reversal or downturn at higher doses is likely to occur. For example, many therapies for humans become counterproductive at high doses.

Address correspondence to Chengjie Xiong, Division of Biostatistics, Campus Box 8067, Washington University in St. Louis, St. Louis, MO, 63110. Telephone: 314.362 .3635 . Email: chengjie@wubios.wustl.edu.

Similarly, in many in vitro mutagenicity assays, experimental organisms may succumb to toxic effects at high doses of the test agents, thereby reducing the number of organisms at risk of mutation and causing a downturn in the doseresponse curve (Collings et. al., 1981; Margolin et al., 1981). These types of non-monotonic dose-response behavior may not be caused by a random effect, but may occur due to an underlying biological mechanism. Mechanistic arguments for non-monotonic dose-response shapes can be found in many medical areas, such as toxicology (Calabrese \& Baldwin, 1998), carcinogenesis (Portier \& Ye, 1998), and epidemiology (Thorogood et al., 1993).

One of the simplest non-monotonic dose-response is the so-called umbrella pattern in which the response increases (decreases) until certain dose level (usually unknown) and then decreases (increases). Ames, McCann and Yamasaki (1975) reported experimental data exhibiting this pattern from three replicate Ames tests in which plates containing Salmonella bacteria of strain TA98 were exposed to various doses of Acid Red 114. The number of visible revertant colonies on each plate was observed. Figure 1 is a scatter plot of the number of visible revertant colonies against dose level, which
clearly indicates an umbrella pattern peaked between the third dose and the fourth dose. This same phenomenon is also observed and discussed by Simpson and Margolin (1986).

When the dose-response curve contains an umbrella pattern, the usual statistical trend tests become inadequate because of their power loss and inherent, and possibly erroneous decisions (Collings et al., 1981; Bretz \& Hothorn, 2001). The statistical test of homogeneity in response against an umbrella alternative has been studied by many authors. Most of these discussions deal with a continuous response variable and assume the normality for the associated distributions. The typical approaches under the assumption of normality are based on the framework of one-way analysis of variance and the simultaneous confidence intervals for umbrella contrasts of mean parameters (Bretz \& Hothorn, 2001; Rom et al., 1994; Shi, 1988; Marcus \& Genizi, 1994; Hayter \& Liu, 1999). Nonparametric approaches have also been considered by several authors (Lim \& Wolfe, 1997; Mack \& Wolfe, 1981 \& 1982, Simpson \& Margolin, 1994).

When data are based on counts such as those reported by Ames, McCann and Yamasaki (1975), however, a more reasonable distributional assumption might be the Poisson distribution. The statistical test of homogeneity against umbrella alternatives in a sequence of Poisson distributions associated with an ordered dose variable has not been addressed in the biostatistics literature to the best of our knowledge. This article studies this problem using an approach based on so-called $C(\alpha)$ statistics as proposed and studied by Neyman (1959) and Bailey (1956). The $C(\alpha)$ statistics are also discussed in more details by Moran (1970) and by Cox and Hinkley (1974) under the more general category of score statistics.

We propose a test similar to the Cochran-Armitage trend test and study the asymptotic null distribution and the power of our test. We also propose an estimator of the vertex point when the umbrella pattern is confirmed and study the performance of the estimator. A real data set reported by Ames, McCann and Yamasaki (1975) pertaining to the number of visible revertant colonies associated with
different doses of test agents in an in vitro mutagenicity assay is used to demonstrate the test and estimation process. We also present results of a simulation study about the proposed test and estimation.

## Methodology

We consider an experiment in which independent random samples are taken from $k$ distinct dose levels. Suppose that the $k$ dose levels are meaningfully ordered. Let $d_{1}, d_{2}, \ldots, d_{k}$ be the scores associated with these dose levels and $d_{1} \leq d_{2} \leq \ldots \leq d_{k}$. We assume that at dose level $i$, the response follows a Poisson distribution with mean $\mu_{i}, i=1,2, \ldots, k$.
Let $n_{i}$ be the sample size associated with dose level $i$ and $n=\sum_{i=1}^{k} n_{i}$. Let $x_{i}$ be the total response in the $i$-th dose level. For each $i$ and $p, \quad 1 \leq i, p \leq k$, let $d_{i}^{p}=\left(d_{i}-d_{p}\right)^{2}$ and $\bar{d}^{p}=\sum_{i=1}^{k} n_{i} d_{i}^{p} / n$. Suppose that the relationship between the mean response and the score takes the form of

$$
\mu_{i}=H\left\lfloor\alpha+\beta\left(d_{i}-d_{p}\right)^{2}\right\rfloor,
$$

where $H$ is a monotonic function that is twice differentiable, $d_{p}$ is the dose level associated with the vertex dose of the umbrella pattern. Notice that when $p=1$ or $k$, this formulation reduces to the monotone trend. We consider the problem of testing $H_{0}: \beta=0$ against the alternative hypothesis $H_{a}: \beta \neq 0$. The likelihood function as a function of $\alpha, \beta$, and $p$ is:

$$
\begin{aligned}
& L(\alpha, \beta, p) \propto \Pi_{i=1}^{k} \\
& \left.\exp ^{( }-n_{i} H\left[\alpha+\beta\left(d_{i}-d_{p}\right)^{2}\right]\right\}\left\{H\left[\alpha+\beta\left(d_{i}-d_{p}\right)^{2}\right]\right\}_{\}_{i}}^{x_{i}} .
\end{aligned}
$$

## p Is Known

When $p$ is given, the test is the same as the trend test based on the redefined dose score $d_{i}^{p}, i=1,2, \ldots, k$. The test is based on the $C(\alpha)$ statistic (Moran, 1970) and is obtained by evaluating the derivative of the loglikelihood with respect to $\beta$ at the maximum likelihood estimate of $\alpha$ under $H_{0}$ :

$$
C(\alpha)=\left.\frac{\partial \log L}{\partial \beta}\right|_{\hat{\alpha}, \beta=0}=\frac{H^{\prime}(\hat{\alpha})}{H(\hat{\alpha})}\left(\sum_{i=1}^{k} x_{i} d_{i}^{p}-\hat{x} \sum_{i=1}^{k} n_{i} d_{i}^{p}\right),
$$

where

$$
\hat{x}=\frac{\sum_{i=1}^{k} x_{i}}{n},
$$

and $\hat{\alpha}=H^{-1}(\hat{x})$. The test statistic for testing $H_{0}: \beta=0$ against the alternative hypothesis $H_{a}: \beta \neq 0$ is obtained after dividing $C(\alpha)$ by its asymptotic standard deviation under $H_{0}$ computed from the information matrix of $(\alpha, \beta)$ (Tarone, 1982):

$$
\begin{equation*}
X_{p}^{2}=\frac{\left[\sum_{i=1}^{k} x_{i} d_{i}^{p}-\hat{x} \sum_{i=1}^{k} n_{i} d_{i}^{p}\right]^{2}}{\hat{x} \sum_{i=1}^{k} n_{i}\left(d_{i}^{p}-\bar{d}^{p}\right)^{2}} \tag{1}
\end{equation*}
$$

Notice that this test statistic does not depend on the choice of the function $H$. Under the null hypothesis, $X_{p}^{2}$ has an asymptotic Chi-square distribution with one degree of freedom. Notice also that this test statistic is identical in formula to the test statistic for testing monotone trend with the redefined score in binomial proportions proposed by Armirage (1955). In addition, Tarone (1982) showed that, like the binomial trend test (Tarone \& Gart, 1980), this Poisson trend test is asymptotically locally optimal against any choice of smooth monotone function $H$ that satisfies

$$
\mu_{i}=H\left\lfloor\alpha+\beta\left(d_{i}-d_{p}\right)^{2}\right\rfloor, i=1,2, \ldots, k
$$

## p Is Unknown

When $p$ is unknown and $H_{0}: \beta=0$ is tested against the alternative hypothesis $H_{a}: \beta \neq 0$, we propose to reject $H_{0}: \beta=0$ when $\quad X^{2}=\max _{1 \leq p \leq k} X_{p}^{2} \quad$ is large. Let $\lambda_{i}=\lim _{n \rightarrow \infty} \frac{n_{i}}{n}$ and assume that $0<\lambda_{i}<1$ for $1 \leq i \leq k$. For $1 \leq p \leq k$, let $d^{p}=\sum_{i=1}^{k} \lambda_{i} d_{i}^{p}$ and $\mu=\sum_{i=1}^{k} \lambda_{i} \mu_{i}$. Let $\Delta$ be the $k$ by $k$ matrix whose ( $i, p$ ) entry is given by

$$
\Delta_{i}^{p}=\frac{d_{i}^{p}-d^{p}}{\sqrt{\mu \sum_{i=1}^{k} \lambda_{i}\left(d_{i}^{p}-d^{p}\right)^{2}}} .
$$

Let $A=\left(a_{i j}\right)$ be the $k$ by $k$ matrix such that $a_{i j}=0$ if $i \neq j$ and $a_{i j}=\mu_{i} \lambda_{i}$ if $i=j$. The following theorem gives the limiting distribution of the proposed test when the null hypothesis is true.

Theorem 1: If $H_{0}$ is true, then for any $x>0$,

$$
\begin{align*}
& \lim _{n \rightarrow \infty} P\left(X^{2} \geq x^{2}\right)= \\
& 1-\int_{-x}^{x} \int_{-x}^{x} \cdots \int_{-x}^{x} \frac{1}{\sqrt{(2 \pi)^{k}\left|\Delta^{\prime} A \Delta\right|}}  \tag{2}\\
& \exp \left[-\frac{X^{\prime}\left(\Delta^{\prime} A \Delta\right)^{-1} X}{2}\right] d x_{1} d x_{2} \ldots d x_{k},
\end{align*}
$$

where $X=\left(x_{1}, x_{2}, \ldots, x_{k}\right)^{\prime}$ and $|\mid$ is the matrix determinant.

The proof of Theorem 1 can be found in Appendix. Notice that the asymptotic null distribution does not depend on the unknown common mean $\mu_{1}=\mu_{2}=\ldots=\mu_{k}$ as the common mean $\mu$ is cancelled out in the integrand. Therefore, $\mu=1$ can always be assumed for the computation. The evaluation of
the integration can be done by the iterative algorithm proposed by Genz (1992). This algorithm begins with a Cholesky decomposition of the covariance matrix and then uses a simple Monte-Carlo algorithm. Another possible way of evaluating the distribution of the test statistic under the null hypothesis is through a large simulation of the test statistic. We point out that the asymptotic null distribution does depend on the unknown proportion $\lambda_{i}, i=1,2, \ldots, k . \frac{n_{i}}{n}$ can be used for $\lambda_{i}$ in the computation based on the consistency of $\frac{n_{i}}{n}$ to $\lambda_{i}$. In addition, according to Šidák and Zbynĕk (1967), under $H_{0}$,

$$
\operatorname{Pr}\left(X^{2} \leq x^{2}\right) \geq[2 \Phi(x)-1]^{k},
$$

where $\Phi$ is the distribution function of the standard normal distribution. Therefore, under $H_{0}$,

$$
\lim _{n \rightarrow \infty} \operatorname{Pr}\left(X^{2} \geq x^{2}\right) \leq 1-[2 \Phi(x)-1]^{k},
$$

which then provides a conservative test of $H_{0}$ against $H_{a}$.

## Estimation of the Vertex Point

If the alternative hypothesis is true, the problem of interest is then the estimation of the true vertex point. To avoid the problem of parameter identification, we assume that the umbrella pattern satisfies

$$
\mu_{1} \leq \mu_{2} \leq \ldots \leq \mu_{l-1}<\mu_{l}>\mu_{l+1} \geq \ldots \geq \mu_{k}
$$

or

$$
\mu_{1} \geq \mu_{2} \geq \ldots \geq \mu_{l-1}>\mu_{l}<\mu_{l+1} \leq \ldots \leq \mu_{k}
$$

i.e., we only consider the case where a single vertex point $l$ exists. Notice that this formulation does not rule out the possibility that the vertex point is on the boundary of the dose interval if a monotone trend is the alternative
hypothesis. We propose to estimate $l$ by $\hat{l}$ such that $X_{\hat{l}}^{2}=\max _{1 \leq p \leq k} X_{p}^{2}$, where $X_{p}^{2}$ is given by (1). Notice that as $n \rightarrow \infty$, for any $1 \leq p \leq k$,

$$
\begin{aligned}
\lim _{n \rightarrow \infty} \frac{X_{p}^{2}}{n} & =\frac{\left[\sum_{i=1}^{k} \lambda_{i} \mu_{i} d_{i}^{p}-\sum_{i=1}^{k} \lambda_{i} \mu_{i} \sum_{i=1}^{k} \lambda_{i} d_{i}^{p}\right]^{2}}{\sum_{i=1}^{k} \lambda_{i} \mu_{i} \sum_{i=1}^{k} \lambda_{i}\left(d_{i}^{p}-d^{p}\right)^{2}} \\
& =\frac{\sum_{i=1}^{k} \lambda_{i}\left(\mu_{i}-\mu\right)^{2}}{\sum_{i=1}^{k} \lambda_{i} \mu_{i}} R_{U^{p}, V}^{2}
\end{aligned}
$$

where $R_{U^{p}, V}^{2}$ is the correlation coefficient between random variables $U^{p}$ and $V$ defined on the sample space $\{1,2, \ldots, k\}$ with the multinomial probability distribution $\left\{\lambda_{1}, \lambda_{2}, \ldots, \lambda_{k}\right\}$, and $U^{p}(i)=d_{i}^{p}, V(i)=\mu_{i}$. Since

$$
\begin{aligned}
& -U^{p}(1) \leq-U^{p}(2) \leq \ldots \leq-U^{p}(p-1)<-U^{p}(p) \\
& =0>-U^{p}(p+1) \geq \ldots \geq-U^{p}(k)
\end{aligned}
$$

and either

$$
\mu_{1} \leq \mu_{2} \leq \ldots \leq \mu_{l-1}<\mu_{l}>\mu_{l+1} \geq \ldots \geq \mu_{k}
$$

or

$$
\mu_{1} \geq \mu_{2} \geq \ldots \geq \mu_{l-1}>\mu_{l}<\mu_{l+1} \leq \ldots \leq \mu_{k}
$$

holds, the proposed estimator to the true vertex point $l$ asymptotically maximizes the square of the correlation coefficient between $U^{p}$ and $V$ over $p=1,2, \ldots, k$.

A Real Example
In in vitro mutagenicity assays, experimental organisms may succumb to toxic effects at high doses of test agents, thereby reducing the number of organisms at risk of
mutation and causing a downturn in the doseresponse curve (Collings et al., 1981; Margolin et al., 1981).

Ames, McCann and Yamasaki (1975) reported experimental data exhibiting this pattern from three replicate Ames tests in which plates containing Salmonella bacteria of strain TA98 were exposed to various doses of Acid Red 114. The number of visible revertant colonies on each plate was observed. We assume a Poisson distribution for the number of visible revertant colonies and test whether an umbrella pattern in the mean exists.

Figure 1 is a scatter plot of the number of visible revertant colonies against dose level, which clearly indicates an umbrella pattern peaked between the third dose and the fourth dose. The test statistic is

$$
\begin{aligned}
X^{2} & =\max (75.71,75.76,75.90,76.20,69.78,55.96) \\
& =76.20 .
\end{aligned}
$$

The conservative test gives a $p$-value less than 0.00001 , indicating a strong evidence that an umbrella pattern exists. Since $\max _{1 \leq p \leq 6} X_{p}^{2}$ is obtained when $p=4$, i.e., when dose $d_{4}=10000(\mu \mathrm{~g} / \mathrm{ml})$ of Acid Red 114 is used, the estimated peak dose is $d_{4}=10000$ $(\mu \mathrm{g} / \mathrm{ml})$.


Figure 1. Visible colonies count against dose

## Simulation Studies

To understand the performance of the proposed test and the estimator for the vertex point when the alternative hypothesis is true, we have carried out a simulation study to evaluate the statistical power of the proposed test and the probability that the vertex point estimator
correctly estimates the true vertex point for a set of selected parameters.

In our simulation, we assume that a total of five independent Poisson distributions associated with five different dose levels $d_{i}=i, i=0,1,2,3,4$. We also assume that the sample size of all 5 groups is the same, i.e., $n_{1}=n_{2}=n_{3}=n_{4}=n_{5}$. Theorem 1 is used to determine the $x^{2}$ which achieves the upper $5 \%$ percentile of the test statistic under the null hypothesis.

The empirical power of the proposed test is computed as the proportion of rejections of the null hypothesis over repeated independent tests with a selected set of umbrella patterns. The performance of the proposed estimator to the vertex point is assessed by computing the empirical probability that the proposed estimator correctly estimates the true vertex point.

Table 1 presents the empirical power of the test and the empirical probability of correct estimation of the vertex point for three different choices of the true umbrella pattern and various sample sizes. Each entry in Table 1 is computed from 10000 independent hypotheses tests and estimations. All the tests assume a significance level of $5 \%$.

The first column in Table 1 is the true mean vector $\left(\mu_{1}, \mu_{2}, \mu_{3}, \mu_{4}, \mu_{5}\right)$. Notice that these umbrella patterns are chosen so that each possible interior vertex point (i.e., $l=2,3,4$ ) within the boundary of the dose interval is considered. Because the monotone trend is included in the alternative hypothesis when the vertex point falls on the boundary of the dose interval, it is of interest to see how our proposed test performs in these alternatives.

This is relevant given the fact that, when an umbrella pattern is likely in the dose-response relationship, the traditional statistical monotone trend tests become inadequate because of their power loss and inherent, and possibly erroneous decisions (Collings et al., 1981; Bretz and Hothorn, 2001). We simulated the statistical power of the proposed test for detecting the monotone trend and compared that to the traditional trend test as discussed by Cochran (1954) and Tarone (1982).

Table 1: Empirical Power and Probability with an Interior Vertex Point.

| Umbrella Pattern | Sample Size Per Dose | Power (\%) | Correct Vertex Estimation (\%) |
| :---: | :---: | :---: | :---: |
| (2,2.5,3,2.5,1.5) | 10 | 51.98 | 68.81 |
|  | 20 | 84.66 | 77.90 |
|  | 30 | 96.56 | 82.85 |
|  | 40 | 99.22 | 87.18 |
|  | 50 | 99.79 | 89.71 |
|  | 80 | 100 | 93.59 |
| (1.5,2,2.5,3,2.5) | 10 | 53.31 | 47.84 |
|  | 20 | 85.97 | 57.63 |
|  | 30 | 96.54 | 62.92 |
|  | 40 | 99.34 | 66.64 |
|  | 50 | 99.86 | 69.15 |
|  | 80 | 100 | 74.32 |
| (2.5,3,2.5,2,1.5) | 10 | 53.23 | 46.85 |
|  | 20 | 85.47 | 58.08 |
|  | 30 | 96.60 | 63.88 |
|  | 40 | 99.24 | 66.64 |
|  | 50 | 99.79 | 68.66 |
|  | 80 | 100 | 74.09 |

Table 2 provides the empirical power and the comparison along with the empirical probability of the correct estimation of the vertex point. The second column in Table 2 is the empirical power based on our proposed test. The third column in Table 2 is the empirical power based on the test by Cochran (1954) and Tarone (1982). Because the vertex point for a monotone trend could be either $l=1$ or $l=5$, the

Another different type of alternative hypothesis is when a flat segment appears in the Poisson mean vector $\left(\mu_{1}, \mu_{2}, \mu_{3}, \mu_{4}, \mu_{5}\right)$.

Table 3 presents the empirical power and the empirical probability of the correct estimation of the vertex points for several different choices of such patterns. Since the vertex point in some of these situations is not unique, the empirical probability of the correct

## Table 2: Empirical Power and Probability with a Boundary Vertex Point.

| Umbrella Pattern | Sample Size <br> Per Dose | Power $^{1}(\%)$ | Power $^{2}(\%)$ | Correct Vertex <br> Estimation (\%) |
| :---: | :---: | :---: | :---: | :---: |
| $(1.5,1.8,2.0,2.3,2.5)$ | 10 | 33.69 | 41.85 | 52.86 |
|  | 20 | 62.45 | 70.45 | 68.05 |
|  | 30 | 80.55 | 86.65 | 78.25 |
|  | 40 | 90.20 | 94.18 | 84.46 |
|  | 50 | 95.58 | 97.45 | 88.14 |
|  | 80 | 99.71 | 99.89 | 94.91 |
|  | 10 | 22.81 | 28.27 | 44.66 |
|  | 20 | 40.99 | 49.29 | 57.09 |
|  | 30 | 56.80 | 65.14 | 66.32 |
|  | 40 | 70.90 | 78.63 | 73.20 |
|  | 50 | 80.05 | 86.83 | 78.59 |
|  | 80 | 94.85 | 97.14 | 88.04 |

${ }^{1}$ Proposed test, ${ }^{2}$ Cochran \& Tarone's test.
empirical probability of the correct estimation to the true vertex points reported in Table 2 refers to the proportion over repeated estimates that either $l=1$ or $l=5$ is correctly estimated. Each entry in Table 2 is also computed from 10000 independent hypotheses tests and estimations.
estimation reported in Table 3 refers to the proportion that one of the possible vertex points is correctly identified over 10000 independent estimates. Data simulations are done using the random number generating function RANPOI

Table 3: Empirical Power and Probability with a Flat Segment in the Pattern

| Umbrella Pattern | Sample Size Per Dose | Power (\%) | Correct Vertex <br> Estimation (\%) |
| :---: | :---: | :---: | :---: |
| (2.5,3.0,3.0,2.5,2.0) | 10 | 26.18 | 77.21 |
|  | 20 | 50.48 | 87.36 |
|  | 30 | 69.76 | 92.19 |
|  | 40 | 82.76 | 95.12 |
|  | 50 | 90.61 | 96.64 |
|  | 80 | 98.95 | 98.91 |
| (2.5,3.0,3.0,3.0,2.5) | 10 | 11.95 | 83.36 |
|  | 20 | 20.81 | 90.08 |
|  | 30 | 30.39 | 93.72 |
|  | 40 | 40.36 | 95.98 |
|  | 50 | 49.78 | 97.37 |
|  | 80 | 71.38 | 99.36 |
| (2.5,2.5,3.0,3.0,2.5) | 10 | 9.71 | 63.19 |
|  | 20 | 15.07 | 70.92 |
|  | 30 | 21.47 | 78.14 |
|  | 40 | 28.10 | 81.78 |
|  | 50 | 34.82 | 85.59 |
|  | 80 | 52.70 | 92.83 |

from Statistical Analysis System (SAS Institute, Inc. 1999).

## Conclusion

When an umbrella pattern is likely in the doseresponse relationship, the usual statistical trend tests become inadequate because of their power loss and inherent, and possibly erroneous decisions (Collings et al., 1981; Bretz \& Hothorn, 2001). We proposed in this paper a test of homogeneity against umbrella alternatives in a sequence of Poisson distributions associated with an ordered dose variable and studied the limiting null distribution and the statistical power.

We also proposed an estimator of the vertex point when the umbrella pattern is confirmed and studied the performance of the estimator. Although the simulation study verifies that the increase of the sample size always increases the statistical power of the test and the probability of the correct estimation to the vertex point, Table 1 seems to indicate that for the selected set of parameters, the proposed estimator to the true vertex point performs better when the vertex point $(l=3)$ is in the middle of the dose interval than when it is away from the middle of the dose interval ( $l=2,4$ ). The statistical power of the proposed test, however, seems to be very comparable wherever the interior vertex is.

Our proposed test not only detects an umbrella pattern effectively based on the simulation results in Table 1, but also possesses reasonable statistical power to detect a monotone trend which is a subset of the alternative hypothesis considered in this paper. In fact, the simulation in Table 2 shows that, although our proposed test does not have as much the statistical power for detecting the monotone trend as the trend test of Cochran (1954), the difference in power between these two tests is relatively small. This is especially promising given the fact that the trend test of Cochran (1954) is asymptotically locally optimal against any choice of smooth monotone function $H$ (Tarone, 1982).

On the other hand, the simulation results reported in Table 3 seem to indicate that the
statistical power of the proposed test deteriorates when a substantial flat segment exists in the mean vector of the Poisson distributions, although the proposed vertex estimator still shows a high probability of correctly identifying one of these multiple vertex points.

Like the similarity on the test statistic for testing a monotone trend between a sequence of binomial distributions and a sequence of Poisson distributions (Armitage 1955; Cochran 1954), the proposed test and estimation techniques can be readily extended to the situation for detecting an umbrella pattern in a sequence of binomial distributions.

## References

Ames, B. N., McCann, J., \& Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the Salmonella/Microsome Mutagenicity test. Mutation Research, 31, 347364.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. Biometrics, 11, 375-386.

Bailey, N. T. J. (1956). Significance tests for a variable chance of infection in chainbinomial theory, Biometrika, 43, 332-336.

Bretz, F., \& Hothorn, L. A. (2001). Testing dose-response relationships with a priori unknown, possibly nonmonotone shape. Journal of Biopharmaceutical Statistics, 11, 3, 193-207.

Calabrese, E. J., \& Baldwin, L. A. (1998). A general classification of U-shape dose-response relationships in Toxicology and their mechanistic foundations. Human and Experimental Toxicology, 17, 353-364.

Cochran, W. G. (1954). Some methods for strengthening the common $\chi^{2}$ tests. Biometrics, 10, 417-451.

Collings, B. J. Margolin, B. H., \& Oehlert, G. W. (1981). Analysis for binomial data, with application to the fluctuation tests for mutagenicity. Biometrics, 37, 775-794.

Cox, D. R., \& Hinkley, D. V. (1974). Theoretical Statistics, London: Chapman and Hall.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics 1, 141-149.

Hayter, A. J., \& Liu, W. (1999). A new test against an umbrella alternative and the associated simultaneous confidence intervals. Computational Statistics \& Data Analysis, 30, 393-401.

Lim, D. H., \& Wolfe, D. A. (1997). Nonparametric tests for comparing umbrella pattern treatment effects with a control in a randomized block design. Biometrics, 53, 410418.

Mack, G. A., \& Wolfe, D. A. (1981). Ksample rank tests for umbrella alternatives. Journal of the American Statistical Association 76, 175-181; Correction 1982, 53, 410-418.

Marcus, R., \& Genizi, A. (1994). Simultaneous confidence intervals for umbrella contrasts of normal means. Computational Statistics \& Data Analysis, 17, 393-407.

Margolin, B. H., Kaplan, N., \& Zeiger, E. (1981). Statistical analysis of the Ames Salmonella / Microsome test. Proceedings of the National Academy of Sciences of the United States of America, 78, 3779-3783.

Moran, P. A. P. (1970). On asymptotically optimal tests of composite hypotheses. Biometrika, 57, 47-55.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses, in Probability and Statistics. U. Grenander (Ed.). New York: John Wiley \& Sons, 213-234.

Portier, C. J., \& Ye, F. (1998). U-shaped dose-response curves for carcinogens. Human and Experimental Toxicology, 17, 705-707.

Rom, D. M., Costello, R. J., \& Connell, L. T. (1994). On closed test procedures for doseresponse analysis. Statistics in Medicine, 13, 1583-1596.

SAS Institute, Inc. (1999). SAS Language (Version 8), Cary, NC.

Shi, N. Z. (1988). A test of homogeneity for umbrella alternatives and tables of the level probabilities. Communications in Statistics-Theory and Methods, 17, 657-670.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of American Statistical Association, 62, 626-633.

Simpson, D. G., \& Margolin, B. H. (1986). Recursive nonparametric testing for dose-response relationship subject to downturns at high doses. Biometrika, 73, 3, 589-596.

Tarone, R. E. (1982). The use of historical control information in testing for a trend in Poisson means. Biometrics, 38, 457462.

Tarone R. E., \& Gart, J. J. (1980). On the robustness of combined tests for trends in proportions. Journal of American Statistical Association, 75, 110-116.

Thorogood, M., Mann, J., \& McPherson, K. (1993). Alcohol intake and the U-shaped curve: Do non-drinkers have a higher prevalence of cardiovascular-related disease? Journal of Public Health Medicine, 15, 61-68.

## Appendix

We give the proof of Theorem 1, which gives the null distribution of $X^{2}$. Let

$$
Y_{p}=\frac{\left[\sum_{i=1}^{k} \frac{x_{i} d_{i}^{p}}{n}-\hat{x} \sum_{i=1}^{k} \lambda_{i} d_{i}^{p}\right.}{\sqrt{\mu \sum_{i=1}^{k} \lambda_{i}\left(d_{i}^{p}-d^{p}\right)^{2}}}=\sum_{i=1}^{k} \frac{x_{i} \Delta_{i}^{p}}{n}
$$

where

$$
\begin{aligned}
d^{p} & =\sum_{i=1}^{k} \lambda_{i} d_{i}^{p} \\
\mu & =\sum_{i=1}^{k} \lambda_{i} \mu_{i} \\
\Delta_{i}^{p} & =\frac{d_{i}^{p}-\sum_{i=1}^{k} \lambda_{i} d_{i}^{p}}{\sqrt{\mu \sum_{i=1}^{k} \lambda_{i}\left(d_{i}^{p}-d^{p}\right)^{2}}}
\end{aligned}
$$

Let $\hat{X}_{i}=\frac{x_{i}}{n}$. Note that

$$
\left(\begin{array}{c}
Y_{1} \\
Y_{2} \\
\ldots \\
Y_{k}
\end{array}\right)=\Delta^{\prime}\left(\begin{array}{c}
\hat{X}_{1} \\
\hat{X}_{2} \\
\ldots \\
\hat{X}_{k}
\end{array}\right)
$$

where $\Delta=\left(\Delta_{i}^{p}\right)$ is the $k$ by $k$ matrix whose $(i, p)$ entry is $\Delta_{i}^{p}$. Since

$$
\sqrt{n}\left[\left(\begin{array}{c}
\hat{X}_{1} \\
\hat{X}_{2} \\
\ldots \\
\hat{X}_{k}
\end{array}\right)-\left(\begin{array}{c}
\lambda_{1} \mu_{1} \\
\lambda_{2} \mu_{2} \\
\ldots \\
\lambda_{k} \mu_{k}
\end{array}\right)\right] \rightarrow N(0, A)
$$

where $A=\left(a_{i j}\right)$ is the $k$ by $k$ matrix such that $a_{i j}=0$ if $i \neq j$ and $a_{i j}=\mu_{i} \lambda_{i}$ if $i=j$, and the limit is in distribution. Therefore,

$$
\sqrt{n}\left[\left(\begin{array}{c}
Y_{1} \\
Y_{2} \\
\ldots \\
Y_{k}
\end{array}\right)-\Delta^{\prime}\left(\begin{array}{c}
\lambda_{1} \mu_{1} \\
\lambda_{2} \mu_{2} \\
\ldots \\
\lambda_{k} \mu_{k}
\end{array}\right)\right] \rightarrow N\left(0, \Delta^{\prime} A \Delta\right)
$$

Theorem 1 follows from the fact that $\sqrt{n}\left(Y_{1}, Y_{2}, \ldots, Y_{k}\right)^{\prime}$ and $\left(X_{1}, X_{2}, \ldots, X_{k}\right)^{\prime}$ are stochastically equivalent under $H_{0}$.

# Type I Error Rates Of Four Methods For Analyzing Data Collected In A Groups vs Individuals Design 

Stephanie Wehry<br>University of North Florida

James Algina<br>University of Florida


#### Abstract

Using previous work on the Behrens-Fisher problem, two approximate degrees of freedom tests, that can be used when one treatment is individually administered and one is administered to groups, were developed. Type I error rates are presented for these tests, an additional approximate degrees of freedom test developed by Myers, Dicecco, and Lorch (1981), and a mixed model test. The results indicate that the test that best controls the Type I error rate depends on the number of groups in the group-administered treatment. The mixed model test should be avoided.


Key words: groups-versus-individuals design, approximate degrees of freedom tests, mixed models

## Introduction

When a groups-versus-individuals design is used to compare two treatments, one treatment is administered to $J$ groups of $n$ participants (for a total of $N_{G}$ such participants) and one treatment is individually administered to $N_{I}$ participants or the individual participants may be in a notreatment control group. For example, psychotherapy researchers investigating the efficacy of group therapy often use a wait-list control group (Burlingame, Kircher, and Taylor, 1994). The therapy is provided to participants in groups because the researchers believe group processes will enhance the effectiveness of the

Stephanie Wehry is Assistant Director for Research and Evaluation for The Florida Institute for Education. Her research interests are in evaluating early childhood education programs, applied statistics, and psychometrics. Address correspondence to Florida Institute for Education, University of North Florida University Center, 12000 Alumni Drive, Jacksonville, FL, 32224. Email: swehry@ unf.edu. James Algina is Professor of Educational Psychology at the University of Florida. His research interests are in applied statistics and psychometrics. Email: algina@ufl.edu.
therapy. Group processes do not affect the participants in the wait-list control group because they do not receive a treatment, much less meet in groups. According to Clarke (1998) the most common design in psychotherapy research involves the use of a randomly assigned control condition, which can feature a variety of no-treatment control schemes.

The groups-versus-individuals design is also used when the purpose is to compare the effectiveness of an active treatment delivered to groups to an active treatment delivered individually. For example Bates, Thompson, and Flanagan (1999) compared the effectiveness of a mood induction procedure administered to groups to the effectiveness of the same procedure administered to individuals. Boling and Robinson (1999) investigated the effects of study environment on a measure of knowledge following a distance-learning lecture. The three levels of study environment included a printed study guide accessed by individuals, an interactive multi-media study guide accessed by individuals, and a printed study guide accessed by cooperative study groups.

A possible model for the data collected in a groups-versus-individuals design consists of two submodels. For participants in the individually administered treatment the submodel is

$$
\begin{equation*}
Y_{i: T_{I}}=\mu_{I}+\varepsilon_{i: T_{I}} \tag{1}
\end{equation*}
$$

where $i: T_{I} \quad\left(i=1, \ldots, N_{I}\right)$ denotes the $i$ th participant within the individually-administered treatment. For participants in the groupadministered treatment

$$
\begin{equation*}
Y_{i: j: T_{G}}=\mu_{G}+\alpha_{j: T_{G}}+\varepsilon_{i: j: T_{G}} \tag{2}
\end{equation*}
$$

where $i: j: T_{G} \quad(i=1, \ldots, n)$ denotes the $i$ th participant within the $j$ th group $(j=1, \ldots, J)$ in the group-administered treatment. An important question is whether to treat the $\alpha_{j: T_{G}}$ as fixed or random. When the researcher views the groups in the group-administered treatment as representative of a larger number of groups, $\alpha_{j: T_{G}}$ should be treated as random. In the remainder of the paper we assume that the groups in the group-administered treatment comprise a random factor with the groups in the study representing an infinitely large number of groups.

Burlingame, Kircher, and Taylor (1994) reported that the independent samples $t$ test, ANOVA, and ANCOVA were the most commonly used methods for analyzing data in group psychotherapy research. It is well known that these procedures require the scores for individuals to be independently distributed both between and within treatments, an assumption that is likely to be violated for the participants in the group-administered treatment when $\alpha_{j: T_{G}}$ is random. It is also well known that these procedures are not robust to violations of the independence assumption (see, for example, Scheffe, 1958). When the groups-versusindividuals design is used, lack of independence is indicated by a non-zero intraclass correlation coefficient for the participants who receive the group-administered treatments. Myers, Dicecco, and Lorch (1981), using simulated data, showed that the Type I error rates for the independent samples $t$ test is above the nominal alpha level when the intraclass correlation is positive. Burlingame, Kircher, and Honts (1994) reported similar results. In passing we note that if the researcher believes it is appropriate to treat the
$\alpha_{j: T_{G}}$ as fixed, if both error terms are normally distributed, and if the error terms have equal variances, the treatments can be compared by using an independent samples ANOVA and testing the hypothesis

$$
\begin{equation*}
H_{0}: \mu_{I}=\mu_{G} \tag{3}
\end{equation*}
$$

but generalization of the results to additional groups is not warranted.

Myers et al. (1981) developed two statistical tests of the hypothesis given in equation (3). These tests take the lack of independence into account and allow generalization of the results to the population of groups represented by the groups in the groupadministered treatment. (In the following, groups will always refer to the groups in the group-administered treatment). One of these procedures used a quasi-F statistic and degrees of freedom approximated by the Satterthwaite (1941) method. Formulated as an approximate degrees of freedom (APDF) $t$ statistic, the Myers et al. test statistic is

$$
\begin{equation*}
t_{\text {APDF }}=\frac{\bar{Y}_{I}-\bar{Y}_{G}}{\sqrt{\frac{M S_{S / T_{I}}}{N_{I}}+\frac{M S_{G / T_{G}}}{N_{G}}}} \tag{3}
\end{equation*}
$$

where $\bar{Y}_{I}=\sum_{i=1}^{N_{I}} Y_{i: T_{I}} / N_{I}$ is the mean of the criterion scores and

$$
\begin{equation*}
M S_{S / T_{I}}=\frac{\sum_{i=1}^{N_{I}}\left(Y_{i: T_{I}}-\bar{Y}_{I}\right)^{2}}{N_{I}-1} \tag{5}
\end{equation*}
$$

is the variance for participants who received the individually administered treatment; $\bar{Y}_{G}=\sum_{j=1}^{J} \sum_{i=1}^{n} Y_{i: j: T_{G}} / N_{G}$ is the mean of the criterion scores for participants who received the group-administered treatment $\left(i: j: T_{G}\right)$ and

$$
\begin{equation*}
M S_{G / T_{G}}=\frac{\sum_{j=1}^{J} n\left(\bar{Y}_{j: T_{G}}-\bar{Y}_{G}\right)^{2}}{J-1} \tag{6}
\end{equation*}
$$

is the between-group mean square for these participants. It can be shown that the squared denominator of $t_{\text {APDF }}$ estimates the sampling variance of the numerator assuming a correct model for the data is given by equations (1) and (2) and $\alpha_{j: T_{G}}$ is random. Assuming that $\varepsilon_{i: T_{I}} \sim N\left(0, \sigma_{I}^{2}\right), \quad \alpha_{j: T_{G}} \sim N\left(0, \tau^{2}\right), \quad$ and $\varepsilon_{\mathrm{i}: \mathrm{j}: T_{G}} \sim N\left(0, \sigma_{G}^{2}\right)$, the estimated approximate degrees of freedom are

$$
\begin{equation*}
\hat{f}_{2}=\frac{\left(M S_{S / T_{I}} / N_{I}+M S_{G / T_{G}} / N_{G}\right)^{2}}{\frac{\binom{M S_{S / T_{I}}}{N_{I}}^{2}}{N_{I}-1}+\frac{\left(M S_{G / T_{G}} / N_{G}\right)^{2}}{J-1}} \tag{7}
\end{equation*}
$$

It should be noted that in using the Satterthwaite method, the distribution of the square of the denominator of $t_{A P D F}$ is approximated as a multiple of a chi-square distribution with degrees of freedom estimated by $\hat{f}_{2}$.

Based on simulated data, Myers et al. (1981) reported estimated Type I error rates for their APDF test, including results for $J=4$ and $J=8$ groups in the group-administered treatment. For both numbers of groups, estimated Type I error rates were very similar to the nominal level. While these results indicate that the APDF has adequate control of the Type I error rate when $J \geq 4$, it leaves open the question of how well the test works with a smaller number of groups and the discussion in Satterthwaite (1941) and results in Scariano and Davenport (1986) suggest the test may not control the Type I error rate for $J \leq 3$.

The discussion in Satterthwaite (1941) implies that the approximation of the square of the denominator of $t_{\text {APDF }}$ by a multiple of a chisquare distribution improves as $J-1$ or $N_{I}-1$ increases and as

$$
\begin{equation*}
\frac{\left(N_{I}-1\right)\left(n \tau^{2}+\sigma_{G}^{2}\right)}{(J-1) \sigma_{I}^{2}} \tag{8}
\end{equation*}
$$

becomes closer 1.0. When there are two groups in the group-administered treatment, $J-1$ is as small as it possibly can be. In addition, calculating the ratio in equation (4) for conditions in which $\sigma_{I}^{2}=\tau^{2}+\sigma_{G}^{2}$ shows that the ratio can be much larger than 1 . Therefore, the discussion in Satterthwaite would lead one to expect that the APDF $t$ test in Myers et al. (1981) would not work well when there are just two groups.

Scariano and Davenport (1986) studied Type I error rates for the APDF $t$ test that Welch (1938) proposed as a solution to the BehrensFisher problem:

$$
\begin{equation*}
t=\frac{\bar{Y}_{a}-\bar{Y}_{b}}{\sqrt{\frac{S_{a}^{2}}{N_{a}}+\frac{S_{b}^{2}}{N_{b}}}} . \tag{9}
\end{equation*}
$$

In $t, \bar{Y}_{a}$ and $\bar{Y}_{b}$ are means for two individually administered treatments, $S_{a}^{2}$ and $S_{b}^{2}$ are the sample variances, and the square of the denominator estimates the sampling variance of the numerator. The distribution of the Welch $t$ can be approximated by a $t$ distribution with degrees of freedom approximated the by the Satterthwaite (1941) method. Thus, the Myers et al. (1981) APDF test and the Welch APDF solution to the Behrens-Fisher problem are both based on the same theoretical approach to approximating the sampling distribution of the test statistic.

Scariano and Davenport (1986) developed an analytic procedure for calculating the Type I error rate of the Welch APDF test and showed its Type I error rate can be seriously inflated when (a) there is a negative relationship between the sampling variances of the means and the degrees of freedom for the estimated sampling variances and (b) the smaller of the two degrees of freedom is small. In the Myers et al. (1981) APDF test, the sampling variances of the means are $\left(n \tau^{2}+\sigma_{G}^{2}\right) / N_{G}$ and $\sigma_{I}^{2} / N_{I}$ and the degrees of freedom for estimates of these
variance are $J-1$ and $N_{I}-1$. When $N_{I}=N_{G}$ and $\sigma_{I}^{2}=\tau^{2}+\sigma_{G}^{2}$, for example, the relationship will be negative and, when $J \leq 3$, the degrees of freedom will be small. Consequently, the APDF test may not work well in these conditions. One purpose of the study is to study Type I error rates when $J$ is small.

Satterthwaite (1941) showed how to approximate the distribution of a sum of two chi-square distributed random variables by another chi-square distribution. He determined the degrees of freedom for the approximating distribution by equating the mean and variance of the sum with the mean and variance of the approximating chi-square distribution. Thus, the Satterthwaite approach is a two-moment approach to determining the degrees of freedom. Scariano and Davenport (1986) developed a four-moment approach and showed analytically that it provides a more conservative test than does the two-moment approach. In the fourmoment approach the estimated approximate degrees of freedom are

$$
\begin{equation*}
\hat{f}_{4}=\frac{\left\{\frac{u^{2}}{J-1}+\frac{1}{N_{I}-1}\right\}^{3}}{\left(\frac{u^{3}}{(J-1)^{2}}+\frac{1}{\left(N_{I}-1\right)^{3}}\right)^{2}} \tag{10}
\end{equation*}
$$

where, in the groups-versus individuals design,

$$
\begin{equation*}
u=\frac{M S_{G / T_{g}} / N_{G}}{M S_{S / T_{I}} / N_{I}} . \tag{11}
\end{equation*}
$$

A second purpose of the present study was to calculate the actual Type I error rate for the fourmoment approach.

In Scariano and Davenport (1986), the two-moment approach was sometimes liberal when the four-moment approach was conservative. As a result, they suggested using an average of the estimated degrees of freedom produced by the two approaches. Thus, a third purpose was to analytically evaluate the actual Type I error rate for this averaged degrees of freedom approach.

An alternative to the preceding approaches is based on a mixed model with a proper inference space (McLean, Sanders, \& Stroup, 1991) and Satterthwaite degrees of freedom. When the restricted maximum likelihood estimate (RMLE) of $\tau^{2}$ is larger than zero and there are an equal number of participants in the groups, the mixed model test is equivalent to the Myers et al. (1981) twomoment test. However, if the RMLE is zero, $M S_{G / T_{G}}$ and $M S_{S / G / T_{G}}$ are pooled and replace $M S_{G / T_{G}}$ in equation (1). This statistic, which is equivalent to the Welch $t$ test, is smaller than $t_{\text {APDF }}$ and may be more conservative than the two-moment test. However, it tends to have larger degrees of freedom, which may make it more liberal than the two-moment test.

When there are an equal number of participants in the groups, the RMLE of $\tau^{2}$ is zero when the method of moments estimate of $\tau^{2}$ is $\leq 0$ (McCulloch \& Searle, 2001). The probability that the method of moments estimate of $\tau^{2}$ is $\leq 0$ is

$$
\begin{align*}
& \operatorname{prob}\left\{M S_{G / T_{G}} \leq M S_{S / T_{\sigma} / T_{G}}\right\}  \tag{12}\\
& =\operatorname{prob}\left\{F[J-1, J(n-1)] \leq \frac{1-\rho_{I C C}}{\rho_{I C C}(n-1)+1}\right\}
\end{align*}
$$

where $\rho_{\text {ICC }}=\tau^{2} /\left(\tau^{2}+\sigma_{G}^{2}\right)$. Figure 1 displays the probability as a function of $J, \rho_{\text {ICC }}$, and $n$. The probability can be quite substantial and, in some conditions, we would expect the mixed model test to perform differently than the twomoment, four-moment, and averaged degrees of freedom tests. Thus, a fourth purpose of the study is to compare these tests to the mixed model test.

The research was carried out in two studies. In the first study, actual Type I error rates were calculated for the two-moment approach, the four-moment approach, and the averaged degrees of freedom approach. In the second study, simulated data were used to estimate the actual Type I error rate for the mixed model approach as well as for the twomoment approach, the four-moment approach, and the averaged degrees of freedom approach. Taken together, the purposes of the studies were


Figure 1. Probability of a Negative Estimate for $\tau^{2}$
to compare Type I error rates for the twomoment, four-moment, averaged degrees of freedom, and mixed model approaches when the number of groups in the group administered treatment is small and to study the influence of the number of groups, number of participants in a group, and intraclass correlation on the Type I error rates for these methods.

> Methodology

## Study 1

Actual Type I error rates were calculated for each condition in a 5 (Number of Groups) $\times 4$ (Intraclass Correlation) $\times 15$ (Number of Participants in a Group) completely crossed factorial design. The levels of the factors were $J=2$ to 6 for the number of groups; $n=3$ and 4 , and 6 to 30 in steps of 2 for the number of participants in a group; and $\rho_{\text {ICC }}=.00, .20, .40$, and .80 for the intraclass correlation. In all conditions, $\left(\tau^{2}+\sigma_{G}^{2}\right) / \sigma_{I}^{2}=1$ and, because the design was balanced across treatments, $N_{I}=J(n)$. For all calculations the nominal alpha level was .05 . In the following,
when we use the term Type I error rate without the actual or nominal modifier, we refer to the actual Type I error rate.

## Calculating Type I Error Rates

Scariano and Davenport (1976) developed a method to calculate Type I error rates for the Welch $t$ test. We applied their method, which we describe below, to the three APDF tests considered in this paper. It should be noted that although the method we applied was developed in the context of the BehrensFisher problem, that is, comparing means of independently distributed scores for two groups when the variance are not equal for the groups, we did not apply the method to the BehrensFisher problem. Rather we applied the method to comparison of means for two groups, when scores are not independently distributed within the sub-groups in the group-administered treatment. Thus, our work is not subject to Sawilowsky's (2002) criticisms of research on the Behrens-Fisher problem.

The Type I error rate for the APDF $t$ test is

$$
\begin{align*}
& \operatorname{Pr}\left[t_{\text {quasi }}^{2}>F_{\alpha, 1, \hat{f}}\right]= \\
& \int_{0}^{\infty} \operatorname{Pr}\left[t_{\text {quasi }}^{2}>F_{\alpha, 1, \hat{f}} \mid u\right] g(u) d u \tag{13}
\end{align*}
$$

where $\hat{f}$ is the two-moment, four-moment, or averaged degrees of freedom and $\alpha$ is the nominal Type I error rate. Cochran (1951) has shown that $t_{\text {quasi }}^{2}$ is the ratio of $Q$ to $C$ where

$$
\begin{align*}
& Q \sim F_{1, m_{1}+m_{2}}, m_{1}=J-1, m_{2}=N_{I}-1, \\
& C=\frac{(1+u)\left(m_{1}+m_{2}\right)}{(1+U)\left(\frac{m_{1} u}{U}+m_{2}\right)}, \tag{14}
\end{align*}
$$

and

$$
\begin{equation*}
U=\frac{\left(n \tau^{2}+\sigma_{G}^{2} / N_{G}\right)}{\sigma_{I}^{2} / N_{I}} \tag{15}
\end{equation*}
$$

To facilitate numerical integration the variable $u$ can be transformed to

$$
\begin{equation*}
s=\frac{\left(\frac{m_{1} u}{m_{2}}\right)}{\left(1+\frac{m_{1} u}{m_{2}}\right)} \tag{16}
\end{equation*}
$$

and the Type I error rate is found by numerically integrating

$$
\begin{equation*}
\int_{0}^{1} \operatorname{Pr}\left[Q>C \times F_{\alpha, 1, \hat{f}} \mid s\right] f(s) d s \tag{17}
\end{equation*}
$$

where

$$
\begin{equation*}
f(s)=\frac{\Gamma\left(\frac{m_{1}+m_{2}}{2}\right) U^{m_{2} / 2}}{\Gamma\left(\frac{m_{1}}{2}\right) \Gamma\left(\frac{m_{2}}{2}\right)} \frac{s^{\left(m_{1}-2\right) / 2}(1-s)^{\left(m_{2}-2\right) / 2}}{[U(1-s)+s]^{\left(m_{1}+m_{2} / 2\right)}} . \tag{18}
\end{equation*}
$$

Numerical integration was performed using the trapezoid rule. For $J=2$ a singularity occurs at $s=0$. Therefore, the limits of integration were
.0001 and 1. The interval was divided into 1000 segments of equal width. For $J=3$ a removable singularity occurs at $s=0$. For $J \geq 3$ the limits of integration were 0 and 1 and this interval was also divided into 1000 segments. As a check on the calculations, Type I error rates were estimated by using simulated data with 100,000 replications. The results from the simulation were consistent with the results determined by numerical integration.

## Results

Study 1
Figures 2 to 6 contain plots of the Type I error rates against size of groups. The five plots are for two, three, four, five, and six groups, respectively. Plots within a figure are organized by the intraclass correlation coefficient. Inspection of Figure 2 indicates that when there are two groups, the four-moment degrees of freedom should be used, except perhaps when $\rho_{\text {ICC }}=0$. Then the averaged degrees of freedom might be used. When there are three groups (see Figure 3), the averaged degrees of freedom might be used at the risk of a slightly liberal test when $\rho_{\text {ICC }}$ is at .20 or greater. The two-moment degrees of freedom results in a test that is too liberal and the four-moment degrees of freedom results in a test that is too conservative. When there are four groups (see Figure 4), the twomoment degrees of freedom provides a test that has a slight liberal tendency that increases as $\rho_{\text {ICC }}$ get larger and as the size of the groups get larger. Use of the averaged degrees of freedom provides a test that is slightly conservative when $\rho_{\text {ICC }}$ is small, but controls the Type I error rate well as it increases. Plots for five or more groups (see Figures 5 and 6) are similar to those for four groups. However, the use of either the two-moment degrees of freedom or and average degrees of freedom provide reasonable control of the Type I error rate. Use of the former can result in a slightly liberal test, whereas use of the latter can result in a slightly conservative test.

## Methodology

Study 2
As noted in the introduction, simulated data were used to compare the three APDF tests
and the mixed model test. The design had four factors: the four tests, the number of groups, size of the groups, and level of the intraclass correlation. There were five levels of the number of groups, $J=2,3,4,5$, and 6 ; five levels of group size, $n=4,8,12,16$, and 20 subjects nested in the groups; and seven levels of intraclass correlation, $\rho_{\text {ICC }}=.00$ to .30 in steps of 05 .

The simulation was carried out using the random number generation functions of SAS, Release 8.2. Scores for simulated participants in the individually administered treatment level were generated using the equation (1), where $\mu_{I}$ was arbitrarily set at 100 and the
$\varepsilon_{i \cdot T_{I}} \mathrm{~S}$ were pseudorandom standard normal deviates generated using RANNOR. Scores for simulated participants in the group-administered treatment level were generated using equation (2), where $\mu_{G}$ was arbitrarily set at $100, \alpha_{j: T_{G}}$ was a pseudorandom normal deviate with mean zero and variance $\tau^{2}$ and $\varepsilon_{i: j ; T_{G}}$ was a pseudorandom normal deviate with mean zero and variance $\sigma_{G}^{2}$. Each of the conditions was replicated 5,000 times and the Type I errors of the four tests were counted over the replications of each condition. The nominal type I error rate was .05 in all conditions.



Figure 2. Plots of Type I Error Rates by Size of Group for Two Groups


Figure 3. Plots of Type I Error Rates by Size of Group for Three Groups.


Figure 4. Plots of Type I Error Rates by Size of Group for Four Groups.


Figure 5. Plots of Type I Error Rates by Size of Group for Five Groups.


Figure 6. Plots of Type I Error Rates by Size of Group for Six Groups.

The mixed model specified in equations (1) and (2) was implemented by using the following is a SAS program. The individually administered treatment is coded 1 on the TRT code.

```
PROC MIXED;
CLASS TRT GROUP;
MODEL SCORE=TRT/SOLUTION
DDFM=SATTERTHWAITE;
RANDOM GROUP/GROUP=TRT;
REPEATED/GROUP=TRT;
PARMS (0) (1) (1) (1)/EQCONS=1
ESTIMATE 'COMP' TRT 1-1;
```

The APDF tests are easily carried out in proc iml as the only required statistics are the means for the two groups, the variance for the treatment administered to individuals, and the mean squares within and between subgroups for the group-administered treatments.

## Results

## Study 2

The analytic results showed that, when there were two groups, the APDF test statistic with the four-moment degrees of freedom provided the best control of the Type I error rate. Figure 7 compares Type I error rate for the fourmoment test and the mixed model test for $\rho_{\text {ICC }}=0.00$ and 0.30 . Results for the APDF test statistic and the two-moment degrees of freedom are also included because the mixed model test is equivalent to the two-moment test when the estimate of $\tau^{2}$ is non-zero. The four-moment degree of freedom test still provides the best control of the Type I error rate. The mixed model test is more conservative than the twomoment test and is substantially more conservative in conditions in which the probability of a zero estimate for $\tau^{2}$ is large.


Figure 7. Type I Error Rates for Two Groups

When there were three groups, the analytic results showed that the APDF test statistic with the averaged degrees of freedom provided the best control of the Type I error rate. Type I error rates for the two-moment test tended to be too large. Figure 8 compares Type I error rates for the mixed model test and the APDF tests with two-moment and averaged degrees of freedom when $\rho_{\text {ICC }}=.00$ and .30 . The results indicate that the averaged degrees of freedom test still provides the best control of the Type I error rate.

According to the analytic results, there were four or more groups, both the two-moment and averaged degrees of freedom tests provided good control over the Type I error rate, with the former test being slightly more liberal. Type I
error rates are depicted in Figure 9 for the twomoment, four-moment test and the mixed model tests for $\rho_{\text {ICC }}=.00$ and .30 .

The results indicate that the mixedmodel test is conservative and less adequate than the other tests when $\rho_{\text {ICC }}$ is zero. Inspection of the results for other values of $\rho_{\text {ICC }}$ indicate that when $\rho_{\text {ICC }}=.10$ the performance of the averaged degrees of freedom and the mixed model tests is very similar and as $\rho_{\text {ICC }}$ increases the Type I error rates for the mixed model test become slightly larger than those for the averaged degrees of freedom test. A similar pattern of results emerged for five or six groups. In particular, when $\rho_{\text {ICC }}$ was near zero the mixed model test was too conservative.


Figure 8. Type I Error Rates for Three Groups


Figure 9. Type I Error Rates for Four Groups

Conclusion
Myers et al. (1981) presented a two-moment approximate degrees of freedom test for use when one treatment is delivered to individual participants and one is delivered to groups of participants. The test was based on results in Satterthwaite (1941). Simulation results indicated that the test provided good control of the Type I error rate for both four groups and eight groups of participants.

Satterthwaite (1941) and Scariano and Davenport (1986) studied a two-moment approximate degrees of freedom test for a design in which both treatments are delivered individually. Discussion in Satterthwaite and results in Scariano and Davenport suggest that the Myers et al. test may not perform well when the number of groups is smaller than four. Using an analytic procedure developed by Scariano and Davenport, we showed that the Myers et al. test could provide relatively poor
control of the Type I error rate when there are two or three groups. Using results presented in Scariano and Davenport, we developed two alternatives to the Myers et al. (1981) test, a four-moment approximate degrees of freedom test and an averaged degrees of freedom test.

Using the analytic procedure developed by Scariano and Davenport, Type I error rates were calculated for all three test in a wide range of conditions in which the design was balanced across the individually administered treatment and the group-administered treatment and across the groups in the group-administered treatment. We also estimated Type I error rates for the mixed model test and the three APDF tests. The results indicated that the four-moment test should be used when the group-administered treatments are delivered to two groups and the averaged degrees of freedom test should be used when the group-administered treatments are delivered to three groups. When there are between four and six groups, we recommend
using the averaged degrees of freedom test. However, because (a) this test is slightly conservative, with a Type I error rate between 0.045 and 0.050 , and (b) the two-moment test is slightly liberal but tends to keep the Type I error rate below 0.06, some may prefer the twomoment test. Even when there are four or more groups, we do not recommend the mixed model test because of its conservative tendency when the intraclass correlation coefficient is small. These recommendations are summarized in the Table 1.

Table 1.
Recommended Tests by the Number of Groups in the Group-Administered Treatment

| Number of <br> Groups | Recommended Test |
| :---: | :---: |
| 2 | Four-Moment Test <br> Averaged Degrees of <br> Freedom Test |
| 3 | Averaged Degrees of <br> Freedom Or Two Moment <br> Test |

When there are two groups in the groupadministered treatment, the four-moment test provides better control of the Type I error rate than do the other tests. Nevertheless researchers should be cautious about using a groups-versusindividuals design with two groups because such designs will provide relatively low power. The true degrees of freedom for the four-moment test is

$$
\begin{equation*}
f_{4}=\frac{\left\{\frac{U^{2}}{J-1}+\frac{1}{N_{I}-1}\right\}^{3}}{\left(\frac{U^{3}}{(J-1)^{2}}+\frac{1}{\left(N_{I}-1\right)^{3}}\right)^{2}} \tag{19}
\end{equation*}
$$

where U is defined in equation (15). Calculations show that $f_{4}$ approaches 1.0 from above as $U$ increases. Thus in many situations the degrees of freedom for the four-moment test will be very small and this will have a negative impact on power. In addition, substituting
population parameters for sample statistics in the Myers et al. (1981) $t$ statistic, we have

$$
\begin{equation*}
\frac{\mu_{I}-\mu_{G}}{\sqrt{\frac{\sigma_{I}^{2}}{N_{I}}+\frac{\sigma_{G}^{2}}{N_{G}}+\frac{\tau^{2}}{J}}} . \tag{20}
\end{equation*}
$$

Therefore even as the two sample sizes increase power will not go to 1.0 if $\tau^{2} \neq 0$. Finally, the fact that the Type I error rate for the fourmoment test declines as $n$ increases suggests power will decline as $n$ increases because the test becomes more conservative. The predicted low power and decline in power as $n$ increases were borne out by simulation studies. For example when $\sigma_{I}^{2}=\sigma_{G}^{2}+\tau^{2}=1, \rho_{I C C}=.2$, and $\mu_{G}-\mu_{I}=.8$, estimated power was $.23, .21$ and .19 as $n$ increased from 6 to 18 in steps of 6 . Comparison of these results to the power of an independent samples $t$ test with the same overall sample size indicates how much lower power is when a groups-versus-individuals design is used. Note that because $\sigma_{I}^{2}=\sigma_{G}^{2}+\tau^{2}=1, \mu_{G}-\mu_{I}=.8$ corresponds to Cohen's large effect size. Also as $n$ increases from 6 to 18 the sample size in a treatment increases from 12 to 36 in steps of 12. For an independent sample $t$ test with an effect size equal to .8 , power is $.47, .77$, and .92 as $n$ increase from 12 to 36 by 12 .

When there are three groups and the averaged degrees of freedom approach is used, power does not decline as $n$ increases, but power can still be quit low and does not increase quickly as $n$ increase. As $n$ increased from 4 to 12 , so that the overall sample size remained the same as in the conditions on which power results were reported for $J=2$, estimated power was $.29, .36$, and .40 when $J$ was 3 .

As suggested by equation (20), power continues to increase as $J$ increases. For example with $J=6$, as $n$ increased from 2 to 6 in steps of 2 estimated power was $.41, .58$, and .68 using the averaged degrees of freedom test. Thus when the groups-versus-individuals test is used, it is important to have as many groups as possible and may be more important to have more groups than to have more participants per group.

At least four lines of additional research are attractive. First, the performance of the tests under non-normality should be investigated and if performance is poor developing the test statistic and degrees of freedom using robust estimates of the means and mean squares is of interest. Second, performance of the four tests when the design is unbalanced across the individually administered treatment and the group-administered treatment, but balanced across groups in the group-administered treatment might be investigated. Third, calculating the averaged degrees of freedom by differentially weighting the two-moment and four-moment degrees of freedom might be investigated when there are four or more groups. Weighting the two-moment degrees of freedom more heavily will reduce the slight conservative tendency of the averaged degrees of freedom test. In general, more extensive studies of power than we have conducted would be worthwhile. Fourth, the three APDF tests should be generalized for use when the design is not balanced across groups in the groupadministered treatment and Type I error rates for these tests and the mixed model test should be investigated.

## References

Bates, G. W., Thompson, J. C., \& Flanagan, C. (1999). The effectiveness of individual versus group induction of depressed mood. The Journal of Psycholog, 33, 245-252.

Boling, N. C., \& Robinson, D. H. (1999). Individual study, interactive multimedia, or cooperative learning: Which activity best supplements lecture-based distance education? Journal of Educational Psychology, 91, 169-174.

Bradley, J.V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Burlingame, G. M., Kircher, J. C., \& Taylor, S. (1994). Methodological considerations in group psychotherapy research: Past, present, and future practices. In A. Fuhriman \& G. Burlingame (Eds.), Handbook of group psychotherapy and counseling: An empirical and clinical synthesis. (pp. 41-80). New York: Wiley.

Burlingame, G. M., Kircher, J. C., \& Honts, C. R. (1994). Analysis of variance versus bootstrap procedures for analyzing dependent observations in small group research. Small Group Research, 25, 486-501.

Clarke, G. N. (1998). Improving the transition from basic efficacy research to effectiveness studies: Methodological issues and procedures. In A. E. Kazdin (Ed.), Methodological issues and strategies in clinical research, (2 ${ }^{\text {nd }}$ ed.) (pp. 541-559). New York: Wiley.

Cochran, W. G. (1951). Testing a linear relationship among variances. Biometrics, 7, 1732.

McCulloch, C. E., \& Searle, S. R. (2001). Generalized, linear, and mixed models. New York: Wiley.

McLean, R. A., Sanders, W. L., \& Stroup, W. W. (1991). A unified approach to mixed linear models. American Statisticia, 45, 54-64.

Myers, J., Dicecco, J., \& Lorch, Jr., J. (1981). Group dynamics and individual performances: Pseudogroup and Quasi-F analyses. Journal of Personality and Social Psychology, 40, 86-98.

Satterthwaite, F. W. (1941). Synthesis of variance. Psychometrik,, 6, 309-316.

Scariano, S. M., \& Davenport, J. M. (1986). A four-moment approach and other practical solutions to the Behrens-Fisher problem. Communications in Statistics: Theory and Methods, 15, 1467-1504.

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

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

Welch, B. L. (1938). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

## Endnote

Tables containing Type I error rates for all conditions in the studies are available at http://plaza.ufl.edu/algina/index.programs.html

# A Nonparametric Fitted Test For The Behrens-Fisher Problem 

Terry Hyslop<br>Department of Medicine<br>Thomas Jefferson University

Paul J. Lupinacci<br>Department of Mathematical Sciences<br>Villanova University

A nonparametric test for the Behrens-Fisher problem that is an extension of a test proposed by Fligner and Policello was developed. Empirical level and power estimates of this test are compared to those of alternative nonparametric and parametric tests through simulations. The results of our test were better than or comparable to all tests considered.

Key words: Behrens-Fisher problem, empirical level and power, Wilcoxon-Mann-Whitney, nonparametric, simulation study

## Introduction

The comparison of the means of two independent populations has traditionally been approached using Student's t-test. The use of this test assumes that the observations come from a normal distribution and that the variances of the two populations are equal. When the homogeneity of variances is not a reasonable assumption the problem has been called the Behrens-Fisher problem.

Lee and Gurland (1975) developed a new method for handling the Behrens-Fisher problem and compared their test to many others that have been proposed for this problem. Their test performed very well regarding size and power. However, their method utilized a large table of critical values to determine the correct region of rejection. Lee and Fineberg (1991) sought to simplify the method proposed by Lee and Gurland. They fit a nonlinear function to the critical values derived by Lee and Gurland so that the critical values could be estimated.

Terry Hyslop is Assistant Professor of Medicine in the Biostatistics Section, Division of Clinical Pharmacology, Department of Medicine at Thomas Jefferson University. Email: Terry.Hyslop@jefferson.edu. Paul J. Lupinacci is Assistant Professor in the Department of Mathematical Sciences at Villanova University. E-mail: Paul.Lupinacci@villanova.edu.

Various authors have also considered the Behrens-Fisher problem when the normality assumption is not appropriate. The usual nonparametric approaches assume that the data are continuous and the distributions are of the same shape. For these tests, such as the Wilcoxon-Mann-Whitney test (Wilcoxon 1945; Mann \& Whitney 1947), the level of the test will not be preserved when the populations have different shapes or variances (Fligner \& Policello 1981; Brunner \& Neumann 1982, 1986; Brunner \& Munzel 2000). Fligner and Policello (1981) and Brunner and Neumann (1982, 1986) considered the problem under the assumption that the independent samples are from continuous distributions without the assumption of equal variances or equal shapes of the distributions. Brunner and Munzel (2000) derived an asymptotically distribution free test without the assumption that the data are generated from a continuous distribution function.

Fligner and Policello developed their alternative nonparametric method for comparing two population medians without the equal variance and equal shape assumptions. To implement their test, one must consult a large table of critical values to determine the correct region of rejection. Their table is parameterized by the test's level of significance and the sample sizes of the two samples. We expand on the approach of Fligner and Policello by proposing a fitted test which eliminates the need for large tables or complicated derivations of critical values. We fit a nonlinear function to the critical
values in their table so that the critical values can be estimated. Motivation for the nonlinear function came from the nonlinear function used by Lee and Fineberg. A complete description of the problem and details of the proposed test follow in the Methodology Section. In that section, our method is demonstrated using a numerical example. Simulation studies are used in the Results Section to compare the fitted test to some of the other parametric and nonparametric tests which have been proposed for the Behrens-Fisher problem.

## Methodology

Let $X_{1}, \ldots, X_{m}$ and $Y_{1}, \ldots, Y_{n}$ be independent random samples from continuous distributions with population medians $\theta_{x}$ and $\theta_{y}$, respectively. We are interested in testing the following hypotheses:

$$
\begin{gathered}
H_{0}: \theta_{x}=\theta_{y} \\
\text { versus } H_{a}: \theta_{x}>\theta_{y}\left[\text { or } \theta_{x}<\theta_{y} \text { or } \theta_{x} \neq \theta_{y}\right] .
\end{gathered}
$$

Let $P_{i}$ represent the number of sample observations, $Y_{j}$, less than $X_{i}$, for $i=1, \ldots, m$. Similarly, let $Q_{j}$ represent the number of sample observations, $X_{i}$, less than $Y_{j}$, for $j=1, \ldots, n$. Compute the average placement for each of the samples,

$$
\begin{gathered}
\bar{P}=\frac{1}{m} \sum_{i=1}^{m} P_{i} \text { and } \bar{Q}=\frac{1}{n} \sum_{j=1}^{n} Q_{j} . \\
\text { Let } V_{1}=\sum_{i=1}^{m}\left(P_{i}-\bar{P}\right)^{2} \text { and } V_{2}=\sum_{j=1}^{n}\left(Q_{j}-\bar{Q}\right)^{2},
\end{gathered}
$$

and calculate the test statistic

$$
\hat{U}=\frac{\sum_{i=1}^{m} P_{i}-\sum_{j=1}^{n} Q_{j}}{2\left(V_{1}+V_{2}+\bar{P} \bar{Q}\right)^{1 / 2}} .
$$

Fligner and Policello presented a test of $\mathrm{H}_{0}$ based on this statistic, where the procedure at an
approximate $\alpha$ level of significance versus the one-sided alternative $\theta_{x}>\theta_{y}$ is

Reject $\mathrm{H}_{0}$ if $\hat{U} \geq u_{\alpha}$; otherwise do not reject.

They provided a table of critical values for $u_{\alpha}$ for various values of $\mathrm{m}, \mathrm{n}$, and $\alpha$. Values outside the range of their table are to be derived or estimated by $Z_{\alpha}$ for large sample sizes, where $z_{\alpha}$ is the $1-\alpha$ percentile of the standard normal distribution.

Implementation of this test would be greatly simplified if the large table of critical values was not required. In addition, sample size combinations that are not provided in their table would require either additional effort for derivation, or an assumption of $u_{\alpha}=z_{\alpha}$. We propose fitting the following function to the critical values in the Fligner and Policello table so that the critical values can be estimated:
$u_{\alpha}=b_{0}+b_{1} / f_{1}+b_{2} / f_{2}+b_{3} /\left(f_{1} f_{2}\right)^{+}+b_{4} / f_{1}{ }^{2}+b_{5} / f_{2}{ }^{2}$,
where $f_{1}=m-1, f_{2}=n-1$, and $b_{0, \ldots,}, b_{5}$ are the parameters of the function. We also propose that the parameters $b_{0}, \ldots, b_{5}$ be estimated by ordinary least squares. 54 values obtained from Fligner and Policello's table of critical values were used in the estimation process. Table 1 presents the parameter estimates obtained for the various $\alpha$ values of $0.10,0.05,0.025$, and 0.01 .

Table 1. Parameter estimates for the F-P fitted test polynomial.

| $\alpha$ | $b_{0}$ | $b_{1}$ | $b_{2}$ | $b_{3}$ | $b_{4}$ | $b_{5}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.10 | 1.34 | -1.39 | 0.16 | -0.03 | 5.20 | 1.17 |
| 0.05 | 1.74 | -0.69 | -0.87 | 12.53 | -4.09 | 2.44 |
| 0.025 | 2.15 | -0.60 | -2.54 | 22.05 | -3.50 | 7.36 |
| 0.01 | 3.16 | -11.43 | -6.75 | 50.15 | 51.87 | 19.20 |

Motivation for this functional form comes from a parametric fitted test for the Behrens-Fisher problem proposed by Lee and Fineberg (1991) as an alternative to Lee and Gurland's (1975) test that also required
extensive tables of critical values. Their proposed function is similar to the one proposed here. Other functional forms were also considered, but none were found which provided a better fit to the critical values. Figure 1 displays the critical and fitted values for a level 0.05 test when $m=4,5, \ldots, 12$ and $n=3,4$, and 5. The fit is good for even these small values of n and becomes more precise as n gets larger. The test based on the fitted critical values will be
referred to as $\hat{U}_{f}$, and the critical value will be referred to as $u_{\alpha}^{(f)}$. For example, Fligner and Policello's critical value for a one-sided, level 0.05 test when both samples are of size 5 is 2.063. Using our parameter estimates when $\alpha=0.05$ and the sample sizes, we obtain an estimated critical value of 2.035 .

Figure 1. Plot of Fligner and Policello’s Critical Values and the Fitted Critical Values for $\mathrm{m}=4(1) 12$ and $n=3,4,5$.


## Numerical Example

The following example uses a data set that originated from the simulation studies that are presented in the Results Section of the manuscript. With this data set, we will test the null hypothesis that the two population medians are the same versus the alternative hypothesis that the median of the first population is greater than that of the second population, that is,

$$
H_{o}: \theta_{x}=\theta_{y} \text { versus } H_{a}: \theta_{x}>\theta_{y}
$$

The data in both groups were simulated from uniform distributions with a mean of 100 . However, the variance of the second distribution was ten times that of the first distribution. The first data set consists of twelve observations while the second data set consists of only five observations. Thus, we are simulating a scenario where the data set with fewer observations
comes from an underlying distribution function with a larger variance. We will utilize this data set to demonstrate the test procedure and to illustrate the need for a simulation study which compares the various methods that are used for analyzing this type of data in terms of their power and ability to hold the level of
significance. We will use the notation as defined in the Methodology Section. The data, as well as the placement values, $P_{i}$ and $Q_{j}$, for each observation in the first and second samples, respectively, are given in Table 2.

Table 2. Data for the Numerical Example

| Group 1 |  |  | Group 2 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Observation | Value | $P_{i}$ | Observation | Value | $Q_{j}$ |
| 1 | 101.673 | 4 | 1 | 103.409 | 12 |
| 2 | 101.550 | 4 | 2 | 98.546 | 1 |
| 3 | 100.410 | 4 | 3 | 97.429 | 0 |
| 4 | 100.203 | 4 | 4 | 96.536 | 0 |
| 5 | 99.906 | 4 | 5 | 95.940 | 0 |
| 6 | 99.875 | 4 |  |  |  |
| 7 | 99.861 | 4 |  |  |  |
| 8 | 99.695 | 4 |  |  |  |
| 9 | 99.535 | 4 |  |  |  |
| 10 | 98.985 | 4 |  |  |  |
| 11 | 98.575 | 4 |  |  |  |
| 12 | 98.461 | 3 |  | $\sum_{i=1}^{5} Q_{j}=13$ |  |
| $\sum_{i=1}^{12} P_{i}=47$ |  |  |  |  |  |

For this example, the sum of the placements in the first data set is 47 and the sum of the placements in the second data set is 13 . this leads to the average placements of:

$$
\bar{P}=\frac{1}{12} \sum_{i=1}^{12} P_{i}=3.917
$$

and

$$
\bar{Q}=\frac{1}{5} \sum_{j=1}^{5} Q_{j}=2.600
$$

for each group. The values of $V_{1}$ and $V_{2}$ are

$$
V_{1}=\sum_{i=1}^{12}\left(P_{i}-\bar{P}\right)^{2}=0.917
$$

and

$$
V_{2}=\sum_{j=1}^{5}\left(Q_{j}-\bar{Q}\right)^{2}=111.200,
$$

and the test statistic is calculated as

$$
\hat{U}=\frac{\sum_{i=1}^{12} P_{i}-\sum_{j=1}^{5} Q_{j}}{2\left(V_{1}+V_{2}+\bar{P} \bar{Q}\right)^{1 / 2}}=1.537 .
$$

For $\mathrm{m}=12$ and $\mathrm{n}=5$, the critical value for the fitted test is $u_{0.05}^{(f)}=1.868$, and the critical value for the Fligner and Policello test is $u_{0.05}=$ 1.923. Therefore, we fail to reject the null hypothesis using both tests. However, the calculation of the Fligner and Policello critical value would have been much more complicated if our sample sizes were not given in their table of critical values. Therefore, we suggest using the critical value based on the fitted test.

Let us also consider how alternative
tests for this type of data would have fared with this data set. Since the data are not coming from a normal distribution, most statisticians would use the nonparametric alternative to Student's t test, the Wilcoxon-Mann-Whitney test, without hesitation. For this example, we compared our results to those of the Wilcoxon-Mann-Whitney test and the nonparametric test proposed by Brunner and Munzel (2000). The Brunner and Munzel test led to the same conclusion as our fitted test, that is, their test failed to reject the null hypothesis. The Brunner and Munzel test statistic was $B=1.437$ and the corresponding $p$ value was 0.112 , which is not significant at the 0.05 level of significance. However, there was a different result if one used the Wilcoxon-MannWhitney test. Its test statistic was $W=28$ and the corresponding p -value was 0.041 , which is significant at the 0.05 level of significance. This conflicting result caught our attention and spurred interest in a simulation study which compares the various methods in terms of size and power.

## Results

For our simulation study, we considered three nonparametric procedures and one parametric procedure. The three nonparametric procedures that were considered were the Wilcoxon-MannWhitney test, denoted $W$, the Brunner and Munzel test, denoted $B$, and our fitted test, denoted $\hat{U}_{f}$. The parametric test that we included in our simulation study was the usual t test using Satterthwaite's approximation for the degrees of freedom. We used $t_{s}$ to denote this test. We decided not to include the Fligner and Policello test in the discussion because its empirical level and power estimates were almost identical to those of our fitted test. This was to be expected since we fitted a function to their critical values and the fit was very good. We simulated data using the normal, contaminated normal, double exponential, uniform, and gamma distributions for estimating both the empirical level and power for the four tests. Since we are interested in determining the effect of different variances on the level and power estimates, we considered distributions which differed in scale by assuming that if

$$
X_{1}, \ldots, X_{m} \sim F(x),
$$

then we let

$$
Y_{1}, \ldots, Y_{n} \sim G(y)=F(y / \sigma)
$$

for values of

$$
\sigma^{2}=\{0.01,0.25,1,4,10\} .
$$

All simulations were run in SAS version 8. The SAS function NORMAL was used to generate random standard normal deviates which were then transformed to simulate the desired normal distribution. The contaminated normal deviates were generated by multiplying a random normal deviate by 9 with probability $\mathrm{p}=$ 0.10 . The double exponential deviates were generated using the method of Martinez and Iglewicz (1984) that transforms a random standard normal deviate into a double exponential deviate using the transformation

$$
D E=Z^{\exp \left\{\frac{0.109 Z^{2}}{2}\right\}},
$$

where $Z$ is a random standard normal deviate. Random uniform and gamma deviates were generated using the SAS functions UNIFORM and RANGAM, respectively.

For a statistical test to be meaningful, it must display adequate power while still maintaining its nominal level. We ran simulations to obtain estimates of the level and power for each of the tests under consideration. To estimate the tests' level, we ran 15,500 simulation iterations. The number of simulations provides that a $95 \%$ confidence interval for the estimated level will be approximately $\pm 0.36 \%$ for $\alpha=0.05$. At each iteration $\mathrm{m}+\mathrm{n}$ deviates of the desired type were generated from distributions where $\theta_{x}=\theta_{y}$. The four tests were performed at each interation testing $H_{o}: \theta_{x}=\theta_{y}$ vs. $H_{a}: \theta_{x}>\theta_{y}$. The proportion of the iterations where the null hypothesis was rejected was recorded for each of the four tests. This proportion is the empirical level estimate. The
empirical levels were mulitplied by 1,000 and these values are reported in Table 3. Table 3 lists the empirical levels for each of the five distributions, for five sample size combinations, at each of the five variance ratios, $\tau$, for each of
the four tests. The standard error was calculated assuming a true nominal level of 0.05 . We indicate an empirical level more than two standard deviations above 0.05 by entering the number into the table in boldface type.

Table 3. Empirical Levels Times 1,000 for $\alpha=0.05$ for Each of the 4 Tests.

| Distribution | $\tau$ | $\mathrm{m}=5, \mathrm{n}=5$ |  |  |  | $\mathrm{m}=12, \mathrm{n}=5$ |  |  |  | $\mathrm{m}=11, \mathrm{n}=10$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | W | $\hat{U}^{f}$ | B | $\mathrm{t}_{\text {s }}$ | W | $\hat{U}^{f}$ | B | $\mathrm{t}_{\text {s }}$ | W | $\hat{U}^{f}$ | B | $\mathrm{t}_{\text {s }}$ |
| Normal | 0.1 | 49 | 49 | 47 | 51 | 22 | 35 | 49 | 48 | 59 | 54 | 49 | 50 |
|  | 0.25 | 49 | 49 | 49 | 48 | 24 | 38 | 49 | 49 | 55 | 54 | 52 | 51 |
|  | 1 | 51 | 51 | 55 | 50 | 40 | 52 | 56 | 49 | 51 | 53 | 52 | 50 |
|  | 4 | 52 | 52 | 53 | 51 | 67 | 61 | 56 | 52 | 55 | 52 | 49 | 47 |
|  | 10 | 52 | 52 | 48 | 51 | 89 | 59 | 52 | 53 | 61 | 55 | 48 | 45 |
| Contaminated Normal | 0.1 | 49 | 32 | 44 | 25 | 28 | 13 | 59 | 73 | 59 | 29 | 49 | 41 |
|  | 0.25 | 48 | 29 | 47 | 21 | 29 | 13 | 57 | 75 | 54 | 25 | 49 | 38 |
|  | 1 | 49 | 28 | 53 | 21 | 43 | 21 | 58 | 73 | 49 | 26 | 51 | 41 |
|  | 4 | 45 | 27 | 46 | 22 | 60 | 27 | 49 | 72 | 56 | 29 | 52 | 40 |
|  | 10 | 49 | 30 | 44 | 21 | 76 | 28 | 44 | 72 | 61 | 29 | 51 | 38 |
| Uniform | 0.1 | 50 | 50 | 47 | 57 | 19 | 32 | 47 | 46 | 61 | 52 | 47 | 50 |
|  | 0.25 | 52 | 52 | 53 | 55 | 23 | 37 | 50 | 51 | 55 | 53 | 50 | 50 |
|  | 1 | 47 | 47 | 49 | 44 | 42 | 54 | 56 | 55 | 48 | 50 | 49 | 49 |
|  | 4 | 48 | 48 | 51 | 52 | 80 | 66 | 59 | 59 | 62 | 56 | 49 | 49 |
|  | 10 | 49 | 49 | 48 | 57 | 103 | 58 | 53 | 59 | 70 | 58 | 50 | 50 |
| Double <br> Exponential | 0.1 | 51 | 51 | 46 | 47 | 22 | 36 | 50 | 49 | 58 | 54 | 50 | 49 |
|  | 0.25 | 49 | 49 | 44 | 47 | 25 | 39 | 50 | 46 | 49 | 50 | 48 | 48 |
|  | 1 | 48 | 48 | 51 | 45 | 42 | 52 | 53 | 48 | 50 | 52 | 51 | 48 |
|  | 4 | 50 | 50 | 51 | 47 | 66 | 61 | 56 | 47 | 56 | 53 | 49 | 47 |
|  | 10 | 46 | 46 | 42 | 42 | 87 | 59 | 52 | 47 | 66 | 58 | 52 | 50 |
| Gamma | 0.1 | 32 | 32 | 30 | 32 | 10 | 19 | 27 | 38 | 33 | 30 | 28 | 37 |
|  | 0.25 | 35 | 35 | 37 | 33 | 17 | 27 | 34 | 44 | 34 | 33 | 31 | 38 |
|  | 1 | 48 | 48 | 51 | 46 | 42 | 53 | 57 | 65 | 51 | 53 | 51 | 50 |
|  | 4 | 87 | 87 | 88 | 85 | 123 | 110 | 100 | 109 | 129 | 122 | 115 | 84 |
|  | 10 | 144 | 144 | 132 | 147 | 238 | 165 | 147 | 163 | 253 | 229 | 209 | 124 |

Notes: Wilcoxon-Mann-Whitney (W), Fitted Test ( $\hat{U}^{f}$ ), Brunner-Munzel (B), and Satterthwaite's t-test $\left(\mathrm{t}_{\mathrm{s}}\right)$. Variance of $\mathrm{X}=1$, Variance of $\mathrm{Y}=\tau$. The right side of this table continues on the page below.

Table 3, continued.

| Distribution | $\tau$ | $\mathrm{m}=25, \mathrm{n}=20$ |  |  |  | $\mathrm{m}=40, \mathrm{n}=40$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | W | $\hat{U}^{f}$ | B | $\mathrm{t}_{\text {s }}$ | W | $\hat{U}^{f}$ | B | $\mathrm{t}_{\text {s }}$ |
| Normal | 0.1 | 51 | 47 | 48 | 48 | 60 | 44 | 48 | 49 |
|  | 0.25 | 45 | 44 | 46 | 49 | 54 | 43 | 48 | 46 |
|  | 1 | 47 | 46 | 48 | 50 | 52 | 46 | 52 | 53 |
|  | 4 | 59 | 47 | 47 | 47 | 58 | 47 | 51 | 52 |
|  | 10 | 71 | 50 | 49 | 48 | 66 | 49 | 52 | 52 |
| Contaminated Normal | 0.1 | 53 | 17 | 48 | 64 | 62 | 19 | 51 | 54 |
|  | 0.25 | 50 | 16 | 49 | 67 | 55 | 17 | 50 | 54 |
|  | 1 | 48 | 18 | 49 | 67 | 50 | 15 | 50 | 48 |
|  | 4 | 55 | 19 | 48 | 66 | 52 | 16 | 49 | 52 |
|  | 10 | 65 | 22 | 49 | 66 | 63 | 20 | 54 | 53 |
| Uniform | 0.1 | 58 | 48 | 49 | 49 | 67 | 44 | 48 | 47 |
|  | 0.25 | 50 | 46 | 48 | 49 | 61 | 47 | 50 | 51 |
|  | 1 | 50 | 49 | 51 | 51 | 51 | 46 | 52 | 50 |
|  | 4 | 63 | 48 | 47 | 48 | 61 | 47 | 50 | 51 |
|  | 10 | 78 | 52 | 50 | 52 | 64 | 43 | 46 | 46 |
| Double <br> Exponential | 0.1 | 52 | 46 | 48 | 48 | 64 | 47 | 52 | 51 |
|  | 0.25 | 48 | 46 | 49 | 49 | 51 | 43 | 47 | 48 |
|  | 1 | 52 | 50 | 53 | 53 | 47 | 43 | 48 | 47 |
|  | 4 | 59 | 48 | 48 | 47 | 55 | 46 | 50 | 54 |
|  | 10 | 70 | 50 | 48 | 46 | 61 | 45 | 49 | 49 |
| Gamma | 0.1 | 22 | 19 | 20 | 40 | 20 | 13 | 15 | 41 |
|  | 0.25 | 25 | 24 | 25 | 44 | 19 | 15 | 17 | 40 |
|  | 1 | 47 | 46 | 48 | 50 | 52 | 48 | 53 | 54 |
|  | 4 | 186 | 158 | 158 | 76 | 245 | 215 | 227 | 72 |
|  | 10 | 408 | 339 | 332 | 108 | 590 | 521 | 535 | 91 |

Notes. Continued from previous page. Wilcoxon-Mann-Whitney (W), Fitted Test ( $\hat{U}^{f}$ ), BrunnerMunzel (B), and Satterthwaite's t-test $\left(\mathrm{t}_{\mathrm{s}}\right)$. Variance of $\mathrm{X}=1$, Variance of $\mathrm{Y}=\tau$.

There are a number of interesting conclusions that can be made from observing the values in Table 3. First, the $t$ test using Satterthwaite's approximation for the degrees of freedom maintained its level when the data were generated from a normal distribution regardless of the sample size combination or the ratio of the variances. This was expected since the primary purpose of this test is to handle these situations. However, when the condition of normality was removed, the test became less predictable. In some cases, such as when the data were
generated from the contaminated normal distribution and the sample sizes were similar, the test was very conservative. In other cases, such as when the data were uniformly distributed and when the sample sizes differed, the test became anti-conservative. The Wilcoxon-Mann-Whitney text generally does not maintain its level, even under the optimal condition of normality. It was conservative in situations where the larger sample size was taken from the population with the larger variance, and it was anti-conservative if the reverse was true. The fitted test generally maintained its level. In most of the situations
when it did not, the test was conservative. The Brunner-Munzel test generally maintained its level under all scenarios tested. All of the tests had trouble maintaining the 0.05 level when the data were simulated using the gamma distribution.

To estimate the tests' power, we ran 1,540 simulation iterations. This number of simulations assures that a $95 \%$ confidence interval for power will be approximately $\pm 0.025$ when power is around $80 \%$. For each iteration, $\mathrm{m}+\mathrm{n}$ deviates of the desired type were generated under the condition $\theta_{x}-\theta_{y}=\delta$, where $\delta=\{1,2,3,4\}$. Again, the proportion of the iterations where the null hypothesis was rejected was recorded for each of the four tests. This proportion is the test's estimated power. Since the Wilcoxon-Mann-Whitney test was anti-
conservative in most scenarios, it was not surprising that the power of this test was greater than the power of the other tests. However, since this power is meaningless in the presence of an inflated nominal level, the Wilcoxon test will be removed from the rest of the discussion.

Figure 2 shows the power of the remaining tests under normality when the variances are not equal and the sample sizes are the same. Under these conditions, most statisticians would use the $t$ test with Satterthwaite's approximation for the degrees of freedom. However, the fitted test and the Brunner-Munzel test demonstrate comparable power.

Figure 3 illustrates the power of the tests under normality with the added complication that the smaller sample size corresponds to the group with the larger variance. Once again, all three tests demonstrate similar power levels.

Figure 2. Plot of the Power for the Various Tests under Normality, Equal Sample Sizes, Ratio of the Variances $\tau=0.1$, and $\alpha=0.05$.


Figure 3. Plot of the Power for the Various Tests under Normality, Different Sample Sizes, Ratio of the Variances $\tau=10$, and $\alpha=0.05$.


When symmetry is removed from the distribution, such as in the case of the contaminated normal distribution, the fitted test and the Brunner-Munzel test demonstrate superiority over Satterthwaite's $t$ test. This is illustrated in Figures 4 and 5. Figure 4 illustrates the power of the three tests when samples of the same size are generated from contaminated normal distributions with the same variance. Figure 5 illustrates the power of the three tests
when samples of different sizes are generated from contaminated normal distributions with different variances. In both of these figures, the fitted test and the Brunner-Munzel test demonstrate similar power. However, the $t$ test using Satterthwaite's approximation for the degrees of freedom has considerably less power than the other tests. This pattern is consistent over all of the results run using the contaminated normal distribution.

Figure 4. Plot of the Power for the Various Tests under the Contaminated Normal Distribution, Equal Sample Sizes, and Ratio of the Variances $\tau=1$, and $\alpha=0.05$.


Figure 5. Plot of the Power for the Various Tests under the Contaminated Normal Distribution, Different Sample Sizes, and Ratio of the Variances $\tau=0.1$, and $\alpha=0.05$.


All three tests exhibited comparable power under the double exponential and uniform distributions. All three tests had increased power when the sample with fewer observations was obtained from the distribution with the smaller
variance. However, all three tests exhibited decreased power when the sample with fewer observations was obtained from the distribution with the larger variance.

## Conclusion

In this paper, we developed a method to test the difference between two population medians. Our fitted test was created by fitting a function to the large table of critical values presented by Fligner and Policello. Through a simulation study, we have determined that our test, and the Brunner-Munzel test, generally maintains the expected level of the test for a variety of underlying density functions.

The usual alternative to Student's t test, the Wilcoxon-Mann-Whitney test, has been shown to be anti-conservative in the simulation study under unequal variances by exhibiting empirical level estimates that are generally greater than the nominal level. Therefore, this test also exhibited artificially high power in the simulation results. Whereas, the fitted test and the Brunner-Munzel test have shown comparable power to the $t$ test using Satterthwaite's approximation for the degrees of freedom under the ideal condition of normality. When symmetry is removed from the distribution function, such as in the contaminated normal distribution, the fitted test and the Brunner-Munzel test have shown improved power over the $t$ test using Satterthwaite's approximation for the degrees of freedom.

All three tests exhibited comparable power in the simulation studies when the data were simulated from the double exponential or the uniform distributions. Statisticians should consider using an alternative to the Wilcoxon-Mann-Whitney test when unequal variances are possible.

## References

Brunner, E., \& Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. Biometrical Journal, 42, 17-25.

Brunner, E., \& Neumann, N. (1982). Rank tests for correlated random variables. Biometrical Journal, 24, 373-389.

Brunner, E., \& Nuemann, N. (1986). Two-sample rank tests in general models. Biometrical Journal, 28, 395-402.

Fligner, M. A., \& Policello, G. E. (1981). Robust rank procedures for the BehrensFisher problem. Journal of the American Statistical Association, 76, 162-168.

Lee, A. F. S., \& Fineberg, N .S. (1991). A fitted test for the Behrens-Fisher problem. Communications in Statistics - Theory and Methods, 20, 653-666.

Lee, A. F. S., \& Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. Journal of the American Statistical Association, 70, 933-941.

Mann, H. B., \& Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50-60.

Martinez, J., \& Iglewicz, B. (1984). Some properties of the Tukey g and h family of distributions. Communications in Statistics Theory and Methods, 13, 353-369.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics, 1, 80-83.

# Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis 

David A. Walker<br>Educational Research and Assessment<br>Northern Illinois University

Denise Y. Young<br>Institutional Research<br>University of Dallas

Many national data sets used in educational research are not based on simple random sampling schemes, but instead are constructed using complex sampling designs characterized by multi-stage cluster sampling and over-sampling of some groups. Incorrect results are obtained from statistical analysis if adjustments are not made for the sampling design. This study demonstrates how the use of weights and design effects impact the results of contingency tables and chi-square analysis of data from complex sampling designs.

Key words: Design effect, chi-square, weighting

Introduction
Many large-scale data sets used in educational research are constructed using complex designs characterized by multi-stage cluster sampling and over-sampling of some groups. Common statistical software packages such as SAS and SPSS yield incorrect results from such designs unless weights and design effects are used in the analysis (Broene \& Rust, 2000; Thomas \& Heck, 2001). The objective of this study is to demonstrate how the use of weights and design effects impact the results of contingency tables and chi-square analysis of data from complex sampling designs.

This article is based upon work supported by the Association for Institutional Research (AIR), National Center for Education Statistics (NCES), and the National Science Foundation (NSF) through fellowship grants awarded to the authors to participate in the 2001 AIR Summer Data Policy Institute on Databases of NCES and NSF. Correspondence for this article should be sent to: David Walker, Northern Illinois University, ETRA Department, 208 Gabel, DeKalb, IL 60115, (815)-753-7886. E-mail him at: dawalker@niu.edu.

Methodology
In large-scale data collection, survey research applies varied sample design techniques. For example, in a single-stage simple random sample with replacement (SRS), each subject in the study has an equal probability of being selected. Thus, each subject chosen in the sample represents an equivalent total of subjects in the population. More befitting, however, is that data collection via survey analysis often involves the implementation of complex survey design (CSD) sampling, such as disproportional stratified sampling or cluster sampling, where subjects in the sample are selected based on different probabilities. Each subject chosen in the sample represents a different number of subjects in the population (McMillan \& Schumacher, 1997).

Complex designs often engender a particular subgroup, due to oversampling or selection with a higher probability, and consequently the sample does not reflect accurate proportional representation in the population of interest. Thus, this may afford more weight to a certain subgroup in the sample than would be existent in the population. As Thomas and Heck (2001) cautioned, "When using data from complex samples, the equal weighting of observations, which is appropriate with data collected through simple random samples, will bias the model's parameter
estimates if there are certain subpopulations that have been oversampled" (p. 521).

The National Center for Education Statistics (NCES) conducts various national surveys that apply complex designs to projects such as the Beginning Postsecondary Students study (BPS), the National Educational Longitudinal Study of 1988 (NELS: 88), or the National Study of Postsecondary Faculty (NSOPF). Some statistical software programs, for instance SPSS or SAS, presuppose that data were accumulated through SRS. These statistical programs tend not to use as a default setting sample weights with data amassed through complex designs, but instead use raw expansion weights as a measure of acceptable sample size (Cohen, 1997; Muthen \& Satorra, 1995). However, the complex sampling designs utilized in the collection of NCES survey data allocates larger comparative importance to some sampled elements than to others. To illustrate, a complex design identified by the NCES may have a sample selection where 1 subject out of 40 is chosen, which indicates that the selection probability is $1 / 40$. The sample weight of 40 , which is inversely proportional to the selection probability, indicates that in this particular case 1 sample subject equals 40 subjects in the population.

Because of the use of complex designs, sample weighting for disparate subject representation is employed to bring the sample variance in congruity with the population variance, which supports proper statistical inferences. The NCES incorporates as part of its data sets raw expansion weights to be applied with the data of study to ensure that the issues of sample selection by unequal probability sampling and biased estimates have been addressed. Relative weights can be computed from these raw expansion weights.

Because the NCES accrues an abundance of its data for analysis via CSD, the following formulae present how weights function. The raw expansion weight is the weight that many statistical software programs use as a default setting and should be avoided when working with the majority of NCES data. Instead, the relative weight should be used when conducting statistical analyses with NCES complex designs.


Notes: $\mathrm{n}=$ sample size, $\mathrm{j}=1$ = subject response, $\mathrm{w}_{\mathrm{j}}=$ raw weight, $\mathrm{x}_{\mathrm{j}}=$ variable value, $\mathrm{N}=$ population size

Furthermore, the lack of sample weighting with complex designs causes inaccurate estimates of population parameters. The existence of variance estimates, which underestimate the true variance of the population, induce problems of imprecise confidence intervals, larger than expected degrees of freedom, and an enhancement of Type I errors (Carlson, Johnson, \& Cohen, 1993; Lee, Forthofer, \& Lorimor, 1989).

Design effect (DEFF) indicates how sampling design influences the computation of the statistics under study and accommodates for the miscalculation of sampling error. As noted previously, since statistical software programs often produce results based on the assumption that SRS was implemented, DEFF is used to adjust for these inaccurate variances. DEFF, as defined by Kish (1965), is the ratio of the variance of a statistic from a CSD to the variance of a statistic from a SRS.

$$
\begin{equation*}
\mathrm{DEFF}=--\frac{S E^{2}}{\mathrm{SE}^{2}} \frac{\mathrm{CSD}}{\text { SRS }} \tag{5}
\end{equation*}
$$

The size of DEFF is affined to conditions such as the variables of interest or the attributes of the clusters used in the design (i.e., the extent of in-cluster homogeneity). A DEFF greater than 1.0 connotes that the sampling design decreases precision of estimate compared to SRS, and a DEFF less than 1.0 confirms that
the sampling design increases precision of estimate compared to SRS (Kalton, 1983; Muthen \& Satorra, 1995). As Thomas and Heck (2001) stated, "If standard errors are underestimated by not taking the complex sample design into account, there exists a greater likelihood of finding erroneously 'significant' parameters in the model that the a priori established alpha value indicates" (p. 529).

Procedures
Three variables were selected from the public-use database of the National Education Longitudinal Study of 1988 to demonstrate the impact of weights and design effects on contingency tables and chi-square analysis. A two-stage cluster sample design was used in NELS: 88, whereby approximately 1,000 eighthgrade schools were sampled from a universe of approximately 40,000 public and private eighthgrade schools (first stage) and 24 eighth-grade students were randomly selected from each of the participating schools (second stage).

An additional 2 to 3 Asian and Hispanic students were selected from each school, which resulted in a total sample of approximately 25,000 eighth-grade students in 1988. Follow-up studies were conducted on subsamples of this cohort in 1990, 1992, 1994, and 2000. Additional details on the sampling methodology for NELS: 88 are contained in a technical report from the U.S. Department of Education (1996).

The three variables used in this example are F2RHMA_C (total Carnegie Units in mathematics taken in high school), RMATH (flag for whether one or more courses in remedial math were taken since leaving high school), and F3TRSCWT (1994 weight to be used with 1992 transcript data). Five categories for the number of Carnegie Units of math taken in high school were created (up through 1.99, 2.00 through 2.99, 3.00 through 3.99, 4.00 through 4.99, 5.00 or more). The other variable of interest was whether a student had taken a postsecondary remedial math course by the time of the 1994 follow-up study. Four chi-square contingency tables were developed for these two variables using SPSS. Differences in the four tables are due to use of weights and DEFF.

Only those observations where RMATH $>0$ and F3TRSCWT > 0 were selected for this
analysis, which resulted in 6,948 students. Although there were 14,915 students in the 1994 follow-up of NELS: 88, only 12,509 had high school transcript data (F3TRSCWT > 0) from which F2RHMA_C was obtained. Of these, 6,948 participated in post-secondary education by the time of the third follow-up in 1994.

Missing values were not a problem with RMATH. Of the 14,915 students in the 1994 follow-up of NELS: 88, 6,943 had a legitimate missing value because they had not participated in postsecondary education (i.e., not of interest for this paper), 16 had missing values, and 7,956 had a value (yes or no) for postsecondary remedial math.

There were some missing values for high school transcript data, but the transcript weight (F3TRSCWT) provided in NELS: 88 takes into account missing transcript data. The Carnegie units of high school math (F2RHMA_C) came from high school transcript data. There were 14,915 students in the 1994 follow-up of NELS: 88; however, only 12,509 had high school transcript data. That is why NCES provides a separate weight (F3TRSCWT) that is to be used specifically with variables from high school transcript data.

This weight has already been adjusted by NCES for missing high school transcript observations. Of the 7,956 students with a value for RMATH, 1,008 did not have high school transcript data. These 1,008 students were not included in the analysis presented here (7,956$1,008=6948$ students for analysis in this paper). After selecting the 7,956 students with a value for RMATH, only those observations with F3TRSCWT>0 were selected. No further adjustment was necessary for missing values since F3TRSCWT had already been adjusted by NCES for missing values.

Effect sizes are reported for each chisquare statistic addressed in the research. For the chi-square statistic, a regularly used effect size is based on the coefficient of contingency (C), which is not a true correlation but a "scaled" chisquared (Sprinthall, 2000). As a caveat with the use of C , it has been noted that its highest value cannot attain 1.00 , as is common with other effect sizes, which makes concordance with akin effect sizes arduous.

In fact, C has a maximum approaching 1.0 only for large tables. In tables smaller than 5 $x$ 5, C may underestimate the level of association (Cohen, 1988; Ferguson, 1966). As an alternative to C, Sakoda’s Adjusted C (C*) may be used, which varies from 0 to 1 regardless of table size. For chi-square related effect sizes, Cohen (1988) recommended that $.10, .30$, 50 represent small, medium, and large effects.

$$
\begin{align*}
& \mathrm{C}=\operatorname{SQRT}\left[\chi^{2} /\left(\chi^{2}+\mathrm{n}\right)\right]  \tag{6}\\
& \mathrm{C}^{*}=\mathrm{C} / \operatorname{SQRT}[(\mathrm{k}-1) / \mathrm{k}] \tag{7}
\end{align*}
$$

$\mathrm{k}=$ number of rows or columns, whichever is smaller.

## Results

A total of 6,948 observations met the selection criteria (i.e., availability of high school transcripts and participation in post-secondary education by the time of the third follow-up in 1994). The first contingency table (Table 1), without any weights or design effects, has a total count of 6,948 and a chi-square value of 130.92 . This table is useful for determining minimum cell sizes, but the percentages in each of the cells and the overall chi-square (130.92) are incorrect because the sample observations were not weighted to represent the population.

Table 1. Contingency Table Without Weights: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^{2}(4)=130.92$, C $=.136,95 \%$ CI (.112, .160), C* $=.192,95 \%$ CI (.168, .216).

|  |  | PSE Remedial Math |  |  |
| :--- | :--- | :---: | :---: | :---: |
| Units HS Math |  | Yes | No | Row Total |
|  |  |  |  |  |
| $0-1.99$ | Count | 84 | 231 | 315 |
|  | \% of Grand Total | $1.2 \%$ | $3.3 \%$ | $4.5 \%$ |
|  |  |  |  |  |
| $2-2.99$ | Count | 215 | 661 | 876 |
|  | $\%$ of Grand Total | $3.1 \%$ | $9.5 \%$ | $12.6 \%$ |
|  |  |  |  |  |
| $3-3.99$ | Count | 495 | 1,875 | 2,370 |
|  | $\%$ of Grand Total | $7.1 \%$ | $27.0 \%$ | $34.1 \%$ |
|  |  |  |  |  |
| $4-4.99$ | Count | 382 | 2,504 | 2,886 |
|  | $\%$ of Grand Total | $5.5 \%$ | $36.0 \%$ | $41.5 \%$ |
|  |  |  |  |  |
| $>=5$ | Count | 43 | 458 | 501 |
|  | $\%$ of Grand Total | $0.6 \%$ | $6.6 \%$ | $7.2 \%$ |
|  |  |  |  |  |
| Column Total | Count | 1,219 | 5,729 | 6,948 |
|  | $\%$ of Grand Total | $17.5 \%$ | $82.5 \%$ | $100.0 \%$ |

Asian and Hispanic students were oversampled in NELS: 88, so the sample contained higher proportions of these ethnic groups than did the reference population. Sampling weights must be applied to the observations to adjust for the over-sampling. In contrast, a chi-square table without weights or design effects is appropriate for a simple random sample because each observation represents the same number of cases in the population.

The variable F3TRSCWT, a raw expansion weight, is used as the weight in Table 2. This is one of several raw expansion weights provided by NCES, and it is the weight that is to be used when analyzing variables from the 1994 follow-up (e.g., RMATH) in conjunction with
high school transcript variables such as F2RHMA_C. The raw expansion weight is the number of cases in the population that the observation represents. Unlike simple random sampling, the weights are not the same for each subject. The weights for these 6,948 observations range from 7 to 12,940 with a mean of 228.50 . The total count of $1,587,646$ in this table represents the number of students from the 1988 eighth-grade cohort that met the selection criteria. This table contains correct population counts and percentages in the cells; however, the overall chi-square $(27,500.88)$ is too high because the cell sizes are overstated. The cell sizes represent counts of the population rather than the sample.

Table 2. Contingency Table With Raw Expansion Weight F3TRSCWT: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^{2}(4)=27,500.88$, C $=.130$, $95 \%$ CI (.128, .132), C* = .184, 95\% CI (.182, .186).

|  |  | PSE Remedial Math |  |  |
| :--- | :--- | :---: | :---: | :---: |
| Units HS Math |  | Yes | No | Row Total |
|  |  |  |  |  |
| $0-1.99$ | Count | 24,353 | 63,532 | 87,885 |
|  | $\%$ of Grand Total | $1.5 \%$ | $4.0 \%$ | $5.5 \%$ |
|  |  |  |  |  |
| $2-2.99$ | Count | 53,767 | 167,485 | 221,252 |
|  | $\%$ of Grand Total | $3.4 \%$ | $10.5 \%$ | $13.9 \%$ |
|  |  |  |  |  |
| $3-3.99$ | Count | 118,230 | 427,763 | 545,993 |
|  | $\%$ of Grand Total | $7.4 \%$ | $26.9 \%$ | $34.4 \%$ |
|  |  |  |  |  |
| $4-4.99$ | Count | 81,325 | 537,884 | 619,209 |
|  | $\%$ of Grand Total | $5.1 \%$ | $33.9 \%$ | $39.0 \%$ |
|  |  |  |  |  |
| $8=5$ | Count | 14,951 | 98,356 | 113,307 |
|  | $\%$ of Grand Total | $0.9 \%$ | $6.2 \%$ | $7.1 \%$ |
|  |  |  |  |  |
| Column Total | Count | 292,626 | $1,295,020$ | $1,587,646$ |
|  | $\%$ of Grand Total | $18.4 \%$ | $81.6 \%$ | $100.0 \%$ |

The relative weight of F3TRSCWT is used in Table 3 to bring the cell counts in Table 2 back into congruence with the sample counts. For each of the 6,948 observations, the relative
weight of F3TRSCWT is computed by dividing F3TRSCWT by 228.50, which is the mean of F3TRSCWT for the 6,948 observations. The total count in Table 3 is 6,947 , which differs
from Table 1 only because of rounding (note: although displayed in whole numbers by SPSS, Table 3 actually contains fractional numbers of observations in each cell). Table 3 contains
correct cell percentages, but the cell sizes and chi-square (120.62) are overstated due to the two-stage clustered sample design of NELS: 88.

Table 3. Contingency Table With Relative Weight = F3TRSCWT / 228.5. Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^{2}(4)=120.62$, C= .131, 95\% CI (.107, .155), C* = .185, 95\% CI (.161, .209).

|  |  | PSE Remedial Math |  |  |
| :--- | :--- | :---: | :---: | :---: |
| Units HS Math |  | Yes | No | Row Total |
|  |  |  |  |  |
| $0-1.99$ | Count | 107 | 278 | 385 |
|  | \% of Grand Total | $1.5 \%$ | $4.0 \%$ | $5.5 \%$ |
|  |  |  |  |  |
| $2-2.99$ | Count | 235 | 733 | 968 |
|  | $\%$ of Grand Total | $3.4 \%$ | $10.6 \%$ | $13.9 \%$ |
|  |  |  |  |  |
| $3-3.99$ | Count | 517 | 1,872 | 2,389 |
|  | $\%$ of Grand Total | $7.4 \%$ | $26.9 \%$ | $34.4 \%$ |
|  |  |  |  |  |
| $4-4.99$ | Count | 356 | 2,354 | 2,710 |
|  | $\%$ of Grand Total | $5.1 \%$ | $33.9 \%$ | $39.0 \%$ |
|  |  |  |  |  |
| $>=5$ | Count | 65 | 430 | 495 |
|  | $\%$ of Grand Total | $0.9 \%$ | $6.2 \%$ | $7.1 \%$ |
|  |  |  |  |  |
| Column Total | Count | 1,280 | 5,667 | 6,947 |
|  | $\%$ of Grand Total | $18.4 \%$ | $81.6 \%$ | $100.0 \%$ |

Table 4 was obtained by dividing the relative weight for F3TRSCWT by the NELS: 88 average DEFF (2.94), extrapolated via Taylor series methods, which resulted in effective cell sizes with correctly weighted cell counts and proportions and the appropriate overall chisquare (40.81) for this clustered design. The counts in Table 4 are the effective sample size after accounting for the clustered sample design (i.e., a sample of 6,948 from this clustered design is equivalent to a sample size of 2,363 randomly selected students). Essentially, a mean DEFF of 2.94 tells us that if a SRS design had been conducted, only $33 \%$ as many subjects when compared against a CSD, would have been necessary to observe the statistic of study.

DEFFs that range between 1.0 and 3.0 tend to be indicative of a well-designed study. The current study's DEFF of 2.94 indicated that
the variance of the NELS: 88 estimates was increased by 194\% due to variations in the weights. The square root of DEFF, the DEFT, yields the degree by which the standard error has been increased by the CSD. The DEFT (1.71) implied that the standard error was 1.71 times as large as it would have been had the present results been realized through a SRS design, or the standard error was increased by $71 \%$. An intra-class correlation coefficient (ICC) of .20 or less is desirable for indicating the level of association between the responses of the members in the cluster. Since an ICC was not used in the computation of the Taylor seriesderived average DEFF for NELS: 88, an estimated, average ICC was calculated from the following formula for determining

$$
\begin{equation*}
\text { DEFF: } 1+\delta(\mathrm{n}-1), \tag{8}
\end{equation*}
$$

where $\delta$ is the ICC and $n$ is the typical size of a cluster (Flores-Cervantes, Brick, \& DiGaetano, 1999). The low ICC (.0844) indicated that the members in the same cluster were only about $8 \%$, on average, more probable of having
corresponding characteristics than if compared to another member selected randomly from the population.

Table 4. Contingency Table With Weight $=($ F3TRSCWT / 228.5) / 2.94: Carnegie Units of High School Math by Postsecondary Education (PSE) Remedial Math. $\chi^{2}(4)=40.81$, C $=.130,95 \%$ CI (.090, .170), $C^{*}=.184,95 \%$ CI (.144, .224).

|  |  | PSE Remedial Math |  |  |
| :--- | :--- | :---: | :---: | :---: |
| Units HS Math |  | Yes | No | Row Total |
|  |  |  |  |  |
| $0-1.99$ | Count | 36 | 95 | 131 |
|  | $\%$ of Grand Total | $1.5 \%$ | $4.0 \%$ | $5.5 \%$ |
|  |  |  |  |  |
| $2-2.99$ | Count | 80 | 249 | 329 |
|  | $\%$ of Grand Total | $3.4 \%$ | $10.5 \%$ | $13.9 \%$ |
|  |  |  |  |  |
| $3-3.99$ | Count | 176 | 637 | 813 |
|  | $\%$ of Grand Total | $7.4 \%$ | $27.0 \%$ | $34.4 \%$ |
|  |  |  |  |  |
| $4-4.99$ | Count | 121 | 801 | 922 |
|  | $\%$ of Grand Total | $5.1 \%$ | $33.9 \%$ | $39.0 \%$ |
|  |  |  |  |  |
| $8=5$ | Count | 22 | 146 | 168 |
|  | $\%$ of Grand Total | $0.9 \%$ | $6.2 \%$ | $7.1 \%$ |
|  |  |  |  |  |
| Column Total | Count | 435 | 1,928 | 2,363 |
|  | $\%$ of Grand Total | $18.4 \%$ | $81.6 \%$ | $100.0 \%$ |

NELS: 88 used a clustered sample design, in which schools were randomly selected, and then students within those schools were randomly selected. Students selected from such a sampling design would be expected to be more homogeneous than students selected from a simple random design across all schools. The chi-square values from SPSS cross-tabulations and SAS Proc Freq tables presume simple random samples. One method for estimating the proper chi-square for the two variables under investigation from NELS: 88 is to divide the relative weight for F3TRSCWT by the average DEFF (2.94), and use the result as the weight in SPSS cross-tabulations or SAS Proc Freq. The results in Table 4 were obtained by such a computation, which yields effective cell sizes and correctly weighted proportions.

Furthermore, the chi-square (40.81) is an appropriate approximation of the true chi-square for this clustered design. These are the values that should be used in a chi-square analysis of Carnegie Units of high school math by whether or not a student took a postsecondary education remedial math course. Notice that the cell counts and the total count in Table 4 are equal to those in Table 3 divided by 2.94. The counts in Table 4 are the effective sample size after accounting for the clustered sample design.

As was found with the chi-square statistics, weighting, or lack thereof, also influenced effect size values. For example, the coefficient of contingency and Sakoda's Adjusted C in Table 1, where the default of no weighting occurred, had higher values than any of the reported C or $\mathrm{C}^{*}$ estimations where a
form of weighting transpired. It should be noted that the C values ranged from . 130 to .136 , or in the case of adjusted C from .184 to .192 , which means that regardless of weighting scheme, or none at all, the practical implication of the chisquare statistics of study was that they had a small effect. Thus, although the chi-square statistics were all statistically significant, they had a small effect, which indicates that the results derived from the chi-square statistics would not be deemed very important practically and also in terms of accounting for much of the total variance of the outcome.

## Conclusion

Some sampling designs over-sample certain groups (i.e., their proportion in the sample is greater than their proportion in the population) in order to obtain sufficiently large numbers of observations in these categories so that statistical analyses can be conducted separately on these groups. When analyzing the entire sample, relative weights should be used to bring the sample proportions back in congruence with the population proportions. When clustered sampled designs are used, then relative weights should be divided by the DEFF to adjust for the fact that a sample from a clustered design is more homogeneous than if a simple random sampling scheme had been employed. The chi-square values from SPSS cross-tabulations and SAS Proc Freq tables presume simple random samples. Design effects must be used with such software in order to obtain an appropriate approximation for the true chi-square, and its accurate effect size, of a clustered design.

## References

Broene, P., \& Rust, K. (2000). Strengths and limitations of using SUDAAN, Stata, and WesVarPC for computing variances from NCES data sets. (U.S. Department of Education, National Center for Education Statistics Working Paper No. 2000-03). Washington, DC: U.S. Department of Education.

Carlson, B. L., Johnson, A. L., \& Cohen, S. B. (1993). An evaluation of the use of personal computers for variance estimation with complex survey data. Journal of Official Statistics, 9, 795-814.

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

Cohen, S. B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistician, 51, 285-292.

Ferguson, G. A. (1966). Statistical analysis in psychology and education (2nd ed.). New York: McGraw-Hill.

Flores-Cervantes, I., Brick, J. M., \& DiGaetano, R. (1999). Report no. 4: 1997 NSAF variance estimation. http://newfederalism.urban. org/nsaf/methodology1997.html

Kalton, G. (1983). Introduction to survey sampling. In J. L. Sullivan \& R. G. Niemi (Series Eds.), Quantitative applications in the social sciences. Beverly Hills, CA: Sage Publications.

Kish, L. (1965). Survey sampling. New York: John Wiley \& Sons.

Lee, E. S., Forthofer, R. N., \& Lorimor, R. J. (1989). Analyzing complex survey data. In M. S. Lewis-Beck (Series Ed.), Quantitative applications in the social sciences. Newbury Park, CA: Sage Publications.

McMillan, J. H., \& Schumacher, S. (1997). Research in education: A conceptual introduction (4th ed.). New York: Longman.

Muthen, B. O., \& Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), Sociological methodology (pp. 267-316). Washington, DC: American Sociological Association.

Sprinthall, R. C. (2000). Basic statistical analysis (6th ed.). Boston, MA: Allyn and Bacon.

Thomas, S. L., \& Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. Research in Higher Education, 42, 517-540.
U.S. Department of Education. (1996). Methodology report, National Education longitudinal study: 1988-1994 (NCES 96-174). Washington, DC: Author.

# Correcting Publication Bias In Meta-Analysis: A Truncation Approach 

Guillermo Montes Bohdan S. Lotyczewski<br>Children's Institute, Inc.

Meta-analyses are increasingly used to support national policy decision making. The practical implications of publications bias in meta-analysis are discussed. Standard approaches to correct for publication bias require knowledge of the selection mechanism that leads to publication. In this study, an alternative approach is proposed based on Cohen’s corrections for a truncated normal. The approach makes less assumptions, is easy to implement, and performs well in simulations with small samples. The approach is illustrated with two published meta-analyses.

Key words: Meta-analysis, methods, truncation, publication bias

Introduction
Publication bias presents possibly the greatest methodological threat to the validity of a meta-analysis. It can be caused by the biased and selective reporting of the results of a given study, or, more seriously, by the selective decision to publish the results of the study in the first place. Undetected publication bias is especially serious owing to the fact that the meta-analysis may not only lead to a spurious conclusion, but the aggregation of data may give the impression, with standard statistical methodology, that the conclusions are very precise. (Cooper \& Hedges, 1994, p. 407).

With these words, Cooper and Hedges (1994) concluded their discussion on the detection and correction of publication bias in meta-analysis. For all its theoretical and practical importance, it is not often that one sees a meta-analysis corrected for publication bias.

Address correspondence to: Guillermo Montes, Children's Institute, Inc., 274 N. Goodman St., Suite D103, Rochester, NY, 14607. E-mail: gmontes@ childrensinstitute.net. Telephone: 585.295.1000 ext 227.

Undoubtedly, the reason is that the methodology available to the address the problem (Vevea \& Hedges, 1995; Hedges \& Vevea, 1996; Cleary \& Casella, 1997) is complex, not easily accessible to the average meta-analyst practitioner and has been unable to make a strong practical case for supporting its use. The problem is difficult because publication bias, by its own nature, is a phenomena we know little about and because it does not suffice to show that, theoretically, a corrected estimate exists. One must show that the correction performs better than the original biased statistics in small samples.

In spite of these practical problems, the struggle against the effects of publication bias should not be abandoned. The presence of publication bias can lead to an erroneous consensus regarding the efficacy of a class of interventions or the importance of a particular factor in a psychological process of interest. Moreover, because one cannot assume that the same level of publication bias exists across meta-analyses, even in related content areas, there is little solid ground on which to base comparisons across meta-analyses.

Not only is the scientific community in danger of conceding to the evidence what the evidence does not warrant; but often social scientists are called to testify to critical allocations of funds and to the implementation of far-reaching social policies. Meta-analytic evidence plays an increasing role in those policy
discussions as legislators and other policy makers demand simple summaries of complex information. Therefore, publication bias can also lead to harm in the public policy arena.

To be widely used, a method for correcting publication bias in meta-analysis must meet the following criteria: 1) It must recover the true population parameters in large samples, 2 ) it must be an improvement over the biased sample statistics in small samples, and 3) it must be relatively easy to calculate and easy to use for the average meta-analytic practitioner.

Modeling Publication Bias: Two Approaches
Traditional approaches to correct metaanalysis require some model for observed effect sizes that incorporates the selection process. Two aspects to such a model are given, the selection model and the effect size model. (Hedges \& Vevea, 1996). Typically, the effect size model has been constructed using the random effects model and assuming a normal compound distribution. The selection process is modeled as a complex weight function of the probability of obtaining significant results based on sample size. This approach is based on the notion that publication bias is directly related to the presence of significant results.

This approach, commonplace in the literature, has a number of problems. First, it is unclear whether significance is the only criteria that impacts publication bias, effect sizes may be equally important, particularly when the sign is unexpected. Second, the selection process is an unknown and complex social phenomenon. Modeling publication bias as a function of a process we know little about seems unwise.

An alternative approach is to use a simple truncation model, based not on statistical significance but on effect size. After all if publication bias is having an impact on the overall results of a meta-analysis is because the bias is systematically truncating one of the tails, typically the left tail, of the distribution of program effects.

Because the standard approach assumes normality, modeling publication bias with a truncated normal model may be a practical alternative to modeling selection processes without imposing additional unverified
assumptions; at least until the selection processes are better understood.

The truncation approach is more practical than the standard approach for three reasons. First, detecting publication bias becomes an exercise in elementary statistics. Is the observed distribution of effects normal or it is missing one of the tails? Both a standard histogram of the observed distribution and the computation of the distance between the median and the mean in standard deviation units can be used to answer this question.

Second, although we provided a rationale for our approach, the truncation model does not require us to specify a selection mechanism or to know how publication bias occurs. All we need to know is that there were no published studies below a particular effect size, and that the observed distribution of effects is skewed to the right. Truncation relies exclusively on the assumption of normality of the effect size model.

Third, it simplifies the correction for publication bias considerably because it uses a long-time developed method already in use in other disciplines as the standard way to deal with the statistics of truncated phenomena. Since 1959 engineers, economists, cosmologists and physicists have used Cohen's (1959) estimates for the population mean and standard deviation of a truncated normal to investigate truncated phenomena.

Cosmologists observe only the brightest stars, engineers observe only products that meet tolerance checks, economists observe only portions of the income distribution and particle physicists observe only the energy signature of higher energy particles. Similarly, highly effective programs are likely to be observed in the published literature while less effective interventions with non-significant or negatively significant results are likely to become unavailable results. Meta-analysis can benefit from the research and development of truncation-related statistics in other fields. These include truncation regression, correction for doubly truncated normals and many others (Greene, 1990).

Correcting for Publication Bias in MetaAnalysis

Assume that the distribution of effects is normal. One can model the distribution of effects in a variety of ways but the simplest method is to posit a compound normal where each study would be a realization of a normal distribution with mean $\Delta$. Where $\Delta$ represents the true effect sizes of each actual intervention. Yet, each true intervention effect size $\Delta$ is itself a random variate of a normal distribution with mean $\mu$. $\mu$ represents the true effect size of a class of interventions.

The resulting distribution of effects is a compound normal distribution:

$$
N(\Delta, \sigma) \underset{\Delta}{\wedge} N\left(\mu, \sigma^{\prime}\right) .
$$

It can be shown (Johnson, Koptz, \& Balakrishnan, 1994) that such distribution is also normal with $N\left(\mu, \sqrt{\sigma^{2}+\sigma^{\prime 2}}\right)$.

Consider now the presence of publication bias. Because of the reasons described above, effect sizes below some level T are unlikely to be published. The resulting observable distribution of effect sizes will be a truncated normal.

Truncation of the left tail of a normal distribution produces the following effects: 1) the sample mean will overestimate the true mean, and 2) the sample standard deviation will underestimate the true standard deviation.

In other words, publication bias will result in the systematic overestimation of average effect sizes and the lowering of the associated standard deviation resulting in the illusion of precision that Cooper and Hedges (1994) described as one the greatest threats to the validity of meta-analysis.

## Correction for truncation

Cohen (1959) first developed estimation procedures to recover the mean and standard deviation from a truncated observed normal distribution. Equations 1-5 describe the process. First, calculate the left-hand side of equation 5, using the minimum observed value in the truncated distribution as a proxy variable for T .

Then solve for $\xi$ and calculate $\theta(\xi)$. There are two ways of making the process less painful. One can look up the value of $\theta(\xi)$ in Cohen's book (1991, Table 2.1.) Alternatively, one can use a numerical solver, now standard in many applications, to numerically solve for $\xi$.

Once $\theta(\xi)$ is known, calculate $\mu_{\mathrm{C}}$ and $\sigma_{C}^{2}$ using equations 1 and 2 . Note that the estimated degree of truncation is simply $\Phi(\xi)$.

$$
\begin{align*}
& \mu_{R}=\bar{x}-\theta(\xi)(\bar{x}-T)  \tag{1}\\
& \sigma_{R}^{2}=s^{2}+\theta(\xi)(\bar{x}-T)^{2}  \tag{2}\\
& \xi=\frac{T-\mu_{R}}{\sigma_{R}}  \tag{3}\\
& \theta(\xi)=\frac{Q(\xi)}{Q(\xi)-\xi} \\
& Q(\xi)=\frac{\phi(\xi)}{1-\Phi(\xi)} \\
& \frac{s^{2}}{(\bar{x}-T)^{2}}=\frac{1-Q(\xi)(Q(\xi)-\xi)}{(Q(\xi)-\xi)^{2}}
\end{align*}
$$

Cohen's formulas to calculate the $95 \%$ confidence interval around the mean and standard deviation are:

$$
\begin{aligned}
& \phi_{11}=1-Q(Q-\xi) \\
& \phi_{12}=Q(1-\xi(Q-\xi)) \\
& \phi_{22}=2+\xi \phi_{12} \\
& \mu_{11}=\frac{\phi_{22}}{\phi_{11} \phi_{22}-\phi_{12}^{2}} \\
& \mu_{22}=\frac{\phi_{11}}{\phi_{11} \phi_{22}-\phi_{12}^{2}} \\
& V\left(\mu_{R}\right)=\frac{\sigma_{R}^{2}}{N} \mu_{11} \\
& V\left(\sigma_{R}\right)=\frac{\sigma_{R}^{2}}{N} \mu_{22} \\
& 95 \% C I \mu_{R}= \\
& \left(\mu_{R}-2 \sqrt{V\left(\mu_{R}\right)}, \mu_{R}+2 \sqrt{V\left(\mu_{R}\right)}\right)
\end{aligned}
$$

where Q is evaluated at $\xi$.

## A Large-Sample Example

To illustrate the process, consider a class of interventions whose true effect size is 0.4 with a 0.8 standard deviation. Because of the large standard deviation, if 2000 studies were performed on this class of interventions one would expect that $20 \%$ of the studies would have negative results, some of them with considerable effect sizes (e.g., -0.8 and below).

Assume that there is no theoretical explanation for a negative effect size, so studies showing negative effect sizes are unlikely to be published. In the random sample we generated, that would leave 1393 publishable studies with some of them reporting non-significant results.

A meta-analysis performed on the 1393 studies would yield a biased sample mean of 0.82 with a biased standard deviation of 0.55 . By all accounts, this class of interventions would be deemed to have large effects. Cohen's corrected estimates are 0.475 [95\% CI (0.3723,0.5776)] for the mean and 0.77 [95\% CI (0.709-0.829)] for the standard deviation. As can be seen, these estimates are quite close to the true values of 0.4 and 0.8 .

As mentioned before, once the original mean and standard deviations have been recovered one can calculate the degree of truncation by simply calculating the value of the cumulative normal with the recovered mean and standard deviation at the truncation point, $\Phi(\xi)$. In this case, the degree of truncation was 26.84\%.

## Behavior of the Estimator in Small Samples

For Cohen's estimates to be useful in the correction of meta-analysis publication bias they need perform adequately in small samples. The standard criteria of using 95\% confidence intervals does not seem appropriate in this small sample context. Some severely truncated samples will have sample sizes of below 15 observations and, therefore we expect that the 95\% confidence interval of the corrected mean will contain the biased sample estimate. Other approaches to correct publication bias have the same problem (Vevea \& Hedges, 1995). Therefore, we studied the direct improvement of using Cohen's formulas in terms of distance to the true parameters.

The population parameters were picked to represent meta-analytic results of importance both for scientific and policy purposes. We chose a large effect (0.8) with a relatively small standard deviation (0.4) and a total sample size of 100 published and unpublished studies (of which only a few will be published under high truncation).

Maxwell and Cole (1995) stated that "simulation studies are experiments and must be described and interpreted in this light". Therefore, we will use the language of experiments to describe our simulations. Table 1 shows the result of an experiment designed to answer seven questions and analyze how the answers vary as the truncation level increases:

- Question 1: What is the average sample bias for $\mu$ ?
- Question 2: What is the average sample bias for $\sigma$ ?
- Question 3: What would be the average number of studies published?
- Question 4: What is the average error in correction for $\mu$ using Cohen's estimates?
- Question 5: What is the average error in correction for $\sigma$ using Cohen's estimates?
- Question 6: On average, by how much do we benefit by performing the correction?
- Question 7: In what percentage of samples would the meta-analyst practitioner benefit from using Cohen's estimates?

To answer these questions we simulated 10,000 samples of a normal distribution of effect sizes mean $=0.8$ and sd $=0.4$. We then truncated it to create an observed distribution. We used the four different points of truncation ranging from almost no truncation (2 standard deviations below the mean) to severe truncation (one standard deviation above the mean). Then we used Cohen's (1959) formulas to estimate the corrected mean and standard deviation.

To answer questions 6 and 7 we defined an improvement measure as the ratio of two distances. The numerator is the distance between
the sample moment from the biased distribution and the corresponding true value. The denominator is the distance between the corrected estimate and the true value. We used the absolute value measure of distance (although in the next section we also ran simulation with the Euclidean distance without substantial differences).

$$
\operatorname{IMPROVE}_{\mu}=\frac{\operatorname{Dist}(\bar{x}, \mu)}{\operatorname{Dist}\left(\mu_{R}, \mu\right)}=\frac{|\bar{x}-\mu|}{\left|\mu_{R}-\mu\right|}
$$

for the mean; and

$$
\operatorname{IMPROVE}_{\sigma}=\frac{\operatorname{Dist}(s, \sigma)}{\operatorname{Dist}\left(\sigma_{R}, \sigma\right)}=\frac{|s-\sigma|}{\left|\sigma_{R}-\sigma\right|}
$$

for the standard deviation.
An improvement factor below one indicates that the correction gets us farther way from the true value, an improvement factor of one indicates that the correction does as badly as
the biased sample moments; finally, improvement factors higher than one indicate how much closer the correction for truncation gets us to the true mean (e.g. a value of 2 would indicate that Cohen's correction gets us two times closer to the real mean than the biased estimates do).

Since the improvement factors are always positive, their distribution is not likely to be symmetric; therefore, we report the median improvement, as the preferred measure of central tendency. This median will be the answer provided to question 6.

Because it is possible to have large average improvements while the majority of the samples would not be improved by using Cohen's corrections, question 7 asks the proportion of the 10,000 that benefit from the correction. Benefit is defined as having an improvement factor strictly higher than one. It is a measure of the risk that an average metaanalyst practitioner incurs by correcting the estimates of her study.

Table 1. Results of Experiment 1.

|  | Almost <br> none | Mild | Serious | Severe |
| :--- | :---: | :---: | :---: | :---: |
| Truncation point (T) | $\mu-2 \sigma=0$ | $\mu-\sigma=0.4$ | $\mu=0.8$ | $\mu+\sigma=1.2$ |
| Truncation level $\Phi(\mathrm{T})$ | 0.023 | 0.1586 | 0.5 | 0.841 |
| Average Observed Sample Size | 97.73 | 84.12 | 50.03 | 15.88 |
| Average Sample Bias (for $\mu$ ) | 0.022 | 0.114 | 0.319 | 0.610 |
| Average Sample Bias(for $\sigma$ ) | -0.024 | -0.084 | -0.161 | -0.227 |
| Average Error in Correction (for $\mu$ ) | -0.017 | -0.040 | -0.150 | 0.116 |
| Average Error in Correction (for $\sigma$ ) | 0.0123 | 0.016 | 0.033 | -0.056 |
| Median Improvement factor (for $\mu)$ | 1.1724 | 2.147 | 1.889 | 4.233 |
| Median Improvement factor (for $\sigma$ ) | 1.375 | 2.378 | 2.202 | 2.597 |
| \% of Samples that benefited $(\mu)$ | $52.83 \%$ | $75.84 \%$ | $72.70 \%$ | $100 \%$ |
| \% of Samples that benefited $(\sigma)$ | $56.68 \%$ | $80.90 \%$ | $80.80 \%$ | $96.4 \%$ |
| 95\% CI range | 0.48 | 0.939 | 0.921 | 11.29 |
| Does CI contain sample mean? | $100 \%$ | $100 \%$ | $100 \%$ | $100 \%$ |
| Does CI contain true mean? | $96.43 \%$ | $96.93 \%$ | $92.08 \%$ | $85.68 \%$ |
| Simulations based on 10,000 <br> level. random samples from the normal | $(0.8,0.4)$ for each truncation |  |  |  |

Answer to Question 1: The overestimation of $\mu$ increases with truncation level ranging from 0.02 to 0.609 .

Answer to Question 2: The underestimation bias of the standard deviation increases as the truncation gets progressively
worse, but it does so at a slower rate than the sample mean. It ranges from -0.02 to -0.22 .

Answer to Question 3: The observed sample size varies form 97 (almost the 100 possible publications) to about 15 in the case of severe truncation. It is, of course, a linear function of the truncation level.

Answer to Question 4: The average errors made in correcting for $\mu$ ranged form 0.02 to 0.15 in absolute value, roughly increasing in a non-linear manner with the truncation level. At all levels of truncation, the average correction error was smaller than the corresponding average bias.

Answer to Question 5: The average error made in correcting for $\sigma$ ranged from 0.01 to 0.05 , roughly increasing in a non-linear manner with the truncation level. Answer to question 5. At all levels of truncation, the average correction error was smaller than the corresponding average bias.

Answer to Question 6: The median improvement from using the correction ranges form 1.17 to 4.23. In other words, Cohen's estimation method got us anywhere from 1.17 to four times closer to the true mean. The improvement function is a nonlinear function of the truncation level, increasing with the truncation level at early stages of truncation, decreasing until past the 0.5 truncation level to quickly ascend again.

The median improvement for the standard deviation ranged from 1.27 to 2.59 . Again, the function is nonlinear with truncation level, although less dramatically non-linear than the improvement for the mean was.

Answer to Question 7: Regardless of the level of truncation, the correction for both $\mu$ and $\sigma$ was beneficial in more than half of the cases. With mild truncation the proportion of samples that benefited from the correction were over $75 \%$, there was a small decrease in the proportion of samples that benefit as truncation nears the 0.5 point and then a dramatic increase so that for serious truncation virtually all samples benefited form Cohen's correction. This
nonlinear risk function was carefully investigated in the next section.

When almost no truncation is present (truncation level of 0.02 ) slightly half of the samples did not benefit from Cohen's correction. At that small level of truncation, however, both the error of the correction and the bias are unlikely to have substantial scientific or policy implications. As truncation increases, both the chances of benefiting from using Cohen's correction and the improvement in terms of distance to the true parameters are sizeable. Therefore, if truncation is detected, the use of Cohen's estimates seems warranted even for small sample sizes.

We now turn our attention to investigating in detail how the proportion of samples that benefit from correction increase as a nonlinear function of truncation level.

Proportion of Samples that Benefit from Cohen's Correction as a Function of Truncation

The experiment of the previous section yielded that the proportion of samples that benefit from Cohen's corrections were nonlinear functions of the truncation level $\Phi(\xi)$. To investigate these nonlinear functions further we generated 1000 random samples ( $\mu=0.8, \sigma=0.4$, $\mathrm{N}=100$ ) for each of 121 levels of truncation ranging from $\mathrm{T}=\mu-2 \sigma=0$ to $\mathrm{T}==\mu+\sigma=1.2$ at 0.01 intervals. We then plotted the percentage of samples that benefit from Cohen's correction for $\mu$ as a function of the truncation level, and proceeded similarly for $\sigma$. We repeated the process for different values of $\mu$ and $\sigma$ but the nonlinear pattern remained essentially unchanged.

We employed the absolute value distance function in our improvement measure as before; but also generated a complete independent set of random samples and calculated the improvement factors using standard Euclidean distance function. Figure 1 and 2 show the results.

Figure 1. Samples Improved By Correction: The Mean

## Samples Improved by Correction



Based on 121 sets of 1000 samples with sample size 100
Distances calculated on two independent runs.

Figure 2: Samples Improved By Correction: The Standard Deviation.

## Samples Improved by Correction

By Type of Distance Used


Based on 121 sets of 1000 samples with sample size 100
Distances calculated on two independent runs.

Note the following patterns: 1) for all truncation levels, the proportion of samples that benefit from Cohen's correction for both $\mu$ and $\sigma$ was over $50 \%, 2$ ) for mild truncation levels, the proportion of samples that benefit from Cohen's correction increases quite rapidly until about $\Phi(\xi)=0.25,3)$ in the case of $\mu$, the proportion of samples decreases until $\Phi(\xi)=0.65$ truncation level to then rise dramatically to $100 \%$, and 4 ) in the case of $\sigma$, the proportion of samples stabilizes at about $80 \%$ until past the $\Phi(\xi)=0.5$ truncation level to then rise dramatically to almost $100 \%$.

Therefore, Cohen's estimates perform adequately in small samples, with over $60 \%$ chance of obtaining a better point estimate through Cohen's estimation method. The correction seems to be particularly beneficial for
the mild levels $(\Phi(\xi) \approx 0.2)$ of truncation commonly believed to be present in metaanalysis.

## Illustrative Examples

To demonstrate the applicability of the method we have chosen two meta-analysis. The meta-analysis were previously published by Psychological Bulletin and contained the necessary data to make the corrections. We are not presenting the corrections as substantive revisions, but simply as illustrations of the method. The two meta-analyses show different levels of truncation.

Example 1: Mild Truncation
The first example is taken from table 3 of Yirmiya, et al. (1998) meta-analysis
comparing theory of mind abilities of individuals with autism, individuals with mental retardation and normally developing individuals. The data used here refers only to the comparison of individuals with autism versus normally developing individuals. The authors report different average statistics because they calculated a weighted average. We had no information to replicate the weights (sample size of the studies).

There were 22 effect sizes, with sample mean 1.1173, standard deviation 0.9667 , median 1.030 and minimum value -0.40 . The authors report different numbers because they used a weighted function to calculate average effect sizes.

The histogram of the observed distribution and the fact that the median was larger than the sample mean revealed mild truncation on the left size. Cohen's corrections are as follows: Corrected $\mu=0.689$, corrected $\sigma$ $=1.258$. Degree of truncation 0.1933. Therefore, in this case the correction would cast some doubt on the average large effect differential between normally developing individuals and those with autism.

## Example 2: No Truncation

The second example is taken from Appendix A of Rind, Tromovitch, and Bauserman's (1998) controversial meta-analysis on the assumed consequences of child sexual abuse using college samples. This is an example of real-world research in which it was easier to explain the lack of significant positive findings by using a number of methodological and theoretical arguments. Because of this, one would expect less truncation to have occurred.

Using the 56 studies, the average effect size is 0.0953 , with a standard deviation of 0.0947 and a minimum observation of -0.25 . The histogram revealed little or no truncation, as did the fact that the median was almost identical to the sample mean. The corrected mean was 0.09531 . The point estimate is essentially identical to the uncorrected mean. The corrected standard deviation is 0.0948 . The estimated degree of truncation was only 0.0001 .

This example illustrates how some meta-analysis may suffer very little from publication bias because negative and positive
results are interpretable in the context of new theories or methodological issues. It is also suggestive that at least part of the controversy regarding diverse meta-analytic findings from several types of studies may be due to the degree of publication bias.

## Conclusion

Publication bias is an important threat to the validity of meta-analysis. It can lead to error regarding the efficacy of classes of interventions or the importance of particular factors in psychological processes. These errors can have a detrimental effect on both scientific knowledge and on public policy. Therefore, it is important to find some correction, even if imperfect, to the problem.

First, modeling publication bias by estimating a selection function of what remains a fundamentally unknown process seems to us unwise. Selection rules are likely to vary depending on the nature of the study, the availability of theoretical and methodological explanations for the unexpected result, the other results in the study, a very complex web of reputation and financial incentives, and the larger context of scientific or popular debate on the content of the study. Therefore, to correct publication bias using a selection approach one either needs to know the complexity of how the publication bias originated or oversimplify the problem substantially by using a simple mechanical rule. In either case, one is likely to impose additional assumptions on the data.

We make a case for using truncation instead of selection as a method to correct for publication bias on practical grounds: truncation does not require any additional assumptions beyond the normality of the effect size distribution; in particular it does not require us to know how the publication selection took place. Truncation is easy to detect in practice by looking at simple statistics like the difference between the median and the mean or plotting a histogram. It can be corrected by well-developed estimators currently in use by other disciplines with the attendant benefits of on-going research and development in the area.

In addition, our simulations demonstrate that in the small samples typical of meta-analytic
studies Cohen's correction performs adequately. In cases of mild truncation (defined as around $20 \%$ ), the proposed correction will, on average, get point estimates that are two times closer to the true parameters, and the correction will benefit over $70 \%$ of the samples. Therefore, the odds favor making the correction. The size of the correction is likely to have a substantial impact on the interpretation of the results.

Certainly, this approach is not perfect. The truncation approach is presented simply as an approximation to the real underlying structure of publication bias. Yet, complicating the statistics in favor of a more accurate portrayal of the underlying structure, given our wide ignorance of the phenomena and the increasing complexity of the statistics, seems to us not be a practical approach to a problem that has important policy ramifications. Given the seriousness of the potential damage publication bias may be doing both to social science and to public policy finding some correction procedure that requires minimal assumption and is easy to use seems to us as a more responsible course of action than ignoring the problem until a complete solution has been found.

## References

Cleary, R. J., \& Casella, G. (1997). An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias. Journal of Educational and Behavioral Statistics, 22, 2, 141-154.

Cohen, A. C. (1991). Truncated and Censored Samples: Theory and Applications. New York, NY: Marcel Dekker Inc.

Cohen, A. C. (1959). Simplified Estimators for the Normal Distribution When Samples are Singly Censored or Truncated. Technometrics, 1, 3, 217-237.

Cooper, H., \& Hedges, L. V. (Eds) (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Greene, W. H . (1990). Econometric Analysis. New York: MacMillan Publishing Company.

Hedges, L. V., \& Vevea, J. (1996). Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model. Journal of Educational and Behavioral Statistics, 21, 4, 299-332.

Johnson, N. L., Kotz, S., \& Balakrishnan, N. (1994). Continuos Univariate Distributions. V. 1. $2^{\text {nd }}$ edition. NY: John Wiley \& sons, Inc.

Maxwell, S. E., \& Cole, D. A. (1995). Tips for Writing (and Reading) Methodological Articles. Psychological Bulletin, 118, 2,193-198.

McEachern, W. A. (1994). Economics: A Contemporary Introduction. Cincinnati, OH : South-Western Publishing Company.

Rind, B., Tromovitch, P., Bauserman, R. (1998). A meta-analytic examination of assumed properties of child sexual abuse using college samples. Psychological Bulletin, 124, 22-53.

Vevea, J. L., \& Hedges, V. (1995). A General Linear Model for Estimating Effect Size in the Presence of Publication Bias. Psychometrika, 60, 3, 419-435.

Yirmiya, N., Osnat, E., Shaked, M., \& Solomonica-Levi, D. (1998). Meta-Analyses Comparing Theory of Mind Abilities of Individuals With Autism, Individuals With Mental Retardation, and Normally Developing Individuals. Psychological Bulletin, 124, 3, 283307.

# Comparison Of Viral Trajectories In Aids Studies By Using Nonparametric Mixed-Effects Models 

Chin-Shang Li<br>Department of Biostatistics<br>St. Jude Children's Research Hospital<br>Ying-Hen Hsieh<br>Department of Applied Mathematics<br>National Chung-Hsing University<br>Taichung, Taiwan

Hua Liang<br>Department of Biostatistics<br>St. Jude Children's Research Hospital<br>Shiing-Jer Twu<br>National Health Research Institutes<br>Taipei, Taiwan

The efficacy of antiretroviral therapies for human immunodeficiency virus (HIV) infection can be assessed by studying the trajectory of the changing viral load with treatment time, but estimation of viral trajectory parameters by using the implicit function form of linear and nonlinear parametric models can be problematic. Using longitudinal viral load data from a clinical study of HIV-infected patients in Taiwan, we described the viral trajectories by applying a nonparametric mixed-effects model. We were then able to compare the efficacies of highly active antiretroviral therapy (HAART) and conventional therapy by using Young and Bowman's (1995) test.

Key words: AIDS clinical trial, HIV dynamics, longitudinal data, kernel regression, nonparametric mixed-effects model, viral load trajectory

## Introduction

Surrogate viral markers, such as the amount of HIV RNA in the plasma (the amount of HIV RNA in the patient's plasma represents the patient's viral load), currently play important

Chin-Shang Li and Hua Liang are Assistant Members, St. Jude Children's Research Hospital, Memphsis, TN, 38105. E-mail: chinshang.li@stjude.org; hua.liang@stjude.org. Ying-Hen Hsieh is Professor, Department of Applied Mathematics, National Chung-Hsing University, Taichung, Taiwan. Shiing-Jer Twu is Visiting Professor, NHRI Forum, National Health Research Institutes, Taipei, Taiwan. The research of C.S. Li and H. Liang was partially supported by Cancer Center Support Grant CA21765 from the National Institutes of Health and by the American Lebanese Syrian Associated Charities (ALSAC). Y.H. Hsieh was supported by grant DOH91-DC-1059 from the CDC of Taiwan. The authors thank Nanyu Wang for preparing the data analyzed.
roles in clinical research evaluating antiviral therapies for the acquired immunodeficiency syndrome (AIDS). Before HIV RNA assays were developed in mid-1990s, CD4+ cell counts served as the primary surrogate marker in AIDS clinical trials. Later, the amount of HIV RNA in the patient's plasma (viral load, measured as the copy number of the viral RNA) was shown to better predict the clinical outcome (Mellors et al., 1995; Mellors et al., 1996; Saag et al., 1996), and thus replaced CD4+ cell counts as the primary surrogate marker used in most AIDS clinical trials.

It is, therefore, important to characterize the trajectory that describes the change in viral load that occurs during antiviral treatment, because it is this trajectory that is commonly used to evaluate the efficacy of the treatment. For example, if the viral load reduces, we may infer that the treatment has successfully suppressed the replication of the virus. The differences between the viral loads resulting from different antiviral treatments may be used to compare the antiviral activities of the treatments. Appropriate analysis of the viral load
is therefore very important in HIV/AIDS drug development. In general, it is believed that the replication of the virus is suppressed at the beginning of an antiviral treatment, but recovery of the virus (called rebound) can occur in later stages of treatment, because of drug resistance or treatment failure. Some parametric models have been developed to describe the progression of AIDS phenomenologically; among the best known of these models are the exponential models (Ho et al., 1995; Wei et al., 1995). More recently, biomathematicians and biologists have proposed a variety of complicated models that include the use of differential equations. The use of these models has led to a deeper understanding of the pathogenesis of AIDS (e.g., Perelson \& Nelson, 1999; Wu and Ding, 1999).

In recent years, the necessity for appropriate models has gained more importance with the widespread use of highly active antiretroviral therapy (HAART) to treat HIV/AIDS (Ghani et al., 2003). Numerous studies have shown that HAART is effective in extending the time taken from the diagnosis of HIV-infection to AIDS or death in HIV-infected patients (e.g., Detels et al., 1998; Tassie et al., 2002) as well as reducing the likelihood of perinatal HIV transmission (Cooper et al., 2002). However, in many clinical practices, combination antiviral therapy has failed to completely and durably suppress HIV replication (e.g., Deeks et al., 1999).

To determine the efficacy of treatments in suppressing HIV replication in patients, the present study focuses on the following questions: (i) Given longitudinal viral load data, how can one identify a common feature of the antiviral activities of each treatment? (ii) How can we compare the antiviral efficacies of two different treatments? If we can answer question (ii), we may be able to demonstrate that the better treatment should be evaluated in a largescale clinical study. However, it may be difficult to answer these questions by using existing parametric or semi-parametric methods. To sufficiently consider all of the information available from the observations, and to avoid the misspecification of parametric modeling, we will use a nonparametric mixed-effects model to analyze the longitudinal viral load data, and we will incorporate the local linear approximation
technique developed by Wu and Zhang (2002). The test statistic proposed by Young and Bowman (1995) will then be used to answer question (ii).

The remainder of this paper is organized as follows. In Section 2, we give details of the proposed model, with the method of estimation, and use the test statistic of Young and Bowman (1995) to determine whether there is a difference between the effects of two treatments. In Section 3, we illustrate the use of the proposed methodology with longitudinal viral load data from 30 HIV-infected patients treated with HAART alone and another 30 patients treated with monotherapy or dual therapy. Some discussion is given in Section 4.

## Methodology

Nonparametric Models and Estimation Methods
We fit the viral load trajectory data of HIV-infected patients receiving a treatment by using a nonparametric mixed-effects (NPME) model:

$$
\begin{align*}
& y_{i}(t)=\log _{10}\left\{V_{i}(t)\right\}=\eta(t)+v_{i}(t)+\varepsilon_{i}(t), \\
& i=1,2, \ldots, n \tag{2.1}
\end{align*}
$$

where $V_{i}(t)$ is the number of copies of HIV-1 RNA per mL of plasma at treatment time $t$ for the $i$ th patient and $y_{i}(t)$ is the corresponding value in $\log _{10}$ scale; $\eta(t)$ is the population mean function, also called the fixed-effects or population curve; $v_{i}(t)$ are individual curve variations from the population curve $\eta(t)$ and these variations are called random-effects curves; and $\varepsilon_{i}(t)$ are measurement errors. We assume that $v_{i}(t)$ and $\varepsilon_{i}(t)$ are independent in which $v_{i}(t)$ can be considered as realizations of a mean 0 process with a covariance function $\gamma(s$, $t)=\mathrm{E}\left(v_{i}(s) v_{i}(t)\right)$, and $\varepsilon_{i}(t)$ can be considered as realizations of an uncorrelated mean 0 process with variance $\sigma^{2}(t)$. The population curve $\eta(t)$ reflects the overall trend or progress of the
treatment process in an HIV-infected population and, hence, can provide an important index of the population's response to a drug or treatment in a clinical or biomedical study, so in this paper we are mainly interested in estimating $\eta(t)$. In addition, an individual curve $s_{i}(t)=\eta(t)+v_{i}(t)$ can represent an individual's response to a treatment in a study, so a good estimate of $s_{i}(t)$ would help the investigator to make better decisions about an individual's treatment management and would enable us to classify subjects on the basis of individual response curves. Similar models have been proposed by Shi et al. (1996) and Zeger and Diggle (1994) to describe CD4+ cell counts.

Let $t_{g i j}, j=1,2, \ldots, n_{g i}$, be the design time points for the $i$ th individual in treatment group $g$. Then, NPME model (2.1) becomes

$$
\begin{align*}
& y_{g i}\left(t_{g i j}\right)=\eta_{g}\left(t_{g i j}\right)+v_{g i}\left(t_{g i j}\right)+\varepsilon_{g i}\left(t_{g i j}\right), \\
& j=1,2, \ldots, n_{g i} ; i=1,2, \ldots, n_{g} ; g=1,2 \tag{2.2}
\end{align*}
$$

Here, $n_{\mathrm{g}}$ is the number of subjects in treatment group $g$, and $n_{\mathrm{gi}}$ is the number of measurements made from subject $i$ in treatment group $g$. We now wish to estimate $\eta_{g}(t)$ and $v_{g i}(t)$ simultaneously, via a local approximation of the NPME model (2.2), by using the local linear mixed-effects model approach of Wu and Zhang (2002), which combines linear mixedeffects (LME) models (Laird \& Ware, 1982) and local polynomial techniques (Fan \& Gijbels, 1996). For this purpose, we assume the existence of the second derivatives of $\eta_{g}(t)$ and $v_{g i}(t)$ at $t$, which are then approximated locally by a polynomial of order 2 as follows:

$$
\eta_{g}\left(t_{g i j}\right) \approx \eta_{g}(t)+\eta_{g}^{\prime}(t)\left(t_{g i h}-t\right) \equiv X_{g i j}^{T} \boldsymbol{\beta}_{g}
$$

and
$v_{g i}\left(t_{g i j}\right) \approx v_{g i}\left(t_{g i j}\right)+v_{g i}^{\prime}(t)\left(t_{g i j}-t\right) \equiv X_{g i j}^{T} \boldsymbol{b}_{g i}$ where
$X_{g i j}=\left(1,\left(t_{g i j}-t\right)\right)^{T}, \quad \boldsymbol{\beta}_{g}=\left(\eta_{g}(t), \quad \eta_{g}^{\prime}(t)\right)^{T}$, and $\boldsymbol{b}_{g i}=\left(v_{g i}(t), v_{g i}^{\prime}(t)\right)^{T}$.

Consequently, the NPME model (2.2) can be approximated by the following model:
$y_{g i j}=X_{g i j}^{T}\left(\boldsymbol{\beta}_{g}+\boldsymbol{b}_{g i}\right)+\varepsilon_{g i j}, \quad j=1,2, \ldots, n_{g i} ;$
$i=1,2, \ldots, n_{g} ; g=1,2$
which is called a LME model. Note that, for simplicity of notation,
$y_{g i j}=y_{g i}\left(t_{g i j}\right), \varepsilon_{g i j}=\varepsilon_{g i}\left(t_{g i j}\right), \boldsymbol{\varepsilon}_{g i}=$
$\left(\varepsilon_{g i 1}, \ldots, \varepsilon_{g i_{g i}}\right)^{T} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{g i}\right)$, and $\boldsymbol{b}_{g i} \sim N\left(\mathbf{0}, D_{\mathrm{g}}\right)$ for $\boldsymbol{\Sigma}_{g i}=\mathrm{E}\left(\boldsymbol{\varepsilon}_{g i} \boldsymbol{\varepsilon}_{g i}^{T}\right)$ and $\mathrm{D}_{\mathrm{g}}=\mathrm{E}\left(\boldsymbol{b}_{\boldsymbol{g i}} \boldsymbol{b}_{\boldsymbol{g i}}^{\boldsymbol{T}}\right)$.

To estimate $\eta_{g}(t)$ and $v_{g i}(t)$, which are the first element of $\boldsymbol{\beta}_{g}$ and $\boldsymbol{b}_{g i}$, respectively, under the standard normality assumptions for $\boldsymbol{b}_{g i}$, we can minimize the following objective function:

$$
\begin{aligned}
& \sum_{i=1}^{n_{g}}\left\{\left(\boldsymbol{y}_{g i}-\boldsymbol{X}_{g i}\left(\boldsymbol{\beta}_{g}+\boldsymbol{b}_{g i}\right)\right)^{T} \boldsymbol{K}_{g i \lambda}^{1 / 2} \Sigma_{g i}^{-1} \boldsymbol{K}_{g i \lambda}^{1 / 2}\right. \\
& \left(\boldsymbol{y}_{g i}-\boldsymbol{X}_{g i}\left(\boldsymbol{\beta}_{g}+\boldsymbol{b}_{g i}\right)\right)+\boldsymbol{b}_{g i}^{T} D_{g}^{-1} \boldsymbol{b}_{g i}+ \\
& \left.\log \left|\Sigma_{g i}\right|\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
& y_{g i}=\left(y_{g i 1}, \ldots, y_{g i_{g i}}\right)^{T} ; \boldsymbol{X}_{g i}=\left(X_{g i 1}, \ldots, X_{g i n_{g i}}\right)^{T} \\
& ; \boldsymbol{K}_{g i \lambda}=\operatorname{diag}\left\{K_{\lambda}\left(t_{g i j}-t\right), \ldots, K_{\lambda}\left(t_{g i n_{g i}}-t\right)\right\}
\end{aligned}
$$

is the kernel weight of the residual term for $K_{\lambda}(\cdot)$ $=K(\cdot / \lambda) / \lambda$, in which $K(\cdot)$ is a kernel function; $\lambda$ is a bandwidth selected by a leave-one-subjectout cross-validation approach (Wu \& Zhang, 2002); and the term $\boldsymbol{b}_{g i}^{T} D_{g}^{-1} \boldsymbol{b}_{g i}$ is a penalty term to account for the random effects $\boldsymbol{b}_{\boldsymbol{g i}}$, taking between-subject variation into account.

Thus, for given $\Sigma_{g i}$ and $D_{g}$, the resulting estimators can be obtained as follows:

$$
\begin{align*}
& \hat{\boldsymbol{\beta}}_{g}= \\
& \left(\sum_{i=1}^{n_{g}} \boldsymbol{X}_{g i}^{T} \boldsymbol{\Omega}_{g i} \boldsymbol{X}_{g i}\right)^{-1}\left(\sum_{i=1}^{n_{g}} \boldsymbol{X}_{g i}^{T} \boldsymbol{\Omega}_{g i} \boldsymbol{y}_{g i}\right) \\
& \hat{\boldsymbol{b}}_{g i}=\left(\boldsymbol{X}_{g i}^{T} \boldsymbol{K}_{g i \lambda}^{1 / 2} \Sigma_{g i}^{-1} \boldsymbol{K}_{g i \lambda}^{1 / 2} \boldsymbol{X}_{g i}+\right. \\
& \left.D_{g}^{-1}\right)^{-1} \boldsymbol{X}_{g i}^{T} \boldsymbol{K}_{g i \lambda}^{1 / 2} \Sigma_{g i}^{-1} \boldsymbol{K}_{g i \lambda}^{1 / 2}\left(\boldsymbol{y}_{g i}-\boldsymbol{X}_{g i} \hat{\boldsymbol{\beta}}_{\boldsymbol{g}}\right) \tag{2.4}
\end{align*}
$$

where
$\Omega_{g i}=\boldsymbol{K}_{g i \lambda}^{1 / 2}\left(\boldsymbol{K}_{\boldsymbol{g i}}^{1 / 2} \boldsymbol{X}_{g i} D_{g} \boldsymbol{X}_{g i}^{T} \boldsymbol{K}_{\boldsymbol{g i}}^{1 / 2}+\boldsymbol{\Sigma}_{g i}\right)^{-1}$ $\boldsymbol{K}_{g i \lambda}^{1 / 2}$. As a result, the estimators of $\eta_{g}(t)$
and
$v_{g i}(t)$ are $\hat{\eta}_{g}(t)=(1,0) \hat{\boldsymbol{\beta}}_{\mathrm{g}}$ and $\hat{v}_{g i}(t)=(1$, 0) $\hat{\boldsymbol{b}}_{g i}$.

The unknown variance-covariance parameters in $D_{\mathrm{g}}$ and $\Sigma_{g i}$ can be estimated by using maximum or restricted maximum likelihood, implemented by using the EM algorithm or the NewtonRaphson method (Davidian \& Giltinan, 1995; Vonesh \& Chinchilli, 1996).

Of particular interest are the comparative effects of the two treatments. Therefore, we need to compare the equality of the two population curves $\eta_{1}(t)$ and $\eta_{2}(t)$. To do this, we fit the model $\eta_{c}(t)+v_{c g i}(t)$ to all data, where $\eta_{c}(t)$ is the fixed-effects (population) curve for the data and $\quad v_{c g i}(t)$ are random-effects curves that deviate from $\eta_{c}(t)$. As is done when estimating $\eta_{g}(t)$ and $v_{g i}(t)$, we can use the local linear approximation approach of Wu and Zhang (2002) to obtain the estimators, $\hat{\eta}_{c}(t)$ and $\hat{v}_{c g i}(t)$, of $\eta_{c}(t)$ and $v_{c g i}(t)$.

Our main concern is how to justify that the difference between the two population
curves is statistically significant. To compare the effects of two treatments, we apply the following test statistic (Young \& Bowman, 1995):

TS
$=\sum_{g=1}^{2} \sum_{j \in T_{g}} \frac{\left\{\hat{\eta}_{g}\left(t_{g j}\right)-\hat{\eta}_{c}\left(t_{g j}\right)\right\}^{2}}{\hat{\sigma}^{2}}$
where $T_{g}=\left\{\right.$ all distinct times $t_{g i}$ in treatment $g$ \} and
$\hat{\sigma}^{2}=\sum_{g=1}^{2} \sum_{i=1}^{n_{g}}\left(n_{g i}-1\right) \hat{\sigma}_{g i}^{2} /\left(n-\sum_{g=1}^{2} n_{g}\right)$ is an estimator of the variance of the measurement error with $n=\sum_{g=1}^{2} \sum_{i=1}^{n_{g}} n_{g i} ; \hat{\sigma}_{g i}^{2}$ are obtained by using the first-order difference approach proposed by Rice (1984), as follows:
$\hat{\sigma}_{g i}^{2}=\frac{1}{2\left(n_{g i}-1\right)} \sum_{j=1}^{n_{g i}-1}\left(y_{g i[j+1]}-y_{g i[k]}\right)^{2}$,
$i=1,2, \ldots, n_{g} ; g=1,2$
If the two population curves are equal; that is, under the null hypothesis $\mathrm{H}_{0}: \eta_{1}(t)=\eta_{2}(t)$, the distribution of the test statistic TS in (2.5) is then approximated by $a \chi^{2}(b)+c$, where $\chi^{2}(b)$ is a chi-squared distribution with $b$ degrees of freedom. Moreover, $a, b$, and $c$ are constants such that the mean, variance, and skewness of $a \chi^{2}(b)+c$ are equal to the corresponding quantities of the test statistic TS, which can be calculated directly. The distribution of $a \chi^{2}(b)+$ $c$ is then used to calculate the $p$-value. The standard error of the difference between the estimates for the two population curves can be computed as
$\operatorname{se}_{\text {diff }}(t)=\operatorname{se}\left\{\hat{\eta}_{1}(t)-\hat{\eta}_{2}(t)\right\}=\sqrt{\operatorname{se}_{1}^{2}(\mathrm{t})+\mathrm{se}_{2}^{2}(\mathrm{t})}$
where $\operatorname{se}_{1}(t)=\operatorname{se}\left\{\hat{\eta}_{1}(t)\right\}$ and $\operatorname{se}_{2}(t)=\operatorname{se}\left\{\hat{\eta}_{2}(t)\right\}$ are the standard errors of the estimates of the population curves, respectively. A reference band whose width is centered at the average of the two estimated curves $\pm 2 \times \mathrm{se}_{\text {diff }}(t)$ can be
used to see how much difference there is between the two treatment groups (Young and Bowman, 1995). Note that, theoretically we should consider correlation when using the approach of Young and Bowman (1995), but we do not just because of mathematical simplicity. Ignoring the correlation may lose some efficiency, however, as you will see, for the reallife data analysis given in the next section there is significant difference between the treatment effects of the two groups even using independent structure. Considering correlation may increase power but seems unnecessary.

## Results

The Analysis of Longitudinal Viral Load Data In this section, we illustrate the practical use of the proposed methodology with longitudinal viral load data from HIV-infected patients. The
data set we are using includes the longitudinal viral load data obtained from 30 HIV-infected patients who received monotherapy or dual therapy and 30 HIV-infected patients who received HAART in several hospitals in Taipei, Taiwan, between 1997 and 2002. These data are subsets of data from a much larger cohort data of 1,195 HIV-infected patients in Taipei. Among the 1,195 HIV-infected patients, most of them received diverse treatments, so, to ensure the validity of the comparison, we chose to use data from the patients treated with HAART who had never been given any other treatment regimen and non-HAART patients who had never been treated with HAART. Treatment durations varied, because patients began receiving treatment at different times during the study period. Figure 1 presents scatter plots of viral load (in $\log _{10}$ scale) against treatment durations for the HIV-1-positive patients.



Figure 1: Scatter plot of viral $\operatorname{load}^{\left(\log _{10} \text { of copy number of HIV RNA in plasma) versus }\right.}$ duration of treatment with HAART (left) or non-HAART (right).

After excluding missing data, we have 208 complete viral load observations in the HAART group, of which 108 have a value less than 400; and we have 164 complete viral load observations in the non-HAART group, of which 69 have a value less than 400 . If we use the criterion that a treatment is considered successful in its antiviral effect when the viral load is below 400, the success rates in the HAART and non-HAART groups are $51.9 \%$ and $42.1 \%$, respectively.

For data analysis, we used the quartic kernel, $K(u)=(15 / 16)\left(1-u^{2}\right)^{2} \mathrm{I}_{(u \mid \leq 1} \leq$. The estimates of the two population curves are depicted in Figure 2. From Figure 2, we can see that the estimates of the two population curves have different patterns although both decrease at the beginning of treatment. The estimated curve for the HAART group shows that the viral load is maintained at a constant level until the end of the treatment, whereas that for the non-HAART group shows that the viral load decreases sharply during the first 480 days, reaching its lowest point on day 480 . However, after 480 days, the
viral load increases, remains constant for a short time, and increases again at the end of the treatment.

A Chi-squared test for the success rates of the two treatments gives a $p$-value of 0.07 . It is hard to say that there is a significant difference between the effects of the two treatments, although the success rate in the HAART group is greater than that in the nonHAART group. Therefore, to look more closely at the difference between the effects of the two treatments, we use the principle described in Section 2. The $p$-value obtained by using this method is less than $10^{-4}$, which indicates that the two population curves for each treatment are substantially different. To confirm this conclusion, we obtained a range of reference values and plotted them with our viral load trajectory estimates in Figure 2. The two estimated population curves deviate from the reference band, and the efficacy of the HAART is seen to be almost significantly superior to that of the conventional therapy that does not include HAART.


Figure 2. Estimate of two population curves.

## Discussion

To determine the efficacy of antiviral treatments by using longitudinal viral load data, we applied nonparametric mixed-effects models to estimate the patterns of the viral trajectories in the two sampled populations. This approach avoids misspecification and, thus, the occurrence of an artificial bias. By combining the betweensubject and within-subject information, the models we have proposed can parsimoniously capture the features of viral response to an antiviral therapy, such that the estimated curve is able to show common features of the antiviral activity.

In implementing the estimation of population curves, we used local linear regression and the bandwidth selection method proposed by Wu and Zhang (2002) to select the bandwidth. Besides the local linear methods applied in this article, the method of regression splines may also be implemented for parameter estimation. The approach of regression splines transforms the models to standard linear mixed-effects models and is easy to implement by using existing software such as SAS and SPLUS.

The result of our illustrative example indicates that HAART has effects that are significantly different from those of treatment that did not include HAART. At the beginning of treatment, non-HAART has strong antiviral activity, which is lacking with HAART. However, during the course of the treatment, the superiority of non-HAART lessens, and this therapy ultimately fails, whereas HAART maintains a constant effect throughout treatment. This maintenance of the viral load at a constant level confirms previous findings and is preferable to the fluctuation of load resulting from non-HAART. This result confirms that HAART is worth continuing, despite its inability to suppress viral replication completely (Deeks \& Martin 2001).

Finally, the reference band covers a wider range of viral loads at the end of treatment, despite the increasing difference between the two estimated curves. This is not surprising because of the smaller sample size resulting from a shorter treatment duration for some patients at that time.

## References

Cooper, E. R., Charurat, M., Mofenson, L., Hanson, I. C., Pitt, J., Diaz, C., Hayani, K., Handelsman, E., Smeriglio, V., Hoff, R., Blattner, W., \& Women and Infants' Transmission Study Group (2002). Combination antiretroviral strategies for the treatment of pregnant HIV-1-infected women and prevention of perinatal HIV-1 transmission. Journal of Acquired Immune Deficiency Syndromes, 29, 484-94.

Davidian, M. \& Giltinan, D. M. (1995). Nonlinear model for Repeated Measurement Data. London: Chapman and Hall.

Deeks, S. G. \& Martin, J. N. (2001). Reassessing the goal of antiretroviral therapy in the heavily pre-treated HIV-infected patient. AIDS, 15, 117-119.

Deeks, S. G., Hecht, F. M., Swanson, M., Elbeik, T., Loftus, R., Cohen, P. T., \& Grant, R.M. (1999). HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: response to both initial and salvage therapy. AIDS, 13, F35-43.

Detels, R., Munoz, A., McFarlane, G., Kingsley, L. A., Margo lick, J. B., Giorgi, J., Schrager, L. K., \& Phair, J. P. (1998). Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators. The Journal of the American Medical Association, 280, 1497-503.

Fan, J. \& Gijbels, I. (1996). Local polynomial modeling and its applications. London: Chapman and Hall

Ghani, A. C., Donnelly, C. A., \& Anderson, R.M. (2003). Patterns of antiretroviral use in the United States of America: analysis of three observational databases. HIV Medicine, 4, 24-32.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., \& Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature, 373, 123-126.

Laird, N. M. \& Ware, J. H. (1982). Random effects models for longitudinal data. Biometrics, 38, 963-974.

Mellors, J. W., Kingsley, L. A., Rinaldo, C. R., Todd, J. A., Hoo, B. S., Kotta, R. P., \& Gupto, P. (1995). Quantitation of HIV-1 RNA in plasma predicts outcome after seroconversion. Annals of Internal Medicine, 122, 573-579.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., \& Kingsley, L. A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science, 272, 1167-1170.

Perelson, A. S. \& Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics in vivo. SIAM Review, 41, 3-44.

Rice, J. (1984). Bandwidth choice for nonparametric regression. The Annals of Statistics 12, 1215-1230.

Saag, M. S., Holodniy, M., Kuritzkes, D. R., O'Brien, W. A, Coombs, R., Poscher, M. E., Jacobsen, D. M., Shaw, G. M., Richman, D. D., \& Volberding, P. A. (1996). HIV viral load markers in clinical practice. Nature Medicine, 2, 625-629

Shi, M., Weiss, R. E., \& Taylor, J. M. G. (1996). An analysis of pediatric CD4+ counts for acquired immune Deficiency syndrome using flexible random curves. Journal of the Royal Statistical Society. Series C, 45, 151-163.

Tassie, J. M, Grabar, S., Lancar, R., Deloumeaux, J., Bentata, M., \& Costagliola, D. (2002). Clinical Epidemiology Group from the French Hospital Database on HIV. Time to AIDS from 1992 to 1999 in HIV-1-infected subjects with known date of infection. Journal of Acquired Immune Deficiency Syndromes, 30, 81-87.

Vonesh, E. F. \& Chinchilli, V. M. (1996). Linear and Nonlinear Models for the Analysis of Repeated Measurements. New York: Marcel Dekker.

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., \& Shaw, G. M. (1995). Viral dynamics in human immunodeficiency virus type-1 infection. Nature, 373, 117-122.

Wu, H. \& Ding, A. A. (1999). Population HIV-1 dynamics in vivo: applicable models and inferential tools for virological data from AIDS clinical trials. Biometrics, 55, 410-418.

Wu, H. \& Zhang, J. (2002). Local polynomial mixed-effects models for longitudinal data. Journal of the American Statistical Association, 97, 883-897.

Young, S. G. \& Bowman, A. W. (1995). Non-parametric analysis of covariance. Biometrics, 51, 920-931.

Zeger, S. L. \& Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4+ cell numbers in HIV seroconverters. Biometrics, 50, 689-699.

# Alphabet Letter Recognition And Emergent Literacy Abilities Of Rising Kindergarten Children Living In Low-Income Families 

Stephanie Wehry<br>Florida Institute of Education<br>The University of North Florida

Alphabet letter recognition item responses from 1,299 rising kindergarten children from low-income families were used to determine the dimensionality of letter recognition ability. The rising kindergarteners were enrolled in preschool classrooms implementing a research-based early literary curriculum. Item responses from the TERA-3 subtests were also analyzed. Results indicated alphabet letter recognition was unitary. The ability of boys and younger children was less than girls and older children. Child-level letter recognition was highly associated with TERA-3 measures of letter knowledge and conventions of print. Classroom-level mean letter recognition ability accounted for most of variance in classroom mean TERA-3 scores.

Key words: Early childhood literacy, alphabet letter knowledge, latent variable modeling, two-level modeling, categorical factor analysis.

Introduction
The No Child Left Behind Act has focused attention on reading instruction in kindergarten through third-grade. Programs such as the Preschool Curriculum Evaluation Research (PCER) and Early Reading First (ERF) grants expand that focus to preschool curricula that support cognitive development including emergent literacy. Literacy researchers are connecting theories about the acquisition of reading and emergent literacy skills and experiences.

The emergent literacy model embodies more than reading readiness and is used to describe the acquisition of literacy on a developmental continuum. The model provides a picture of the acquisition of literacy that occurs from early childhood rather than beginning at kindergarten and further suggests literacy skills develop concurrently and interdependently.

Whitehurst and Lonigan (1998) listed vocabulary, conventions of print, emergent writing, knowledge of graphemes, graphemephoneme correspondence, and phonological

Stephanie Wehry is the Associate Director for Research at the Florida Institute of Education, University of North Florida. Email her at swerhy@unf.edu.
awareness as the skill and knowledge base of emergent literacy. They further suggested emergent literacy consists of outside-in processes that include the context in which reading and writing occurs and inside-out processes that include the knowledge and skills associated with the alphabetic principle, emergent writing, and cognitive processes. Specific examples of outsidein processes include oral language, conceptual skills, and concepts of print. The inside-out processes are letter knowledge, phonological processing skills, and syntax awareness. A study by Whitehurst et al. (1999) of 4 -year-old Head Start children indicated inside-out processes were much stronger influences on first- and secondgrade reading outcomes than outside-in processes.

Historically, reading has been defined in two ways; code breaking and meaning making (Riley, 1996) or as decoding and comprehension (Gough, Juel, \& Griffin, 1992; Mason, 1980; Perfetti, 1984). Two stages of reading acquisition relative to the code breaking definition were originally proposed and those models were often refined to include three stages (Frith, 1985; Gough \& Hillinger, 1980; Gough, Juel, \& Griffith, 1992; Mason, 1980; Sulzby, 1992).

The first stage involves the association of a spoken word with some visual feature of the corresponding printed word. The second stage involves cryptanalysis of printed words or
phonological processing involving the correspondence of graphemes and phonemes, and the third stage involves orthographic processing involving the correspondence of spelling patterns and printed words. Baker, Torgeson, and Wagner (1992) studied the role of phonological and orthographic processing and determined that orthographic skills make an independent contribution to reading achievement. Goswami (1993) saw these stages as cyclical where orthographic skills enhance phonological skills, which in turn enhance orthographic skills.

Mason (1980) suggested alphabet knowledge initiates the first level of reading acquisition by facilitating the breaking down of words into letters. Later, in a critique of five studies of children's alphabet knowledge, Ehri (1983) went further and suggested children's knowledge of the alphabet is the main skill that enables them to move from the first stage to the alphabetic or phonological stage of reading acquisition and that it is difficult to separate children's letter-sound knowledge from other emergent literacy skills. Chall (1983) summarized 17 studies of the relationship between knowledge of the alphabet and future reading achievement. Although causation was not claimed, knowledge of the letters of the alphabet was seen as an important predictor of reading achievement.

Sulzby (1983) suggested children's lettername ability is integrated into a more complex set of early literacy skills and that children attempt to use some mechanism as they learn to associate letter names with their visual forms. Children learn these skills from exposure to books, songs, blocks, and learning to write their names. Sulzby (1992) further suggested alphabet letter knowledge precedes understanding the concept of word and comprehension; however, these stages reinforce each other. Bialystok (1991) suggested that children who can identify letters in non-alphabetic order and understand that letters symbolize sounds are on their way to code breaking. Riley (1996) proposed the link between alphabet letter knowledge and concepts of print is the key to why alphabet letter knowledge is such a powerful predictor of reading achievement.

Moreover, recent studies of emergent literacy have focused on the relationships between phonological awareness and later reading. But children's letter knowledge is associated in some
manner with their phonological sensitivity (Bowley, 1994; Stahl \& Murray, 1994). Stahl and Murray suggested children's letter knowledge enables them to manipulate initial sounds - a skill that leads to word recognition.

Researchers have also found measures of phonological awareness independently predicted measures of word recognition and decoding (McGuiness, McGuiness, \& Donohue, 1995), and that among preschool children from low-income families, measures of phonological sensitivity were associated with measures of letter knowledge (Lonigan, Burgess, Anthony, \& Barker, 1998). Whitehurst et al. (1999) found that reading ability in early elementary school was strongly related to measures of preschool children's skills that included items requiring them to name a pictured letter and to identify initial letters and sounds of pictured and named objects - tasks that measure grapheme-phoneme relationships. Lonigan, Burgess, and Anthony (2000), in a longitudinal study, found letter knowledge was independent of phonological sensitivity, environmental print, and decoding, and that $54 \%$ of the variation in kindergarten and first grade children's reading skills was accounted for by preschool phonological sensitivity and letter knowledge.

As Adams (1990) suggested, a child's level of phonological processing is irrelevant if the child cannot identify the letters of the alphabet. If a beginning reader cannot identify the letters then the reader cannot associate sounds with letters (Bond \& Dykstra, 1967; Chall, 1967; Mason, 1980). Moreover, orthographic competency depends on the ability to visually identify and discriminate the individual letters of the alphabet. How children acquire this ability falls in the domain of perceptual learning theory.

There are two prevalent theories (Adams, 1990; Gibson \& Levin, 1975); the template and the feature theories. In the template theory, the brain stores templates of the most typical representation of the letters and stimuli are compared to the stored templates. In the feature theory, the letters of the alphabet are considered a group of symbols that share common distinct features. The brain stores the common features of different letters and matches features of stimuli to the stored list. Both theories involve search and comparison.

Studies of children's alphabet letter knowledge span more than four decades, involve preschool to third-grade children from low- and middle-income families, and use either all or a sample of the letters. Sulzby (1983) suggested knowledge of the alphabet measured in kindergarten, not later, is the predictor of reading achievement. However, Early Childhood Longitudinal Study-Kindergarten researchers reported $66 \%$ of children entering kindergarten for the first time recognized most of the letters of the alphabet (Zill \& West, 2001).

In recent studies of children's alphabet knowledge, Whitehurst et al. (1999) studied Head Start children and used a sample of letters embedded as items in another measure; Lonigan et al. (1998) studied preschool children from lowincome families and used all uppercase letters; Lonigan et al. (2000) studied preschool children from middle- to upper-income families and used all uppercase letters; and Roberts (2003) studied preschool children whose primary language was not English and used a sample of letters.

Studies of children from low-income families are especially important because one third of American children experience reading difficulties in school (Adams, 1990), and children from low-income families have comparatively lower levels of emergent literacy (Whitehurst \& Lonigan, 1998). Because individual differences in emergent literacy at entry into kindergarten are stable or increase over school years (Baydar, Brooks-Gunn, \& Furstenberg, 1993; Juel, 1988; Stevenson \& Newman, 1986), the impact of lower levels of emergent literacy follows preschool children through school. For these reasons, this study analyzed responses from rising kindergarten children from low-income families using all upper- and lowercase letters of the alphabet and other items measuring emergent literacy abilities.

Moreover, the children studied were nested in classrooms nested in locations. Head Start researchers (Westat, 1998) found significant variation in program quality across Head Start programs, centers, and classrooms with the largest variation occurring at the classroom level. Whitehurst et al. (1999) also found the performance of Head Start children differed across centers. Violating the assumption of independent observations across experimental units is a major concern with the use of nested data. In most cases,
correlations between observations nested in groups are positive resulting in inflated Type I error rates in significance testing.

Further research is needed to estimate the magnitudes of intraclass correlations in preschool achievement data. In this study, classrooms were studied because of the large number of singleclassroom locations in the data and because Head Start researchers found most of the variance in program quality occurred at the classroom-level. A two-level model was used to estimate the size of the intraclass correlations; however, a two-level study confounds the effects classrooms and sites for sites with more than one classroom.

## Purposes Of This Study

The primary purpose of this study was to analyze the alphabet letter recognition ability of rising kindergarten children from low-income families and determine if the ability was unitary or if it divided along the perceptual learning or instructional features (Adams, 1990; Gibson \& Levin, 1975). A second purpose of this study was to investigate the relationship between recognition of the letters of the alphabet and other measures of emergent literacy using methodology that developed an interval measurement scale and acknowledged the nested nature of the data. The three research questions about responses from rising kindergarten children from low-income families are

1. Is the ability to recognize upper- and lowercase letters of the alphabet unitary or multidimensional?
2. Does a latent trait model of children's responses on the three Test of Early Reading Ability (TERA-3) subtests confirm the test publisher's three-factor structure?
3. Using children's two-parameter normal ogive scores on alphabet letter recognition and TERA-3 subtests in a two-level model:
a. What is the relationship between children's alphabet letter knowledge and the TERA-3 subtest abilities?
b. Do these relationships differ by the age and/or gender of the children?
c. What portion of the individual differences in the children's scores is accounted for the by the classrooms in which they learn?
d. Are differences in the classroom means of TERA-3 subtest scores predicted by classroom mean alphabet letter recognition scores?

## Methodology

## Participants

Data were collected from 1,299 4-year-old children during a one-month period from April 15, 2002 to May 17, 2002. All children were eligible to attend public school kindergarten the following year. Birth dates were available for 1,025 of the children and their ages as of September 1 of the school year ranged from 48 to 65 months with the average and median ages of 54.7 and 55 months, respectively. Gender was reported for 1001 children: $530(53 \%)$ were boys. The average (median) ages for boys and girls were 54.7 (55) and 54.6 (55) months, respectively. Ethnicity data were not collected; however, nearly all of the children were African American.

Classroom Context
The children were from low-income families; therefore, were considered at risk for academic failure. They were attending Head Start, faith-based, subsidized, and early intervention preschool programs located in six counties in southeastern United States. Most of the children attended classrooms in urban settings; however, a few classrooms were located in small towns. Children with complete scores and gender information were enrolled in 121 classrooms at 76 locations.

Fifty-five of the locations were singleclassroom sites, 16 of the locations were two- or three-classroom sites, and the remaining five locations had four or more classrooms at each site. All children in the study experienced at least one semester of an intensive early literacy curriculum. Classroom teachers explicitly taught the inside-out early literacy skills in classroom contexts that provided outside-in early literacy experiences (Whitehurst \& Lonigan, 1998). Agencies funding participation in the literacy curriculum provided materials, teaching strategies, and weekly coaching for preschool teachers as they explicitly
taught children alphabet letter knowledge, phonemic awareness, and print concepts. Teachers also used dialogic reading (Valdez-Menchara, \& Whitehurst, 1992; Whitehurst, Arnold, Epstein, Angell, Smith, \& Fischel, 1994) and provided opportunities for emergent writing, reading, and comprehension. All instruction occurred in printrich environments with labeled furniture and word walls. The evaluation of the literacy curriculum used measures of alphabet letter recognition and other emergent literacy abilities in a pretest/posttest design. Data used in this study were the posttest data of that evaluation.

## Measurement

Data were collected on the children's ability to recognize the 52 upper- and lowercase letters of the alphabet and from Form A of the Test of Early Reading Ability (TERA-3) (Reid, Hresko, \& Hammill, 2001a). Trained examiners collected responses from children in school settings in age appropriate one-on-one sessions. The children's responses were recorded on scannable forms.

## Alphabet Letter Recognition

Uppercase letter flashcards, arranged in a fixed non-alphabetic order, were presented one at a time to each child. The child was asked to name the letter. Following presentation of the 26 uppercase letters, lowercase letter flashcards, also arranged in a fixed non-alphabetic order, were presented one at a time.

## TERA-3

The TERA-3 is composed of three subtests measuring unique but related early literacy skills. Items within each subtest are arranged by difficulty and each subtest has a stopping mechanism. All children began testing with the first item in each subtest. According to Reid, Hresko, and Hammill (2001b), the Alphabet subtest measures graphophomenic knowledge, the Conventions subtest measures knowledge of conventions of English print, and the Meaning subtest measures ability to comprehend meaning of print. Published validity and reliability information indicates Cronbach Alpha coefficients of internal consistency for 4 -year old children (5-year-old children) for the Alphabet, Conventions, and Meaning subtests are .94 (.93), .88 (.86), and .94 (.84), respectively.

## Data Analysis

Data were analyzed using Mplus 2.13 (Muthén \& Muthén, 2003). The flexibility of Mplus permits latent variable modeling with categorical indicators. The use of raw scores formed by summing correct item responses assumes all items are equally important in measuring the underlying construct and that intervals between scores are uniform across the ability continuum. In contrast, measurement modeling within the latent variable context permits a distinction between observed item scores and the underlying construct, and the continuous latent variables are free from measurement error.

Categorical confirmatory factor analyses (CFAs) were conducted using the item responses from the alphabet letter recognition and the three TERA-3 subtests. The analyses produced twoparameter normal ogive item response theory (IRT) models. The CFAs resulted in error free continuous latent variables; however, Mplus does not have the capability to use these results directly in multilevel models. Factor scores, which are estimated as in IRT modeling, were used as continuous variables in the two-level model. This procedure reintroduced some measurement error.

## Results <br> Alphabet Letter Recognition

Distribution of Items and Summed Scores
Item responses were available from 1,299 rising kindergarten children. Correct responses were coded one and incorrect responses were coded zero. Table 1 shows alphabet letter item means and standard deviations. Additionally, three scores were formed by summing responses; one for uppercase letters, one for lowercase letters, and one for total of the upper- and lowercase scores. The means (standard deviations) for each of these summed scores were 16.41 (9.11), 13.69 (8.89), and 30.08 (17.74), respectively.

Adams (1990) suggested alphabet letter recognition instruction begins with the uppercase letters for preschool children, and the mean scores indicated rising kindergarten children recognized more uppercase than lowercase letters and more
than $22 \%$ of the children recognized all uppercase letters. Calfee, Cullenbine, DePorcel, and Royston (cited in Mason, 1980) found the distribution of children's uppercase letter recognition ability was bimodal with most children either recognizing less than eight or more than 20 letters. Figure 1 shows the distribution of the children's upper- and lowercase letter recognition summed scores. Data pile up on both extremes of the distribution (ceiling and floor effects) as previously determined. The pattern at both extremes is more obvious in the distribution of lowercase letter responses.

Dimensionality of Alphabet Letter Recognition: Classical Test Theory

Traditional methods of assessing test dimensionality use factor analytic methods and coefficients of internal consistency as indicators. Cronbach's Alpha, a measure of internal consistency, for the 52 items was .98 indicating items consistently measured a unitary construct. Factor analysis of the alphabet letter recognition data produced four eigenvalues greater than 1.00 ; 26.49, 1.97, 1.11, and 1.06 explaining 50.94, 3.79, 2.14 , and 2.04 percent of the variance in the observations, respectively. These eigenvalues suggested the presence of one central factor with possibly up to three additional minor or difficulty factors.

Dimensionality of Alphabet Letter Recognition: Item Response Theory

Latent variable modeling permits a measurement model of data that is error free, weighs the relative importance of each item, and places measurement on an interval scale. Several theoretical measurement models of alphabet letter recognition ability were evaluated using categorical CFA.

Alphabet letter recognition often begins with the uppercase letters as they are more visually distinct than the lowercase letters (Tinker, 1931). Therefore, Model I was a two-factor model with one factor representing the uppercase letters and one representing the lowercase letters. Model I was based on instructional strategy.

Table 1. Summary Statistics and Model VII Factor Loadings for Items Measuring Recognition of the Upper- and Lowercase Letters of the Alphabet

|  | Uppercase letters |  |  | Lowercase letters |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Variable | Mean | Standard <br> deviation | Factor <br> loading | Mean | Standard <br> deviation | Factor <br> loading |
| Aa | .75 | .43 | .90 | .51 | .50 | .86 |
| Bb | .81 | .39 | .82 | .43 | .50 | .84 |
| Cc | .70 | .46 | .90 | .68 | .47 | .91 |
| Dd | .65 | .48 | .89 | .32 | .47 | .77 |
| Ee | .65 | .48 | .90 | .58 | .49 | .90 |
| Ff | .57 | .50 | .93 | .42 | .49 | .91 |
| Gg | .56 | .50 | .91 | .39 | .49 | .89 |
| Hh | .61 | .50 | .88 | .43 | .50 | .86 |
| Ii | .59 | .49 | .89 | .60 | .49 | .87 |
| Jj | .59 | .49 | .89 | .56 | .50 | .89 |
| Kk | .66 | .47 | .84 | .64 | .48 | .84 |
| Ll | .59 | .49 | .90 | .31 | .46 | .79 |
| Mm | .57 | .50 | .84 | .55 | .50 | .86 |
| Nn | .57 | .50 | .88 | .39 | .49 | .82 |
| Oo | .85 | .36 | .87 | .82 | .38 | .86 |
| Pp | .65 | .48 | .91 | .54 | .50 | .86 |
| Qq | .60 | .49 | .86 | .36 | .48 | .81 |
| Ss | .65 | .47 | .88 | .63 | .48 | .89 |
| Tt | .64 | .48 | .88 | .56 | .50 | .89 |
| Uu | .52 | .50 | .89 | .43 | .50 | .85 |
| Vv | .45 | .50 | .87 | .46 | .50 | .86 |
| Ww | .63 | .48 | .76 | .64 | .48 | .75 |
| Xx | .71 | .45 | .75 | .72 | .45 | .76 |
| Yy | .59 | .49 | .85 | .55 | .50 | .86 |
| Zz | .65 | .48 | .88 | .63 | .48 | .86 |

Perceptual learning theory suggests other models. One theory suggests children holistically perceive the letters and form templates in their memories for each letter learned. Another theory suggests children recognize letters by a set of distinctive visual features stored in their memories. The feature theory is more mentally efficient than the template theory.

Gibson and Levin (1975) reported that both children and adults sorted the uppercase letters of the alphabet by whether or not they have only
straight-line features or have curved features in possible combination with straight-line segments. The secondary sort was by whether or not the letters with curved features have places of intersections such as B and P , or look round such as O and Q . The tertiary sort was by whether letters with straight-line features have diagonal segments such as $M$ and $Z$, or not such as $E$ and $F$.

Figure 1.The distribution of simple summed upper- and lowercase alphabet letter recognition scores.


Several models involving the distinct features of the letters were investigated. Model II was a two-factor model with one factor representing letters whose visual representation is composed of diagonal line segments with no curved features (AKMNVWXYZkvwxyz) and a factor representing the remaining letters (BCDEFGHIJLOPQRSTUabcdefghijlmnopqrstu). Model III was a two-factor model with one factor representing letters whose visual representation is composed only of line segments (AEFHIKLMNTVWXYZikltvwxyz) and one representing the remaining letters (BCDGJOPQRSUabcdefghjmnopqrsu). Model IV was a two-factor model with one factor representing letters whose visual representation exhibits line symmetry (ABCDEHIMOTUVWXYZclotvwxz) and one representing the remaining letters (FGJKLNPQRSabdefghijkmnpqrsuy).

Roberts (2003) used explicit instruction to teach alphabet letter recognition to preschool children and suggested there are 44 distinct abstract symbols children must learn. She reasoned the upper- and lowercase forms for $\mathrm{C}, \mathrm{O}$, S, U, V, W, X, and Z are the same. Model V was a two-factor model with one factor representing these eight pairs (COSUVWXZcosuvwxz) and one factor representing the remaining letters (ABDEFGHIJKLMNPQRTYabdefghijklmnpqrty).

Rotated exploratory factor analysis of the data suggested four highly correlated factors with one primary factor. Therefore, a unitary model, Model VI, was fit. Additionally there are at least seven letters whose upper- and lowercase visual forms are identical ( $\mathrm{C}, \mathrm{O}, \mathrm{S}, \mathrm{V}, \mathrm{W}, \mathrm{X}$, and Z ) and four more whose upper- and lowercase visual forms are nearly identical ( $\mathrm{K}, \mathrm{P}, \mathrm{U}$, and Y ); therefore, another unitary model with errors for these eleven pairs of letters freed to correlate was also fit, Model VII.

Categorical confirmatory factor analysis of the seven models was conducted using Mplus. A matrix of 1,299 observations, each observation having 52 binary items, was analyzed. Weighted least squares estimation (WLSM) was used to estimate model parameters. Five fit statistics are available for WLSM estimation: the comparative fit index (CFI), the Tucker-Lewis index (TLI), root mean square error approximation (RMSEA), weighted room mean square residual (WRMR), and standardized root mean square residual (SRMR). Guidelines for good fit of categorical models suggested CFI >.95, TLI >.95, RMSEA < .06 , WRMR <. 90, and SRMR <. 08 (Hu \& Bentler, 1999; Yu \& Muthén, 2002). Table 2 shows fit statistics for each of the seven models.

All seven models had CFI, TLI, and SRMR fit statistics within limits established for good fit. None of the seven models had WRMR within limits established by Yu and Muthén (2002). The RMSEA fit statistic of Model VII was the only one within limits and Model VII had the lowest WRMR. Therefore Model VII, a unitary model, exhibited the best overall fit and is supported by classical test theory and parsimony. Table 1 shows factor loadings for Model VII, and factor scores from Model VII were used in the two-level model.

## TERA-3

TERA-3 is composed of three subtests measuring graphophemic knowledge (Alphabet), knowledge of conventions of English print (Conventions), and the ability to comprehend meaning of print (Meaning), and is designed for use with children whose ages are between three years six months and eight years six months. There are 29 Alphabet items, 21 Conventions items, and 30 Meaning items. Any subtest item whose mean was less than .05 was not used in this study. TERA-3 was administered to 1009 children in one-on-one settings by trained examiners. Correct responses were coded one and incorrect responses were coded zero. Table 3 shows TERA-3 item means and standard deviations.

## Subtest Alphabet

Twenty-two Alphabet items were included in the study, and these items required children to identify pictured upper- and lowercase named letters, to name identified pictured upper- and lowercase letters, to identify initial letters and sounds of text and named words, and to choose the correct text corresponding to a pictured object. Cronbach Alpha coefficient for the Alphabet subtest items used in the study was 93 .

Table 2. Fit Indices and Factor Correlations for Seven Measurement Models of Alphabet Letter Recognition.

| Model | CFI | TLI | RMSEA | WRMR | SRMR | Correlations |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| I | $.99^{*}$ | $.99^{*}$ | .09 | 2.52 | $.05^{*}$ | .77 |
| II | $.99^{*}$ | $.99^{*}$ | .09 | 2.39 | $.05^{*}$ | .71 |
| III | $.99^{*}$ | $.99^{*}$ | .09 | 2.48 | $.05^{*}$ | .71 |
| IV | $.99^{*}$ | $.99^{*}$ | .09 | 2.51 | $.05^{*}$ | .80 |
| V | $.99^{*}$ | $.99^{*}$ | .09 | 2.34 | $.05^{*}$ | .78 |
| VI | $.99^{*}$ | $.99^{*}$ | .09 | 2.56 | $.05^{*}$ | - |
| VII | $1.00^{*}$ | $1.00^{*}$ | $.04^{*}$ | 1.31 | $.03^{*}$ | $.14-.37$ |

Note. *Denotes the value indicates model fit.

## Subtest Conventions

Twelve Conventions items were included in the study, and these items required children to identify pictured books that were oriented correctly for reading, to distinguish pictured text from other pictured line markings, to match pictured uppercase with corresponding lowercase letters, to distinguish between text, title, author's name, and illustrations when presented pictured first pages of a story, to identify the first and last words of a pictured paragraph, and to follow (by pointing) pictured text as it was read indicating knowledge that text is read from left to right, top to bottom, and when to turn a pictured page. Cronbach Alpha coefficient for the Conventions subtest items used in the study was .80 .

## Subtest Meaning

Ten Meaning items were included in the study, and these items required children to identify pictured product labels corresponding to named product categories, to identify pictured upper- and lowercase text placed adjacent to named pictured objects, and to identify pictured text corresponding to named pictured objects when presented amongseveral sets of pictured objects with corresponding text. Cronbach Alpha coefficient for the Meaning subtest items used in the study was. 74 .

Confirmatory factor analysis of these 45 items was performed using Mplus. Items were
restricted to measuring TERA-3 subtests suggested by test developers. However, one Conventions item, C3, involved alphabet letter knowledge; therefore, it was freed to load on both the Alphabet and Conventions latent variables. Figure 2 provides a visual representation of the model, and, as can be seen, C3 was more strongly associated with the Alphabet latent variable. Model parameters were estimated using WLSM, and fit indices were $\mathrm{CFI}=.99, \mathrm{TLI}=.99$, RMSEA $=.05$, and $\mathrm{WRMR}=1.50$. Three of the indices, CFI, TFI, and RMSEA, indicated model fit (Yu \& Muthén, 2002). The three latent factors were correlated with the strongest correlation occurring between Alphabet and Conventions. Table 3 shows factor loadings for the TERA-3 model, and factor scores from the model were used in the two-level model.

Two-Level Path Analysis of the Alphabet Letter Recognition and TERA-3: Emergent Literacy Abilities of the Rising Kindergartners

Alphabet letter recognition Model VII factor scores (Letters) and the TERA-3 subtest factor scores (Alphabet, Conventions, and Meaning) were used in a two-level path analysis. The within-level used the child-level data and the between-level used the classroom-level data. Table 4 shows summary statistics for the 986 child-level and the 121 classroom-level factor scores of the four variables.

The analysis in multilevel terms involved the following variables and notations:
$i$ is the $i^{\text {th }}$ child of the 986 children studied, $j$ is the $j^{\text {th }}$ classroom of the 121 classrooms studied,
Subtest ${ }_{i j}$ is the TERA-3 subtest factor score of the $i^{\text {th }}$ child in the $j^{\text {th }}$ classroom,
Letters $_{i j}$ is the alphabet letter recognition Model VII factor score
of the $i^{\text {th }}$ child in the $j^{\text {th }}$ classroom,
Gender $_{i j}$ is the gender (girls coded 0 and boys coded 1 ) of the $i^{\text {th }}$ child
in the $j^{\text {th }}$ classroom, and
Age ${ }_{i j}$ is the age in months on September 1 of the $i^{\text {th }}$ child in the $j^{\text {th }}$ classroom.

All three TERA-3 subtests were simultaneously analyzed. The analysis in multilevel terms involved the following child-level and classroom-level equations:

> Child-Level

$$
\text { Subtest }_{i j}=\beta_{0 j}+\beta_{1 j}\left(\text { Letters }_{i j}\right)+\beta_{2 j}\left(\text { Gender }_{i j}\right)+\beta_{3 j}\left(\text { Age }_{i j}\right)+r_{i j}
$$

Classroom-Level

$$
\beta_{0 j}=\gamma_{00}+\gamma_{01}\left(\text { Letters. }_{\mathrm{j}}\right)+u_{j}
$$

where,
$\beta_{0 j}$ is the mean TERA-3 subtest factor score of the $j^{\text {th }}$ classroom,
$\beta_{1 j}$ is the expected change in children's TERA-3 subtest factor scores associated with a change in their alphabet letter recognition factor scores,
$\beta_{2 j}$ is the expected difference in boys' TERA-3 subtest factor scores,
$\beta_{3 j}$ is the expected difference in children's TERA-3 subtest factor scores associated with a difference in their age,
$r_{i j}$ is the unaccounted for individual differences in children's TERA-3 subtest ability,
$u_{j}$ is the unaccounted for classroom differences in TERA-3 factor score classroom means,
$\gamma_{00}$ is the grand mean of the TERA-3 subtest factors scores, and
$\gamma_{01}$ is the expected change in TERA-3 subtest classroom mean factor scores associated with a change in the classroom mean alphabet letter recognition factor scores.

This set of equations was replicated for each of the three TERA-3 subtest factor scores. Figure 3 shows the child-level and classroom-level path models and results. Parameters for the multilevel path analysis were estimated using Muthén's maximum likelihood estimator for balanced data (MUMLM). The fit indices for the model were CFI $=1.00, \mathrm{TFI}=.99$, RMSEA $=.02$, and SRMR $<.01$ for the within model ( .04 for the classroom-level model): all indicated good fit. The intraclass correlations were $.19, .21, .15$, and .17 for Letters, Alphabet, Conventions, and Meaning, respectively.

The analyses indicated that alphabet letter knowledge predicted all three TERA-3 subtest abilities. Not surprisingly, the strongest influence was on the Alphabet subtest scores. Both age and gender influenced the Alphabet subtest scores directly and indirectly through the Letters variable. Boys and younger children had lower Alphabet subtest ability than girls and older children. The child-level model accounted for almost $70 \%$ of the child-level variance in the Alphabet subtest scores.

Alphabet letter recognition ability also influenced the Conventions subtest scores with the strength of association about two thirds as large as in the Alphabet subtest scores. Following the same pattern found with the Alphabet subtest scores, age

Table 3. Summary Statistics and CFA Factor Loadings for TERA-3 Alphabet, Conventions, and Meaning Subtests

| Variable | Mean | Standard <br> deviation | Factor <br> loading | Variable | Mean | Standard <br> deviation | Factor <br> loading |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| A1 | .90 | .30 | .82 | M1 | .94 | .24 | .29 |
| A2 | .78 | .41 | .72 | M2 | .95 | .22 | .52 |
| A3 | .71 | .46 | .60 | M3 | .92 | .28 | .61 |
| A4 | .75 | .43 | .86 | M4 | .78 | .41 | .94 |
| A5 | .55 | .50 | .61 | M5 | .79 | .41 | .93 |
| A6 | .57 | .50 | .78 | M6 | .81 | .39 | .74 |
| A7 | .43 | .50 | .70 | M7 | .89 | .31 | .87 |
| A8 | .43 | .50 | .93 | M8 | .25 | .43 | .66 |
| A9 | .36 | .48 | .87 | M9 | .46 | .50 | .77 |
| A10 | .36 | .48 | .90 | M10 | .09 | .29 | .53 |
| A11 | .40 | .49 | .94 | C1 | .62 | .49 | .66 |
| A12 | .38 | .49 | .89 | C2 | .52 | .50 | .43 |
| A13 | .33 | .47 | .91 | C3 | .72 | .45 | .20 |
| A14 | .21 | .41 | .78 | C4 | .68 | .47 | .75 |
| A15 | .24 | .43 | .88 | C5 | .22 | .41 | .69 |
| A16 | .28 | .45 | .90 | C6 | .45 | .50 | .58 |
| A17 | .25 | .43 | .94 | C7 | .16 | .36 | .85 |
| A18 | .09 | .29 | .79 | C8 | .29 | .46 | .80 |
| A19 | .16 | .37 | .87 | C9 | .12 | .32 | .75 |
| A20 | .17 | .38 | .89 | C10 | .08 | .28 | .82 |
| A21 | .09 | .29 | .85 | C11 | .13 | .34 | .99 |
| A22 | .11 | .31 | .83 | C13 | .09 | .29 | .90 |
| C3 | .72 | .50 | .64 |  |  |  |  |

Note. $\mathrm{n}=1,009$ rising kindergarten children; A1-A22 are Alphabet Subtest items; $\mathrm{C} 1-\mathrm{C} 11$, and C 13 are Conventions Subtest items; and M1-M10 are Meaning Subtest items.
and gender influenced the Conventions subtest scores both directly and indirectly through the Letters variable. Boys and younger children had lower Conventions subtest ability than girls and older children. The child-level model accounted for almost $36 \%$ of the child-level variance in the Conventions subtest scores.

Alphabet letter recognition knowledge also influenced the Meaning subtest scores with the strength of the influence more than one fourth as large as in the Alphabet subtest scores. Age influenced the Meaning subtest scores both directly and indirectly through the Letters variable; older children had higher Meaning subtest ability than younger children. Gender influenced Meaning subtest scores only indirectly through the
letters variable. The child-level model accounted for almost $19 \%$ of the child-level variance in the Meaning subtest scores.

The classroom means of the Letters variable predicted the classroom means of the Alphabet, Conventions, and Meaning subtest scores. Residuals of classroom means of all three subtest scores were significantly different from zero indicating the need for the multilevel model. The proportion of variance in TERA-3 subtest classroom means accounted for by the classroom mean ability to recognize the letters of the alphabet was 88,60 , and 27 percent, respectively for the Alphabet, Conventions, and Meaning subtests.

Figure 2.The confirmatory factor analysis measurement model of the TERA-3 subtest items. All pictured correlations were statistically significant at $\alpha=.05$. The $t$ statistics ranged from a low value of 2.81 for Conventions measured by C 3 to a high value of 48.35 for Alphabet measured by A4. The complete set of factor loadings is presented in Table 3.


Figure 3.The two-level path analysis of the child- and classroom-level TERA-3 and alphabet letter recognition (Letters) factor scores. The pictured child-level correlations were all statistically significant at $\alpha=.05$. The child-level $t$ statistics ranged from a low of 1.98 for Alphabet by gender to a high value of 42.14 for Letters regressed on Alphabet. Additionally, the classroom-level $t$ statistics ranged from a low value of 4.67 for Mean Letters regressed on Mean Meaning to a high value of 16.51 for Mean Letters regressed on Mean Alphabet.


## Conclusion

Participating classrooms were sponsored by agencies that were either recruited by curriculum developers for participation or whose sponsoring agencies requested participation and funded some extent of their participation. However, the participating children form a large, mostly urban, African American population of children from low-income families who attended a variety of preschool programs.

## Child-Level Path Analysis

The path analyses indicated that alphabet letter knowledge predicted all three TERA-3 subtest abilities. The TERA-3 items measured alphabet letter knowledge, conventions of print, and emergent comprehension.

The findings from this study indicated the ability to recognize the upper- and lowercase letters in non-alphabetic order in classroom environments suggested by Lonigan et al. (1998) was also highly associated with measures of graphophemic knowledge, conventions of print, and knowledge of environmental print. Moreover, the classroom mean ability to recognize the letters of the alphabet accounted for differences in classroom mean measures of other emergent literacy abilities.

What is more, the link between phonological sensitivity and alphabet knowledge is especially problematic for boys from lowincome families. McGuiness et al. (1995) found that deficits in phonological awareness were more problematic to future reading achievement for boys than girls. The results of this study suggest these deficits for boys from low-income families
may begin at the point of learning to recognize the letters of the alphabet.

The residuals of the child-level Alphabet and Conventions subtest were correlated. Both Bialystok (1991) and Sulzby (1992) suggested the influence of alphabet letter knowledge is linked to concept of word. The relationship between the alphabet letter recognition variable and the Conventions subtest scores may reflect the influence of letter recognition ability on those Conventions items requiring children to use their concept of word to respond to items that required them to follow pictured text as it was read to them or to point to various words.

## Classroom-Level Path Analysis

The classroom-level model used four variables, the Letters, Alphabet, Conventions, Meaning variables aggregated at the classroom level. The classroom mean of the Letters variable predicted the classroom means of the Alphabet, Conventions, and Meaning subtest scores. The intraclass correlations for Letters, Alphabet, Conventions, and Meaning were $.19, .21, .15$, and .17, respectively. These intraclass correlations are relatively large for a homogeneous population. For instance, in heterogeneous populations, Bryk and Raudenbush (1992) estimated $18 \%$ of the variance in math achievement scores of children in the 1982 High School and Beyond Survey was betweenschools and Goldstein (1987) estimated $9 \%$ and $13 \%$ of the variance in reading achievement of elementary school children was between-schools and between-classes, respectively.

A possible explanation for these relatively large intraclass correlations is instruction of some of the subtest constructs is more readily adapted to the use of explicit instruction to enhance child learning. In fact, historical evaluation of the literacy curriculum used with preschool children indicated the greatest increases in mean TERA-3 subtest scores occurred with the Alphabet subtest scores. Additionally, the percent of available subtest items used in this study (items with means greater than .05 ) were 76,57 and 33 percent for the Alphabet, Conventions, and Meaning subtests, respectively, and 88 and 60 percent of the classroom-level variance in the Alphabet and Conventions subtest means was accounted for by the classroom mean ability of the children to recognize the letters of the alphabet. The children
in this study could correctly respond to a much greater percent of the Alphabet and Conventions than Meaning items which suggests higher ability in those areas. That ability was directly related to their classrooms' combined ability to identify the upper- and lowercase letters of the alphabet.

Because of this evidence and the explicit teaching of letter knowledge among other skills, classroom mean letter knowledge is seen as a measure of the implementation of the literacy curriculum, especially because participation was not uniformly implemented across sites in terms of the length of involvement during the school year or in terms of previous literacy curriculum experience of classroom teachers. Some teachers were new to the curriculum having worked with it less than a semester and other teachers had worked with it for several years. Supporting this implementation explanation is the fact that of the classrooms with the 16 lowest mean Letters scores, 12 were new sites with teachers new to the curriculum and with participation beginning after the winter holidays. The remaining four classrooms were early intervention special education classrooms. The implications of this explanation suggest mean classroom letter recognition ability may be simple measure of the quality of emergent literacy curricula and experiences.

Perceptual Learning Theory of Alphabet Letter Recognition

Inspection of Table 1 indicates the most frequently recognized letters were uppercase A, B, and C and upper- and lowercase X and O . This coupled with the alphabet letter summed scores depicted in Figure 1 suggests rising kindergarten children recognized more of the uppercase letters; however, it cannot be determined from this study whether this is because the uppercase letters are more visually distinct and therefore more easily recognized (Tinker, 1931) or whether the uppercase letters are taught first to preschool children (Adams, 1990). The path analysis also indicated boys' ability to recognize letters of the alphabet was lower than girls and older children's ability was higher than younger children.

These findings are limited by the lack of experimental design, but the size of the sample indicates these are areas for further research. The fact that alphabet letter knowledge is an integral
part of a broader set of emergent literacy skills and is frequently learned in conjunction with broader skills enhances Sulzby's (1983) call for a better understanding of how children learn letter names and the processes they use to recognize the letter forms. If children, in fact, recognize letters of the alphabet by their distinctive features, a more controlled study is needed in which data are collected earlier in the learning process and at several time points with instructional strategy modeled into the design. Perhaps as children actively engage in learning to recognize the letters of the alphabet, the construct changes from a multidimensional to a unitary one.

## References

Adams, M. J. (1990). Learning about print: The first steps. In Beginning to read: Thinking and learning about print (pp. 333-367). Cambridge, MA: MIT Press.

Baker, T. A., Torgenson, J. K., \& Wagner, R. K. (1992). The role of orthographic processing skills on five different reading tasks. Reading Research Quarterly, 27, 334-345.

Baydar, N., Brooks-Gunn, J., \& Furstenberg, F. F. (1993). Early warning signs of functional illiteracy: Predictors in childhood and adolescence. Child Development, 64, 815-829.

Bialystok, E. (1991). Letters, sounds, and symbols: Changes in children's understanding of written language. Applied Psycholinguistics 12, 75-89.

Bond, G. L., \& Dykstra, R. (1967). The cooperative research program in first-grade reading instruction. Reading Research Quarterly, 2, 5-142.

Bowey, J. A. (1994). Phonological sensitivity in novice readers and nonreaders. Journal of Experimental Child Psychology, 58, 134-159.

Bryk, A. S., \& Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analyses methods. Newbury Park, CA: Sage.

Chall, J. S. (1983). The ABC's of reading: Is the alphabet necessary? In Learning to read: The great debate (pp.140-159). NY: McGraw Hill.

Ehri, L. C. (1983). A critique of five studies related to letter-name knowledge and learning to read. In L. M. Gentile, M. L. Kamil, and J. S. Blanchard (Eds.) Reading Research Revisited (143-153).Columbus, OH: Charles E. Morrill.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, \& M. Coltheart (Eds.), Surface dyslexia (pp. 301-330). London: Erlbaum.

Gibson, E. J., \& Levin, H. (1975). A theory of perceptual learning and its relevance for understanding reading. In The psychology of reading. Cambridge, MA: MIT Press.

Goldstein, H. (1987). Multilevel models in educational and social research. New York: Oxford University Press.

Goswami, U. (1993). Orthographic analogies and reading development. Bulletin of the British Psychological Society, 6, 312-316.

Gough, P. B., \& Hillinger, M. L. (1980). Learning to read: An unnatural act. Bulletin of the Orton Society, 30, 171-176.

Gough, P. B., Juel, C., \& Griffin, P.L. (1992). Reading, spelling, and orthographic cipher. In P.B. Gough, L. C. Ehri, \& R. Treiman (Eds.), Reading acquisition (pp. 35-48). Hillsdale, NJ: Erlbaum.

Hu, L. T. \& Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 155.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. Journal of Educational Psychology, 80, 437-447.

Lonigan, C. J., Burgess, S. R., \& Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent-variable longitudinal study. Developmental Psychology, 36, 596-613.

Lonigan, C. J., Burgess, S. R., Anthony, J. L., \& Barker, T. A. (1998). Development of phonological sensitivity in 2- to 5-year-old children. Journal of Educational Psychology, 90, 294-311.

Mason, J. M. (1980). When do children begin to read: An exploration of four-year-old children's letter and word reading competencies. Reading Research Quarterly, 13, 203-327.

McGuiness, D., McGuiness, C., \& Donohue, J. (1995). Phonological training and the alphabet principle: Evidence for reciprocal causality. Reading Research Quarterly, 30, 830852.

Muthén, B. O., \& Muthén, L. (2003). Mplus statistical analysis with latent variables Version 2.13. Los Angeles: Muthén \& Muthén.

Perfetti, C. A. (1984, November). Reading acquisition and beyond: Decoding includes cognition. American Journal of Education, 40-57.

Reid, D. K., Hresko, W. P., \& Hammill, D. D. (2000a). The Test of Early Reading Ability Third Edition. Austin, TX: Pro-Ed.

Reid, D. K., Hresko, W. P., \& Hammill, D. D. (2000b). TERA-3 examiners manual. Austin, TX: pro•ed.

Riley, J. L. (1996). The ability to label letters of the alphabet at school entry: A discussion on its value. Journal of Research in Reading 19, 87-101.

Roberts, T. A. (2003). Effects of alphabetletter instruction on young children's word recognition. Journal of Educational Psychology, 95, 41-51.

Stahl, S. A., \& Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. Journal of Educational Psychology, 86, 221-234.

Stevenson, H. W., \& Newman, R. S. (1986). Long-term predictions of achievement and attitudes in mathematics and reading. Child Development, 57, 646-659.

Sulzby, E. (1983). A commentary on Ehri's critiques of five studies relating to lettername knowledge and learning to read: Broadening the question. In L. M. Gentile, M. L. Kamil, and J. S. Blanchard (Eds.) Reading Research Revisited (pp. 143-153). Columbus, OH: Charles E. Morrill.

Sulzby, E. (1992). Research directions: Transition from emergent to conventional writing. Language Arts, 69, 291-297.

Tinker, M. A. (1931). The influence of form of type on perception of words. Journal of Applied Psychology, 16, 167-174.

Valdez-Menchara, M. C., \& Whitehurst, G. J. (1992). Accelerating language development through picture book reading: a systematic extension to Mexican day care. Developmental Psychology, 28, 1106-1114.

Westat, Inc., Rockville, MD., Abt Associates, Inc., Bethesda, MD., CDM Group, I., \& Ellsworth Associates, Mclean, VA. (1998). Head Start Program Performance Measures. Second Progress Report. Head Start Research. U.S.; District of Columbia.

Whitehurst, G. (1999). Measurement of emergent literacy and literacy outcomes. Report to the Advisory Committee on Head Start Research and Evaluation. Retrieved January 2002 from http:lwww2.acf.dhhs.gov/programs/hsb/hsreac/jun 99/issue4.html.

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, J., \& Fischel, J. E. (1994). A picture book reading intervention in day care and home for children from low-in-come families. Developmental Psychology, 30, 542-555.

Whitehurst, G. J. \& Lonigan, C. J. (1998). Child development and emergent literacy. Child Development, 68, 848-872.

Whitehurst, G. J., Zevenbergen, A. A., Crone, D. A., Schultz, M. D., Velting, O. N., \& Fischel, J. E. (1999). Outcomes of an emergent literacy intervention from Head Start through second grade. Journal of Educational Psychology, 91, 261-272.

Yu, C.-Y. \& and Muthén, B. O. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Technical report in preparation.

Zill, N., \& West, J. (2001). Entering Kindergarten: A Portrait of American Children When They Begin School. Findings from the Condition of Education, 2000. (Rep. No. NCES2001035). U.S.; District of Columbia: ED Pubs.

# Deconstructing Arguments From The Case Against Hypothesis Testing 

Shlomo S. Sawilowsky<br>Educational Evaluation and Research<br>Wayne State University

The main purpose of this article is to contest the propositions that (1) hypothesis tests should be abandoned in favor of confidence intervals, and (2) science has not benefited from hypothesis testing. The minor purpose is to propose (1) descriptive statistics, graphics, and effect sizes do not obviate the need for hypothesis testing, (2) significance testing (reporting p values and leaving it to the reader to determine significance) is subjective and outside the realm of the scientific method, and (3) Bayesian and qualitative methods should be used for Bayesian and qualitative research studies, respectively.

Key words: Hypothesis testing, bracketed intervals, significance testing, effect size, Bayes, qualitative

## Introduction

There has been an increasing amount of journal space given to the case against hypothesis testing over the past quarter of a century. The ensuing debate has taken many directions and has been graced with many forms of argumentation (see, e.g., Sawilowsky, 2003a; Knapp \& Sawilowsky, 2001). Two styles of attack against hypothesis testing are contested here.

The first is the proposition that hypothesis testing should be abandoned in favor of confidence intervals. (I prefer the term "bracketed" instead of "confidence" interval for reasons noted in Sawilowsky, 2003a.) Ancillary to this attack is the proposition that hypothesis testing is tolerable if and only if it is (a) buttressed with a report of effect sizes, (b) accompanied by graphical displays, or (c) Bayesian.

The second style of attack is that hypothesis testing should be abandoned due to philosophical arguments. An example is embodied in the question if science has benefited by hypothesis testing.

Shlomo Sawilowsky is Professor of Education and Wayne State University Distinguished Faculty Fellow. Email: shlomo@wayne.edu. The author gratefully acknowledges discussions on the ether with Rabbi Chaim Moshe Bergstein.

The "Confidence" Interval Attack
Neyman (1934), who discovered the bracketed interval, equated the probabilities associated with its lower and upper bound with "the ordinary concept of probability" (1934, p. 590). Initially, he seemed to equate it with the fiducial argument promulgated by Fisher (1930). The presumed lack of difference in the derivation of bracketed intervals and fiducial probabilities was the focus of the discussion subsequent to the reading of Neyman's (1934) paper before the Royal Statistical Society. Bowley (1934) raised the question and presented his answer, "I am not at all sure that the 'confidence' is not a 'confidence trick'... Does it really take us any further?... I think it does not" (p. 609). He considered bracketed intervals to be nothing more than ordinary probabilities expressed in a new form.

Neyman (1934) replied that "questions raised in the discussion on the confidence intervals would require too much space. In fact, to clear up the matter entirely, a separate publication is needed...[and] this is in preparation" (p. 623). He alluded to the nature of the response that would follow: "It has been suggested in the discussion that I used the term 'confidence coefficient' instead of the term 'fiducial probability'. This is certainly a misunderstanding" (p. 623). Did Neyman differentiate between his proposed bracketed interval and the venerable hypothesis test?

No. Neyman (1935) immediately disabused readers of the statistical literature of this notion. He stated, "The problem of estimation in its form of confidence intervals stands entirely within the bound of the theory of probability" (p. 116), as does hypothesis testing. How, then, did the claim that bracketed intervals are superior and preferred eventually arise as a weapon in the arsenal of the camp attempting to make a case against hypothesis testing?

Neyman (1941) reviewed the development of the bracketed interval, which is translated from the Polish "przedzial ufności." He mentioned this phrase in 1930 in lectures at the University of Warsaw and the Central College (Agriculture) in Warsaw, Poland. Prior to the redaction of the theory, Pytkowksi (1932) published a practical application.

Neyman (1941) recounted that he had noticed numerical similarities obtained with his method and that of the fiducial argument. As a result, he had initially assumed the two paradigms were identical. Neyman was satisfied with considering the bracketed interval as an extension of the fiducial argument because Fisher (1930) had priority.

Eventually, Neyman (1934) became estranged from the fiducial argument. He no longer considered the two theories interchangeable. He left the reasons unstated in his opening presentation before the Society.

Fisher (1934) attended the reading as a discussant. Historical accounts of the exchange were varied. Some expressed chagrin with Fisher, who offered minimal comments on the new methodology, and instead concentrated on the relative merits of random vs purposive sampling selection. Others, in noting Bowley's (1934) comment that the paper was difficult to understand, assumed that Fisher might have neglected to read Neyman's paper prior to the reading and simply didn't follow it. Still others proposed that this was Fisher's feeble attempt at blocking his baton from being passed to Neyman, just as Karl Pearson had tried in vain two decades prior with Fisher.

These reports misrepresented Fisher's response. Most of his comments were directed to the sampling problem because that was the primary thesis of Neyman's (1934) paper. Moreover, a careful review of the published
discussion indicates that Fisher understood the paper's implication quite well. His response was a terse defense of the fiducial argument as the explanation of ordinary probability.

Neyman (1941) was surprised! Fiducial probability and the fiducial distribution of a parameter were "more or less, lapsus linguae, difficult to avoid in the early stages of a new theory" (p. 129). The fiducial argument was vague, misconceived, and vacuous in explaining ordinary probability.

The aftermath took the form of considerable and animated debate in the literature on the fiducial argument. Many mathematical statisticians, regardless of theoretical persuasion, joined in the fray by publishing their support or concern. Wald (1939), Wald and Wolfowitz (1939), and Welch (1939) sided with the bracketed interval. Fisher (1935), Starkey (1938), Sukhatme (1938), and Yates (1939) defended the fiducial argument. Pitman (1939) opined that the two theories were essentially the same, as did Bartlett (1939) to a lesser extent.

Bartlett $(1936,1939)$ also escalated the debate with the contention that where results diverge, the fault lies within the fiducial argument. As can be imagined, Fisher (1937, 1939a, 1939b) and Yates (1939) accepted the gauntlet. Jeffreys (1940) attempted to restore calm in claiming that the bracketed interval and the fiducial argument were both subsumed under inverse probability in the system of Bayes. This had no effect on the debate, of course, because few of the combatants were Bayesian. The controversy would only die with Fisher.

Neyman (1941) succinctly described the relationship between the two theories: "There is none" (p. 130) because "the theories of fiducial argument and of confidence intervals differ in their basic conceptions" (p. 149). He was:
inclined to think that the literature on the theory of fiducial argument was born out of ideas similar to those underlying the theory of confidence intervals. These ideas, however, seem to have been too vague to crystallize into a mathematical theory. Instead, they resulted in misconceptions of 'fiducial probability' and 'fiducial distribution of a
parameter'... In this light, the theory of fiducial inference is simply non-existent. (p. 149)

Return to the "confidence" interval attack against hypothesis testing. Fisher's fiducial argument as the explanation of probability was challenged and defeated. However, the ordinary understanding of probability, even in its application to Fisher's F test, was never challenged, much less defeated. Those who have raised the bracketed interval attack against hypothesis testing are merely exploiting Fisher's discredited nomenclature and explanation of probability as he applied it to hypothesis testing.

Ordinary probability is synonymous in the theories of hypothesis testing and bracketed intervals. Certainly, this was Neyman's (1934) view. That is why we concluded, "There is an illogical swagger associated with criticizing hypothesis testing and subsequently advocating CIs [confidence intervals]" (Compton \& Sawilowsky, 2003, p. 584).

Philosophical Attack
"Has science benefited from hypothesis testing?" The question is silly. No reputable quantitative physical, behavioral, or social scientist would overlook the breadth and depth of scholarly knowledge and its impact on society that has accrued from over a century of hypothesis testing. The definitive evidence: William Sealy Gosset created the $t$ test to make better beer.

In an invited paper in this issue of Journal of Modern Applied Statistical Methods, Professor Dayton addresses alternative strategies to hypothesis testing. The motivating reference, Carver (1978), championed the case against hypothesis testing. Carver's (1978) attack was based on a variant of the philosophical attack: speculation and assertion. "Even if properly used in the scientific method, educational research would still be better off without statistical significance testing" (p. 398). Carver (1993) offered an "Einstein" gambit:

An example from the history of science will help to illustrate this point. Michelson and Morley (1887) collected
data relevant to the speed of light, testing the hypothesis that light travels through a medium called luminiferous ether. If this ether existed, then light should travel faster when moving in the same direction as the motion of the earth - similar to a boat traveling faster when going downstream compared with upstream. Michelson and Morley interpreted their published data, without tests of significance, as indicating that light traveled the same speed no matter what direction it was traveling. However, I subjected their published data to a simple analysis of variance (ANOVA) and found statistical significance associated with the direction the light was traveling ( $\mathrm{p} .<.01$ ).

It is interesting to speculate how the course of history might have been changed if Michelson and Morley had been trained to use this corrupt form of the scientific method, that is, testing the null hypothesis first. They might have concluded that there was evidence of significant differences in the speed of light associated with its direction and that therefore there was evidence for luminiferous ether. If this ether existed, then light should travel faster when moving in the same ether. That conclusion would have set back Einstein's ideas many years, because his notions about relativity are based on light traveling in every direction at the same speed. Fortunately, Michelson and Morley did not corrupt the scientific method by testing the null hypothesis before they interpreted their data with respect to their research hypothesis. (p. 288)

The best research articles are those that include no tests of statistical significance. In a single study, these tests can be replaced with estimates of effect size and of sampling error, such as standard errors and confidence intervals. Better still, by conducting multiple studies, replication of results can replace statistical significance testing. (p. 289290)

Responses to Carver's (1993) claims appear below. In order to understand these remarks, it is necessary to preface with a description of interferometer data. Carver (1993) claimed the results were null. Indeed, the 1887 Michelson-Morley experiment is nearly unanimously touted as the most famous experiment that produced a null result. (See, e.g., Feynman, Leighton, \& Sands, 1963.)

The interferometer was invented by Michelson to estimate the speed of light. It was refined by Michelson (1881) and by Michelson and Morley (1887a, 1887b) in an attempt to acquire evidence on the medium of propagation of light called ether proposed by Aristotle. The hypothesized value, equal to the Earth's orbital velocity, was approximately $30 \mathrm{~km} / \mathrm{s}$.

Michelson and Morley (1887a) did not use hypothesis tests (which had yet to be invented, not withstanding allegations regarding the dating of the sign test). Initially, they presented "the results of the observations... graphically" (p. 333). Visual inspection led to the conclusion there was an observed fringe shift, although it was less than what would be expected if the ether existed as hypothesized. They wrote, "It seems fair to conclude from the figure that if there is any displacement due to the relative motion of the earth and the luminiferous ether, this cannot be much greater than 0.01 of the distance between the fringes" (Michelson \& Morley, 1887a, p. 333).

Next, they presented descriptive statistics. This led to the conclusion that "the ether is probably less than one sixth the earth's orbital velocity, and certainly less than one fourth" (p. 341). Values probably less than 5 $\mathrm{km} / \mathrm{s}$ and certainly less than $7.5 \mathrm{~km} / \mathrm{s}$ are not null, although different from the expected value of $30 \mathrm{~km} / \mathrm{s}$. Some results on interferometer experiments conducted from 1887-1935 are compiled in Table 1.

The only null results via interferometry were obtained by Kennedy in 1926. His results were criticized by Illingsworth (1927), who found the equipment suffered from a "reduced optical system" (p. 692). Múnera (1998) noted that the Kennedy experiment was unclear regarding the local solar time of the initial orientation of the interferometer, which may have been at one of the four times per day that

Table 1. A Sampling Of Interferometry Results.

| Experimenter | Date | Velocity (k/s) |
| :---: | :---: | :---: |
| Michelson \& - |  |  |
| Morley | 1887 | $5-\leq 7.5$ |
| Morley \& Miller | 1902-4 | $8.7 \pm 0.6$ |
| Morley \& Miller | 1905 | 7.5 |
| Miller | 4/1/1925 | $10.1 \pm .33$ |
| Miller | 8/1/1925 | $11.2 \pm .33$ |
| Miller | 9/15/1925 | $9.6 \pm .33$ |
| Miller | 9/23/1925 | 8.22 |
| Miller | 2/8/26 | $9.3 \pm .33$ |
| Picard \& Stahel | 1926 | 6.9 |
| Picard \& Stahel | 1927 | $1.45 \pm .007$ |
| Illingworth | 1927 | <3-5 |
| Michelson, Pease, \& Pearson | 1929 | 20 |
| Joos | 1930 | < 1.5 |
| Kennedy \& Thornkike | 1932 | 24 |
| Michelson, Pease, \& Pearson | 1935 | 20 |

the expected shift tends to zero. Subsequent experiments conducted by Illingsworth (1927) with Kennedy's equipment, but with resilvered mirrors, presented nonnull results.

A variety of technical corrections were introduced to account for the non-null results. Experiments were carefully designed to rule out rival hypotheses, such as temperature, drift, sign of displacement, diurnal variation, and intersession averaging. Nevertheless, no study produced null results.

Most interferometer experiments were conducted by Miller (1933). He took more than 200,000 readings from 1902-1927 based on 12,500 turns of the interferometer, including a joint effort with Morley in the early 1900s. (In comparison, Michelson and Morley made 36 turns in four days, and Piccard and Stahel made 96 turns in Belgium and 60 turns in Brussels.) Yet, Miller never obtained a null result.

Shankland (et al., 1955) was Miller's assistant, and subsequently was Professor of physics at Case Western Reserve University (where Morley was Professor of chemistry until 1906). After the death of his boss, he criticized Miller's work on the ether, notably with
assistance from Albert Einstein. DeMeo (2000, 2001) strenuously defended Miller against Shankland's criticisms. (The reader interested in the dissident literature on ether should read DeMeo, 2000, 2001; and Múnera, 1998). Later, Shankland (1973, p. 2283) cited a letter received from Einstein dated 31 August 1954:

I thank you very much for sending me your careful study about the Miller experiments. Those experiments, conducted with so much care, merit, of course, a very careful statistical investigation. This is more so as the existence of a not trivial positive effect would affect very deeply the fundament of theoretical physics as it is presently accepted.

You have shown convincingly that the observed effect... has nothing to do with 'ether-wind', but has to do with differences of temperature.

Einstein's letter is instructive for many reasons. First, he believed the interferometer experiments on the ether "merit, of course, a very careful statistical analysis" [emphasis added]. Second, as late as the year of his death, Einstein still believed that the interferometer experiments were a threat to his special theory of relativity. Third, he had not updated his knowledge many years after the specter of temperature as a confounding variable was first raised. The Cleveland Plain Dealer (27 January 1926) published an exchange between Einstein and Miller, with the latter concluding,
"The trouble with Prof. Einstein is that he knows nothing about my results," Dr. Miller said. "He has been saying for thirty years that the interferometer experiments in Cleveland showed negative results. We never said they gave negative results, and they did not in fact give negative results. He ought to give me credit for knowing that temperature differences would affect results. He wrote to me in November suggesting this. I am not so simple as to make no allowance for temperature."

In his experiments in 1923, and from 1925-1926 at Mt. Wilson, Miller took many steps to control for the effects of temperature. The results were consistent with earlier measurements. Similarly, Miller (cited in Joos \& Miller, 1934) noted, "when Morley and Miller designed their interferometer in 1904 they were fully cognizant of this... Elaborate tests have been made... especially with artificial heating, for the development of methods which would be free from this effect [of temperature]" (p. 114). The Cleveland Plain Dealer (27 January 1926) added, "Speaking before scientists at the University of Berlin, Einstein said the ether drift experiments [were null in the Michelson-Morley experiment but] on Mount Wilson they showed positive results", although he attributed it to temperature and altitude.

## Einstein Gambit Declined

There were thousands of interferomic studies conducted by dozens of physicists since 1887, and in all but one experiment the results were demonstrably non-null. The only known null result was subsequently determined to be caused by a miscalibrated instrument. When the instrument was resilvered, and the experiment replicated in the same location, the results were about $4 \mathrm{~km} / \mathrm{s}$.

Carver (1993) conducted a simple analysis of variance (ANOVA) and found statistical significance ( $p<.01$ ). These results are tenable, assuming the null hypothesis was the observations did not differ from zero. Nevertheless, Carver's (1993) analysis suffers from a bewildering array of questions, such as:

- What data set was used? Was it from the noon readings, the afternoon readings, or a combination of readings? Was it from July $8^{\text {th }}, 9^{\text {th }}, 11^{\text {th }}$, or $12^{\text {th }}$ of 1887 ; or perhaps some combination of days? Did it include all 36 turns of the interferometer, or some subset?
- What was the value of F?
- What were the degrees of freedom?
- Were the underlying assumptions of independence, homoscedasticity, and normality considered?
- Were covariates such as diurnal variation or drift considered?
- How was intersession averaging based on different calibration curves handled?
- According to Carver's (1993) advice and recommendation, why did he fail to present summary statistics or a graphic display of the results (either prior to the ANOVA or afterwards)?

Carver (1993) claimed that this significant result from the hypothesis test would have set Einstein back many years. This is unwarranted speculation. In his lecture in Berlin, Einstein rejected the 1887 Michelson-Morley results as being nonnull, despite the evidence contained within their descriptive statistics and graphs. Similarly, he would have ignored the outcome of a hypothesis test.

Einstein's theory was not based on any experimental evidence. At various times throughout his career, Einstein reminisced that it was based on the principles of Maxwell and Lorentz, and he had not relied on the MichelsonMorley experiment. Holton (1969, 1988) suggested that not only did the interferometer experiments have little or no impact, but there is evidence that Einstein was unaware of the Michelson-Morley experiment prior to developing the special theory of relativity.

Interferometer experimenters presented graphical displays, from simple scatter grams and histograms to more complex time series charts and hodograms. All pictorial representations substantiated nonzero results. Some of the latter interferometer experimenters reported standard errors. (Obviously, those who did not were remiss.) Many of the latter experimenters also reported bracketed intervals, and zero was not in them. Múnera (1998) summarized the bulk of interferometer studies with a bracketed interval, and zero was not in it. If statistical tests had been invented by 1887, it would have been easy to confirm the data were statistically significantly different from zero. Even Shankland (et al., 1955; 1973) was forced to admit this.

Carver (1993) reported an effect size (eta squared) of .005. He concluded "if Michelson and Morley had been forced ... to do a test of statistical significance, they could have
minimized its influence by reporting this effect size measure indicating that less that $1 \%$ of the variance in the speed of light was associated with its direction" (p. 289). The fallacy of his analysis is Michelson and Morley's (1887a, 1887b) experiment obtained results of 5 to 7.5 $\mathrm{km} / \mathrm{s}$. Regardless of what percent of variance it represents, how can anyone call a speed that exceeds the Earth's satellite orbital velocity "null" and "seek to minimize its influence"?

Of paramount importance, however, Carver (1993) tested the wrong hypothesis. Data inspection and graphs demonstrated interferomic data did not support the static model of luminiferous ether as a medium of propagation for light. Should a hypothesis test be desired, the correct test is whether the data were statistically significantly different - not from zero - but rather, from the hypothesized value of $30 \mathrm{k} / \mathrm{s}$.

Carver (1993) described the process of conducting hypothesis tests prior to examining descriptive data as a corruption of the scientific method. This is a straw-person argument. Who promotes conducting hypothesis tests as a first step in the analysis of data? Who objects to examining raw data (e.g., for data entry errors, outliers), computing descriptive statistics, and inspecting graphics prior, or as a follow-up, to conducting hypothesis tests?

Carver (1993) stated the best research articles are those that contain no hypothesis tests. This regressive approach would truly set quantitative physical, behavioral, and social science back more than a century. Reasonable people have different expectations of what constitutes a rare event vs what constitutes a common event expected by chance alone. This is true with a single study, and all the more so with many replications of a study. The debate is diminished, and possibly vanishes, with the simple agreement on a threshold (i.e., nominal alpha level) prior to conducting an experiment.

Carver's (1993) reliance on reporting effect sizes as a panacea is naïve. Effect sizes are sensitive to their own underlying assumptions. In addition, the process of enclosing effect sizes in a bracketed interval relies on the same probabilities as does the obtained value of a hypothesis test. Carver (1993) also recommended the practice of reporting an effect size whether the hypothesis
test "is significant or not" (p. 288). This leads to the "trouble with trivials" problem (see e.g., Sawilowsky, 2003b, 2003c).

Currently, it is a popular slogan among effect size enthusiasts to warn against "becoming stupid in another metric." Yet, Carver (1993) interpreted an eta squared of . 005 as null to minimize the study outcome. The experimental results Carver (1993) sought to minimize were speeds of over 16,750 miles per hour!

The Next Generation of Arguments
As soon as these two lines of attack against hypothesis testing falter, three more assaults are quickly proffered. This is not the place to elaborate on them, but they are parried briefly below.

The first is to replace hypothesis testing with significance testing. P values are reported and it is left to the reader to decide if it is significant. Aside from being outside the realm of the scientific method, subjective significance testing is, in my view, a recipe for disaster (Knapp \& Sawilowsky, 2001). (Note that Carver's, 1978, 1993, attack is actually against hypothesis testing, although he calls it a case against significance testing.)

The second is to abandon the frequentist approach and conduct a Bayesian analysis. I strongly promote the method of Bayes in selecting a pinch hitter in baseball because of the plethora of informative priors. However, in the absence of definitive objective priors, a condition that pervades most of physical, behavioral, and social science, Bayesian methods are not likely to be optimal.

The third is to abandon quantitative methodology altogether in favor of qualitative techniques. I discussed this option elsewhere (Sawilowsky, 1999). Qualitative methods should be used when the research hypothesis is qualitative, not because of some perceived limitation of a quantitative method in pursuing a quantitative research question.

## References

Bartlett, M. S. (1936). The information available in small samples. Proceedings of the Cambridge Philosophical Society, 32, 560-566.

Bartlett, M. S. (1939). Complete simultaneous fiducial distributions, Annals of Mathematical Statistics, 10, 129-138.

Bowley, A. L. (1934). Discussion on Dr. Neyman's paper. The Journal of the Royal Statistical Society, 97, 607-610.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Compton, S., \& Sawilowsky, S. (2003). Do not discourage the use of p values. Annals of Emergency Medicine, 41(4), p. 584.

DeMeo, J. (2001). Dayton Miller's ether-drift experiments: A fresh look. Infinite Energy Magazine, 38, 72-82.

DeMeo, J. (2002). Dayton Miller's ether-drift experiments: A fresh look. http://www.orgonelab.org/miller.htm.

Feynman, R. P., Leighton, R. B., \& Sands, M. (1963). The Feynman lectures on physics: Mainly mechanics, radiation, and heat. Vol 1. Reading, MA: Addison-Wesley, 15-5.

Fisher, R. A. (1930). Inverse probability. Proceedings of the Cambridge Philosophical Society, 26, 528-535.

Fisher, R. A. (1934). Discussion on Dr. Neyman's paper. The Journal of the Royal Statistical Society, 97, 614-619.

Fisher, R. A. (1935). The fiducial argument in statistical inference. Annals of Eugenics, 6, 391-398.

Fisher, R. A. (1937). On a point raised by M. S. Bartlett on fiducial probability. Annals of Eugenics, 7, 370-375.

Fisher, R. A. (1939a). The comparison of samples with possibly unequal variances. Annal of Eugenics, 9, 174-180.

Fisher, R. A. (1939b). A note on fiducial inference. Annals of Mathematical Statistics, 10, 383-388.

Jeffries, H. (1940). Note on the BehrensFisher formula. Annals of Eugenics, 10, 48-51.

Joos, G., \& Miller, D. (1934, January 15). Letters to the Editor. Physical Review, 45, 114.

Holton, G. (1969). Einstein, Michelson, and the 'crucial' experiment. Isis, 60 (1969).

Holton, G. (1988), Thematic origins of scientific thought, Kepler to Einstein. (Revised ed.). Cambridge, MA: Harvard University Press.

Illingworth, K. K. (1927) A repetition of the M-M experiment using Kennedy's refinement. Physics Review, 30, 692-696.

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

Michelson, A. A. (1881). The relative motion of the earth of the luminiferous aether. American Journal of Science, 22(S3), 120-129.

Michelson, A. A., \& Morley, E. W. (1887a). On the relative motion of the Earth and the luminiferous ether. American Journal of Science, 34(S3), 333-345.

Michelson, A. A., \& Morley, E. W. (1887b). On the relative motion of the earth and the luminiferous aether. Philosophical Magazine, 24(151) S5, 449-463.

Miller, D. (1933, July). The ether-drift experiment and the determination of absolute motion of the Earth. Review of Modern Physics, 5(2), 203-242.

Múnera, H. A. (1998). MichelsonMorley experiments revisited: Systematic errors, consistency among different experiments, and compatibility with absolute space. Apeiron, 5, 37-54.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive sampling. The Journal of the Royal Statistical Society, 97, 558-625.

Neyman, J. (1935). On the problem of confidence intervals. Annals of Mathematical Statistics, 6, 111-116.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. Biometrika, 32, 128-150.

Pitman, E. J. G. (1939). The estimation of the location and scale parameters of a continuous population of any given form. Biometrika, 30, 391-421.

Pytkowski, W. (1932). The dependence of the income in small farms upon their area, the outlay and the capital invested in cows. Warsaw: Bibljoteka Pulawska.

Sawilowsky, S. (1999). Quasiexperimental design: The legacy of Campbell and Stanley. In (Bruno D. Zumbo, Ed.) Social indicators/quality of life research methods: Methodological developments and issues, Yearbook 1999. Norwell, MA: Kluwer.

Sawilowsky, S. (2003a). A different future for social and behavioral science research. Journal of Modern Applied Statistical Methods, 2(1), 128-132.

Sawilowsky, S. (2003b). You think you've got trivials? Journal of Modern Applied Statistical Methods, 2(1), 218-225.

Sawilowsky, S. (2003c). Trivials: The birth, sale, and final production of meta-analysis. Journal of Modern Applied Statistical Methods, 2(1), 242-246.

Shankland, R., McCuskey, S. W., Leone, F.C., \& Kuerti, G. (1955). New analysis of the interferometer observations of Dayton C. Miller. Review of Modern Physics, 27(2), 167178.

Shankland, R. (1973). Michelson's role in the development of relativity. Applied Optics, 12(10), 2280-2287.

Starkey, D. M. (1938). A test of the significance of the difference between means of samples from two normal populations without assuming equal variances. Annals of Mathematical Statistics, 9, 201-213.

Sukhatme, P. V. (1938). On Fisher and Behrens' test of significance of the difference in means of two normal samples. Sankhyā, 4, 3948.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, Annals of Mathematical Statistics, 10, 299-326.

Wald, A., \& Wolfowitz, J. (1939). Confidence limits for continuous distribution functions, Annals of Mathematical Statistics, 10, 105-118.

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location, Annals of Mathematical Statistics, 10, 58-69.

Yates, F. (1939). An apparent inconsistency arising from tesets of significance based on fiducial distributions of unknown parameters. Proceedings of the Cambridge Philosophical Society, 35, 579-591.

## Brief Reports

# A Note On MLEs For Normal Distribution Parameters Based On Disjoint Partial Sums Of A Random Sample 

W. J. Hurley<br>Royal Military College of Canada

Maximum likelihood estimators are computed for the parameters of a normal distribution based on disjoint partial sums of a random sample. It has application in the disaggregation of financial data.

## Introduction

## Motivation

The Canadian Forces conducts much of its army individual training at the Combat Training Center (CTC) in eastern Canada. Over the 2001-2002 Training Year, 97 serials (a "serial" is an instance of a "course") were run for a total of 2008 students. The overall expenditure on ammunition was $\$ 28.8$ million. The Commander, CTC, was interested in developing a model of the ammunition dollar cost for each type of course in order to help him assess the risk of over-expending his annual ammunition budget for a given slate of serials. At the point of budgetary deliberations for a given fiscal year, the ammunition cost for any serial is uncertain due primarily to uncertain course enrollments, uncertain student failure rates, and uncertain weather (ranges are closed when it gets dry due to the threat of forest fires).

As a first pass, we conceptualized the ammunition cost of a course as a normal random variable. To estimate its mean and variance, it would be reasonable to use historical data. For some courses this is what we did. However there were some high demand courses where a number of serials were run each year, and unfortunately, ammunition expenditures for these individual serials were aggregated into a single number for the year.

Bill Hurley is Professor of Business Administration. His research interests are military operations research, decision analysis, game theory, logistics modeling and the application of OR to problems in sport. Contact him at hurley-w@rmc.ca

The expenditures for individual serials were not tracked. Hence, for these high demand courses, the problem was to estimate the normal distribution parameters using this aggregated data.

With this background in mind, suppose the ammunition cost for a particular course is a normal random variable with mean $\mu$ and variance $\sigma^{2}$. Let

$$
X=\left\{X_{1}, X_{2}, \ldots, X_{n}\right\}
$$

be an iid sample from this distribution. Unfortunately we cannot observe individual elements of this sample. Rather, we can only observe a sample of disjoint partial sums. Suppose the sample is partitioned into sets $S_{1}, S_{2}, \ldots, S_{m}$ with cardinalities $k_{1}, k_{2}, \ldots, k_{m}$ where

$$
\begin{aligned}
& S_{1} \cup S_{2} \cup \ldots \cup S_{m}=X \\
& S_{i} \cap S_{j}=0 \quad \text { for all } i \neq j \quad \text { and } \\
& k_{1}+k_{2}+\ldots+k_{m}=n .
\end{aligned}
$$

Let $\kappa(i)$ be the set of indices of the elements of $S_{i}$. For instance if $S_{2}=\left\{X_{2}, X_{3}, X_{7}\right\}$, then $\kappa(i)=\{2,3,7\}$. Then we observe the set of partial sums, $Y=\left\{Y_{1}, Y_{2}, \ldots, Y_{m}\right\}$, where

$$
Y_{i}=\sum_{j \in \kappa(i)} X_{j} \quad \text { for } i=1,2, \ldots, m
$$

We want to compute MLEs for $\mu$ and $\sigma^{2}$ using $Y$ rather than $X$.

There has been a lot of research on grouped and combined datasets. See, for example, the work of Rao (1973). However, to my knowledge, the estimation problem described above has not been mentioned in the literature.

## Solution

Note first that $Y_{i}$ is normally distributed with mean $k_{i} \mu$ and variance $k_{i} \sigma^{2}$. Also, the $Y_{i}$ are independent since the partial sums are disjoint. Hence the likelihood function is
$L\left(\mu, \sigma^{2}\right)=\prod_{i} \frac{1}{\sqrt{2 \pi k_{i} \sigma^{2}}} \exp \left[-\sum_{i} \frac{1}{2} \frac{\left(y_{i}-k_{i} \mu\right)^{2}}{k_{i} \sigma^{2}}\right]$.
Maximizing the $\ln$ of this likelihood function gives
$\hat{\mu}=\bar{y}=\frac{\sum_{i} y_{i}}{\sum_{i} k_{i}}=\frac{\sum_{i} y_{i}}{n}$
and
$\hat{\sigma}^{2}=\frac{1}{m} \sum_{i=1}^{m} \frac{\left(y_{i}-k_{i} \bar{y}\right)^{2}}{k_{i}}$.
Note that for the special case $m=n$ (we are working at the level of the iid sample), the last equation returns the usual MLE for variance.

As for the properties of these estimators, the MLE for the mean is unbiased,
$E(\bar{Y})=\mu$,
but, not surprisingly, the estimator for the variance is biased:

$$
E\left(\frac{1}{m} \sum_{i=1}^{m} \frac{\left(Y_{i}-k_{i} \bar{Y}\right)^{2}}{k_{i}}\right)=\frac{m-1}{m} \sigma^{2} .
$$

Hence, the estimator
$\hat{\sigma}^{2}=\frac{1}{m-1} \sum_{i=1}^{m} \frac{\left(y_{i}-k_{i} \bar{y}\right)^{2}}{k_{i}}$
is an unbiased estimate of the variance.
Another aspect of this problem is how to revise these estimates as new data becomes available. At the CTC, this new data will not be aggregated. Suppose the new sample is $Z=\left\{Z_{1}, Z_{2}, \ldots, Z_{p}\right\}$. What now are the maximum likelihood estimates (MLEs) of $\mu$ and $\sigma^{2}$ based on $Y$ and $Z$ ? The answer is a straightforward application of the previous development. We simply think of $Z_{i}$ as an additional element of $Y$ having cardinality $k_{i}=1$. Hence we have that

$$
\hat{\mu}_{\mathrm{Y} \cup Z}=\bar{y}^{*}=\frac{\sum_{i} y_{i}+\sum_{i} z_{i}}{\sum_{i} k_{i}+p}=\frac{\sum_{i} y_{i}+\sum_{i} z_{i}}{n+p}
$$

and

$$
\hat{\sigma}_{Y \cup Z}^{2}=\frac{1}{m-p}\left[\sum_{i=1}^{m} \frac{\left(y_{i}-k_{i} \bar{y}^{*}\right)^{2}}{k_{i}}+\sum_{j=1}^{p}\left(z_{j}-\bar{y}^{*}\right)^{2}\right] .
$$

Another Example
Returning to the CTC problem, suppose we have the following data set for a given course:

| Fiscal Year | \#Serials | Total <br> Ammunition <br> Dollars <br> Expended |
| :---: | :---: | :---: |
| 2001 | 3 | 713,316 |
| 2002 | 2 | 486,345 |
| 2003 | 3 | 728,408 |
| 2004 | 3 | 700,843 |
| 2005 | 2 | 462,004 |

The MLEs for the mean and standard deviation are

$$
\hat{\mu}=\frac{\sum_{i} y_{i}}{\sum_{i} k_{i}}=\frac{3,090,916}{13}=237,763
$$

and

$$
\begin{aligned}
\hat{\sigma}^{2} & =\frac{1}{m-1} \sum_{i=1}^{m} \frac{\left(y_{i}-k_{i} \bar{y}\right)^{2}}{k_{i}} \\
& =11,691
\end{aligned}
$$

respectively.

## Discussion

This analysis suggests that it would be easy to find maximum likelihood estimators for the parameters of other underlying distributions. The main requirement is to identify the distributions of sums of these random variables.

An interesting extension would be to calculate maximum likelihood estimators in the case where the partial sums overlapped. In this case the $Y_{i}$ are no longer independent, and hence the likelihood function is more difficult to calculate.

## References

Rao, C. R. (1973). Linear statistical inference and its applications. NY: Wiley

# On Treating A Survey Of Convenience Sample As A Simple Random Sample 

W. Gregory Thatcher<br>Department of Health<br>University of West Florida

J. Wanzer Drane<br>Department of Epidemiology \& Biostatistics<br>University of South Carolina

Threat of bias has kept many from using data gathered in less than optimal conditions. We maintain that when convenience sampling represents race and gender at nearly correct proportions and can be beneficial, as these two variables are quite often used as stratification variables. We compared a convenience sample with a proven sample. Race and Sex were nearly proportional as was found in the proven sample. We conclude that the convenience sample can be used as though it is simple random.

Key words: Simple random sampling, convenience sampling

## Introduction

From the first semester of Introduction to Statistics through our career as scientists by whatever names, we are warned of the sampling and non-sampling errors and how to overcome them. Recently a question was asked: "May I treat my convenience sample as a simple random sample?" To answer the question we employed a sample of known qualities, SCYRBS99, the South Carolina Youth Risk Behavior Survey of 1999.

Representative coverage of Gender and Race is paramount, if the sample is to be instructive when formulating health policy, and we know that SCYRBS99 and earlier YRBS samples are constructed so that the estimates of prevalence among these two variables, as well as others, are nearly unbiased (CDC, 1999).

If we can show that the estimates of the percentages of gender and race are nearly the same in the convenience sample as are in the
W. Gregory Thatcher, Department of Health, Leisure and Exercise Science. University of West Florida, 11000 University Parkway Pensacola, FL 32514. Phone: 850-474-2598, Fax: 850-474-2106. Email: wthatcher@uwf.edu
weighted estimates of the YRBS sample of the same year, then we can at least increase our confidence in the treatment of our sample as simple random. Such a comparison does not, nor will it ever, PROVE the convenience sample to be totally unbiased and simple random, but it will go a long way toward our believing the prevalence calculated are nearly unbiased.

## Results

The estimates of gender and race prevalence will be compared to those obtained from the SCYRBS99 sample, which are treated as population constants. Tables 1 and 2 display those values together with the estimates from the convenience sample.

Remembering that $X^{2}$ is directly proportional to the sample size, which is 4421 in this case, then a Chi-square of 9.43 is not large at all. In order to reach a significance of only 0.05 , N had to be at least (4421/9.43)*3.84) $=$ 1800. This is a case in which we have too much power. From an administrative point of view we would require alpha to be equivalent to about four standard errors or 0.0001. Therefore, we are able to accept a difference of 46.66$44.36=2.30 \%$ as non-significant and administratively not important. Further, we can treat this sample as a simple random sample.

Table 1: SCMS (Convenience Sample) with expected percentages and numbers obtained from SCYRBS99. Variable $=$ GENDER. Expected F $=\mathrm{P}(\mathrm{F} \mid$ SCYRBS99 $) * 4733 . \mathrm{X}^{2}=(2409-2376.91)^{2} / 2376.91$ $+(2324-2356.09)^{2} / 2356.09=0.87, \mathrm{df}=1, \mathrm{p}$-value $=0.35$.

|  | SCMS <br> GENDER | SCYRBS99 <br> Percent <br> Count |
| :---: | :---: | :---: |
| F | 50.90 | Percent |
|  | 2409 | 50.22 |
| M | 49.10 | 2376.91 |
|  | 2324 | 49.78 |
| Total | 4733 | 2356.09 |
|  |  | 4733 |

Table 2: SCMS (Convenience Sample) with expected percentages and numbers obtained from SCYRBS99. Variable $=$ RACE. Expected B $=\mathrm{P}(\mathrm{B} \mid$ SCYRBS99 $) * 4733 . \mathrm{X}^{2}=(1961-2022.41)^{2} / 2022.41+$ $(2460-2310.65)^{2} / 2310.65+(312-399.94)^{2} / 399.94=30.85 \mathrm{df}=2$, $p$-value $=0.0000002$.

|  | SCMS | SCYRBS99 |
| :---: | :---: | :---: |
| RACE | Percent | Percent |
|  | Count | Expected |
| B | 41.43 | 42.73 |
|  | 1961 | 2022.41 |
| W | 51.98 | 48.82 |
|  | 2460 | 2310.65 |
| O | 6.59 | 8.45 |
|  | 312 | 399.94 |
| Total | 4733 | 4733 |

Table 3: A repeat of Table 2 with the O category excluded. Expected $\mathrm{B}=\mathrm{P}(\mathrm{B} \mid$ SCYRBS99 $) * 4421 . \mathrm{X}^{2}=$ $(1961-2062.84)^{2} / 2062.84+(2460-2358.16)^{2} / 2358.16=9.43, \mathrm{df}=1, \mathrm{p}$-value $=0.0021$.

|  | SCMS | SCYRBS99 |
| :---: | :---: | :---: |
| RACE | Percent | Percent |
|  | Count | Expected |
| B | 44.36 | 46.66 |
|  | 1961 | 2062.84 |
| W | 55.64 | 53.34 |
|  | 2460 | 2358.16 |
| Total | 4421 | 4421 |

Between female and male distribution the convenience sample is right on target, but the p -value of the chi-square among the three racial groups indicates a noticeable difference. An examination of actual count versus the expectations show there is an excess of white students at the expense of those captured as ' O ' or other than Black or White. If those are omitted, as usually is the case because of small numbers in more complex analyses, we have the results in Table 3.

## Conclusion

The convenience sample has nearly the same gender and racial compositions as is estimated from the SCYRBS99 data. It can then be treated as a simple random sample. For the skeptic or purist, caution should be used when generalizing across racial lines when using the SCMS data.

If stratification is made along the four categories (B,F), (B,M), (W,F) and (W,M), estimates within category should be nearly unbiased. From those four strata, comparisons could still be made without hesitation. If you insist on a larger alpha, then the RACE variable should not appear in a regression, linear or logistic, in conjunction with a set of risk and confounder variables.

References
Centers for Disease Control and Prevention. (1999). Division of Adolescent and School Health Youth Risk Behavior Survey, 1999. http://www.cdc.gov/nccdphp/dash/yrbs/

# Early Scholars <br> Conventional And Robust Paired And Independent-Samples $t$ Tests: Type I Error And Power Rates 

Katherine Fradette and H. J. Keselman<br>University of Manitoba<br>Department of Psychology

Lisa Lix<br>University of Manitoba<br>Department of Community Health Sciences

James Algina<br>University of Florida<br>Department of Educational Psychology

Rand R. Wilcox<br>University of Southern California<br>Department of Psychology

Monte Carlo methods were used to examine Type I error and power rates of 2 versions (conventional and robust) of the paired and independent-samples $t$ tests under nonnormality. The conventional (robust) versions employed least squares means and variances (trimmed means and Winsorized variances) to test for differences between groups.

Key words: Paired $t$ test, independent $t$ test, robust methods, Monte Carlo methods

## Introduction

It is well known that the paired-samples $t$ test has more power to detect a difference between the means of two groups as the correlation between the groups becomes larger. That is, as the population correlation coefficient, $\rho$, increases, the standard error of the difference between the means gets smaller, which in turn increases the magnitude of the $t$ statistic (Kirk, 1999). Equation 1, the population variance of the difference between mean values, demonstrates how the standard error of the difference between the means ( $\sigma_{\bar{x}_{1}-\bar{X}_{2}}$ ) is reduced as the value of $\rho$ increases.

Katherine Fradette (umfradet@cc.umanitoba.ca) is a graduate student in the Department of Psychology. H. J. Keselman is a Professor of Psychology, email: famoy1kf@cmich.edu. Lisa Lix (lisa_lix@cpe.umanitoba.ca) is an Assistant Professor of Community Health Sciences. James Algina (algina@ufl.edu) is a Professor of Educational Psychology. Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology.

$$
\begin{equation*}
\sigma_{\bar{X}_{1}-\bar{X}_{2}}^{2}=\sigma_{\bar{X}_{1}}^{2}+\sigma_{\bar{X}_{2}}^{2}-2 \rho \sigma_{\bar{X}_{1}} \sigma_{\bar{X}_{2}}, \tag{1}
\end{equation*}
$$

where $\sigma_{\bar{X}_{j}}^{2}=\sigma_{j}^{2} / n_{j}$ is the population variance of the mean for group $j(j=1,2)$.

It must be kept in mind, however, that the independent-samples $t$ test has twice the degrees of freedom of the paired-samples $t$ test. Generally, an increase in degrees of freedom is accompanied by an increase in power. Thus, considering the loss of degrees of freedom for the paired-samples test, there is the question of just how large $\rho$ must be in order for the pairedsamples $t$ test to achieve more power than the independent-samples $t$ test.

Vonesh (1983) demonstrated that the paired-samples $t$ test is more powerful than the independent-samples test when the correlation between the groups is .25 or larger. Furthermore, Zimmerman (1997) observed that many authors recommend the paired-samples $t$ test only if "the two groups are highly correlated" and recommend the independent samples test if "they are uncorrelated or only slightly correlated" (p. 350). Zimmerman argued, however, that such authors often fail to take into account an important consequence of the use of the independent $t$ test on dependent
observations. Namely, Zimmerman (1997) noted that the independence assumption is violated when the independent-samples $t$ test is performed on groups that are correlated, even to a very small degree, and such a violation of the independence assumption distorts both Type I and Type II error rates.

Zimmerman (1997) compared the Type I error and power performance of the paired and independent-samples $t$ tests for normally distributed data, varying the magnitude of $\rho$. He found that a correlation as small as .1 seriously distorted Type I error rates of the independentsamples $t$ test. Thus, according to Zimmerman, the practice of employing the independentsamples $t$ test when groups are slightly correlated fails to protect against distortion of the significance level and concluded that "a correlation coefficient of .10 or .15 is not sufficient evidence of independence, not even for relatively small sample sizes" (p. 359). Zimmerman also demonstrated an example in which, even when the correlation between two groups was as low as .1 , the paired $t$ test was more powerful than the independent-samples $t$ test. Consequently, contrary to the recommendations of the authors he cites (e. g., Edwards, 1979; Hays, 1988; Kurtz, 1965), Zimmerman advocates the use of the pairedsamples $t$ test even when groups are only correlated to a very small degree (i.e., .1), when distributions are normal.

The question regarding how large $\rho$ should be in order for the paired-samples $t$ test to achieve more power than the independentsamples $t$ test, when data are not normally distributed has not been examined (Wilcox, 2002). Evaluating the performance of statistics under nonnormality is important, given that psychological data are often not normal in shape (Micceri, 1989; Wilcox, 1990). Hence, the goal of this study was to extend Zimmerman's (1997) work by examining the Type I error and power rates of both the paired-samples and the independent-samples $t$ tests when distributions were nonnormal, again varying the magnitude of $\rho$.

An investigation of the performance of both the paired and independent-samples $t$ tests under nonnormality raises a problem, however. Both tests assume normally distributed data in
the population. Violation of the normality assumption leads to distortion of Type I error rates and can lead to a loss of power to detect a difference between the means (MacDonald, 1999; Wilcox, 1997). Thus, in addition to an examination of the performance of the conventional (least squares) versions of the paired and independent-samples $t$ tests, the performance of a robust version of each of the tests was also investigated.

The robust versions of the paired and independent-samples $t$ tests involve substituting robust measures of location and scale for their least squares counterparts. Specifically, the robust versions of the tests substitute trimmed means for least squares means, and Winsorized variances for least squares variances. Calculation of the trimmed mean, which is defined later in Equation 7, involves trimming a specified percentage of the observations from each tail of the distribution (for symmetric trimming), and then computing the average of the remaining observations. The Winsorized variance, which is defined later in Equation 8, is computed by first Winsorizing the observations (see Equation 5), which also involves removing the specified percentage of observations from each end of the distribution. However, in this case the eliminated observations are replaced with the smallest and largest observation not removed from the left and right side of the distribution, respectively. The Winsorized variance is then computed in the same manner as the conventional least squares variance, using the set of Winsorized observations.

Numerous studies have shown that, under nonnormality, replacing least squares means and variances with trimmed means and Winsorized variances leads to improved Type I error control and power rates for independent groups designs (e.g., Keselman, Kowalchuk \& Lix, 1998; Keselman, Wilcox, Kowalchuck \& Olejnik, 2002; Lix \& Keselman, 1998; Yuen, 1974), as well as dependent groups designs (e.g., Keselman, Kowalchuk, Algina, Lix \& Wilcox, 2000; Wilcox, 1993). In particular, Yuen (1974) was the first to propose that trimmed means and Winsorized variances be used with Welch's (1938) heteroscedastic statistic in order to test for differences between two independent groups, when distributions are nonnormal and variances
are unequal. Thus, Yuen's method helps to protect against the consequences of violating the normality assumption and is designed to be robust to variance heterogeneity. Yuen's method reduces to Welch's (1938) heteroscedastic method when the percentage of trimming is zero (Wilcox, 2002). Yuen's method can also be extended to dependent groups.

It is important to note that while the conventional paired and independent-samples $t$ statistics are used to test the hypothesis that the population means are equal ( $H_{0}: \mu_{1}=\mu_{2}$ ), the robust versions of the tests examine the hypothesis that the population trimmed means are equal ( $H_{0}: \mu_{t 1}=\mu_{t 2}$ ). Although the robust versions of the procedures are not testing precisely the same hypotheses as their conventional counterparts, both the robust and conventional versions test the hypothesis that measures of the typical score are equal. In fact, according to many researchers, the trimmed mean is a better measure of the typical score than the least squares mean, when distributions are skewed (e.g., Keselman et al., 2002).

This study compared (a) the conventional (i.e., least squares means and variances) paired-samples $t$ test, (b) the conventional independent-samples $t$ test, (c) the robust (trimmed means and Winsorized variances) paired-samples $t$ test, and (d) the robust independent-samples $t$ test, based on their empirical rates of Type I error and power. As in Zimmerman's (1997) study with normal data, it was expected that as the size of the correlation between the groups increased, both the conventional and robust versions of the pairedsamples $t$ tests would perform better than their independent-samples counterparts, in terms of their ability to maximize power while maintaining empirical Type I error rates close to the nominal $\alpha$ level. It was also expected, based on previous findings (e.g., Keselman, et al., 1998; Keselman, et al., 2000; Keselman et al., 2002; Lix et al., 1998; Wilcox, 1993; Yuen, 1974), that the robust versions of both the paired and independent-samples $t$ tests would perform better in terms of Type I error and power rates than the corresponding conventional versions.

## Methodology

Definition of the Test Statistics
Conventional Methods
Suppose that $n_{j}$ observations, $X_{1 j}, X_{2 j}, \ldots, X_{n_{j} j}$, are sampled from population $j(j=1,2)$. In order to compute the conventional independent-samples $t$ test, let $\bar{X}_{j}=\sum_{i} X_{i j} / n_{j}$ be the $j^{\text {th }}$ sample mean ( $i=1, \ldots, n_{j} ; N=\sum_{j} n_{j}$ ). Also let $S_{j}^{2}=\sum_{i}\left(X_{i j}-\bar{X}_{j}\right)^{2} /\left(n_{j}-1\right)$ be the $j^{\text {th }}$ sample variance. The estimate of the common (i.e., pooled) variance is

$$
\begin{equation*}
S_{p}^{2}=\frac{\left(n_{1}-1\right) S_{1}^{2}+\left(n_{2}-1\right) S_{2}^{2}}{n_{1}+n_{2}-2} \tag{2}
\end{equation*}
$$

The test statistic for the conventional independent-samples $t$ test is

$$
\begin{equation*}
T=\frac{\bar{X}_{1}-\bar{X}_{2}}{\sqrt{S_{p}^{2}\left(\frac{1}{n_{1}}+\frac{1}{n_{2}}\right)}}, \tag{3}
\end{equation*}
$$

which is distributed as a $t$ variable with $v=n_{1}+n_{2}-2$ degrees of freedom, assuming normality and homogeneity of variances.
In order to compute the conventional pairedsamples $t$ test, which assumes that the two groups are dependent, let $S_{\bar{X}_{j}}^{2}=S_{j}^{2} / n_{j}$, where $S_{\bar{X}_{j}}$ is the estimate of the standard error of the mean of group $j$. An estimate of the correlation between the two groups is also needed to compute the paired-samples $t$ statistic. The correlation is defined as $r=S_{12} / S_{1} S_{2}$, where

$$
S_{12}=\sum_{i}\left(X_{i 1}-\bar{X}_{1}\right)\left(X_{i 2}-\bar{X}_{2}\right) /(n-1),
$$

and $n$ represents the total number of pairs. The paired-samples test statistic is

$$
\begin{equation*}
T_{(\text {PAREED })}=\frac{\bar{X}_{1}-\bar{X}_{2}}{\sqrt{S_{\bar{X}_{1}}^{2}+S_{\bar{X}_{2}}^{2}-2 r S_{\bar{X}_{1}} S_{\bar{X}_{2}}}} \tag{4}
\end{equation*}
$$

which is distributed as a $t$ variable with $v=n-1$
degrees of freedom, assuming normality.

## Robust Methods

Suppose, again, that $n_{j}$ observations, $X_{1 j}, X_{2 j}, \ldots, X_{n_{j} j}$, are sampled from population $j$. For both the independent-samples and pairedsamples $t$ tests, first let $X_{(1) j} \leq X_{(2) j} \leq \cdots \leq X_{\left(n_{j}\right) j}$ be the ordered observations of group $j$, and let $\gamma$ be the percentage of observations that are to be trimmed from each tail of the distribution. Also let $g_{j}=\left[\gamma n_{j}\right]$, where $[x]$ is the largest integer $\leq x$. To calculate the robust versions of both statistics we must first Winsorize the observations by letting

$$
\begin{align*}
Y_{i j} & =X_{\left(g_{j}+1\right) j} \text { if } X_{i j} \leq X_{\left(g_{j}+1\right) j} \\
& =X_{i j} \text { if } X_{\left(g_{j}+1\right) j}<X_{i j}<X_{\left(n_{j}-g_{j}\right) j} .  \tag{5}\\
& =X_{\left(n_{j}-g_{j}\right) j} \text { if } X_{i j} \geq X_{\left(n_{j}-g_{j}\right) j}
\end{align*}
$$

The sample Winsorized mean is defined as

$$
\begin{equation*}
\bar{Y}_{W j}=\frac{1}{n_{j}} \sum_{i=1}^{n_{j}} Y_{i j} . \tag{6}
\end{equation*}
$$

The sample trimmed mean for the $j^{\text {th }}$ group is also required to compute the robust versions of the paired and independent-samples $t$ tests and is defined as

$$
\begin{equation*}
\bar{X}_{t j}=\frac{1}{h_{j}} \sum_{i=g_{j}+1}^{n_{j}-g_{j}} X_{(i) j}, \tag{7}
\end{equation*}
$$

where $h_{j}=n_{j}-2 g_{j}$. The sample Winsorized variance for the robust independent-samples $t$ test is

$$
\begin{equation*}
S_{W j}^{2}=\frac{1}{n_{j}-1} \sum_{i=1}^{n_{j}}\left(Y_{i j}-\bar{Y}_{W j}\right)^{2}, \tag{8}
\end{equation*}
$$

where $Y_{i j}$ and $\bar{Y}_{W j}$ are defined in Equations 5 and 6, respectively. Finally, let

$$
\begin{equation*}
d_{j}=\frac{\left(n_{j}-1\right) S_{W_{j}}^{2}}{h_{j}\left(h_{j}-1\right)} . \tag{9}
\end{equation*}
$$

Then the robust independent-samples $t$ test is

$$
\begin{equation*}
T_{Y}=\frac{\bar{X}_{t 1}-\bar{X}_{t 2}}{\sqrt{d_{1}+d_{2}}} \tag{10}
\end{equation*}
$$

which is approximately distributed as a $t$ variable with degrees of freedom

$$
\begin{equation*}
v_{Y}=\frac{\left(d_{1}+d_{2}\right)^{2}}{d_{1}^{2} /\left(h_{1}-1\right)+d_{2}^{2} /\left(h_{2}-1\right)} . \tag{11}
\end{equation*}
$$

To compute the robust paired-samples $t$ test, as enumerated by Wilcox (2002), the paired observations must first be Winsorized, as in Equation 5. It is important to note that when Winsorizing the observations for the pairedsamples $t$ statistic, care must be taken to maintain the original pairing of the observations. The sample size for the robust version of the paired-samples $t$ test is $h=n-2 g$, where $n$ is the total number of pairs. Let

$$
\begin{equation*}
d_{j}=\frac{1}{h(h-1)} \sum_{i}\left(Y_{i j}-\bar{Y}_{W_{j}}\right)^{2}, \tag{12}
\end{equation*}
$$

and

$$
\begin{equation*}
d_{12}=\frac{1}{h(h-1)} \sum_{i}\left(Y_{i 1}-\bar{Y}_{W 1}\right)\left(Y_{i 2}-\bar{Y}_{W 2}\right), \tag{13}
\end{equation*}
$$

where $Y_{i j}$ and $\bar{Y}_{W j}$ are defined in Equations 6 and 7 , respectively. The test statistic for the robust paired-samples $t$ test is

$$
\begin{equation*}
T_{Y(\text { PAIRED })}=\frac{\bar{X}_{t 1}-\bar{X}_{t 2}}{\sqrt{d_{1}+d_{2}-2 d_{12}}}, \tag{14}
\end{equation*}
$$

which is approximately distributed as a $t$ variable with $v=h-1$ degrees of freedom.

## Simulation Procedures

Empirical Type I error and power rates were collected for the conventional and robust versions of the paired and independent-samples $t$ tests using a Monte Carlo procedure. Thus, a total of four tests were investigated: (a) the conventional paired-samples $t$ test, (b) the
conventional independent-samples $t$ test, (c) the robust paired-samples $t$ test, and (d) the robust independent-samples $t$ test. Two-tailed tests were performed on each of the four procedures.

Four variables were manipulated in the study: (a) sample size, (b) magnitude of the population correlation coefficient, (c) magnitude of the difference between groups, and (d) population distribution. Following Zimmerman (1997), four sample sizes ( $N$ ) were investigated: $10,20,40$, and 80 , and population correlations ( $\rho$ ) ranging from -.5 to .5 , in increments of .1 , were induced.

The difference in the mean (trimmed mean) value for the two populations was also manipulated. When empirical Type I error rates were investigated, there was no difference between the groups. When empirical power rates were investigated, three values of the effect size were investigated; the difference between the groups was set at $.25, .5$, and .75 . These values were chosen in order to avoid ceiling and floor effects, a practice that has been employed in other studies (e.g., Keselman, Wilcox, Algina, Fradette, \& Othman, 2003).

There were two population distribution conditions. Data for both groups were generated either from an exponential distribution or a chisquared distribution with one degree of freedom $\left(\chi_{1}^{2}\right)$. Skewness and kurtosis values for the exponential distribution are $\gamma_{1}=2$ and $\gamma_{2}=6$, respectively. Skewness and kurtosis values for the $\chi_{1}^{2}$ distribution are $\gamma_{1}=\sqrt{8}$ and $\gamma_{2}=12$, respectively.

For the robust versions of both the paired and the independent-samples $t$ tests, the percentage of trimming was $20 \%$; thus, $20 \%$ of the observations from each tail of the distribution were removed. This proportion of trimming was chosen because it has been used in other studies (e.g., Keselman et al., 1998; Keselman et al., 2000; Keselman et al., 2002; Lix et al., 1998) and because $20 \%$ trimming has previously been recommended (e.g., Wilcox, 1997).

In order to generate the data for each condition, the method outlined in Headrick and Sawilowsky (1999) for generating correlated multivariate nonnormal distributions was used. First, the SAS generator RANNOR (SAS

Institute, 1989) was used to generate pseudorandom normal variates, $Z_{i}(i=1, \ldots, N)$. Next, the $Z_{i}$ s were modified using the algorithm

$$
\begin{equation*}
Y_{i j}=r Z_{i}+\sqrt{1-r} E_{i j}, \tag{15}
\end{equation*}
$$

where the $E_{i j} \mathrm{~s}$ are pseudo-random normal variates. In the case of this study, the $E_{i j}$ s were also generated by the SAS generator RANNOR. The variable $r$ is determined as in Headrick and Sawilowsky (1999), and is dependent on the final desired population correlation ( $\rho$ ). Both $Y_{i 1}$ and $Y_{i 2}$ are random normal deviates with a correlation of $r^{2}$. Finally, the $Y_{i j} \mathrm{~s}$ generated for the study were further modified in order to obtain nonnormally distributed observations, via the algorithm

$$
\begin{equation*}
Y_{i j}^{*}=a+b Y_{i j}+(-a) Y_{i j}^{2}+d Y_{i j}^{3}, \tag{16}
\end{equation*}
$$

where $a, b$, and $d$ are constants that depend on the desired values of skewness ( $\gamma_{1}$ ) and kurtosis ( $\gamma_{2}$ ) of the distribution, and can be determined by solving equations found in Fleishman (1978, p. 523). The resultant $Y_{i j}^{*} \mathrm{~s}$ are nonnormal deviates with zero means and unit variances, and are correlated to the desired level of $\rho$, which is specified when determining $r$.

Observations with mean $\mu_{j}$ (or $\mu_{t j}$ ) and variance $\sigma_{j}^{2}$ were obtained via $X_{i j}=\mu_{j}+\sigma_{j} \times Y_{i j}^{*}$. The means (trimmed means) varied depending on the desired magnitude of the difference between the two groups. In order to achieve the desired difference, constants were added to the observations in each group. The value of the constants, corresponding to each of the four difference conditions investigated, were (a) 0,0 , (b) . 25,0 , (c) $.5,0$, and (d) $.75,0$. These values were added to each observation in the first and second group, respectively. Thus, $\mu_{j}$ ( $\mu_{t j}$ ) represents the value of the constants corresponding to a given desired difference. Variances were set to $\sigma_{j}^{2}=1$ in all conditions. When using trimmed means, the empirically
determined population trimmed mean $\mu_{t}$ was subtracted from the $Y_{i j}^{*}$ variates before multiplying by $\sigma_{j}$ (see Keselman et al., 2002 for further discussion regarding the generation of variates to be used with trimming). Ten thousand replications of the data generation procedure were performed for each of the conditions studied.

## Results

## Type I Error Rates

Each of the four investigated tests was evaluated based on its ability to control Type I errors, under conditions of nonnormality. In the case of the two versions of the independentsamples $t$ tests, the independence assumption was also violated when $\rho$ was not equal to zero.

In order for a test to be considered robust, its empirical rate of Type I error ( $\hat{\alpha}$ ) had to be contained within Bradley's (1978) liberal criterion of robustness: $0.5 \alpha \leq \hat{\alpha} \leq 1.5 \alpha$. Hence, for this study, in which a five percent nominal significance level was employed, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the $.025-.075$ interval. A test was considered to be nonrobust in a particular condition if $\hat{\alpha}$ fell outside of this interval. Tables 1 and 2 display the range of Type I errors made by each of the investigated tests across all samples sizes ( $N=$ $10,20,40,80$ ), as a function of $\rho$. We felt it was acceptable to enumerate a range across all sample sizes investigated because at all values of $N$, a similar pattern of results was observed.

Table 1: Range of Proportion of Type I Errors for All Tests Under the Exponential Distribution

## Exponential Distribution

| Rho ( $\rho$ ) | Conventional Procedure |  | Robust Procedure |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Independent | Paired | Independent | Paired |
| -0.5 | . $116-.143$ | . $060-.093$ | . 103 - . 108 | . $051-.057$ |
| -0.4 | . $100-.128$ | . $056-.085$ | . 092 - . 099 | . $052-.054$ |
| -0.3 | . $089-.116$ | . $055-.083$ | . $078-.092$ | . 047 - . 054 |
| -0.2 | . $081-.108$ | . $059-.086$ | . $070-.080$ | . $049-.057$ |
| -0.1 | . $071-.091$ | . $059-.078$ | . 062 - . 067 | . $049-.053$ |
| 0 | . 042 - . 048 | . $039-.049$ | . $038-.046$ | . $035-.045$ |
| 0.1 | . $035-.043$ | . 042 - . 053 | . $031-.038$ | . $031-.049$ |
| 0.2 | . 025 - . 029 | . 044 - . 050 | . 024 - . 031 | . $030-.052$ |
| 0.3 | . $019-.021$ | . 042 - . 053 | . $017-.021$ | . 028 - . 048 |
| 0.4 | . $011-.012$ | . $039-.052$ | . 012 - . 016 | . $03-.044$ |
| 0.5 | . $006-.007$ | . $04-.047$ | . $006-.01$ | . $028-.045$ |

Table 2: Range of Proportion of Type I Errors for All Tests Under the $\chi_{1}^{2}$ Distribution

| Rho ( $\rho$ ) | Conventional Procedure |  | Robust Procedure |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Independent | Paired | Independent | Paired |
| -0.5 | . $120-.171$ | . 068 - . 129 | . $102-.107$ | . $056-.073$ |
| -0.4 | . $100-.161$ | . $060-.125$ | . $090-.096$ | . $056-.068$ |
| -0.3 | . 093 - . 145 | . 063 - . 118 | . 079 - . 089 | . $051-.066$ |
| -0.2 | . 087 - . 135 | . 067 - . 114 | . $070-.082$ | . $052-.067$ |
| -0.1 | . $075-.114$ | . $064-.102$ | . 063 - . 068 | . $052-.058$ |
| 0 | . $038-.046$ | . $034-.046$ | . 026 - . 045 | . $025-.042$ |
| 0.1 | . $031-.041$ | . $033-.049$ | . $023-.036$ | . $022-.042$ |
| 0.2 | . 026 - . 029 | . $035-.046$ | . $020-.030$ | . $023-.044$ |
| 0.3 | . $020-.021$ | . $033-.052$ | . 018 - . 023 | . $023-.043$ |
| 0.4 | . $011-.015$ | . $035-.051$ | . $015-.018$ | . $022-.046$ |
| 0.5 | . 006 - . 011 | . $035-.045$ | . $009-.013$ | . $020-.042$ |

Table 1 displays the range of empirical Type I error rates for each test, as a function of $\rho$, under the exponential distribution condition. It is apparent from the table that both versions of the paired-samples $t$ test maintained Type I errors near the nominal level of significance, $\alpha$. In fact, only 6 of 44 values fell outside the range of Bradley's .025-. 075 interval for the conventional paired $t$ test; none did for the robust paired $t$ test. Thus, for data that follow an exponential distribution, the robust paired $t$ test was insensitive to nonnormality at every value of $\rho$. A comparison of the conventional and robust versions of the paired $t$ test in Table 1 reveals that, in particular, the robust version was more effective at controlling Type I errors when the population correlation ( $\rho$ ) between the groups was negative.

Table 1 also shows that the independentsamples tests were not as robust, overall, as their
paired-samples counterparts. In fact, the total number of values that fell outside of the range of Bradley's liberal criterion was 30 and 26 (out of 44) for the conventional and robust versions of the independent $t$ test, respectively. Thus, the robust independent $t$ test was indeed slightly more robust, overall, than the conventional independent $t$ test. Both versions of the independent-samples $t$ test were effective at controlling Type I errors when the population correlation ( $\rho$ ) was zero; however, this control was reduced the more that $\rho$ deviated from zero.

An inspection of Table 2, which displays the range of Type I errors for the tests for the $\chi_{1}^{2}$ distribution, reveals a pattern of results similar to that for the exponential distribution. However, all of the tests were somewhat less robust under the $\chi_{1}^{2}$ distribution than the exponential distribution condition. That
is, nonrobust liberal values were greater in value for $\chi_{1}^{2}$ data than for exponentially distributed data. Specifically, the total number of values that fell outside of Bradley's liberal interval for the conventional versions of the paired and independent-samples $t$ tests were 12 and 31 (out of 44), respectively. The total number of nonrobust values for the robust versions of the paired and independent-samples $t$ tests were five and 28 , respectively.
Power Rates
The four tests were also evaluated based on empirical power rates. Therefore, each test was judged on its ability to detect a true difference between the trimmed means of the
groups (in the case of the robust tests), or the least squares means of the groups (in the case of the conventional tests). Figures 1, 2, and 3 display the power of each of the investigated tests to detect a true difference between the (trimmed) means of the groups, as a function of the magnitude of the difference between the (trimmed) means. The results portrayed in these figures were averaged over all sample sizes. While the power rates of the tests increased as the size of $N$ increased, again, we felt it was acceptable to collapse over the sample size conditions because the tests showed a similar pattern of results in relation to one another for all values of $N$.


Figure 1. Probability of rejecting $H_{0}$ for the conventional and robust paired and independent-samples $t$ tests; $\rho=0$.

Exponential Distribution


Figure 2. Probability of rejecting $H_{0}$ for the conventional and robust paired and independent-samples $t$ tests; $\rho=0.3$.


Figure 3. Probability of rejecting $H_{0}$ for the conventional and robust paired and independent-samples $t$ tests; $\rho=-0.3$.

Figure 1 displays the power rates of the tests for both the $\chi_{1}^{2}$ and the exponential distributions when $\rho=0$. The upper portion of the figure reveals that when data followed an exponential distribution, the power functions of the four tests were quite similar, with the empirical power of the robust versions only slightly higher than the corresponding power of the conventional versions. However, an inspection of the lower portion of Figure 1 indicates that under the $\chi_{1}^{2}$ distribution, the power functions of the robust tests were considerably higher than those of both conventional versions. In addition, Figure 1 shows that when no correlation existed between the groups, the power functions of the independent-samples $t$ tests were slightly higher than their paired-samples counterparts.

Figure 2 shows the power functions of the tests for both the $\chi_{1}^{2}$ and exponential distributions when $\rho=3$. The upper portion of Figure 2 indicates that when the data were exponentially distributed and positively correlated, the power functions of both versions of the paired-samples $t$ test were higher than those of the independent-samples tests. The lower portion of the figure, which displays power for the $\chi_{1}^{2}$ distribution for this same value of $\rho$, demonstrates that while the power function of each of the paired-samples $t$ tests was higher than its respective independent-samples counterpart, the power rates of both robust tests were higher than those of the conventional tests.

Figure 3 displays the power rates of the tests for the $\chi_{1}^{2}$ and exponential distributions when $\rho=-.3$. Unlike the results obtained for positively correlated data, the paired-samples $t$ tests showed no apparent power advantage over the independent-samples $t$ tests when the groups were negatively correlated, for either the
exponential or the $\chi_{1}^{2}$ distributions. In fact, the figure shows that the power functions of the independent-samples $t$ tests were higher than their paired-samples counterparts under both distributions. The lower portion of Figure 3 shows that under the $\chi_{1}^{2}$ distribution, while the power functions of both versions of the independent-samples $t$ test were higher than their corresponding versions of the pairedsamples test, the power rates of both robust tests were higher than the conventional tests, as was the case with the other levels of $\rho$.

## Conclusion

Four different statistics for testing the difference between two groups were investigated based on their power to detect a true difference between two groups and their ability to control Type I errors. The primary objective for conducting the study was to determine which of the tests would perform best when the data for the two groups were correlated and the assumption of a normal distribution of the responses was violated.

Although empirical Type I error and power rates are two separate measures of a test's effectiveness, in order to evaluate the overall performance of the investigated procedures, power and Type I error rates must be considered concomitantly. The reason for this is that if a test does not maintain the rate of Type I errors at or around the nominal $\alpha$ level, this can cause a distortion in power. Figures 4 and 5 provide a summary of the results for the exponential and $\chi_{1}^{2}$ distributions, respectively. These figures were included to allow the reader to easily examine the Type I error and power rates of each of the distributions concurrently.

Exponential Distribution


Figure 4. Probability of rejecting $H_{0}$ as a function of $\rho$ and the magnitude of the difference between (trimmed) means for the conventional and robust paired and independent-samples $t$ tests exponential distribution.

## Chi-Squared Distribution (One df)




Figure 5. Probability of rejecting $H_{0}$ as a function of $\rho$ and the magnitude of the difference between (trimmed) means for the conventional and robust paired and independent-samples $t$ tests under the $\chi_{1}^{2}$ distribution.

As the results indicated, the only time the independent tests maintained the Type I error rate close to the nominal level was when there was no correlation between the groups; this ability grew worse as $\rho$ got larger. In fact, the Type I error control of the independent $t$ tests began to break down when the correlation between the groups was as small as $\pm .1$. Thus, with the exception of the $\rho=0$ condition, both the robust and the conventional versions of the independent $t$ test were quite poor at controlling Type I errors. Because of this distortion of the Type I error rate, the powers of the independent tests are not interpretable (Zimmerman, 1997) when $\rho$ is not equal to zero.

Both versions of the paired $t$ test, however, did a much better job of controlling Type I errors than their independent-samples counterparts when there was a correlation between the groups, for nonnormal data. Because the paired-samples $t$ tests maintained Type I errors close to the nominal level, the empirical power rates of the paired $t$ tests, unlike those of the independent tests, can be taken to accurately represent their ability to detect a true difference between the groups. Thus, as expected, when power and Type I error rates are both taken into account, it can be said that the paired $t$ tests were more effective than their independent samples counterparts when groups were correlated, even when this correlation was low (i.e., $\pm .1$ ). This finding agrees with Zimmerman's (1997) results for normally distributed data.

Furthermore, the robust paired-samples $t$ test was more effective, in terms of Type I error control, than the conventional paired test. The robust paired test was also consistently more powerful than the conventional version, and this power advantage increased as skewness and kurtosis in the population increased. Therefore, as expected, the robust version of the pairedsamples $t$ test performed better than the conventional version of the test, for nonnormal data. This result is supported by many other studies involving trimmed means and Winsorized variances (e.g., Keselman, et al., 1998; Keselman, et al., 2000; Keselman et al., 2002; Lix et al., 1998; Wilcox, 1993; Yuen, 1974).

In conclusion, there need only be a small positive or negative correlation between two groups in order for the paired $t$ test to be more effective than the independent $t$ test when the data are nonnormal. In fact, although Vonesh (1983) showed that there needs to be a correlation of at least .25 in the population for the paired $t$ test to be more powerful than the independent test, when the distortion of Type I error rates, resulting from the application of the independent-samples $t$ test on dependent data, was taken into account, the paired-samples $t$ tests performed best when the correlation was as low as $\pm .1$. Thus, just as Zimmerman (1997) cautions when dealing with normal data, researchers should take care to ensure that their data is not correlated in any way when using the independent $t$ test on nonnormal data, lest the existence of even a slight dependence alters the significance level of the test. In addition, given that the population distributions were not normal in shape, the robust version of the paired $t$ test performed the best under all the conditions that were studied. Thus, based on the results of this investigation, it is recommended that researchers use the robust paired-samples $t$ test, which employs trimmed means and Winsorized variances, when dealing with nonnormal data.

## References

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Edwards, A. L. (1979). Multiple regression and the analysis of variance and covariance. New York: Freeman.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43, 521-532.

Hays, W. L. (1988). Statistics (4 $4^{\text {th }}$ ed.). New York: Holt, Rinehart, \& Winston.

Headrick, T. C., \& Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. Psychometrika, 64, 25-35.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., \& Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. British Journal of Mathematical and Statistical Psychology, 53, 175-191.

Keselman, H. J., Kowalchuk, R. K., \& Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. Psychometrika, 63, 145-163.

Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., Othman, A. R. (2003). A power comparison of robust test statistics based on adaptive estimators. Submitted for publication.

Keselman, H. J., Wilcox, R. R., Kowalchuk, R. K., \& Olejnik, S. (2002). Comparing trimmed or least squares means of two independent skewed populations. Biometrical Journal, 44, 478-489.

Kirk, R. E. (1999). Statistics: An introduction. Orlando, FL: Harcourt Brace.

Kurtz, K. H. (1965). Foundations of psychological research. Boston: Allyn \& Bacon.

Lix, L. M., \& Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 58, 409-429.

MacDonald, P. (1999). Power, Type I error, and Type III error rates of parametric and nonparametric statistical tests. Journal of Experimental Education, 67, 367-380.

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

SAS Institute Inc. (1989). SAS/IML software: Usage and reference, version 6 ( $1^{\text {st }}$ ed.). Cary NC: Author.

Vonesh, E. F. (1983). Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. Communications in Statistics A: Theory and Methods, 12, 289-301.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350-362.

Wilcox, R. R. (1990). Comparing the means of two independent groups. Biometrical Journal, 36, 259-273.

Wilcox, R. R. (1993). Analysing repeated measures or randomized block designs using trimmed means. British Journal of Mathematical and Statistical Psychology, 46, 63-76.

Wilcox, R. R. (1997). Introduction to robust estimation and testing. San Diego: Academic Press.

Wilcox, R. R. (2002). Applying contemporary statistical methods. San Diego: Academic Press.

Yuen, K. K. (1974). The two-sample trimmed $t$ for unequal population variances. Biometrika, 61, 165-170.

Zimmerman, D. W. (1997). A note on the interpretation of the paired-samples $t$ test. Journal of Educational and Behavioral Statistics, 22, 349-360.

# Fitting Generalized Linear Mixed Models For Point-Referenced Spatial Data 

Armin Gemperli Penelope Vounatsou<br>Swiss Tropical Institute<br>Basel, Switzerland

Non-Gaussian point-referenced spatial data are frequently modeled using generalized linear mixed models (GLMM) with location-specific random effects. Spatial dependence can be introduced in the covariance matrix of the random effects. Maximum likelihood-based or Bayesian estimation implemented via Markov chain Monte Carlo (MCMC) for such models is computationally demanding especially for large sample sizes because of the large number of random effects and the inversion of the covariance matrix involved in the likelihood. We review three fitting procedures, the Penalized Quasi Likelihood method, the MCMC, and the Sampling-Importance-Resampling method. They are assessed in terms of estimation accuracy, ease of implementation, and computational efficiency using a spatially structured dataset on infant mortality from Mali.

Key words: Geostatistics, infant mortality, kriging, Markov chain Monte Carlo (MCMC), penalized quasi likelihood (PQL), risk mapping, sampling-importance-resampling (SIR)

## Introduction

Point referenced spatial data arise from observations collected at geographical locations over a fixed continuous space. Proximity in space introduces correlations between the observations rendering the independence assumption of standard statistical methods invalid. Ignoring spatial correlation will result in underestimation of the standard error of the parameter estimates, and therefore liberal inference as the null hypothesis is rejected too often. A wide range of analytical tools within the field of geostatistics have been developed concerning with the description and estimation of spatial patterns, the modeling of data in the presence of spatial correlation and the kriging, that is the spatial prediction at unobserved locations.

Armin Gemperli is completing his PhD at the Biostatistics Unit. Penelope Vounatsou is Senior Statistician. We are grateful for assistance from Macro International Inc., and acknowledge discussions with Tom Smith and Marcel Tanner. This work was supported by the Swiss National Science Foundation grant Nr. 3200-057165.99.

Statistical inference of point referenced data often assumes that the observations arise from a Gaussian spatial stochastic process and introduce covariate information and possibly trend surface specification on the mean structure while spatial correlation on the variancecovariance matrix $\boldsymbol{\Sigma}$ of the process. Under second order stationarity, $\boldsymbol{\Sigma}$ determines the well-known variogram. When isotropy is also assumed, the elements of $\boldsymbol{\Sigma}$ are modeled by parametric functions of the separation between the corresponding locations. For non-Gaussian data, the spatial correlation is modeled on the covariance structure of location-specific random effects introduced into the model and assumed to arise from a Gaussian stationary spatial process.

For Gaussian data, the generalized least squares (GLS) approach can be used iteratively to obtain estimates $\hat{\boldsymbol{\beta}}$ of the regression coefficients conditional on the covariance parameters. The covariance parameters $\boldsymbol{\theta}$ can be estimated conditional on $\hat{\boldsymbol{\beta}}$ by fitting the semivariogram empirically or by maximum likelihood or restricted maximum likelihood methods (Zimmerman and Zimmerman, 1991).

Statistical estimation for non-Gaussian data is based on the theory of generalized linear mixed models (GLMM). A common approach is
to integrate out the random effects and proceed with maximum likelihood based approaches for estimating the covariate and covariogram parameters. This integration can be implemented numerically (Anderson and Hinde, 1998; Preisler, 1988; Lesaffre and Spiessens, 2001) when dimensionality is low or via approximations. Breslow and Clayton (1993) show, that for known covariance parameters, the Laplace approximation leads to the same estimator for the fixed and random effects parameters as the one arising by maximizing the penalized quasi-likelihood (PQL). Implementation of this approach requires iterating between iterated weighted least squares for estimating the fixed and random effects parameters and maximizing the profile likelihood for estimating the covariance parameters. An extension of the PQL procedure is discussed by Wolfinger and O'Connell (1993). The PQL approach is implemented in some statistical packages due to its relative simplicity, however it provides biased estimates when the number of random effects increases (McCulloch, 1997; Booth and Hobert, 1999) or when the data are far from normal.

The generalized estimating equation methods developed by Liang and Zeger (1986) and Zeger and Liang (1986) estimate covariate effects under the assumption of independence, but correct their standard error to account for the spatial dependence. The method is unable to estimate the spatial random effects. The EM algorithm (Dempster, Laird and Rubin, 1977) has been implemented in model fit by treating the spatial random effects as "missing" data. The intractable integration of the random effects which is required in the E-step is overcome by simulation, such as Metropolis-Hastings algorithm (McCulloch, 1997) or importance sampling/rejection sampling method (Booth and Hobert, 1999). For spatial settings, particular Pseudo-Likelihood approaches have been established which capture solely the site to site variation between pairs or groups of observations (Besag, 1974). For the special case of a binary outcome, Heagerty and Lele (1998) have proposed a thresholding model using a composite likelihood approach.

A drawback of the maximum likelihoodbased methods employed in geostatistical modeling is the large sample asymptotic inference. For a spatial stochastic process $\{\mathbf{Y}(\mathbf{u}) ; \mathbf{u} \in D\}$, with $D \subset R^{2}$ the asymptotic concept can be applied either to the sample size within a fixed space $D$ (infill asymptotics) or to the space $D$ (increasing domain asymptotics). In the latter, observations are spaced far enough to be considered uncorrelated. The results can differ, depending on the type of asymptotics used (see e.g. Tubilla, 1975).

Bayesian hierarchical geostatistical models implemented via Monte Carlo methods avoid asymptotic inference as well as many computational problems in model fitting and prediction. Diggle et al. (1998) suggest inference on the posterior density via Markov chain Monte Carlo (MCMC). This iterative approach requires repeated inversions of the covariance matrix of the spatial process, which is involved in the likelihood. The size of this matrix increases with the number of locations. Inversion of large matrices can drastically slow down the running time of the algorithm and cause numerical instabilities affecting the accuracy of the estimates. To overcome this problem Gelfand et al. (1999) suggest non-iterative simulation via the Sampling-Importance-Resampling (SIR) algorithm (Rubin, 1987). The quality of SIR hinge on the ability to formulate an easy-to-draw-from importance-density, which comes as close as possible to the true joint posterior distribution of the parameters.

In this article, we review three fitting procedures; the maximum likelihood-based PQL method, the MCMC simulation and the SIR. We assess these methods in terms of estimation accuracy, ease of implementation and computational efficiency using a spatially structured dataset on infant mortality from Mali collected over 181 locations. A description of the dataset and the applied questions which motivated this work are given in the next section. Then we describe the model as well as the three fitting approaches. A discussion on the ease of implementation of each approach and a comparison of the inferences obtained is given in the conclusion section.

## Data

The data that motivated this work were collected under the Demographic and Health Surveys (DHS) program. The aim of the program is to collect and analyze reliable demographic and health data for regional and national family and health planning. Data are commonly collected in developing countries. DHS is funded by the U.S. Agency for International Development (USAID) and implemented by Macro International Inc. The standard DHS methodology involves collecting complete birth histories from women of childbearing age, from which a record of age and survival can be computed for each child. The data are available to researchers via the internet (www.measureDHS.com).

Birth histories corresponding to 35,906 children were extracted from the data of the DHS-III 1995/96 household survey carried out in Mali. Additional relevant covariates extracted were the year of birth, residence, mothers education, infant's sex, birth order, preceding birth interval and mothers age at birth. Using location information provided by Macro International, we were able to geo-locate 181 distinct sites by using digital maps and databases, such as the African data sampler (1995) and the Geoname Gazetteer (1995). The objective of data analysis was to assess the effect of birth and socio-economic parameters on infant mortality and produce smooth maps of mortality risk in Mali. These maps will help identifying areas of high mortality risk and assist child mortality intervention programs.

## Methodology

Let $Y_{i j}$ be a binary response corresponding to the mortality risk of child $j$ at site $S_{i}$, $i=1, \ldots, n$ taking value 1 if the child survived the first year of life and 0 otherwise, and let $\mathbf{X}_{i j}$ be the vector of associated covariates. Within the generalized linear model framework (GLM), we assume $Y_{i j}$ are i.i.d. Bernoulli random variables with $E\left(Y_{i j}\right)=\pi_{i j}$ and model predictors as $g\left(\pi_{i j}\right)=\mathbf{X}_{i j}^{t} \boldsymbol{\beta}$ where $g(\cdot)$ is a link function
such as logit in our mortality risk application. However the spatial structure of the data renders the independence assumption of $Y_{i j}$ invalid, leading to narrower confidence intervals for $\boldsymbol{\beta}$ and thus to overestimation of the significance of the predictors.

One approach to take into account spatial dependence is via the generalized linear mixed model (GLMM) reviewed by Breslow and Clayton (1993). In particular, we introduce the unobserved spatial variation by a latent stationary, isotropic Gaussian process $\mathbf{U}$ over our study region $D$, such that $\mathbf{U}=\left(U_{1}, U_{2}, \ldots, U_{n}\right) \sim N(0, \Sigma)$, where $\Sigma_{i j}$ is a parametric function of the distance $d_{i j}$ between locations $\mathbf{s}_{i}$ and $\mathbf{s}_{j}$. Conditional on the random term $U_{i}$, we assume that $Y_{i j}$ are independent with $E\left(Y_{i j} \mid U_{i}\right)=\pi_{i j}$. The $U_{i}$ enters the model on the same scale as the predictors, that is

$$
\begin{equation*}
g\left(\pi_{i j}\right)=\mathbf{X}_{i j}^{t} \boldsymbol{\beta}+U_{i} \tag{1}
\end{equation*}
$$

and captures unmeasured geographical heterogeneity (small scale variation).

A commonly used parameterization for the covariance $\Sigma$ is $\Sigma_{i j}=\sigma^{2} \rho\left(\phi ; d_{i j}\right)$ where $\sigma^{2}$ is the variance of the spatial process and $\rho\left(\phi ; d_{i j}\right)$ a valid correlation function with a scale parameter $\phi$ which controls the rate of correlation decay with increasing distance. In most applications a monotonic correlation function is chosen i.e. the exponential function which has the form $\rho\left(\phi ; d_{i j}\right)=\exp \left(-\phi d_{i j}\right)$. Ecker and Gelfand (1997) propose several other parametric correlation forms, such as the Gaussian, Cauchy, spherical and the Bessel.

A separate set of location-specific random effects, $\mathbf{W}=\left(W_{1}, W_{2}, \ldots, W_{n}\right)^{t}$ is often added in Equation 1 to account for unexplained non-spatial variation (Diggle et al., 1998), where $W_{i}, \quad i=1, \ldots, n$ are considered to be independent, arising from a Normal distribution, $W_{i} \sim N\left(0, \tau^{2}\right)$. The $\tau^{2}$ is known in
geostatistics as the nugget effect and introduces a discontinuity at the origin of the covariance function:

$$
\Sigma_{i j}=\tau^{2} 1(i=j)+\sigma^{2} \rho\left(\phi ; d_{i j}\right) .
$$

A large number of repeated samples at the same location make the nugget identifiable, otherwise its use in the model is not justifiable because the extra binomial variation is already accounted for by the spatial random effect.

Parameter estimation
The above GLMM is highly parameterized and maximum likelihood methods can fail to estimate all parameters simultaneously. The estimation approach starts by integrating out the random effects and estimating the other parameters using the marginal likelihood

$$
\int p\left(\mathbf{Y} \mid \mathbf{U}, \boldsymbol{\beta}, \sigma^{2}, \phi\right) p\left(\mathbf{U} \mid \sigma^{2}, \phi\right) d \mathbf{U} .
$$

However, this integral has analytical solution only for Gaussian data. For non-Gaussian data the integrand can be approximated using a firstorder Taylor series expansion around its maximizing value, after which the integration is feasible. This approach, known as the Laplace approximation, results in the penalized quasilikelihood (PQL) estimator (Breslow and Clayton, 1993), which was shown in various simulation studies to produce biased results (Browne and Draper, 2000; Neuhaus and Segal, 1997). Breslow and Lin (1995) determined the asymptotic bias in variance component problems for first- and second-order approximations in comparison to McLaurin approximations.

Following the Bayesian modeling specification, we need to adopt prior distributions for all model parameters. We chose non-informative Uniform priors for the regression coefficients, i.e. $p(\boldsymbol{\beta}) \propto \mathbf{1}$, and vague inverse Gamma priors for the $\sigma^{2}$ and $\phi$ parameters: $\quad p(\phi)=\operatorname{IG}\left(a_{1}, b_{1}\right) \quad$ and $p\left(\sigma^{2}\right)=\operatorname{IG}\left(a_{2}, b_{2}\right)$. Bayesian inference is based on the joint posterior distribution

$$
\begin{aligned}
& p\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right) \propto L(\boldsymbol{\beta}, \mathbf{U} ; \mathbf{Y}) \times \\
& p(\boldsymbol{\beta}) p\left(\mathbf{U} \mid \sigma^{2}, \phi\right) p\left(\sigma^{2}\right) p(\phi)
\end{aligned}
$$

where $p\left(\mathbf{U} \mid \sigma^{2}, \phi\right)$ is the distribution of the spatial random effects, that is $p\left(\mathbf{U} \mid \sigma^{2}, \phi\right) \equiv N(0, \boldsymbol{\Sigma})$.

Markov chain Monte Carlo estimation
Diggle et al. (1998) suggest Markov chain Monte Carlo and in particular Gibbs sampling for fitting GLMM for point-referenced data. The standard implementation of the Gibbs algorithm requires sampling from the full conditional posterior distributions which in our application have the following forms:

$$
\begin{align*}
& p\left(\beta_{k}, \mid \boldsymbol{\beta}_{-k}, \mathbf{U}, \mathbf{Y}\right) \propto \\
& \prod_{i=1}^{n} \prod_{j=1}^{n_{i}} \frac{\exp \left(\mathrm{X}_{i j k} \beta_{k} Y_{i j}\right)}{1+\exp \left(\mathbf{X}_{i j}^{t} \boldsymbol{\beta}+U_{i}\right)} \tag{2}
\end{align*}
$$

$$
\begin{equation*}
\left.\left(\sigma^{2}-\boldsymbol{\Sigma}_{i,-i} \boldsymbol{\Sigma}_{-i}^{-1} \boldsymbol{\Sigma}_{-i, i}\right)^{-1}\right) \tag{3}
\end{equation*}
$$

$$
p\left(\phi \mid \mathbf{U}, \sigma^{2}\right) \propto|\mathbf{\Sigma}|^{-1 / 2} \times
$$

$$
\exp \left(-\frac{1}{2}\left(\mathbf{U}^{t} \boldsymbol{\Sigma}^{-1} \mathbf{U}+b_{1} / \phi\right)\right) \phi^{-\left(a_{1}+1\right)}
$$

$$
p\left(\sigma^{2} \mid \mathbf{U}, \phi\right) \sim
$$

InverseGamma $\left(a_{2}+n / 2\right.$,

$$
\begin{gather*}
\left.b_{2}+\frac{1}{2} \mathbf{U}^{t} \mathbf{R}^{-1} \mathbf{U}\right) \\
R_{k l}=\rho\left(\phi ; d_{k l}\right) \tag{5}
\end{gather*}
$$

where

$$
\begin{aligned}
& \boldsymbol{\beta}_{-k}=\left(\beta_{1}, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_{K}\right)^{t}, \\
& \mathbf{U}_{-i}=\left(U_{1}, \ldots, U_{i-1}, U_{i+1}, \ldots, U_{n}\right)^{t}, \\
& \boldsymbol{\Sigma}_{-i, i}=\boldsymbol{\Sigma}_{i,-i}^{t}=\operatorname{Cov}\left(\mathbf{U}_{-i}, U_{i}\right) \text { and }
\end{aligned}
$$

$$
\mathbf{\Sigma}_{-i}=\operatorname{Cov}\left(\mathbf{U}_{-i}, \mathbf{U}_{-i}^{t}\right)
$$

Samples from $p\left(\sigma^{2} \mid \mathbf{U}, \phi\right)$ can be drawn easily as this is a known distribution. The conditionals of the other parameters do not have standard forms and a random walk Metropolis algorithm with a Gaussian proposal density having mean equal to the estimate from the previous iteration and variance derived from the inverse second derivative of the log-posterior could be employed for simulation.

The likelihood calculations in Equations 3, 4, and 5 require inversions of the ( $n-1) \times(n-1)$ matrices $\boldsymbol{\Sigma}_{-i}, i=1, \ldots, n$ and the $n \times n$ matrix $\Sigma$, respectively. Matrix inversion is an order 3 operation, which has to be repeated for evaluating the conditional distribution of all $n$ random effects $U_{i}$ and that of the $\phi$ parameter, within each Gibbs sampling iteration. This leads to an enormous demand of computing capacity and makes implementation of the algorithm extremely slow (or possibly infeasible), especially for large number of locations.

## Sampling-Importance-Resampling

Gelfand et al. (1999) propose Bayesian inference for point-referenced data using noniterative Sampling-Importance-Resampling (SIR) simulation. They replace matrix inversion with simulation by introducing a suitable importance sampling density $g(\cdot)$ and re-write the joint posterior as

$$
\begin{align*}
& p^{*}\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right)= \\
& \frac{p\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right)}{g\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi ; \mathbf{Y}\right)} g\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi ; \mathbf{Y}\right) . \tag{6}
\end{align*}
$$

They construct the importance sampling density (ISD) by

$$
\begin{align*}
& g\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi ; \mathbf{Y}\right) \\
& =g_{s}(\boldsymbol{\beta} \mid \mathbf{U} ; \mathbf{Y}) g_{s}\left(\mathbf{U} \mid \sigma^{2}, \phi\right) g_{s}\left(\sigma^{2}, \phi\right) \tag{7}
\end{align*}
$$

which is easy to simulate from and then resample from $g\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi ; \mathbf{Y}\right)$ according to the importance weights

$$
\begin{equation*}
w\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi\right)=\frac{p\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right)}{g\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi ; \mathbf{Y}\right)} \tag{8}
\end{equation*}
$$

The density $g_{s}\left(\sigma^{2}, \phi\right)$ of the ISD could be taken as a product of independent inverse Gamma distributions $g_{s}\left(\sigma^{2}\right) g_{s}(\phi)$. It is however preferable to adopt a bivariate distribution which accounts for interrelations between the two parameters and thus it approximates closer the $p\left(\sigma^{2}, \phi \mid \mathbf{Y}\right)$. We considered a bivariate t-distribution on $\log \left(\sigma^{2}\right)$ and $\log (\phi)$ with low degrees of freedom and mean around the maximum likelihood estimates of $\log \left(\sigma^{2}\right)$ and $\log (\phi)$. The spatial random effects can be simulated from a multivariate normal distribution,

$$
g_{s}\left(\mathbf{U} \mid \sigma^{2}, \phi\right) \equiv N\left(0, \sigma^{2} \rho(\phi, \cdot)\right)
$$

This step requires matrix decomposition of $\sigma^{2} \rho(\phi, \cdot)$, repeatedly at every iteration. This is an operation of order 2 and the most expensive numerical part of the simulation from the ISD. The density $g_{s}(\boldsymbol{\beta} \mid \mathbf{U} ; \mathbf{Y})$ can be a Normal distribution, $\quad g_{s}(\boldsymbol{\beta} \mid \mathbf{U} ; \mathbf{Y}) \equiv N\left(\hat{\boldsymbol{\beta}}_{\mathbf{U}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right)$, with $\hat{\boldsymbol{\beta}}_{\mathbf{U}}$ equal to the regression coefficients estimated from an ordinary logistic regression with offset $\mathbf{U}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ equal to the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathbf{U}}$.

When the ISD approximates well the posterior distribution, one expects that the standardized importance weights are Uniformly distributed. When this is not the case, the ISD would give rise to very few dominant weights leading to an inefficient and wrong sampler. A possible remedy would be to embed the Sampling-Importance-Resampling simulation in an iterative scheme which refines the initial
guesses of the ISD and allows after few iterations more uniform weights.

Point estimates of the parameters should preferably be calculated from the importance weights using all sampled values, rather than from the re-sampled values, what leads to smaller bias. For example the mean and variance of $\beta_{i}$ is estimated by $\bar{\beta}_{i}=\sum_{k} w_{k} \beta_{i}^{(k)} / \sum_{k} w_{k}$ and $\quad \sum_{k} w_{k}\left(\beta_{i}^{(k)}-\bar{\beta}_{i}\right)^{2} / \sum_{k} w_{k} \quad$ respectively, where $\beta_{i}^{(k)}$ is the $k$ th sampled value of $\beta_{i}$ from the ISD.

## Spatial Prediction

Modeling point-referenced data is not only useful for identifying significant covariates but for producing smooth maps of the outcome by predicting it at unsampled locations. Spatial prediction is usually refereed as kriging.

Let $\mathbf{Y}_{0}$ be a vector of the binary response at new, unobserved locations $s_{0 i}$, $i=1, \ldots, n_{0}$. Following the maximum likelihood approach, the distribution of $\mathbf{Y}_{0}$ is given by:

$$
\begin{align*}
& P\left(\mathbf{Y}_{0} \mid \hat{\boldsymbol{\beta}}, \hat{\mathbf{U}}, \hat{\sigma}^{2}, \hat{\boldsymbol{\phi}}\right)=  \tag{9}\\
& \int P\left(\mathbf{Y}_{0} \mid \hat{\boldsymbol{\beta}}, \mathbf{U}_{0}\right) P\left(\mathbf{U}_{0} \mid \hat{\mathbf{U}}, \hat{\sigma}^{2}, \hat{\phi}\right) d \mathbf{U}_{0}
\end{align*}
$$

where $\hat{\boldsymbol{\beta}}, \quad \hat{\sigma}^{2}$ and $\hat{\phi}$ are the maximum likelihood estimates of the corresponding parameters. In PQL, $\hat{\mathbf{U}}$ is derived as part of the iterative estimation process (Breslow and Clayton, 1993). $P\left(\mathbf{Y}_{0} \mid \hat{\boldsymbol{\beta}}, \mathbf{U}_{0}\right)$ is the Bernoullilikelihood at new locations and $P\left(\mathbf{U}_{0} \mid \hat{\mathbf{U}}, \hat{\sigma}^{2}, \hat{\phi}\right)$ is the distribution of the spatial random effects $\mathbf{U}_{0}$ at new sites, given $\hat{\mathbf{U}}$ at observed sites and is Normal

$$
\begin{align*}
& P\left(\mathbf{U}_{0} \mid \hat{\mathbf{U}}, \hat{\sigma}^{2}, \hat{\phi}\right)=  \tag{10}\\
& N\left(\boldsymbol{\Sigma}_{01} \boldsymbol{\Sigma}_{11}^{-1} \hat{\mathbf{U}}, \boldsymbol{\Sigma}_{00}-\boldsymbol{\Sigma}_{01} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{10}\right)
\end{align*}
$$

with $\quad \boldsymbol{\Sigma}_{11}=E\left(\mathbf{U U}^{t}\right), \quad \boldsymbol{\Sigma}_{00}=E\left(\mathbf{U}_{0} \mathbf{U}_{0}^{t}\right) \quad$ and $\boldsymbol{\Sigma}_{01}=\boldsymbol{\Sigma}_{10}^{t}=E\left(\mathbf{U}_{0} \mathbf{U}^{t}\right)$. The mean of the Gaussian distribution in (10) is the classical kriging estimator (Matheron, 1963).

The Bayesian predictive distribution of $\mathbf{Y}_{0}$ is given by:

$$
\begin{align*}
& P\left(\mathbf{Y}_{0} \mid \mathbf{Y}\right)=\int P\left(\mathbf{Y}_{0} \mid \boldsymbol{\beta}, \mathbf{U}_{0}\right) P\left(\mathbf{U}_{0} \mid \mathbf{U}, \sigma^{2}, \phi\right) \times \\
& P\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right) d \boldsymbol{\beta} d \mathbf{U}_{0} d \mathbf{U} d \sigma^{2} d \phi \tag{11}
\end{align*}
$$

$P\left(\boldsymbol{\beta}, \mathbf{U}, \sigma^{2}, \phi \mid \mathbf{Y}\right)$ is the posterior distribution of the parameters and obtained by the Gibbs sampler or the SIR approach. Simulation-based Bayesian spatial prediction is performed by consecutive drawing samples from the posterior distribution, the distribution of the spatial random effects at new locations and the Bernoulli-distributed predicted outcome. In SIR, drawing is performed from the set of all sampled parameters with weighting given in Equation (8).

The maximum-likelihood predictor (Equation 9) can be interpreted as the Bayesian predictor (Equation 11), with parameters fixed at their maximum-likelihood estimates. In contrast to Bayesian kriging, classical kriging does not account for uncertainty in estimation of $\boldsymbol{\beta}$ and the covariance parameters.

## Results

A generalized linear mixed model was fitted to the infant mortality data in Mali using the three estimation approaches discussed in the methodology-section, PQL, MCMC and SIR together with an ordinary logistic regression (GLM) which did not account for spatial dependence. The purpose of the analysis was to assess the effect of maternal and socio-economic factors on infant mortality, produce a smooth map of mortality risk in Mali and compare the results obtained from the above procedures. Univariate analysis based on the ordinary logistic regression revealed that the following variables should be included in the model: child's birthday, region type, mother's degree of education, sex, birth order, preceding birth interval and mother's age at birth.

We fitted the non-spatial logistic model (GLM) in SAS (SAS Institute Inc., Cary, NC, USA) using Proc Logistic. The spatial model with the PQL estimation method was also fitted in SAS using the \%GLIMMIX-macro. This macro is based on the approach of Wolfinger and O'Connell (1993) and does subsequent calls of Proc Mixed to iteratively estimate mixed models for non-normal data. It is supported by a collection of spatial correlation functions, such as the exponential, Gaussian, linear, power and spherical. In our application, we have chosen the exponential function. MCMC and SIR estimation were implemented in software written
by the authors in FORTRAN 95 (Compaq Visual Fortran v6.6) and run on an Unix AlphaServer 8400. For small number of locations the freeware software WinBUGS (www.mrc-bsu.cam.ac.uk/bugs) can also be used to obtain MCMC simulation-based estimates. Proc Mixed for normal data supports Bayesian modeling by allowing specification of prior distributions for the parameters and MCMC simulation. However, this possibility is currently available only for variance component models and not for spatial covariances, which holds for the \%GLIMMIX macro, too.

Table 1: Comparison of the computational costs for the Bayesian, simulation based approaches.

| Model | Initial <br> sample <br> size | Final <br> sample <br> from <br> posterior | No. of batches and <br> size | Iterations to <br> convergence | Thinning* | Time per <br> 1,000 <br> iterations |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| MCMC | 50,000 | 1,720 | - | 7,000 | 25 | 7 hrs 14 <br> min |
| SIR | 400,000 | 1,600 | 800 batches with <br> 500 values (2 <br> batches per draw) | 0 | 0 | 1 hr 23 <br> min |

*Minimum lag at which autocorrelation was not significant.
Table 2: Comparison of parameter estimates from the binary spatial model using different estimation strategies. The binary outcome is the survival of the first year of life.

|  |  |  | Birth year |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Model | Estimate | $\sigma^{2}$ | $\phi$ | Intercept | $1966-71$ | $1972-77$ | $1978-83$ | $1984-89$ |
| GLM | MLE | - | - | 1.81 | -0.18 | 0.04 | 0.09 | 0.12 |
|  | 95\% CI | - | - | $1.43,2.11$ | $-0.44,0.09$ | $-0.22,0.29$ | $-0.16,0.34$ | $-0.13,0.37$ |
| PQL | MLE | 1.05 | 2.07 | 2.59 | -0.19 | 0.03 | 0.09 | 0.12 |
|  | 95\% CI | $0.72,1.81$ | $0.54,4.63$ | $1.43,3.74$ | $-0.48,0.11$ | $-0.26,0.31$ | $-0.19,0.37$ | $-0.17,0.40$ |
| MCMC | Mean | 1.32 | 0.07 | 1.76 | -0.20 | 0.01 | 0.07 | 0.10 |
|  | Median | 0.91 | 0.04 | 1.75 | -0.21 | 0.01 | 0.07 | 0.09 |
|  | 95\% CI | $0.22,3.89$ | $0.008,0.24$ | $1.47,2.09$ | $-0.46,0.08$ | $-0.25,0.27$ | $-0.19,0.33$ | $-0.16,0.36$ |
| SIR | Mean | 0.91 | 0.005 | 1.77 | -0.19 | 0.03 | 0.08 | 0.11 |
|  | Median | 0.61 | 0.03 | 1.73 | -0.18 | 0.03 | 0.08 | 0.11 |
|  | $95 \%$ CI | $0.22,2.62$ | $0.0004,0.015$ | $0.34,3.25$ | $-0.44,0.06$ | $-0.21,0.27$ | $-0.16,0.31$ | $-0.13,0.34$ |



Convergence of the PQL approach to the global mode of the likelihood was highly dependent on the starting values. We suggest to compare the results by running the procedure with several starting values. Computationally, the PQL is fast in comparison to the simulation-based procedures, MCMC and SIR, but it runs quickly out of workspace for larger dataset. A comparison of the computational time required for the MCMC and SIR algorithms is given in table 1. MCMC estimation was applied using a single chain. Convergence was assessed using Geweke's (1992) criterion. The algorithm converged after 7,000 iterations. A final sample from the posterior distribution of size 1,720 was obtained by sampling every 25th iterations after convergence was reached. The SIR algorithm
required extensive fine tuning in order to derive good estimates. We ran the sampler several times and adjusted the degrees of freedom and mean parameter in the bivariate $t$-distribution $g_{s}\left(\sigma^{2}, \phi\right)$, according to those values leading to large weights. Instead of resampling from the whole sequence of parameters according to their weights, we obtained better results by dividing the generated parameters into batches and drawing an equal number of samples with replacement from every batch. The implementation of the SIR algorithm was found to be difficult. Despite the effort applied to improve the SIR estimator, the derived weights show a highly skewed distribution, with a few dominating values (Figure 1).

Figure 1: Distribution of the weights in the Sampling-Importance-Resampling (SIR) procedure.


Table 2 gives the parameter estimates obtained by the four approaches. The fixed effect coefficients $\boldsymbol{\beta}$ show no fundamental difference in their point estimates between the competitive models, with the exception of the intercept coefficient. The PQL estimate of the intercept is higher than from the other estimators. The standard error of $\boldsymbol{\beta}$ estimated from GLM is narrower than in the spatial models, as we were expecting. Discrepancies between the fitting approaches are observed in the estimates of the covariance parameters $\sigma^{2}$ and $\phi$. The posterior density of $\sigma^{2}$ obtained from MCMC simulation was found to be highly skewed to the left. PQL overestimates $\phi$ suggesting a lower spatial variation than the Bayesian approaches. This confirms known results about bias in the PQL estimates especially for the covariance parameters $\sigma^{2}$ and $\phi$ due to the bad quality of the first-order approximation of the integrand. The SIR estimates are similar to those obtained from MCMC.

Figure 2: Variogram cloud of the residuals in a non-spatial model.


Figure 2 shows three plots of the semivariogram cloud based on the Anscombe residuals obtained after fitting the GLM model. The semivariogram cloud is a plot of half the squared difference of the residuals versus the distance between their sample locations. The mean of the squared differences at each lag gives an estimator of the semivariogram. The three plots correspond to the $5 \%, 50 \%$ and $95 \%$ quartile of the squared difference of the residuals. The semivariogram cloud shows high variability and an increasing trend from the origin indicating lag-dependent variation. For a stationary spatial process, the semivariogram relates to the covariance of the random effects. Therefore we expect high variability in the covariance parameters.

Figure 3 depicts different semivariogram estimators. The classical estimator by Matheron (1963) was calculated by

$$
\hat{\gamma}(h)=\frac{1}{|N(h)|} \sum_{N(h)}\left(Z\left(\mathbf{s}_{i}\right)-Z\left(\mathbf{s}_{j}\right)\right)^{2}
$$

where $Z\left(\mathbf{s}_{i}\right)$ is the Anscombe residual at location $\mathbf{s}_{i}$,

$$
N(h)=\left\{\left(\mathbf{s}_{i}, \mathbf{s}_{j}\right):\left\|\mathbf{s}_{i}-\mathbf{s}_{j}\right\|=h \pm \varepsilon\right\}
$$

and $|N(h)|$ is its cardinality. This estimator is sensitive to outliers and a robust version was
proposed by Cressie and Hawkins (1980), which is displayed in Figure 3, too. The MCMC, SIR and PQL based estimators were calculated by replacing the estimates of $\sigma^{2}$ and $\phi$ obtained from the three approaches in

$$
\gamma(h)=\sigma^{2}(1-\exp (-\phi \cdot h)) .
$$

The MCMC and SIR estimators appear to be between the two other empirical semivariogram estimators. Because we have omitted the nugget term, they pass through the origin. Nevertheless, their values fit nicely into the graph. The PQL estimate does not capture the correlation present at large lags. It represents the classical semivariogram estimator well, but it is far off the robust version.

Regarding our application, Figure 4 displays the locations of the DHS surveys and the observed infant mortality risk in Mali. The risk factors which were found to be statistically significant related to infant mortality (table 2) confirm findings made by other authors. The negative association between maternal education and mortality has been described by Farah and Preston (1982) and Cleland and Ginneken (1989). Higher education may result in higher health awareness, better utilization of health facilities (Jain, 1988), higher income and ability to purchase goods and services which improves infants health (Schultz, 1979).

Figure 3: Semivariogram estimators: Classical semivariogram estimator by Matheron (circles), Robust version by Cressie and Hawkins (triangles), MCMC (long dashed line), SIR (short dashed line) and PQL (line) fit.


The observed time trend, with higher infant survival for more recent years, was found not statistically significant. Longer birth intervals and low birth order reduce the risk of infant death. Mortality was related to the residency and sex of the infant with girls and urbanites being at lower risk of dying during the first year of life. The impact mothers age has on infant mortality shows the typical J-shape
(Kalipeni, 1993) with lowest risk for age around thirty. The higher risk in young women may be explained by not fully developed maternal resources and that in older women by the effect of ageing. The MCMC-based estimate of the $\phi$ parameter revealed strong spatial correlation which reduces to less than $5 \%$ for distances longer than 75 km .

Figure 4: Observed mortality in 36,906 infants from the DHS surveys conducted in the years 1995 and 1996 at 181 distinct locations in Mali.


Figure 5: Predicted spatial random effects from the infant mortality model using MCMC. The darker the shading, the lower the survival.


Predictions of the child mortality risk using the MCMC approach were made at 600,000 new locations on a regular grid, covering the whole area of Mali south of 18 degrees latitude north. Because the covariates are infant-specific and can not be extrapolated for the new locations, we predict the random effects only. The map with prediction is displayed in Figure 5. The map indicates a higher infant mortality risk mainly in the Northern part of the Niger delta. This region has low population density and water availability is seasonal. The many lakes in this region are preferred breeding site for the malaria mosquito. Low mortality is predicted in North-Western Mali at the border to Mauritania and Senegal. In this region, the population is more active in migrating to other countries for business purposes, bringing money to the region. Health facility coverage is also reflected in the predictive map, where the coverage is low in the Northern Niger delta and high in the North-East.

## Conclusion

Generalized linear mixed models for large pointreferenced spatial data are highly parameterized and their estimation is hampered by computational problems. Reliable estimation methods that can be applied in standard software or algorithms that can accurately estimate the model parameters within practical time constraints do not exist. In this paper we compared a few recent developments using a real dataset on infant mortality in Mali.

The advantage of the PQL method is that it can be applied in standard statistical software package. However estimates are biased especially those for the covariance parameters. The algorithm depends highly on the starting values and can easily converge to a local mode. For medium to large number of locations implementations of this algorithm is impeded by computer memory problems.

Bayesian methods can provide flexible ways of modeling point-referenced data, give unbiased estimates of the parameters and their standard error and have computational advantages for problems larger than the ones the maximum likelihood methods can handle. However, for very large number of locations, an
implementation may be infeasible due to long computing time. The SIR runs considerably faster than MCMC, but it requires tedious tuning. Finding an ISD which approximates well the posterior distribution is difficult to develop and application-specific. Rigorous methods for evaluating the suitability of the ISD do not exist. This increases the possibility of drawing misleading inference.

MCMC is the most practical and, when it comes to prediction, accurate approach to date for fitting geostatistical problems. However, it is computationally intensive, especially for dataset with large number of locations. More research is required in ways of improving the convergence of the algorithm and the inversion of large matrices. Gilks and Roberts (1996), Mira and Sargent (2000) and Haran et al. (2001) have proposed general MCMC algorithms for improving convergence. Rue (2000) and Pace and Barry (1997) have applied innovative numerical methods using sparse matrix solvers for fitting areal data. In future, similar approaches need to be adapted and assessed for modeling point-referenced spatial data.

## References

Anderson, D. A., \& Hinde, J. P. (1988). Random effects in generalized linear models and the EM algorithm. Communication in Statistics: Theory and Methods, 17, 3847-3856.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B, 36, 192-236.

Booth, J. G., \& Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. Journal of the Royal Statistical Society, Series B, 61, 265-285.

Breslow, N. E., \& Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9-25.

Breslow, N. E., \& Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. Biometrika, 82, 81-91.

Browne, W. J., \& Draper, D. (2000). A comparison of Bayesian and likelihood methods for fitting multilevel models. Submitted.

Cleland, J., \& van Ginneken, J. K. (1989). Maternal education and child survival in developing countries: The search for pathways of influence. Social Science and Medicine, 27, 1357-1368.

Cressie, N., \& Hawkins, D. M. (1980). Robust estimation of the variogram. Mathematical Geology, 12, 115-125.

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, 39, 138.

Diggle, P. J.,Tawn, J. A., \& Moyeed, R. A. (1998). Model-based geostatistics. Journal of the Royal Statistical Society, Series C, 47, 299350.

Ecker, M., \& Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process. Journal of Agricultural, Biological and Environmental Statistics, 2, 347369.

Farah, A. A., \& Preston, S. H. (1982). Child mortality differentials in Sudan. Population and Development Review, 8, 365383.

GDE Systems Inc. (1995). Geoname Digital Gazetteer, Version I, CD-ROM.

Gelfand. A. E.,Ravishanker. N., \& Ecker, M. (1999). Modeling and inference for point-referenced binary spatial data. In D. Dey,S. Ghosh \& B. Mallick (Eds.), Generalized linear models: a bayesian perspective. 373-386. Marcel Dekker Inc.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger,A.P. Dawid \& A. F. M. Smith (Eds.), Bayesian statistics, 4, 169-193. Oxford University Press.

Gilks, W. R., \& Roberts, G. O. (1996). Strategies for improving MCMC. In W. R. Gilks, S. Richardson, \& D. J. Spiegelhalter (Eds.), Markov chain Monte Carlo in practice. 89-114. London: Chapman and Hall.

Haran, M., Hodges, J. S., \& Carlin, B. P. (2001). Accelerating computation in Markov random field models for spatial data via structured MCMC. Journal of Computaional and Graphical Statistics, 12, 249-264.

Heagerty, P. J., \& Lele, S. R. (1998). A composite likelihood approach to binary spatial data. Journal of the American Statistical Association, 93, 1099-1111.

Jain, A. (1988). Determinants of regional variation in infant mortality in rural India. In A. Jain, \& L. Visaria (Eds.), Infant mortality in India: Differentials and determinants, 127-167. Sage Publications.

Kalipeni, E. (1993). Determinants of infant mortality in Malawi: a spatial perspective. Social Science and Medicine, 37, 183-198.

Lesaffre, E., \& Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. Journal of the Royal Statistical Society, Series C, 50, 325-335.

Liang, K. Y., \& Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13-22.

Matheron, G. (1963). Principles of geostatistics. Economic Geology, 58, 12461266.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association, 92, 162-170.

Mira, A., \& Sargent, D. J. (2000). Strategies for speeding Markov chain Monte Carlo algorithms. Technical Report, University of Insubria, Varese.

Neuhaus, J. N., \& Segal, M. R. (1997). An assessment of approximate maximum likelihood estimators in generalized linear mixed models. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, \& R. D. Wolfinger (Eds.), Modeling longitudinal and spatially correlated data: Lecture notes in statistics, 122, 11-22.

Pace, K. R., \& Barry, R. (1997). Quick computation of the regressions with spatially autoregressive dependent variable. Geographical Analysis, 29, 232-247.

Preisler. H. K. (1988). Maximum likelihood estimates for binary data with random effects. Biometrical Journal, 3, 339-350.

Rubin, D. B. (1987). Comment on: The calculation of posterior distributions by data augmentation. by M. A. Tanner and W. H. Wong. Journal of the American Statistical Association, 82, 543-546.

Rue, H. (2000). Fast sampling of gaussian markov random fields. Journal of the Royal Statistical Society, Series B, 48, 233-243.

Schultz, T. P. (1979, June 19-25). Interpretation of relations among mortality, economics of the household and the health environment. Proceedings of the meeting on socio-economic determinants and consequences of mortality, Mexico City. Geneva: World Health Organization.

Tubilla, A. (1975). Error convergence rates for estimates of multidimensional integrals of random functions. Technical Report No. 72, Department of Statistics, Stanford University, Stanford, CA.

Wolfinger, R., \& O’Connell, M. (1993). Generalized linear mixed models: A pseudolikelihood approach. Journal of Statistical Computation and Simulation, 48, 233-243.

World Resources Institute (1995) Africa Data Sampler. Edition I, CD-ROM.

Zeger, S. L., \& Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrika, 42, 121-130.

Zimmerman, D. L., \& Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometrics, 33, 77-91.

# Bootstrapping Confidence Intervals For Robust Measures Of Association 

Jason E. King<br>Baylor College of Medicine

A Monte Carlo simulation study compared four bootstrapping procedures in generating confidence intervals for the robust Winsorized and percentage bend correlations. Results revealed the superior resiliency of the robust correlations over $r$, with neither outperforming the other. Unexpectedly, the bootstrapping procedures achieved roughly equivalent outcomes for each correlation.

Key words: Robust methods, bootstrapping, percentage bend correlation, Winsorized correlation

## Introduction

A number of "robust" (Box, 1953) analogs to traditional estimators, population parameters, and hypothesis-testing methods have seen development during the past 40 years. Robust procedures typically retain the statistical interpretations associated with classical procedures, but are more resistant to distributional non-normalities and outliers. The Pearson product-moment correlation is without question the most commonly used measure of linear association, yet is not robust to departures from normality, especially when the bivariate surface is non-normal and dependence exists (King, 2003).

Two new robust alternatives to $r$ appear promising. The Winsorized correlation (Devlin, Gnanadesikan, \& Kettenring, 1975; Gnanadesikan \& Kettenring, 1972; Wilcox, 1993) and the percentage bend correlation (Wilcox, 1994, 1997) yield interpretations analogous to $r$ and asymptotically equal zero under bivariate independence, yet possess properties that curb the influence of distributional non-normalities.

This article was based on the doctoral dissertation by Jason E. King. The author acknowledges Professor Bruce Thompson and the doctoral committee for their contributions. Email address: jasonk@bcm.tmc.edu.

The Winsorized correlation $\left(r_{\mathrm{w}}\right)$ is computed in an identical fashion to $r$ except that a specified proportion of extreme scores in each tail are first Winsorized, that is, deleted and set equal to the most extreme score remaining in the tail of the distribution. The percentage bend correlation ( $r_{\mathrm{pb}}$ ) is based on the percentage bend measures of location and midvariance and is less intuitive. See Wilcox $(1994,1997)$ for the relevant equations.

Yet few researchers have explored these newer correlations, notably with respect to estimating confidence intervals and defining their sampling distributions. For statistics with no known sampling distribution, Efron's (1979, 1982) bootstrap has proven to be effective in a variety of contexts. The conjecture is that the sampling distribution of a statistic can be approximated by the distribution of a large number of resampled estimates of the statistic obtained from a single sample of observations.

The distribution of resampled estimates forms an empirically-derived sampling distribution from which confidence intervals or other indices may be estimated, either for inferential or descriptive purposes (Thompson, 1993). The usefulness of bootstrapping is evident because an increasing number of disciplines are now encouraging or requiring the reporting of confidence intervals (Thompson, 2002; Vacha-Haase, Nilsson, Reetz, Lance, \& Thompson, 2000; Wilkinson \& APA Task Force on Statistical Inference, 1999).

An "almost bewildering array" (Hall, 1988, p. 927) of bootstrapping procedures is now available. These vary in the accuracy with which the bootstrap-generated interval spans the true interval. Accuracy is also contingent on the
type of statistic under examination. At the current level of knowledge, it is unknown which bootstrapping procedure produces the most accurate confidence intervals for $r_{\mathrm{pb}}$ and $r_{\mathrm{w}}$. Although Wilcox $(1993,1994,1997)$ compared Type I error rates for these robust correlations, only two studies (Wilcox, 1997; Wilcox \& Muska, 2001) have examined the accuracy of bootstrapped confidence intervals for $r_{\mathrm{pb}}$, and none for $r_{\mathrm{w}}$. Clearly, more research is needed.

The goal of this simulation study was to compare various means of bootstrapping confidence intervals for $r_{\mathrm{w}}$ and $r_{\mathrm{pb}}$ across a variety of conditions. The study compared four bootstrapping procedures, each of which has proven useful in some contexts: the ordinary percentile bootstrap (Efron, 1979), an adjusted bootstrap (Strube, 1988), the bias-corrected bootstrap (BC; Efron, 1981, 1982, 1985), and the bias-corrected and accelerated bootstrap ( $\mathrm{BC}_{\mathrm{a}}$; Efron, 1987). The Pearson $r$ and Fisher's inverse hyperbolic tangent transformation of $r$, $r_{z}$, were included for comparative purposes, although the latter frequently fails to produce even asymptotically correct results (Duncan \& Layard, 1973).

## Methodology

The simulation procedure began by randomly generating $1,000,000$ observations from a population with known characteristics, serving as a derived population. This step was necessary because the Winsorized and percentage bend correlation parameters ( $\rho_{\mathrm{w}}$ and $\rho_{\mathrm{pb}}$ ) will not necessarily exactly equal $\rho$ under dependence conditions. The second step involved drawing $m$ $=100$ samples, each of size $n$, from the derived population and calculating sample estimates for each of the four correlational measures. Lastly, $B=500$ bootstrap samples were drawn by sampling with replacement from each of the $m$ samples and $95 \%$ confidence intervals calculated via each of the four bootstrapping procedures. Gamma ( $\gamma$ ) and beta ( $\beta$ ) are two constants that must be fixed in computing the Winsorized and percentage bend correlations, respectively. These were each set to .2 for all simulations.

Real data often demonstrate excessive distributional non-normality (Bradley, 1977, 1978; Micceri, 1989; Rasmussen, 1986; Stigler, 1973; Wilcox, 1990) and such can moderate the
accuracy of a bootstrapping procedure for a given statistic (Hall, 1988; Wilcox, 1997). Thus, the present study compared bootstrapped correlations across a wide range of conditions including nine distributional shape variations, one contaminated distribution, six mixed distributions, three independence and dependence conditions (i.e., population correlations of $.0, .4, .8$ ), and four sample sizes (i.e., $n s$ of $20,50,100,250$ ).

Four indices served as points of comparison for the bootstrapped correlations: Type I error rate, bias, efficiency, and interval width. The latter was constructed by modifying a ratio proposed by Efron (1988) such that the width of each bootstrap-estimated interval was divided by the width of a "true" (i.e., Monte Carlo-estimated) confidence interval. This required drawing an additional 10,000 samples, each of size $n$, from each simulated population to create the "true" sampling distributions.

Simulation studies typically compare Type I error rates and other indices in an informal manner; however, a more formal analysis is useful for processing the large number of indices obtained in the present study. Analysis of Variance (ANOVA) is well suited for quantifying sources of variation. This procedure allowed for partitioning the systematic variance components affecting the indices (viz., correlational measure, bootstrapping procedure, distributional shape and type, sample size, and strength of bivariate relationship).

## Results

Tables 1-5 and Figures 1-2 display representative results averaged across distributional shape. Disaggregated data and fuller explanations are available in King (2000). Efficiency varied little across the correlational measures and is not presented.

Comparisons Among Bootstrapping Procedures
As regards Type I error rate (see Tables 1, 2, and Figure 1) and bias (see Tables 3, 4, and Figure 2), no bootstrapping procedure emerged as unmistakably superior across a majority of conditions for either robust correlation (e.g., a Bootstrap by Correlation effect is absent in Tables 2 and 4).

Table 1. Type I Error Rates Averaged Across All Distributional Conditions

|  | $n=20$ |  |  |  | $n=50$ |  |  |  | $n=100$ |  |  |  | $n=250$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\mathrm{pb}}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\mathrm{pb}}$ |
| $\rho=0$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 06 | . 06 | . 03 | . 04 | . 07 | . 07 | . 06 | . 07 | . 05 | . 05 | . 05 | . 05 | . 07 | . 07 | . 04 | . 04 |
| Adjusted | . 06 | . 06 | . 03 | . 04 | . 07 | . 07 | . 06 | . 07 | . 05 | . 05 | . 05 | . 05 | . 07 | . 07 | . 04 | . 04 |
| BC | . 05 | . 05 | . 03 | . 05 | . 07 | . 07 | . 06 | . 06 | . 05 | . 05 | . 06 | . 05 | . 06 | . 06 | . 03 | . 04 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 05 | . 04 | . 03 | . 05 | . 07 | . 07 | . 05 | . 06 | . 06 | . 06 | . 06 | . 05 | . 07 | . 07 | . 04 | . 04 |
| $\rho=.4$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 11 | . 11 | . 03 | . 07 | . 08 | . 08 | . 04 | . 06 | . 07 | . 07 | . 04 | . 04 | . 08 | . 08 | . 05 | . 05 |
| Adjusted | . 11 | . 11 | . 04 | . 08 | . 08 | . 08 | . 04 | . 06 | . 08 | . 08 | . 04 | . 04 | . 08 | . 08 | . 05 | . 05 |
| BC | . 08 | . 08 | . 03 | . 06 | . 08 | . 08 | . 03 | . 06 | . 08 | . 08 | . 04 | . 04 | . 09 | . 09 | . 05 | . 05 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 09 | . 08 | . 04 | . 07 | . 09 | . 09 | . 04 | . 06 | . 10 | . 10 | . 04 | . 03 | . 11 | . 11 | . 04 | . 05 |
| $\rho=.8$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 09 | . 09 | . 06 | . 07 | . 06 | . 06 | . 06 | . 06 | . 06 | . 06 | . 06 | . 04 | . 07 | . 07 | . 05 | . 05 |
| Adjusted | . 15 | . 15 | . 09 | . 12 | . 06 | . 06 | . 06 | . 06 | . 08 | . 08 | . 07 | . 07 | . 08 | . 08 | . 04 | . 05 |
| BC | . 10 | . 10 | . 05 | . 05 | . 06 | . 06 | . 06 | . 07 | . 07 | . 07 | . 06 | . 04 | . 08 | . 08 | . 06 | . 05 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 12 | . 12 | . 06 | . 07 | . 07 | . 07 | . 05 | . 06 | . 10 | . 10 | . 06 | . 04 | . 09 | . 09 | . 06 | . 06 |

Note. Italicized values are greater than two standard errors beyond the nominal .05 level.
Table 2. Analysis of Variance for Type I Error Rate by Correlation and Bootstrapping Procedure

| Source | $d f$ | $F$ | $p$ | $\eta^{2}$ |
| :--- | :---: | :---: | :---: | :---: |
| Model | 15 | 11.028 | $<.001$ | .088 |
| CORR | 3 | 50.511 | $<.001$ | .081 |
| BOOT | 3 | 2.735 | .042 | .004 |
| CORR * BOOT | 9 | .631 | .772 | .003 |
| Error | 1712 | $(.002)$ |  |  |
| Total | 1727 |  |  |  |

Note. Mean square error enclosed in parentheses.

Figure 1. Mean Type I error rate by correlation and bootstrapping procedure. Reference line indicates the nominal alpha rate of .05 .


Table 3. Interval Bias Averaged Across All Distributional Conditions

|  | $n=20$ |  |  |  | $n=50$ |  |  |  | $n=100$ |  |  |  | $n=250$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\mathrm{pb}}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\mathrm{pb}}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\mathrm{pb}}$ |
| $\rho=0$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 33 | . 33 | . 30 | . 31 | . 24 | . 24 | . 22 | . 22 | . 17 | . 17 | . 16 | . 16 | . 11 | . 11 | . 10 | . 10 |
| Adjusted | . 36 | . 36 | . 35 | . 34 | . 24 | . 24 | . 23 | . 23 | . 17 | . 17 | . 16 | . 16 | . 11 | . 11 | . 10 | . 10 |
| BC | . 32 | . 32 | . 31 | . 31 | . 23 | . 23 | . 22 | . 22 | . 16 | . 16 | . 16 | . 16 | . 11 | . 11 | . 10 | . 10 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 34 | . 33 | . 31 | . 31 | . 24 | . 24 | . 22 | . 22 | . 17 | . 17 | . 16 | . 16 | . 11 | . 11 | . 10 | . 10 |
| $\rho=.4$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 36 | . 36 | . 33 | . 33 | . 24 | . 24 | . 20 | . 20 | . 21 | . 21 | . 15 | . 15 | . 15 | . 15 | . 10 | . 09 |
| Adjusted | . 38 | . 38 | . 37 | . 37 | . 24 | . 24 | . 21 | . 21 | . 21 | . 21 | . 15 | . 15 | . 15 | . 15 | . 10 | . 09 |
| BC | . 36 | . 36 | . 33 | . 32 | . 24 | . 24 | . 21 | . 20 | . 21 | . 21 | . 15 | . 14 | . 16 | . 16 | . 10 | . 10 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 38 | . 37 | . 33 | . 33 | . 26 | . 25 | . 21 | . 20 | . 23 | . 23 | . 15 | . 15 | . 17 | . 17 | . 10 | . 10 |
| $\rho=.8$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 26 | . 26 | . 25 | . 23 | . 17 | . 17 | . 14 | . 13 | . 13 | . 13 | . 09 | . 08 | . 10 | . 10 | . 06 | . 05 |
| Adjusted | . 31 | . 31 | . 31 | . 30 | . 17 | . 17 | . 15 | . 15 | . 13 | . 13 | . 09 | . 09 | . 10 | . 10 | . 06 | . 06 |
| BC | . 26 | . 26 | . 28 | . 24 | . 17 | . 17 | . 14 | . 14 | . 14 | . 14 | . 09 | . 09 | . 11 | . 11 | . 06 | . 06 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 28 | . 28 | . 28 | . 25 | . 18 | . 18 | . 15 | . 15 | . 15 | . 15 | . 09 | . 09 | . 12 | . 12 | . 06 | . 06 |

Table 4. Analysis of Variance for Bias by Correlation and Bootstrapping Procedure

| Source | $d f$ | $F$ | $p$ | $\eta^{2}$ |
| :--- | :---: | :---: | :---: | :---: |
| Model | 15 | 3.497 | $<.001$ | .030 |
| CORR | 3 | 15.558 | $<.001$ | .026 |
| BOOT | 3 | 1.551 | .003 |  |
| CORR * BOOT | 9 | .125 | .999 | .001 |
| Error | 1712 | $(.010)$ |  |  |
| Total | 1727 |  |  |  |

Note. Mean square error enclosed in parentheses.

Figure 2. Mean bias by correlation and bootstrapping procedure.


Table 5. Confidence Interval Ratios Averaged Across All Distributional Conditions

|  | $n=20$ |  |  |  | $n=50$ |  |  |  | $n=100$ |  |  |  | $n=250$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ | $r$ | $r_{z}$ | $r_{\text {w }}$ | $r_{\text {pb }}$ |
| $\rho=0$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 91 | . 91 | 1.11 | 1.02 | . 91 | . 91 | 1.04 | 1.00 | . 92 | . 92 | 1.03 | 1.00 | . 93 | . 93 | 1.01 | . 99 |
| Adjusted | . 98 | . 98 | 1.20 | 1.10 | . 94 | . 94 | 1.07 | 1.03 | . 94 | . 94 | 1.04 | 1.02 | . 94 | . 94 | 1.01 | 1.00 |
| BC | . 92 | . 92 | 1.11 | 1.02 | . 91 | . 91 | 1.04 | 1.00 | . 92 | . 92 | 1.03 | 1.00 | . 93 | . 93 | 1.01 | . 99 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 92 | . 92 | 1.11 | 1.02 | . 92 | . 92 | 1.04 | 1.00 | . 93 | . 93 | 1.02 | 1.00 | . 94 | . 94 | 1.01 | . 99 |
| $\rho=.4$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 84 | . 84 | 1.09 | 1.00 | . 84 | . 84 | 1.04 | . 99 | . 92 | . 92 | 1.03 | 1.00 | . 86 | . 86 | 1.01 | 1.00 |
| Adjusted | . 91 | . 91 | 1.17 | 1.08 | . 86 | . 86 | 1.07 | 1.02 | . 93 | . 93 | 1.04 | 1.02 | . 87 | . 87 | 1.02 | 1.00 |
| BC | . 86 | . 86 | 1.11 | 1.02 | . 84 | . 84 | 1.05 | 1.00 | . 91 | . 91 | 1.02 | 1.00 | . 86 | . 86 | 1.01 | 1.00 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 85 | . 86 | 1.11 | 1.02 | . 84 | . 84 | 1.05 | 1.01 | . 92 | . 92 | 1.03 | 1.00 | . 86 | . 86 | 1.01 | 1.00 |
| $\rho=.8$ |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Percentile | . 81 | . 81 | 1.08 | . 98 | . 85 | . 85 | 1.05 | 1.02 | . 81 | . 81 | 1.00 | . 98 | . 84 | . 84 | 1.01 | 1.00 |
| Adjusted | . 87 | . 87 | 1.16 | 1.06 | . 88 | . 88 | 1.08 | 1.05 | . 83 | . 83 | 1.01 | . 99 | . 84 | . 84 | 1.01 | 1.00 |
| BC | . 85 | . 85 | 1.17 | 1.04 | . 87 | . 87 | 1.08 | 1.04 | . 82 | . 82 | 1.01 | . 99 | . 84 | . 84 | 1.01 | 1.00 |
| $\mathrm{BC}_{\mathrm{a}}$ | . 83 | . 86 | 1.16 | 1.05 | . 85 | . 87 | 1.10 | 1.06 | . 81 | . 82 | 1.02 | 1.01 | . 84 | . 85 | 1.01 | 1.01 |

Note: Ratios greater than 1.0 indicate a bootstrap-estimated interval wider than the "true" interval, and conversely.

Under a few conditions, the BC and ordinary percentile procedures procured slightly more accurate intervals than did the $\mathrm{BC}_{\mathrm{a}}$. In addition, the adjusted bootstrap intervals were, by and large, unacceptable, regardless of the robust measure under examination.

Regarding the width of the estimated intervals (see Table 5), no bootstrapping procedure clearly bettered the others. For small sample size conditions the adjusted bootstrap averaged relatively wider intervals. This widening effect improved accuracy for the narrow $r$ - and $r_{z}$-generated intervals, but
penalized $r_{\mathrm{w}}$ and $r_{\mathrm{pb}}$. The $\mathrm{BC}_{\mathrm{a}}$ intervals frequently ran short, the BC intervals shorter still, and the percentile bootstrap the shortest of the four. These trends were slight and not unexpected (e.g., it is widely known that the percentile bootstrap tends to produce narrow intervals).

## Comparisons Among Correlations

Confidence intervals formed for $r_{\mathrm{w}}$ and $r_{\mathrm{pb}}$ generally outperformed those for $r$ and $r_{z}$ for both Type I error rate and bias. Although the present paper does not depict the data
disaggregated by distributional shape, predictable results surfaced. Under normality all four correlations produced similar Type I error rates, although the Pearson $r$ and its transform saw slightly lower levels of bias, at least under small sample conditions. However, as distributional shape diverged from normality or included contaminated or mixed distributions, the robust correlations surpassed $r$. As an aside, the bias index generally produced neater, more theoretically consistent results than did Type I error rate. This is probably due to the dichotomous nature of the latter, that is, a given interval either does or does not enclose the parameter of interest and cause a Type I error, whereas bias is measured on the more sensitive ratio scale of measurement.

Regarding interval width, $r$ and its transform consistently underestimated the "true" endpoints, more so under non-normal conditions. At times, such intervals were little more than half the "true" width. Bootstrapped confidence intervals for the percentage bend correlation closely mimicked the "true" intervals in almost every instance, and intervals for the Winsorized correlation tended to run slightly wide.

## Conclusion

This study confirmed that the Winsorized and percentage bend correlations are useful alternatives to the Pearson correlation and are preferred when resilience to distributional nonnormality is needed. Results for three of the four comparative indices (efficiency was virtually a constant) confirmed the robustness of the two robust measures under non-normal, mixed, and contaminated distributional conditions, with neither outperforming the other. The percentage bend and Winsorized correlations reduced bias, more accurately reflected theoretical Type I error probabilities, and more faithfully reproduced the width of true (Monte Carlo simulated) intervals. The robust measures compared favorably to $r$ even under the bivariate normal conditions.

Interestingly, across a wide range of simulation conditions the four bootstrapping procedures achieved roughly equivalent outcomes as applied to either robust correlation. The complex BC and $\mathrm{BC}_{\mathrm{a}}$ procedures failed to offer sizeable improvements in interval accuracy
over the percentile bootstrap, and the "adjusted" bootstrap may have even inflated bias and Type I error rate. While this finding may be interpreted as disappointing because the more elaborate procedures did not offer increased accuracy, researchers can be more confident that the ordinary percentile bootstrap is capable of delivering relatively precise confidence intervals for these robust measures.

It may be that the more complicated procedures did not surpass the percentile bootstrap due to the technical specifications of the simulation. The original study design entailed drawing 1,000 samples for each condition, but this number was reduced to 100 given excessive computational demands. Even though the goal in this component of the simulation procedure is not to fully reproduce a sampling distribution, more samples may be necessary to achieve stable asymptotic dynamics. Similarly, the number of bootstrap samples had to be reduced considerably (e.g., setting $B$ to 3,000 produced only 25 samples in eight hours due to the large number of simulated conditions and the involved calculations for $r_{\mathrm{pb}}$ ). However, for this simulation component, the objective is indeed to model a theoretical sampling distribution, $F(\theta)$, via a bootstrapped sampling distribution, $\hat{F}\left(\hat{\theta}^{*}\right)$. Five hundred bootstrap samples may be sufficient for estimating standard errors (Efron, 1987; Efron \& Tibshirani, 1993; but cf. Booth \& Sarkar, 1998), but not for forming tight confidence bands (Lunneborg, 2000). Follow-up studies should increase these quantities if possible.

The study also revealed that Fisher's transformation of $r$ did not appreciably improve either Type I error rate or bias. When bootstrapping the Pearson correlation, it seems that the $r$-to-z transformation merely increases computational time without concomitantly affecting accuracy, as supported by Seivers (1996) in his conclusions about $r_{z}$.

In sum, the robust measures may be recommended for general use when it is desired to quantify the linear association underlying the majority of the sample observations, while excluding outliers. Each of the bootstrapping procedures reviewed maintained similar levels of accuracy and may be applied in estimating confidence intervals for the robust correlations, excepting the adjusted bootstrap.

## References

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika, 40, 318-335.

Booth, J. G., \& Sarkar, S. (1998). Monte Carlo approximation and bootstrap variances. American Statistician, 52, 354-357.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. American Statistician, 31, 147-50.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Devlin, S. J., Gnanadesikan, R., \& Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika, 62, 531-545.

Duncan, G. T., \& Layard, M. W. (1973). A Monte-Carlo study of asymptotically robust tests for correlation. Biometrika, 60, 551-558.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1-26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. Canadian Journal of Statistics, 9, 139-172.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. Biometrika, 72, 45-58.

Efron, B. (1987). Better bootstrap confidence intervals. Journal of the American Statistical Association, 82, 171-185.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? Psychological Bulletin, 104, 293-296.

Efron, B., \& Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman \& Hall.

Gnanadesikan, R., \& Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. Annals of Statistics, 16, 927-953.

Huber, P. J. (1981). Robust Statistics. New York: Wiley.

King, J. E. (2000). Bootstrapping robust measures of association (Doctoral dissertation, Texas A\&M University, 2000). Dissertation Abstracts International, 61(11), 5948B.

King, J. E. (2003, February). What have we learned from 100 years of robustness studies on $r$ ? Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Lunneborg, C. E., (2000). Data analysis by resampling: Concepts and applications. Pacific Grove, CA: Brooks/Cole.

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

Rasmussen, J. L. (1986). An evaluation of parametric and non-parametric tests on modified and non-modified data. British Journal of Mathematical and Statistical Psychology, 39, 213-220.

Sievers, W. (1996). Standard and bootstrap confidence intervals for the correlation coefficient. British Journal of Mathematical and Statistical Psychology, 49, 381-396.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. Journal of the American Statistical Association, 68, 872-879.

Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. Psychological Bulletin, 104, 290-292.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, 31(3), 24-31.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., \& Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. Theory \& Psychology, 10, 413-425.

Wilcox, R. R. (1990). Comparing variances and means when distributions have non-identical shapes. Communications in Statistics-Simulation and Computation, 19, 155173.

Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. British Journal of Mathematical and Statistical Psychology, 46, 339-349.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. Psychometrika, 59, 601-616.

Wilcox, R. R. (1997). Tests of independence and zero correlations among $p$ random variables. Biometrical Journal, 39, 183193.

Wilcox, R. R., \& Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. British Journal of Mathematical \& Statistical Psychology, 54, 3947.

Wilkinson, L., \& APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594604.

# JMASM Algorithms and Code <br> JMASM8: Using SAS To Perform Two-Way Analysis Of Variance Under Variance Heterogeneity 

Scott J. Richter<br>Department of Mathematical Sciences<br>University of North Carolina at Greensboro

Mark E. Payton<br>Department of Statistics<br>Oklahoma State University

We present SAS code to implement the method proposed by Brunner et al. (1997) for performing twoway analysis of variance under variance heterogeneity.

Key words: ANOVA, heteroscedasticity, SAS

Introduction
Brunner et al. (1997) suggested a method to perform tests on main effects and interaction in two-way analysis of variance that allows variance heterogeneity. Their approach is to use a generalization of chi-square approximations dating back to Patnaik (1949) and Box (1954). Their statistic is identical to the classical ANOVA F-statistic, and thus their method can be regarded as a robust extension of the classical ANOVA to heteroscedastic designs. They recommend that their method should always be preferred (even in the homoscedastic case) to the classical ANOVA. Richter and Payton (2003) found that the performance of their statistic compares favorably to that of the usual ANOVA F-statistic for sample sizes of at least $\mathrm{n}=7$ per factor combination.

In this article, we present a SAS program (SAS Institute, Cary, N.C.) for implementing the Brunner et al. (1997) method.

Scott Richter is an Assistant Professor in the Mathematical Sciences Department at the University of North Carolina at Greensboro. His email address is sjricht2@uncg.edu. Mark Payton is a Professor in the Department of Statistics at Oklahoma State University. His email address is mpayton@okstate.edu.

Brunner Method
The method of Brunner et al. (1997) is a small sample adjustment to the well-known Wald statistic, which permits heterogeneous variance but is known to have inflated Type I error rates for small sample sizes. Consider a two-way layout with $a$ levels of factor $A$ and $b$ levels of factor $B$. Assume a set of independent random variables

$$
X_{i j} \sim N\left(\mu_{i}, \sigma_{i}^{2}\right), i=1, \ldots, a b
$$

Let $\boldsymbol{\mu}=\left(\mu_{1}, \mu_{2}, \ldots, \mu_{a b}\right)^{\prime}$ denote the vector containing the $a \bullet b$ population means. Then the hypotheses of no main effects and interaction can be written as

$$
\begin{gathered}
H_{0}(A): \mathbf{M}_{A} \boldsymbol{\mu}=0 \\
H_{0}(B): \mathbf{M}_{B} \boldsymbol{\mu}=0 \\
H_{0}(A B): \mathbf{M}_{A B} \boldsymbol{\mu}=0,
\end{gathered}
$$

where

$$
\begin{aligned}
& \mathbf{M}_{A}=\mathbf{P}_{a} \otimes \frac{1}{b} \mathbf{J}_{b} \\
& \mathbf{M}_{B}=\frac{1}{a} \mathbf{J}_{a} \otimes \mathbf{P}_{b} \\
& \mathbf{M}_{A B}=\mathbf{P}_{a} \otimes \frac{1}{b} \mathbf{J}_{b} .
\end{aligned}
$$

Here $\mathbf{P}_{c}=\mathbf{I}_{c}-\frac{1}{c} \mathbf{J}_{c}$, where $\mathbf{I}_{c}$ is a $c \times c$ identity matrix, $\mathbf{J}_{c}$ a $c \times c$ matrix of 1 's, and the symbol $\otimes$ represents the Kronecker product of the matrices. The vector of observed cell means is denoted by $\overline{\mathbf{X}}=\left(\bar{X}_{1}, \ldots, \bar{X}_{a b}\right)^{\prime}$ and the estimated covariance matrix is given by $\hat{\mathbf{S}}_{N}=N \bullet \operatorname{diag}\left\{\frac{S_{1}^{2}}{n_{1}}, \ldots, \frac{S_{a b}^{2}}{n_{a b}}\right\}$, where $S_{i}^{2}$ is the $i^{\text {th }}$ sample variance and $N=\sum_{i=1}^{a b} n_{i}$.

For a complete cross-classification, the test statistic is

$$
F B=\frac{N \bullet \overline{\mathbf{X}}^{\prime} \mathbf{M} \overline{\mathbf{X}}}{\frac{1}{(n-1)} \operatorname{tr}\left(\hat{\mathbf{S}}_{N}\right)},
$$

which has an approximate $F$ distribution with

$$
f_{n u m}=\frac{\frac{1}{(n-1)^{2}} \bullet\left[\operatorname{tr}\left(\hat{\mathbf{S}}_{N}\right)\right]^{2}}{\operatorname{tr}\left(\mathbf{M} \hat{\mathbf{S}}_{N} \mathbf{M} \hat{\mathbf{S}}_{N}\right)}
$$

numerator and

$$
f_{\operatorname{den}}=\frac{\left[\operatorname{tr}\left(\hat{\mathbf{S}}_{N}\right)\right]^{2}}{\operatorname{tr}\left(\hat{\mathbf{S}}_{N}^{2} \boldsymbol{\Lambda}\right)}
$$

denominator degrees of freedom, where

$$
\mathbf{\Lambda}=\operatorname{diag}\left\{\frac{1}{n_{1}-1}, \ldots, \frac{1}{n_{a b}-1}\right\}
$$

(Brunner, 1997).

## References

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. The Annals of Mathematical Statistics, 25, 290-302.

Brunner, E., Dette, H., \& Munk, A. (1997). Box-type approximations in nonparametric factorial designs. Journal of the American Statistical Association, 92, 14941502.

Patnaik, P. B. (1949). The noncentral $\chi^{2}$ and F-distributions and their applications. Biometrika, 36, 202-232.

Richter, S. J., \& Payton, M. E. (2003). Performing two-way analysis of variance under variance heterogeneity. Journal of Modern Applied Statistical Methods, 2(1), 152-160.

## Appendix

The following program can be used to perform the FB test described above.

```
/* Program to compute unadjusted and Box-adjusted F-ratios
and p-values for a two-way layout. */
data rcht;
input a b RESP @@;
datalines;
<data>
;
proc sort;
by a b;
```

```
/* Run Proc Mixed to calculate variances */
proc mixed data=rcht; class a b;
model RESP=a|b;
repeated/type=un(1) group=a*b;
```

/* Create data sets for covariances, means and class levels for input
to Proc IML */
ods listing exclude covparms; ods output covparms=tempcov;
ods listing exclude classlevels; ods output classlevels=levels;
ods listing exclude dimensions; ods output dimensions=sizes;
/* Suppress printing of Proc Mixed tables */
ods listing exclude fitstatistics;
ods listing exclude reml;
ods listing exclude ConvergenceStatus;
ods listing exclude IterHistory;
ods listing exclude lrt;
ods listing exclude modelinfo;
ods listing exclude tests3;
/* Use Proc GLM to calculate and output Type III F-statistics */
proc glm data = rcht noprint outstat=fstats;
class a b; model RESP=a|b/ss3; run;
/* Use Proc Means to calculate and output cell means and sample sizes
*/
proc means data=rcht noprint;
var RESP; by a b;
output mean=means $n=n$ out=tempss;
/* Begin Proc IML to calculate adjusted df and p-values */
proc iml;
/* Create matrices from data sets created above */
use levels;
read point 1 var \{levels\} into nla;
read point 2 var \{levels\} into nlb; use sizes;
read point 1 var \{value\} into parms;
read point 8 var \{value\} into nobs; use tempss;
read all var $\{n\}$ into ni; use tempcov;
read all var \{estimate\} into tsighat;
sighat=diag(tsighat);
do i=1 to parms;
sighat[i,i]=sighat[i,i]/ni[i];
end;
shat=nobs*sighat; use tempss;
read all var \{means\} into Xbar;
results=j(3,4,0); use fstats;
read point 2 var $\{F\}$ into FHA;
read point 3 var $\{F\}$ into $F H B$;
read point 4 var $\{F\}$ into $F H A B$;

```
read point 2 var {df} into dfa;
read point 3 var {df} into dfb;
read point 4 var {df} into dfAB;
read point 1 var {df} into dfe;
RESULTS[1,1]=FHA;
RESULTS[2,1]=FHB;
RESULTS[3,1]=FHAB;
/* Calculations for Box-type adjustment */
MA=(i(nla)-(1/nla)*j(nla))@(1/nlb*j(nlb));
DMA=diag(MA);
denA=trace(DMA*Shat);
dmas=dma*shat;
QA=nobs*Xbar`*MA*Xbar;
FNA=QA/denA;
RESULTS[1,3]=FNA;
MB=((1/nla)*j(nla))@(i(nlb)-(1/nlb)*j(nlb));
DMB=diag(MB);
denb=trace(DMb*Shat);
Qb=nobs*Xbar`*Mb*Xbar;
FNB=Qb/denb;
RESULTS[2,3]=FNB;
MAB=(i(nla)-(1/nla)*j(nla))@(i(nlb)-(1/nlb)*j(nlb));
DMAB=diag(MAB);
denAb=trace(DMAb*Shat);
QAb=nobs*Xbar`*MAb*Xbar;
FNAB=QAb/denAb;
RESULTS[3,3]=FNAB;
Lambda=DIAG(1/NI);
/* Calculate adjusted df */
fA=((trace(DMA*Shat))**2)/(trace((MA*Shat)*(MA*Shat)));
foA=(trace(DMA*Shat))**2/(trace(DMA**2*Shat**2*Lambda));
fB=((trace(DMB*Shat))**2)/(trace((MB*Shat)*(MB*Shat)));
foB=(trace(DMB*Shat))**2/(trace(DMB**2*Shat**2*Lambda));
fAB=((trace(DMAB*Shat))**2)/(trace((MAB*Shat)*(MAB*Shat)));
foAB=(trace(DMAB*Shat))**2/(trace(DMAB**2*Shat**2*Lambda));
/* Calculate p-values */
adjpvalA=1-probf(FnA,fA,foA);
RESULTS[1,4]=ADJPVALA;
adjpvalB=1-probf(FnB,fB,foB);
RESULTS[2,4]=ADJPVALB;
adjpvalAB=1-probf(FnAB,fAB,foAB);
RESULTS[3,4]=ADJPVALAB;
pvala=1-probf(fHa,dfa,dfe);
RESULTS[1,2]=PVALA;
pvalb=1-probf(fHb,dfb,dfe);
RESULTS[2,2]=PVALB;
```

pvalab=1-probf(fHab,dfab,dfe);
RESULTS $[3,2]=$ PVALAB;

```
/* Print results */
```

headings $=\left\{\right.$ ANOVA $\mathrm{F}^{\prime}$ ' P-value' ' Adjusted F' ' Pvalue'\};
EFFECT $=\{\mathrm{A}, \mathrm{B}, \mathrm{AB}\}$; mattrib RESULTS colname=HEADINGS
FORMAT=7.3 label='Results';
print effect RESULTS;
quit;

# JMASM9: Converting Kendall’s Tau For Correlational Or Meta-Analytic Analyses 

David A. Walker<br>Educational Research and Assessment<br>Northern Illinois University

Expanding on past research, this study provides researchers with a detailed table for use in meta-analytic applications when engaged in assorted examinations of various $r$-related statistics, such as Kendall's tau ( $\tau$ ) and Cohen's d, that estimate the magnitude of experimental or observational effect. A program to convert from the lesser-used tau coefficient to other effect size indices when conducting correlational or meta-analytic analyses is presented.

Key words: Effect size, meta-analysis

## Introduction

There is a heightened effort within the social and behavioral sciences to report effect sizes with research findings (APA, 2001; Henson \& Smith, 2000; Knapp, 1998). Effect sizes show the strength and magnitude of a relationship and account for the total variance of an outcome. The American Psychological Association (APA) encouraged recently, "Always provide some effect size estimate when reporting a p value" (Wilkinson \& The APA Task Force on Statistical Inference, 1999, p. 599). An analysis of effect sizes allows researchers to evaluate the importance of the result and not just the probability of the result (Kirk, 1996; Shaver, 1985).

Furthermore, effect sizes fall into two categories: d and $r$. The d group encompasses measures of effect size in terms of mean difference and standardized mean difference. Cohen (1988) defined the values of effect sizes for this group as small $\mathrm{d}=.20$, medium $\mathrm{d}=.50$, and large $\mathrm{d}=.80$.

David Walker is an Assistant Professor of Educational Research and Assessment, Northern Illinois University, ETRA Department, 208 Gabel, DeKalb, IL 60115, 815-753-7886. E-mail him at: dawalker@niu.edu.

The $r$ group can be considered as based on the correlation between treatment and result (Levin, 1994). For this group, "Effect size is generally reported as some proportion of the total variance accounted for by a given effect" (Stewart, 2000, p. 687). Cohen (1988) suggested that values of $.01, .06$, and .14 serve as indicators of small, medium, and large effect sizes for this group. However, it is at the discretion of the researcher to note the context in which small, medium, and large effects are being defined when using d and $r$ related indices.

A review of the literature indicated that researchers have discussed the merits, or lack thereof, of employing correlation coefficients, such as Kendall's tau ( $\tau$ ), to assist in conducting meta-analytic studies or other forms of correlational and/or experimental analyses (Cooper \& Hedges, 1994; Ferguson \& Takane, 1989; Gibbons, 1985; Gilpin, 1993; Roberts \& Kunst, 1990; Smithson, 2001; Wolf, 1987). Indeed, the reporting of tau in research studies has not been as prevalent as found with Spearman's rho ( $\rho$ ) or Pearson's $r$. However, tau has been emphasized recently as a substitute for $r$ in various research contexts. For instance, Rupinski and Dunlap (1996) looked at the accuracy of formulas for estimating $r$ from tau.

## Methodology

## Purpose

Tables for transforming correlation coefficients, such as Pearson's $r$ to Spearman's rho, have been produced in the recent past (Gilpin, 1993; Strahan, 1982). Expanding upon Gilpin and Strahan's research, this study will provide researchers with a detailed table for use in meta-analytic applications when engaged in assorted examinations of various $r$-related statistics, such as Fisher's $\mathrm{Z}_{r}$ and Cohen's d, that estimate the magnitude of experimental or observational effect. In addition, the table will be expanded to measure values in increments of .001 of a percent from .001 to 1.000 , add more commonly utilized effect size variants of $r$ not found in the original, and, most importantly, provide SPSS syntax to convert from the lesserused tau coefficient to other effect size indices when conducting correlational or meta-analytic analyses.

Assumptions
This research study is not intended to be an exhaustive study of effect sizes, but serves as a prologue to impart contextual reference to the internal matrix table being presented. Also, it is presupposed that researchers understand that tau and rho apply distinct metrics, which means that they cannot be likened to one another due to a great difference between their absolute values (cf. Kendall, 1970; Strahan, 1982). As noted by researchers (Kendall, 1970; Gilpin, 1993), as the values of $\tau$ and $\rho$ increase, their numerical similitude decreases greatly. The same trend holds for these two correlation coefficients' squared indices, where " $\tau{ }^{2}$ does not reflect at all adequately the proportion of shared variance..." (Strahan, 1982, p. 764).

A further assumption is that the table produced by means of this procedure reflects accurate values when a normal distribution is present, as well as a relatively large sample size. Also, the true values of the squared indices will be non-negative and the non-squared index values are symmetrical (i.e., $\tau<0$ ), thus remaining the same numerically when negative.

## Results

In the accompanying SPSS data set, we are given a value for tau ranging from . 001 to 1.000 . With a presented value of tau, the table that ensues can be created in total as an internal matrix via the SPSS syntax program provided in Appendix A or as individual, selected conditions by way of the subsequent formulas.

With a presented value of tau, we can calculate a Pearson's $r$ using Kendall's formula (1970, p. 126).

$$
\begin{equation*}
r=\sin (.5 \pi \tau) \tag{1}
\end{equation*}
$$

COMPUTE $r=$ SIN (3.141592654* $\tau^{*} .5$ ). EXECUTE.

Further, with a known value for $\tau$ and $r$, we can compute a Spearman's rho statistic using Gilpin's formula (1993, p. 91),

$$
\begin{equation*}
\rho=3\left[\tau \sin ^{-1}(r / 2)\right] / \sin ^{-1} r \tag{2}
\end{equation*}
$$

COMPUTE $p=3^{*} \tau^{*} \operatorname{ARSIN}(r / 2) / \operatorname{ARSIN}(r)$. EXECUTE.

To compute a Fisher's $\mathrm{Z}_{r}$ statistic from a given value of $r$, derived from tau, we can apply the following formula (Rosenthal, 1994, p. 237).

$$
\begin{equation*}
\mathrm{Z}_{r}=1 / 2 \log _{\mathrm{e}}[(1+r) /(1-r)] \tag{3}
\end{equation*}
$$

COMPUTE z = .5*LN((1+r)/(1-r)). EXECUTE.

To calculate a Cohen's d from a known $r$, derived from tau, the subsequent formula was employed (Rosenthal, 1994, p. 239). Note that for small to medium sample sizes, this formula will yield positively biased estimates. To correct for this, if presented with this situation, the expression should be multiplied by the factor $[(n-1) / n]^{5}$

$$
\begin{equation*}
d=2 r /\left[\left(1-r^{2}\right)^{5}\right] \tag{4}
\end{equation*}
$$

```
COMPUTE d = 2*r/SQRT(1-r**2).
EXECUTE.
```

To calculate a Cohen's f statistic from a given d, derived from tau, the ensuing formula was utilized (Cohen, 1988, p.276).

$$
\begin{equation*}
f=1 / 2 d \tag{5}
\end{equation*}
$$

```
COMPUTE f = .5*d.
EXECUTE.
```

To determine the amount of variance accounted for with correlation coefficients, such as $r$, $\rho$, or $\tau$, we square their value, which yields the extent of the effect in terms of "how much of the variability in the dependent variable(s) is associated with the variation in the independent variable(s)" (Snyder \& Lawson, 1993, p. 338). A caution should be noted, however, that when using squared indices to determine effect size, a loss of directionality is an issue and also the power affiliated with these indices is often distorted when reporting research findings (Rosenthal, 1994).

With $r$-related squared indices, it should be mentioned that eta-squared $\left(\eta^{2}\right)$ and $r^{2}$ are identical numerically and $r^{2}$ and $\mathrm{f}^{2}$ are related monotonically. Cohen (1994) determined that $\eta^{2}$ was a population correlation ratio that could be "... computed on samples and its population value estimated therefrom" (p. 281). Effect size estimates of this order have been called epsilonsquared $\left(\xi^{2}\right)$ and omega-squared ( $\omega^{2}$ ). Thus, this type of effect size tends to measure the proportion of variance in the population due to a particular effect. Cohen's formula for $\eta^{2}$ (1994, p. 281) can be used if $r^{2}$ is not preferred, where $\mathrm{f}^{2}=\mathrm{d}^{2} / 4$.

$$
\begin{equation*}
\eta^{2}=f^{2} /\left(1+\mathrm{f}^{2}\right) \tag{6}
\end{equation*}
$$

```
COMPUTE eta2 = f**2/(1+f**2).
EXECUTE.
```


## Conclusion

Methodological appropriateness is a consequential area within research that should be nearly perfect. Concepts such as effect size need to be
addressed correctly in a research study to make reliable, justifiable decisions and have these decisions, based on the statistics of study, authenticated by others. Converting from the lesser-used tau coefficient to other effect size indices, the presented SPSS syntax program can create an internal matrix table and new data set to assist researchers in determining the size of an effect for commonly utilized $r$-related indices when engaging in correlational and metaanalytic analyses.

## References

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

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

Cohen, J. (1994). The earth is round (p <.05). American Psychologist, 49, 997-1003.

Cooper, H., \& Hedges, L. V. (Eds.). (1994). The handbook of research synthesis. New York: Russell Sage Foundation.

Ferguson, G. A., \& Takane, Y. (1989). Statistical analysis in psychology and education (6th ed.). New York: McGraw-Hill.

Gibbons, J. D. (1985). Nonparametric methods for quantitative analysis (2nd ed.). Columbus, OH: American Sciences Press, Inc.

Gilpin, A. R. (1993). Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. Educational and Psychological Measurement, 53, 87-92.

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.

Kendall, M. G. (1970). Rank correlation methods (4th ed.). London: Charles Griffin \& Co.

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

Knapp, T. R. (1998). Comments on the statistical significance testing articles. Research in Schools, 5, 39-41.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. Educational Psychology Review, 6, 231-243.

Roberts, D. M., \& Kunst, R. E. (1990). A case against continuing use of the Spearman formula for rank-order correlation. Psychological Reports, 66, 339-349.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper \& L. V. Hedges (Eds.), The handbook of research synthesis (pp. 213-244). New York: Russell Sage Foundation.

Rupinski, M. T., \& Dunlap, W. P. (1996). Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. Educational and Psychological Measurement, 56, 419-429.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67, 57-60.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. Educational and Psychological Measurement, 61, 605-632.

Snyder, P., \& Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Stewart, D. W. (2000). Testing statistical significance testing: Some observations of an agnostic. Educational and Psychological Measurement, 60, 685-690.

Strahan, R. F. (1982). Assessing magnitude of effect from rank-order correlation coefficients. Educational and Psychological Measurement, 42, 763-765.

Wilkinson, L., \& The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594604.

Wolf, F. M. (1987). Meta-analysis: Quantitative methods for research synthesis. Beverly Hills, CA: Sage Publications.

## Appendix

Notes: To produce an internal matrix output and a complete table as a new data set in SPSS, access the embedded Tau Data Set and then run Tau Syntax. Create the data set in SPSS with a variable (1 column by 1000 rows) containing the values .001-1.000 (.001), or contact the author to obtain copies of the data set and the syntax. An example of the tabled values appears on the following page.

```
compute r = SIN(3.141592654 * t * .5).
compute rs = 3 * t * ARSIN(r / 2) / ARSIN(r).
compute zr = .5 * LN((1 + r) / (1 - r)).
compute d = 2 * r / SQRT(1 - r ** 2).
compute f = d*.5.
compute r2 = r **2.
compute f2 = d**2/4.
compute eta2 = (f2) / (1 + f2).
execute.
* FINAL REPORTS *.
FORMAT r to eta2 (f9.4).
VARIABLE LABELS t 'Tau' /r 'Pearsons r' /rs 'Spearmans Rank' /zr 'Fishers
z' /d 'Cohens d' /f 'f (Related to d as an SD of Standardized Means when k
= 2 and n = n)'/ r2 'R Square' / eta2 'Eta Square'.
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
    /VARIABLES=t r rs zr
    /TITLE "Proportion of Variance-Accounted-For Effect Sizes: Measures of
Relationship".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
    /VARIABLES=d f /TITLE "Standardized Mean Difference Effect Sizes".
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
    /VARIABLES= r2 eta2
    /TITLE "Proportion of Variance-Accounted-For Effect Sizes: Squared
Indices".
```

An Example Of Tabled Values

| tau | r | p | Zr | d | f | R2 | eta2 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.001 | 0.0016 | 0.0015 | 0.0016 | 0.0031 | 0.0016 | 0.0000 | 0.0000 |
| 0.002 | 0.0031 | 0.0030 | 0.0031 | 0.0063 | 0.0031 | 0.0000 | 0.0000 |
| 0.003 | 0.0047 | 0.0045 | 0.0047 | 0.0094 | 0.0047 | 0.0000 | 0.0000 |
| 0.004 | 0.0063 | 0.0060 | 0.0063 | 0.0126 | 0.0063 | 0.0000 | 0.0000 |
| 0.005 | 0.0079 | 0.0075 | 0.0079 | 0.0157 | 0.0079 | 0.0001 | 0.0001 |
| 0.006 | 0.0094 | 0.0090 | 0.0094 | 0.0189 | 0.0094 | 0.0001 | 0.0001 |
| 0.007 | 0.0110 | 0.0105 | 0.0110 | 0.0220 | 0.0110 | 0.0001 | 0.0001 |
| 0.008 | 0.0126 | 0.0120 | 0.0126 | 0.0251 | 0.0126 | 0.0002 | 0.0002 |
| 0.009 | 0.0141 | 0.0135 | 0.0141 | 0.0283 | 0.0141 | 0.0002 | 0.0002 |
| 0.010 | 0.0157 | 0.0150 | 0.0157 | 0.0314 | 0.0157 | 0.0002 | 0.0002 |
| 0.011 | 0.0173 | 0.0165 | 0.0173 | 0.0346 | 0.0173 | 0.0003 | 0.0003 |
| 0.012 | 0.0188 | 0.0180 | 0.0189 | 0.0377 | 0.0189 | 0.0004 | 0.0004 |
| 0.013 | 0.0204 | 0.0195 | 0.0204 | 0.0408 | 0.0204 | 0.0004 | 0.0004 |
| 0.014 | 0.0220 | 0.0210 | 0.0220 | 0.0440 | 0.0220 | 0.0005 | 0.0005 |
| 0.015 | 0.0236 | 0.0225 | 0.0236 | 0.0471 | 0.0236 | 0.0006 | 0.0006 |
| 0.016 | 0.0251 | 0.0240 | 0.0251 | 0.0503 | 0.0251 | 0.0006 | 0.0006 |
| 0.017 | 0.0267 | 0.0255 | 0.0267 | 0.0534 | 0.0267 | 0.0007 | 0.0007 |
| 0.018 | 0.0283 | 0.0270 | 0.0283 | 0.0566 | 0.0283 | 0.0008 | 0.0008 |
| 0.019 | 0.0298 | 0.0285 | 0.0298 | 0.0597 | 0.0299 | 0.0009 | 0.0009 |
| 0.020 | 0.0314 | 0.0300 | 0.0314 | 0.0629 | 0.0314 | 0.0010 | 0.0010 |
| 0.021 | 0.0330 | 0.0315 | 0.0330 | 0.0660 | 0.0330 | 0.0011 | 0.0011 |
| 0.022 | 0.0346 | 0.0330 | 0.0346 | 0.0691 | 0.0346 | 0.0012 | 0.0012 |
| 0.023 | 0.0361 | 0.0345 | 0.0361 | 0.0723 | 0.0361 | 0.0013 | 0.0013 |
| 0.024 | 0.0377 | 0.0360 | 0.0377 | 0.0754 | 0.0377 | 0.0014 | 0.0014 |
| 0.025 | 0.0393 | 0.0375 | 0.0393 | 0.0786 | 0.0393 | 0.0015 | 0.0015 |
| 0.026 | 0.0408 | 0.0390 | 0.0409 | 0.0817 | 0.0409 | 0.0017 | 0.0017 |
| 0.027 | 0.0424 | 0.0405 | 0.0424 | 0.0849 | 0.0424 | 0.0018 | 0.0018 |
| 0.028 | 0.0440 | 0.0420 | 0.0440 | 0.0880 | 0.0440 | 0.0019 | 0.0019 |
| 0.029 | 0.0455 | 0.0435 | 0.0456 | 0.0912 | 0.0456 | 0.0021 | 0.0021 |
| 0.030 | 0.0471 | 0.0450 | 0.0471 | 0.0943 | 0.0472 | 0.0022 | 0.0022 |
| 0.031 | 0.0487 | 0.0465 | 0.0487 | 0.0975 | 0.0487 | 0.0024 | 0.0024 |
| 0.032 | 0.0502 | 0.0480 | 0.0503 | 0.1006 | 0.0503 | 0.0025 | 0.0025 |
| 0.033 | 0.0518 | 0.0495 | 0.0519 | 0.1038 | 0.0519 | 0.0027 | 0.0027 |
| 0.034 | 0.0534 | 0.0510 | 0.0534 | 0.1069 | 0.0535 | 0.0028 | 0.0028 |
| 0.035 | 0.0550 | 0.0525 | 0.0550 | 0.1101 | 0.0550 | 0.0030 | 0.0030 |
| 0.036 | 0.0565 | 0.0540 | 0.0566 | 0.1132 | 0.0566 | 0.0032 | 0.0032 |
| 0.037 | 0.0581 | 0.0555 | 0.0582 | 0.1164 | 0.0582 | 0.0034 | 0.0034 |
| 0.038 | 0.0597 | 0.0570 | 0.0597 | 0.1195 | 0.0598 | 0.0036 | 0.0036 |
| 0.039 | 0.0612 | 0.0585 | 0.0613 | 0.1227 | 0.0613 | 0.0037 | 0.0037 |
| 0.040 | 0.0628 | 0.0600 | 0.0629 | 0.1258 | 0.0629 | 0.0039 | 0.0039 |
| 0.041 | 0.0644 | 0.0615 | 0.0644 | 0.1290 | 0.0645 | 0.0041 | 0.0041 |
| 0.042 | 0.0659 | 0.0630 | 0.0660 | 0.1321 | 0.0661 | 0.0043 | 0.0043 |
| 0.043 | 0.0675 | 0.0645 | 0.0676 | 0.1353 | 0.0676 | 0.0046 | 0.0046 |
| 0.044 | 0.0691 | 0.0660 | 0.0692 | 0.1385 | 0.0692 | 0.0048 | 0.0048 |
| 0.045 | 0.0706 | 0.0675 | 0.0707 | 0.1416 | 0.0708 | 0.0050 | 0.0050 |
| 0.046 | 0.0722 | 0.0690 | 0.0723 | 0.1448 | 0.0724 | 0.0052 | 0.0052 |
| 0.047 | 0.0738 | 0.0705 | 0.0739 | 0.1479 | 0.0740 | 0.0054 | 0.0054 |
| 0.048 | 0.0753 | 0.0719 | 0.0755 | 0.1511 | 0.0755 | 0.0057 | 0.0057 |
| 0.049 | 0.0769 | 0.0734 | 0.0770 | 0.1542 | 0.0771 | 0.0059 | 0.0059 |
| 0.050 | 0.0785 | 0.0749 | 0.0786 | 0.1574 | 0.0787 | 0.0062 | 0.0062 |

# Letters To The Editor 

Additional Reflections On Significance Testing

Knapp, T.R. (2002). Some reflections on significance testing. Journal of Modern Applied Statistical Methods, 1(2), 240-242.

Knapp (2002) raised good points concerning significance testing. The bottom line, "If you have hypotheses to test ... [then] test them", not only makes sense, but in fact is an argument that I have often made when consulting, although the closest I have come to this issue in the literature is an allusion (Berger, 2000, Section 2.1).

Yet, the support for this assertion, based on refuting the statement that confidence intervals provide identical or non-conflicting inferences with significance tests, might benefit from elaboration. In fact, the confidence interval constructed by Knapp (2002) is but one of several that could have been constructed. To argue that it is not the best among these is to argue that its discredit cannot serve as a simultaneous discredit to the class it purports to represent (albeit not very well).

It would be a simple matter to construct a confidence region as precisely the set of parameter values which, when serving as the null hypothesis, lead to significance tests that cannot be rejected. With this definition of a confidence set (which often, but not always, reduces to a confidence interval), it is a tautology that the confidence set cannot contradict the results of the significance test.

Why, then, should hypotheses be tested? Because it is problematic to base policy decisions, that affect the public, on apparent directions of effect when the study is conducted by a party with a vested interest in the outcome. Requiring statistical significance is one reasonable way to operationalize the need for a preponderance of evidence, and raise the hurdle, in such a case. If an alpha level were chosen strategically, perhaps based on safety, convenience, and cost in a medical study, then the results of the significance test of efficacy would correspond to the optimal decision. Clearly, there are other ways to raise the hurdle.

Vance W. Berger, Biometry Research Group, National Cancer Institute. E-mail: vb78c@nih.gov.

## Reference

Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials. Statistics in Medicine 19, 1319-1328.

Predictor Importance In Multiple Regression
Whittaker, T.A., Fouladi, R.T., \& Williams, N.J. (2002). Determining predictor importance in multiple regression under varied correlational and distributional conditions. Journal of Modern Applied Statistical Methods, 1(2), 354-366.

William Kruskal may have been right in noting that the relative importance of predictors in a regression analysis is meaningful to researchers, but I'm not so sure it should always be the case.

My principal concerns about the Whittaker et al. (2002) article are these:

1. Multiple regression analysis is used for prediction and for causal analysis. When a user asks: "What are the most important variables in this regression?", the answer depends upon the purpose of the analysis. Whittaker et al. failed to distinguish sufficiently between the two purposes and seem to advocate a "one size fits all" method for determining the relative importance of regressors (apparently Budescu's dominance analysis, perhaps augmented by the Johnson index).
2. I do not see the need for the Monte Carlo approach to the problem. In his text, Darlington (1990) provided a mathematical explanation for the equivalence with respect to rank-ordering of importance of their Methods 3 (the $t$ statistics for the betas), 5 (the squared partials) and 6 (the squared semi-partials), along with two others (the p -values for the betas and the changes in Rsquare from the reduced model with the variable deleted to the full model with the variable
included). [They do cite Darlington (1968), but do not cite his later text.] On page 364 of their article they speculated as to why Methods 3, 5 , and 6 "all performed identically". Those three methods must perform identically.

As far as the other five methods are concerned, Method 1 (squared zero-order correlations, or unsquared zero-order correlations, for that matter) can be dismissed out of hand, because the other regressors are not statistically controlled.

I can't see any reason why anyone would ever use Method 2 (the betas). For one thing, the betas aren't restricted to the -1 to +1 range, so although they are standardized they are awkward to compare.

Method 4 (the beta-times-r products) has been criticized in the past (see, for example, Darlington, 1968). The fact that those products sum to the over-all R-square is a poor basis for variance partitioning and for the determination of relative importance (some of those products can even be negative--for suppressor variables).

That leaves Methods 7 (Budescu) and 8 (Johnson). I have no doubt that similar, nonMonte Carlo-based, arguments could be made regarding those methods for determining the relative importance of regressors, but even if such arguments were necessary, Whittaker et al. (2002) were interested in comparing all eight methods, not just those two.
3. Two different meanings of the word "dominance" was confusing. One of the meanings, "dominance analysis", is associated with Budescu's method. The other meaning, identifying the "dominant predictor" (p. 358), was apparently the criterion for determining which methods were best. When I first read the article I thought that the Budescu method was
used as one of the methods AND as the goodness criterion, which would of course "stack the deck" in its favor. The confusion with the two meanings, however, remains.

I have a couple of other lesser concerns:

1. Their definition of "the dominant predictor" is a bit strange. In what sense is an independent variable that correlates .40 - . 60 with the dependent variable dominant over other independent variables that correlate .30 with the dependent variable?
2. Their "Nursing Facility Consumer Satisfaction Survey" example is a poor example. The data are for seven-point Likert-type scales with ridiculously high means and there is an inherent regressor/regressand contamination since the three predictors are concerned with specific satisfactions and the dependent variable is over-all satisfaction.

Thomas R. Knapp, Professor Emeritus, University of Rochester \& The Ohio State University.

Note: I would like to thank Richard Darlington for his very helpful suggestions regarding an earlier version of this critique.

## References

Darlington, R.B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69(3), 161-182.

Darlington, R.B. (1990). Regression and linear models. New York: McGraw-Hill.

## Statistical Pronouncements II

"There are some who appear to pride themselves on their absence of knowledge of mathematics. I never understood why it should be a matter of pride" - Arthur L. Bowley, (1934, Discussion, The Journal of the Royal Statistical Society, 97, p. 607).
"Do we know more than was known to Todhunter?" - Arthur L. Bowley, (ibid, p. 609).
"To try and state mathematics without either chalk or with a minimum of chalk [is] perhaps a hopeless task" - L. Isserlis (ibid, p. 614).
"The advantage of excluding by severe mathematical requirements many quacks is bought too dear if it shuts out a single John Graunt" - Major Greenwood (1939, Journal of the Royal Statistical Society, 102, p. 552).
"Many important applications of statistics, while employing elementary statistical techniques, demand thorough knowledge and long experience in the applied field" - William G. Cochran (1945, Training at the professional level for statistical work in agriculture and biology, Journal of the American Statistical Association, 40, p. 163).
"Youth is the time to learn mathematics" - William G. Cochran (1946, Graduate training in statistics, American Mathematics Monthly, 53(4), p. 199).
"Statistics depends primarily on mathematics and mathematicians for its future development...Such mathematicians need not be regarded as lost or strayed from the fold" William G. Cochran (ibid, p. 199).
"The missing link is that we do not know which of the theoretical non-normal distributions that have been studied are typical of the error distributions that turn up in practice" - William G. Cochran (1947, Some consequences when the assumptions for the analysis of variance are not satisfied, Biometrics, 3(1), p. 25).
"No valid sampling error formula exists unless the selection of the sample [is] through the use of an objective method of randomization" - William G. Cochran (1947, Recent developments in sampling theory in the United States, Proceedings of the International Statistical Institute, 3(A), p. 41).
"The student should be warned that he cannot expect miracles to be wrought by the use of statistical tools" - Quinn McNemar (1949, Psychological statistics, Wiley, p. 3.)
"Much ingenuity is shown by investigators in concocting possible explanations of the discrepancies among the results of different workers" - William G. Cochran (1950, The present status of biometry, Bulletin of the International Statistical Institute, 32(2), p. 133).
"Any estimate made from a sample is subject to error" - William G. Cochran (1951, Modern methods in the sampling of human populations, American Journal of Public Health, 41(6), p. 647).
"The principle that governs modern sampling practice is the familiar economic maxim that one should get the most for one's money" - William G. Cochran (ibid, p. 648).
"Any theory is at best approximately true, but nevertheless, if we are going to reject a theory, we do so because it does not fit the data we have, not because it would not fit a much larger sample of data that we do not have" William G. Cochran (ibid, p. 336).
"Sampling is all too often taken far too lightly" - William G. Cochran, (1954, Principles of sampling, Journal of the American Statistical Association, 49, p. 13).
"[There are] unwarranted shotgun marriages between the quantitatively unsophisticated idea of sample as 'what you get by grabbing a handful' and the mathematically precise notion of a "simple random sample"" William G. Cochran, (ibid, p. 13).
"The scientist tends to think of seeing a statistician when he has some problem... mostly when something had gone wrong with the experiment or survey... As a result, statisticians... see a sorry collection of the wrecks of research projects" - William G. Cochran, (1955, Research techniques in the study of human beings, Milbank Memorial Fund Quarterly, 33(2), p. 122).
"In statistical training centers, something is done to teach young statisticians how to get along with scientists" - William G. Cochran (ibid, p. 123).
"The statistician is a poor marriage risk, and may be suffering from marital strain" William G. Cochran (ibid, p. 124).
"The ability to do experiments is one of the most powerful weapons man has for making advances in his understanding of the world" William G. Cochran (1957, The philosophy underlying the design of experiments, Proceedings of the $1^{\text {st }}$ Conference on the Design of Experiments in Army Research, Development and Testing, p 1).
"All mathematical methods are oversimplifications" - William G. Cochran (1961, The role of mathematics in the medical sciences, New England Journal of Medicine, 265, p. 176).
"Many of the standard results in theoretical statistics were obtained without encountering really difficult mathematics" William G. Cochran (ibid, p. 230).
"Electronic machines... can free us from overdependence on the assumption of normality and from confinement to approximate linear solutions to nonlinear problems" - William G. Cochran (ibid, p. 232-232.)
"Nonparametric theory is elegant" Jaroslav Hajek (1969. A course in nonparametric statistics, Holden-Day, preface.)
"As regard the rejection of observations [as outliers], I distrust any slick formal rule" David J. Finney (1970, Discussion, Statistics in endocrinology, MIT Press, p. 72).
"Time is perhaps the most mysterious thing in a mysterious universe." - Maurice Kendall (1976, Time-series, p. 1).
"The modern theory of statistics began with the realization that, although individuals might not behave deterministically, aggregates of individuals were themselves subjects to laws which could often be summarized in fairly simple mathematical terms" - Maurice Kendall (ibid, p. 4).
"Some data are not worth analyzing, even when we have the big guns of a mathematical arsenal ready for attack" attributed to William G. Cochran by Frederick Mosteller in the Forward to Contribution to Statistics: William G. Cochran, Wiley, 1982, p. vii).
"There were just two qualifications for membership [in the national statistics society] first you had to have $\$ 5$ [and] second you had to be willing to give it to the society" - attributed to William G. Cochran by Frederick Mosteller (ibid, p. xi).
"You usually can't follow papers at meetings" - Frederick Mosteller (Contribution to Statistics: William G. Cochran, 1982, p. xii).
"It has been said that more time has been spent generating and testing random numbers than using them" - C. A. Whitney (1984, Generating and testing pseudo-random numbers, BYTE, October, 9(11), p. 128).
"Someone told me that each equation I included in the book would halve the sales" Stephen Hawkins (1988, A brief history of time: From the big bang to black holes, Bantam, p. vi.).

## NEW IN 2004

## significance

## The new magazine of the Royal Statistical Society

Edited by Helen Joyce
Significance is a new quarterly magazine for anyone interested in statistics and the analysis and interpretation of data. It aims to communicate and demonstrate, in an entertaining and thought-provoking way, the practical use of statistics in all walks of life and to show how statistics benefit society.

Articles are largely non-technical and hence accessible and appealing, not only to members of the profession, but to all users of statistics.

As well as promoting the discipline and covering topics of professional relevance, Significance contains a mixture of statistics in the news, casestudies, reviews of existing and newly developing areas of statistics, the application of techniques in practice and problem solving, all with an international flavour.

## Special Introductory Offer: 25\% discount on a new personal subscription Plus Great Discounts for Students!



Further information including submission guidelines, subscription information and details of how to obtain a free sample copy are available at

## Power Analysis and Sample Size Software from NCSS

$\boldsymbol{P A S S}$ performs power analysis and calculates sample sizes. Use it before you begin a study to calculate an appropriate sample size (it meets the requirements of government agencies that want technical justification of the sample size you have used). Use it after a study to determine if your sample size was large enough. PASS calculates the sample sizes necessary to perform all of the statistical tests listed below.

A power analysis usually involves several "what if" questions. PASS lets you solve for power, sample size, effect size, and alpha level. It automatically creates appropriate tables and charts of the results.
PASS is accurate. It has been extensively verified using books and reference articles. Proof of the accuracy of each procedure is included in the extensive documentation.

PASS is a standalone system. Although it is integrated with NCSS, you do not have to own NCSS to run it. You can use it with any statistical software you want.


PASS comes with two manuals that contain tutorials, examples, annotated output, references, formulas, verification, and complete instructions on each procedure. And, if you cannot find an answer in the manual, our free technical support staff (which includes a PhD statistician) is available.

## System Requirements

PASS runs on Windows 95/98/ME/NT/ 2000/XP with at least 32 megs of RAM and 30 megs of hard disk space.

PASS sells for as little as $\mathbf{\$ 4 4 9 . 9 5}$.

## Analysis of Variance

Factorial AOV
Fixed Effects AOV
Geisser-Greenhouse
MANOVA*
Multiple Comparisons*
One-Way AOV
Planned Comparisons
Randomized Block AOV
New Repeated Measures AOV*
Regression / Correlation
Correlations (one or two)
Cox Regression*
Logistic Regression
Multiple Regression
Poisson Regression*
Intraclass Correlation
Linear Regression

## Proportions

Chi-Square Test
Confidence Interval
Equivalence of McNemar* Equivalence of Proportions Fisher's Exact Test Group Sequential Proportions Matched Case-Control McNemar Test
Odds Ratio Estimator One-Stage Designs*
Proportions - 1 or 2
Two Stage Designs (Simon's)
Three-Stage Designs*
Miscellaneous Tests
Exponential Means - 1 or 2*
ROC Curves -1 or $2^{*}$
Variances - 1 or 2

T Tests
Cluster Randomization
Confidence Intervals
Equivalence T Tests
Hotelling's T-Squared*
Group Sequential T Tests
Mann-Whitney Test
One-Sample T-Tests
Paired T-Tests
Standard Deviation Estimator
Two-Sample T-Tests Wilcoxon Test
Survival Analysis
Cox Regression*
Logrank Survival -Simple Logrank Survival - Advanced* Group Sequential - Survival Post-Marketing Surveillance ROC Curves -1 or $2^{*}$

## PASS Beats the Competition!

 No other program calculates sample sizes and power for as many different statistical procedures as does PASS. Specifying your input is easy, especially with the online help and manual.PASS automatically displays charts and graphs along with numeric tables and text summaries in a portable format that is cut and paste compatible with all word processors so you can easily include the results in your proposal.
Choose PASS. It's more comprehensive, easier-to-use, accurate, and less expensive than any other sample size program on the market.

## Trial Copy Available

You can try out PASS by downloading it from our website. This trial copy is good for 30 days. We are sure you will agree that it is the easiest and most comprehensive power analysis and sample size program available.

## PASS 2002 adds power analysis and sample size to your statistical toolbox

## WHAT'S NEW IN PASS 2002?

Thirteen new procedures have been added to PASS as well as a new home-base window and a new Guide Me facility.

## MANY NEW PROCEDURES

The new procedures include a new multifactor repeated measures program that includes multivariate tests, Cox proportional hazards regression, Poisson regression, MANOVA, equivalence testing when proportions are correlated, multiple comparisons, ROC curves, and Hotelling's T-squared.

## TEXT STATEMENTS

The text output translates the numeric output into easy-to-understand sentences. These statements may be transferred directly into your grant proposals and reports.

## GRAPHICS

The creation of charts and graphs is easy in PASS. These charts are easily transferred into other programs such as MS PowerPoint and MS Word.

## NEW USER'S GUIDE II

A new, 250-page manual describes each new procedure in detail. Each chapter contains explanations, formulas, examples, and accuracy verification.

The complete manual is stored in PDF format on the CD so that you can read and printout your own copy.
GUIDE ME
The new Guide Me facility makes it easy for first time users to enter parameter values.
The program literally steps you through those options that are necessary for the sample size calculation.

## NEW HOME BASE

A new home base window has been added just for PASS users. This window helps you select the appropriate program module.
COX REGRESSION
A new Cox regression procedure has been added to perform power analysis and sample size calculation for this important statistical technique.

## REPEATED MEASURES

A new repeated-measures analysis module has been added that lets you analyze designs with up to three grouping factors and up to three repeated factors. The analysis includes both the univariate F test and three common multivariate tests including Wilks Lambda.

## RECENT REVIEW

In a recent review, 17 of 19 reviewers selected $\boldsymbol{P A S S}$ as the program they would recommend to their colleagues.

## PASS calculates sample sizes for...

Please rush me my own personal license of PASS 2002. Qty

PASS 2002 Deluxe (CD and User's Guide): $\$ 499.95$ . $\qquad$ PASS 2002 CD (electronic documentation): \$449.95 .......... \$ _____

PASS 2002 5-User Pack (CD \& 5 licenses): \$1495.00........ \$ $\qquad$
PASS 2002 25-User Pack (CD \& 25 licenses): \$3995.00.... \$ $\qquad$
PASS 2002 User's Guide II (printed manual): \$30.00......... \$ $\qquad$
PASS 2002 Upgrade CD for PASS 2000 users: $\$ 149.95$....... \$ $\qquad$
Typical Shipping \& Handling: USA: \$9 regular, \$22 2-day, \$33 $\qquad$
overnight. Canada: \$19 Mail. Europe: \$50 Fedex....................... \$ $\qquad$
FOR FASTEST DELIVERY, ORDER ONLINE AT WWW.NCSS.COM
Email your order to sales@ncss.com Fax your order to (801) 546-3907
NCSS, 329 North 1000 East, Kaysville, UT 84037 (800) 898-6109 or (801) 546-0445

COMPANY

ADDRESS

## My Payment Options:

___ Check enclosed
___ Please charge my: __VISA __MasterCard __Amex __ Purchase order enclosed

Card Number
Expires $\qquad$
Signature
Please provide daytime phone:
( )
Ship my PASS 2002 to:

NAME

CITY/STATE/ZIP
COUNTRY (IF OTHER THAN U.S.)

## Announcing NCSS 2004

## Seventeen New Procedures

NCSS 2004 is a new edition of our popular statistical NCSS package that adds seventeen new procedures.

New Procedures
Two Independent Proportions
Two Correlated Proportions
One-Sample Binary Diagnostic Tests
Two-Sample Binary Diagnostic Tests
Paired-Sample Binary Diagnostic Tests
Cluster Sample Binary Diagnostic Tests
Meta-Analysis of Proportions
Meta-Analysis of Correlated Proportions
Meta-Analysis of Means
Meta-Analysis of Hazard Ratios
Curve Fitting
Tolerance Intervals
Comparative Histograms
ROC Curves
Elapsed Time Calculator
T-Test from Means and SD's
Hybrid Appraisal (Feedback) Model

## Documentation

The printed, 330-page manual, called NCSS User's Guide V, is available for $\$ 29.95$. An electronic (pdf) version of the manual is included on the distribution CD and in the Help system.

## Two Proportions

Several new exact and asymptotic techniques were added for hypothesis testing (null, noninferiority, equivalence) and calculating confidence intervals for the difference, ratio, and odds ratio. Designs may be independent or paired. Methods include: Farrington \& Manning, Gart \& Nam, Conditional \& Unconditional Exact, Wilson's Score, Miettinen \& Nurminen, and Chen.

## Meta-Analysis

Procedures for combining studies measuring paired proportions, means, independent proportions, and hazard ratios are available. Plots include the forest plot, radial plot, and L'Abbe plot. Both fixed and random effects models are available for combining the results.

## Curve Fitting

This procedure combines several of our curve fitting programs into one module. It adds many new models such as Michaelis-Menten. It analyzes curves from several groups. It compares fitted models across groups using computerintensive randomization tests. It computes bootstrap confidence intervals.

## Tolerance Intervals

This procedure calculates one and two sided tolerance intervals using both distribution-free (nonparametric) methods and normal distribution (parametric) methods. Tolerance intervals are bounds between which a given percentage of a population falls.

## Comparative Histogram

This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups.

## Random Number Generator Matsumoto's Mersenne Twister random number generator (cycle length > $10 * * 6000)$ has been implemented.

## Binary Diagnostic Tests

Four new procedures provide the specialized analysis necessary for diagnostic testing with binary outcome data. These provide appropriate specificity and sensitivity output. Four experimental designs can be analyzed including independent or paired groups, comparison with a gold standard, and cluster randomized.

## ROC Curves

This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUC's and partial AUC's for paired and independent sample designs.

## Hybrid (Feedback) Model

This new edition of our hybrid appraisal model fitting program includes several new optimization methods for calibrating parameters including a new genetic algorithm. Model specification is easier. Binary variables are automatically generated from class variables.

> Statistical Innovations Products
> Through a special arrangement with Statistical Innovations (S.I.), NCSS customers will receive $\$ 100$ discounts on: Latent GOLD® - latent class modeling SI-CHAID® - segmentation trees GOLDMineR® - ordinal regression For demos and other info visit: www.statisticalinnovations.com

Please rush me the following products:
Qty
__ NCSS 2004 CD upgrade from NCSS 2001, \$149.95 $\qquad$ \$
$\$$
__ NCSS 2004 User's Guide V, \$29.95
\$ $\qquad$
NCSS 2004 CD, upgrade from earlier versions, \$249.95 $\qquad$
NCSS 2004 Deluxe (CD and Printed Manuals), \$599.95 ........... \$
\$ $\qquad$
Latent Gold® from S.I., \$995-\$100 NCSS Discount = \$895..... \$ $\qquad$
GoldMineR® from S.I., \$695-\$100 NCSS Discount = \$595 ..... \$ $\qquad$
CHAID® Plus from S.I., $\$ 695-\$ 100$ NCSS Discount $=\$ 595 \ldots . . \$$
\$

Approximate shipping--depends on which manuals are ordered (U.S: \$10 ground, \$18 2-day, or \$33 overnight) (Canada \$24) (All other countries \$10) (Add \$5 U.S. or \$40 International for any S.I. product) ........ \$ $\$$

Total. . $\qquad$
TO PLACE YOUR ORDER
CALL: (800) 898-6109 FAX: (801) 546-3907
ONLINE: www.ncss.com
MAIL: NCSS, 329 North 1000 East, Kaysville, UT 84037
My Payment Option:
Check enclosed
Please charge my: __VISA __ MasterCard

$\qquad$
Purchase order attached
$\qquad$ Exp $\qquad$
Signature

## Telephone:

( ) $\qquad$
Email:
Ship to:

NAME $\qquad$
ADDRESS $\qquad$
ADDRESS $\qquad$

ADDRESS $\qquad$
CITY $\qquad$ STATE

ZIP/POSTAL CODE COUNTRY $\qquad$


## Statistical and Graphics Procedures Available in NCSS 2004

Analysis of Variance / T-Tests
Analysis of Covariance
Analysis of Variance
Barlett Variance Test
Crossover Design Analysis
Factorial Design Analysis
Friedman Test
Geiser-Greenhouse Correction
General Linear Models
Mann-Whitney Test
MANOVA
Multiple Comparison Tests
One-Way ANOVA
Paired T-Tests
Power Calculations
Repeated Measures ANOVA
T-Tests - One or Two Groups
T-Tests - From Means \& SD's
Wilcoxon Test
Time Series Analysis
ARIMA / Box - Jenkins
Decomposition
Exponential Smoothing
Harmonic Analysis
Holt - Winters
Seasonal Analysis
Spectral Analysis
Trend Analysis

| Regression / Correlation | Survival / Reliability |
| :--- | :--- |
| All-Possible Search | Accelerated Life Tests |
| Canonical Correlation | Cox Regression |
| Correlation Matrices | Cumulative Incidence |
| Cox Regression | Exponential Fitting |
| Kendall's Tau Correlation | Extreme-Value Fitting |
| Linear Regression | Hazard Rates |
| Logistic Regression | Kaplan-Meier Curves |
| Multiple Regression | Life-Table Analysis |
| Nonlinear Regression | Lognormal Fitting |
| PC Regression | Log-Rank Tests |
| Poisson Regression | Probit Analysis |
| Response-Surface | Proportional-Hazards |
| Ridge Regression | Reliability Analysis |
| Robust Regression | Survival Distributions |
| Stepwise Regression | Time Calculator* |
| Spearman Correlation | Weibull Analysis |
| Variable Selection |  |
|  |  |
| Quality Control | Multivariate Analysis |
| Xbar-R Chart | Cluster Analysis |
| C, P, NP, U Charts | Correspondence Analysis |
| Capability Analysis | Discriminant Analysis |
| Cusum, EWMA Chart | Factor Analysis |
| Individuals Chart | Hotelling's T-Squared |
| Moving Average Chart | Item Analysis |
| Pareto Chart | Item Response Analysis |
| R \& R Studies | Loglinear Models |
|  | MANOVA |
|  | Multi-Way Tables |
|  | Multidimensional Scaling |
|  | Principal Components |

Plots / Graphs
Bar Charts
Box Plots
Contour Plot
Dot Plots
Error Bar Charts
Histograms
Histograms: Combined*
Percentile Plots
Pie Charts
Probability Plots
ROC Curves*
Scatter Plots
Scatter Plot Matrix
Surface Plots
Violin Plots
Experimental Designs
Balanced Inc. Block
Box-Behnken
Central Composite
D-Optimal Designs
Fractional Factorial
Latin Squares
Placket-Burman
Response Surface
Screening
Taguchi

Curve Fitting
Bootstrap C.I.'s*
Built-In Models
Group Fitting and Testing*
Model Searching
Nonlinear Regression
Randomization Tests*
Ratio of Polynomials
User-Specified Models

Miscellaneous
Area Under Curve
Bootstrapping
Chi-Square Test
Confidence Limits
Cross Tabulation
Data Screening
Fisher's Exact Test
Frequency Distributions
Mantel-Haenszel Test
Nonparametric Tests
Normality Tests
Probability Calculator Proportion Tests
Randomization Tests
Tables of Means, Etc.
Trimmed Means
Univariate Statistics

Meta-Analysis*
Independent Proportions*
Correlated Proportions*
Hazard Ratios*
Means*
Binary Diagnostic Tests*
One Sample*
Two Samples*
Paired Samples*
Clustered Samples*
Proportions
Tolerance Intervals*
Two Independent*
Two Correlated*
Exact Tests*
Exact Confidence Intervals*
Farrington-Manning*
Fisher Exact Test
Gart-Nam* Method
McNemar Test
Miettinen-Nurminen ${ }^{\star}$
Wilson's Score* Method Equivalence Tests*
Noninferiority Tests*
Mass Appraisal
Comparables Reports
Hybrid (Feedback) Model*
Nonlinear Regression
Sales Ratios

## Qualitative research has come a long way...

## from this...



## NS

 Wryive THE LATEST PRODUCTS HAVE ARRIVED www.gsrinternational.comWORLD LEADING PRCOUCTS FROM THE NUD'IST LINE CF SOFTNQRE

## Read more about QSR software in this edition of JMASM.

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."

- Antoine de Saint Exupery

F is a carefully crafted subset of the most recent version of Fortran, the world's most powerful numeric language.


Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- $F$ is easier to use than other popular programming languages
- F compilers are free and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran-modules and data abstraction, for example-but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.
F encourages Module-Oriented Programming.
It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.
The Fortran Company
11155 E. Mountain Gate Place, Tucson, AZ 85749 USA $+1-520-256-1455+1-520-760-1397$ (fax) http://www.fortran.com info@fortran.com

## GGUM2000

## Item Response Theory Models for Unfolding



The GGUM2000 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2000 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected a posteriori (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. The software is accompanied by a detailed user's manual. GGUM2000 is free and can be downloaded from:

## http://www.education.umd.edu/EDMS/tutorials

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2000 today!


## Are you involved in Data Modeling or Data Mining?

## Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology (www.digarch.com). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent $60 \%$ $80 \%$ of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

## Visit our website and see for yourself! >>>> www.digarch.com

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others








## Announcing the highly-anticipated new Numerical Recipes products

## Numerical Recipes in C++

The Art of Scientific Computing Second Edition
William H. Press, Saul A. Teukolsky, William T.Vetterling, and Brian P. Flannery
"This monumental and classic work is beautifully produced and of literary as well as mathematical quality. It is an essential component of any serious scientific or engineering library."

## -Computing Reviews

This new version incorporates completely new $\mathrm{C}++$ versions of the more than 300 Numerical Recipes Second Edition routines widely recognized as the most accessible and practical basis for scientific computing, in addition to including the full mathematical and explanatory contents of Numerical Recipes in C.

## Key Features:

- Includes linear algebra, interpolation, special functions, random numbers, nonlinear sets of equations, optimization, eigensystems, Fourier methods and wavelets, statistical tests, ODEs and PDEs, integral equations, and inverse theory.
- The routines, in ANSI/ISO C++ source code, can be used with almost any existing $\mathrm{C}++$ vector/ matrix class library, according to user preference
0-521-75033-4, Hardback, $\$ 70.00$
Visit us.cambridge.org/numericalrecipes for more intormation on the complete line of Numerical Recipes products.

Available in bookstores or from

us.cambridge.org/mathematics

## Other new Numerical Recipes products for your library...

Numerical Recipes Example Book $[\mathrm{C}++]$
$0-521-75034-2$, Paperback, $\$ 35.00$
Numerical Recipes in C and C++ Source Code CDROM with Windows, DOS, or Macintosh Single Screen License
0-521-75037-7, CD-ROM, $\$ 50.00$
Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License
Source Code for Numerical Recipes in C, C ++ , Fortran 77 , Fortran 90, Pascal, BASIC, lisp and Modula 2 plus many extras $0.521-75036-9$, CD-ROM, $\$ 150.00$

Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License
Source Code for Numerical Recipes in C, C ++ Fortran 77 , Fortran 90 , Pascal, BASIC, Lisp and Modula 2 plus many extras $0-521-75035-0$, CD-ROM, $\$ 90.00$

## Numerical Recipes in Fortran from Cambridge University Press

## Numerical Recipes in Fortran 77

Volume 1 of Fortran Numerical Recipes
Second Edition
William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery
"This reviewer knows of no other single source of so much material of this nature. Highly recommended."
-Choice
"...a valuable resource for those with a specific need for numerical software. The routines are prefaced with lucid, selfcontained explanations....highly recommended for those who require the use and understanding of numerical software."
-SIAM Review
1992992 pp. 0-521-43064-X Hardback $\$ 70.00$

## Highlights include:

- A chapter on integral equations and inverse methods
- Multigrid and other methods for solving partial differential equations
- Improved random number routines
- Wavelet transforms
- The statistical bootstrap method
- A chapter on less-numerical" algorithms including compression coding and arbitrary precision arithmetic.

Numerical Recipes in Fortran 77 Example Book<br>Second Edition<br>William T. Vetterling, Saul A.Teukolsky, William H. Press, and Brian P. Flannery<br>1992256 pp . 0-521-43721-0 Paperback \$35.00

## Numerical Recipes in Fortran 90

The Art of Parallel Scientific Computing
Volume 2 of Fortran Numerical Recipes

## Second Edition

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery
"This present volume will contribute decisively to a significant breakthrough, as it provides models not only of the numerical algorithms for which previous editions are already famed, but also of an excellent Fortran 90 style."
-From the Foreword by Michael Metcalf, one of Fortran 90 's original designers and author of
FORTRAN 90 Explained
"This book is a classic and is essential reading for anyone concerned with the future of numerical calculation. It is beautifully produced, inexpensive for its content, and a must for any serious worker or student."
-Computing Reviews
Contains a detailed introduction to the Fortran 90 language and to the basic concepts of parallet programming, plus source code for all routines from the second edition of Numerical Recipes.
1996576 pp. 0-521-57439-0 Hardback \$50.00

## Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras
$2002 \quad 0.521-75036-9 \quad$ CD-ROM $\$ 150.00$

## Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License <br> Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras <br> $2002 \quad 0-521-75035-0 \quad$ CD-ROM $\quad \$ 90.00$

Visit us.cambridge.org/numericalrecipes for more information on the complete line of Numerical Recipes products.

## @ XML PLUG-IN FOR SAS IMPORT XML-FORMATTED DATA INTO SAS EXPORT SAS DATA AS XML-FORMATTED DATA

Now, an easy way to move CDISC and other XML-formatted data into or out of your SAS-based systems. You don't have to know perl, XSLT, Xpath, Java ${ }^{\text {a }}$, or exotic languages. The remarkable Tekoam ${ }^{\text {XML }}$ plug-in does it all for you.
Provided free of charge to any SAS user currently wrestling with XML. Developed by Zurich Biostatistics, the pioneer in SAS/XML integration.
FREE. EASY. AND IT WORKS.
Just e-mail Michael Palmer (mcpalmer@zbi.net) and receive the fully-functional, proven Tekoa XML plug-in by e-mail.
No charge. No obligation. No hassle. (We even support the tool. Imagine.)


Tekoa XML Technology is a service mark of Zurich Biostatistics, Inc. SAS is a registered trademark of SAS Institute Inc. Java is a registered trademark of Sun Microsysterns, Inc.

XML is easy if you know how. And we do.

# Resampling Stats for Excel 



Select the data you want to resample, select "resample" or "shuffle," then specify an output range for the resampled data. Calculate a statistic of interest, select "Repeat \& Score," and the resampling operation will be repeated thousands (or tens of thousands) of times, and each time the value of your statistic of interest will be recorded. Does not use Excel's random number generator.

View complete user guide and download free 30-day trial at <www.resample.com>
\$249 commercial •\$149 personal/academic • \$89 student

612 N Jackson St., Arlington, VA 22201
Tel 703-522-2713 • Fax 703-522-5846

- Resampling stats@resample.com


## JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to Psychological Methods or Psychological Assessment (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional \$18)
- The Score - the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: $\$ 38$ (APA membership not required); student membership is only $\$ 8$
For further information, please contact the Division's Membership Chair, Yossef Ben-Porath (ybenpora@kent.edu) or check out the Division's website:
http://www.apa.org/divisions/div5/

## ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the Special Interest Group - Educational Statisticians of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.

We also publish a twice-yearly electronic newsletter.
Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at http://orme.uark.edu/edstatsig.htm

To join SIG-ES you must be a member of AERA. Dues are $\$ 5.00$ per year.
For more information, contact Joan Garfield, President of the SIG-ES, at jbg@umn.edu.

Position Available: Top bio-tech company seeks a seasoned statistical manager to hire, develop and lead a team of applied statisticians. Primary role is to integrate statistical methodology and practice into product/process development, manufacturing operations and quality. This key leader will provide linkage between manufacturing, engineering, development and biostatistics. MS in statistics or related field.

Research Statistician: Established clinical group adding staff to provide dedicated preclinical support to a development center. Interact and support scientists with formulation, stability testing, bioanalytics and bio assays. PhD w/3 yrs or MS w/ 6 years industry experience required along with expertise in complicated design methods. Northeast location.

Contact Information: Eve Kriz, Smith Hanley Associates, 99 Park Avenue, New York, NY 10016, 212-687-9696 ext. 228, ekriz@smithhanley.com.


# Statistics Through Monte Carlo Simulation 

## With Fortran

Shlomo S. Sa wilowsky and Gail F. Fahoome Copyright ${ }^{\circ} 2003$ ISBN: 0-9740236-0-4

Purchase Email, CD, \& Softcover Versions Online Via Secure Paypal At
http://tbf.coe.wayne.edu/jmasm

## Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. JMASM uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@edstat.coe.wayne.edu. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e--mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format 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.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font. If the technical expertise is available, submit the manuscript in two column format.
6. 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.
7. The manuscript should contain an Abstract with a 50 word maximum, following by 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 leftjustified, indent optional.
8. 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.
9. 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.
10. 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 spacebar twice after a period.

## Print Subscriptions

Print subscriptions including postage for professions is US $\$ 60$ per year; graduate students is US $\$ 30$ per year; and libraries, universities, and corporations is US $\$ 195$ per year. Subscribers outside of the US and Canada pay a US $\$ 10$ surcharge for additional postage. Online access is currently free at http://tbf.coe.wayne.edu/jmasm. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@edstat.coe.wayne.edu.

## Notice To Advertisers

Send requests for advertising information to jmasm@edstat.coe.wayne.edu.

SOFTWARE SOLUTIONS for Science \& Engineering

# Lahey/Fujitsu Fortran 

The standard for Fortran programming from the leader in Fortran language systems

LF95 Fortran for Linux and Windows

Full Fortran 95/90/77 support Unsurpassed diagnostics Intel and AMD optimizations

IMSL compatible
Fujitsu SSL2 math library
Wisk graphics package

LF Fortran for the Microsoft ${ }^{\circledR}$. NET Framework - Coming Soon!

Visual Studio integration Windows / Web Forms designer Project and code templates

On-line integrated help XML Web services
ADO.NET support

Visit www.lahey.com for more information


Lahey Computer Systems, Inc. 865 Tahoe Blvd - P.O. Box 6091
Incline Village, NV 89450 USA
1-775-831-2500
www.lahey.com


## Two Years in the Making...

## Inte ${ }^{\circledR}$ Visual Fortran 8.0

The next generation of Visual Fortran is here! Intel Visual Fortran 8.0 was developed jointly by Intel and the former DEC/Compaq Fortran engineering team.


| Visual Fortran Timeline |
| :---: |
|  |
| Intel Visual Fortran 8.0 |
| - CVF front-end + Intel back-end <br> - Better performance <br> - OpenMP Support <br> - Real*16 |

## Performance

Outstanding performance on Intel architecture including Intel ${ }^{\ominus}$ Pentium ${ }^{\circledR}$ 4, Intel ${ }^{\oplus}$ Xeon ${ }^{\text {™ }}$ and Intel Itanium ${ }^{\oplus} 2$ processors, as well as support for Hyper-Threading Technology.

## Compatibility

- Plugs into Microsoft Visual Studio* .NET
- Microsoft PowerStation4 language and library support
- Strong compatibility with Compaq* Visual Fortran


## Support

1 year of free product upgrades and Intel Premier Support
"The Intel Fortran Compiler 7.0 was first-rate, and Intel Visual Fortran 8.0 is even better. Intel has made a giant leap forward in combining the best features of Compaq Visual Fortran and Intel Fortran. This compiler... continues to be a 'must-have' tool for any Twenty-First Century Fortran migration or software development project."

—Dr. Robert R. Trippi<br>Professor Computational Finance University of California, San Diego

## FREE trials available at: <br> programmersparadise.com/intel

## Phoghamuner's Paradise.

To order or request additional information call:
800-423-9990
Email: intel@programmers.com

