12-1-2017

Vol. 16, No. 2 (Full Issue)

JMASM Editors

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

Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Editors, JMASM (2017) "Vol. 16, No. 2 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 16 : Iss. 2 , Article 36.

DOI: 10.22237/jmasm/1509494400

Available at: <http://digitalcommons.wayne.edu/jmasm/vol16/iss2/36>

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

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

Journal of Modern Applied Statistical Methods

Invited Articles

Michael Harwell, Nidhi Kohli, Yadira Peralta,
and Rand Wilcox

Vol. 16, No. 2 • November, 2017

Journal of Modern Applied Statistical Methods

Shlomo S. Sawilowsky

SENIOR EDITOR

College of Education
Wayne State University

Harvey Keselman

ASSOCIATE EDITOR EMERITUS

Department of Psychology
University of Manitoba

Alan Klockars

ASSISTANT EDITOR EMERITUS

Educational Psychology
University of Washington

Bruno D. Zumbo

ASSOCIATE EDITOR

Measurement, Evaluation,
& Research Methodology
University of British Columbia

Vance W. Berger

ASSISTANT EDITOR

Biometry Research Group
National Cancer Institute

Todd C. Headrick

ASSISTANT EDITOR

Educational Psychology
& Special Education
So. Illinois University–
Carbondale

Clayton Hayes

EDITORIAL ASSISTANCE

Lofti Kerzabi

EDITORIAL ASSISTANCE

Joshua Neds-Fox

EDITORIAL ASSISTANCE

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

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

Journal of Modern Applied Statistical Methods

Vol. 16, No. 2

❧ November 2017 ❧

Table of Contents

Invited Articles

3 – 28	M. HARWELL N. KOHLI Y. PERALTA	Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences
29 – 38	R. WILCOX	The Regression Smoother LOWESS: A Confidence Band That Allows Heteroscedasticity And Has Some Specified Simultaneous Probability Coverage

Regular Articles

40 – 80	A. ALKHALAF B. D. ZUMBO	The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression
81 – 98	A. D. WU B. D. ZUMBO	Using Pratt's Importance Measures in Confirmatory Factor Analyses
99 – 123	T. T. SAJOBI L. M. LIX	Robust Measures of Variable Importance for Multivariate Group Designs
124 – 137	D. K. GHOSH S. AHUJA	On Variance Balanced Designs
138 – 156	J. KUMAR A. CHATURVEDI U. AFIFA S. YOUSUF S. KUMAR	Unit Root Test for Panel Data AR(1) Time Series Model With Linear Time Trend and Augmentation Term: A Bayesian Approach

157 – 185	H. J. TURNER P. NATESAN R. K. HENSON	Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha
186 – 199	L. TURPIN	Effectively Comparing Differences in Proportions
200 – 214	A. M. GOULD J. E. JOULLIÉ	'Parallel Universe' or 'Proven Future'? The Language of Dependent Means <i>t</i> -Test Interpretations
215 – 230	P. M. BROWN J. A. EZEKOWITZ	Power and Sample Size Estimation for Nonparametric Composite Endpoints: Practical Implementation using Data Simulations
231 – 255	H. Y. ELEDUM	A Monte Carlo Study of the Effects of Variability and Outliers on the Linear Correlation Coefficient
256 – 276	M. AJAMI S. M. A. JAHANSHAHI	Parameter Estimation In Weighted Rayleigh Distribution
277 – 309	A. A. MITANI K. P. NELSON	Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS
310 – 321	M. FAIZAN Z. HAQUE M. A. ANSARI	Characterizations of Distributions by Expected Values of Lower Record Statistics with Spacing
322 – 346	L. N. B. DAKHN M. A. H. EBRAHEM O. EIDOUS	Semi-Parametric Method to Estimate the Time-to-Failure Distribution and its Percentiles for Simple Linear Degradation Model
347 – 358	S. LIPOVETSKY	Prediction of Percent Change in Linear Regression by Correlated Variables <i>s</i>
359 – 377	G. REKABDAR R. CHINIPARDAZ	Approximating the Distribution of Indefinite Quadratic Forms in Normal Variables by Maximum Entropy Density Estimation
378 – 402	P. ZANINOTTO A. SACKER	Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques

403 – 417	R. GRINE H. ZEGHDOUDI	On Poisson Quasi-Lindley Distribution and its Applications
418 – 438	E. A. MAHMOOD S. RANA H. MIDI A. G. HUSSIN	Detection of Outliers in Univariate Circular Data using Robust Circular Distance
439 – 442	H. MESSAADIA H. ZEGHDOUDI	Around Gamma Lindley Distribution
443 – 457	R. P. ABREU J. R. SCHAFFER	A Double EWMA Control Chart for the Individuals Based on a Linear Prediction
458 – 484	F. YE L. DANIEL	The Impact of Inappropriate Modeling of Cross-Classified Data Structures on Random-Slope Models

Brief Reports

486 – 488	Z. D. BAI C. R. RAO M. T. TSAI	A Remark for the Admissibility of Rao's U-Test
-----------	---	--

Algorithms and Code

490 – 505	M. SHAFIQ W. M. AMIR N. S. ZAFAKALI	JMASM 46: Algorithm for Comparison of Robust Regression Methods In Multiple Linear Regression By Weighting Least Square Regression (SAS)
506 – 539	I. LI Y.-H. CHEN Y. WANG P. R. DE GIL T. PHAM D. NGUYEN E. S. KIM J. D. KROMREY	JMASM 47: ANOVA_HOV: A SAS Macro for Testing Homogeneity of Variance in One-Factor ANOVA Models (SAS)

540 – 546	D. A. WALKER	JMASM 48: The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS)
547 – 575	M. ÖNER İ. D. KOCAKOÇ	JMASM 49: A Compilation of Some Popular Goodness of Fit Tests for Normal Distribution: Their Algorithms and MATLAB Codes (MATLAB)
576 – 579	S. MAGGIO G. BHANDARI S. S. SAWILOWSKY	JMASM 50: A Web-based Shiny Application for Conducting a Two Dependent Samples Maximum Test (R)

Emerging Scholars

581 – 603	Y. F. LOH J. ARASAN H. MIDI M. R. A. BAKAR	Inferential Procedures for Log Logistic Distribution with Doubly Interval Censored Data
604 – 617	R. S. NOVAK J. M. MARQUES	Study Evaluating the Alterations Caused in an Exploratory Factor Analysis when Multivariate Normal Data is Dichotomized
618 – 638	O. R. OLANIRAN W. B. YAHYA	Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique

Invited Articles

Experimental Design and Data Analysis in Computer Simulation Studies in the Behavioral Sciences

Michael Harwell
University of Minnesota,
Twin Cities
Minneapolis, MN

Nidhi Kohli
University of Minnesota,
Twin Cities
Minneapolis, MN

Yadira Peralta
University of Minnesota,
Twin Cities
Minneapolis, MN

Treating computer simulation studies as statistical sampling experiments subject to established principles of experimental design and data analysis should further enhance their ability to inform statistical practice and a program of statistical research. Latin hypercube designs to enhance generalizability and meta-analytic methods to analyze simulation results are presented.

Keywords: simulation, experimental design, data analysis

Introduction

Computer simulation studies represent an important tool for investigating statistical procedures difficult or impossible to study using mathematical theory or real data. Descriptors of these studies vary (e.g., statistical experiment, Monte Carlo simulation, computer experiment), but the examples of Hoaglin and Andrews (1975) and Hauck and Anderson (1984) are followed here with use of the term simulation studies. Extensive descriptions of simulation studies can be found in Lewis and Orav (1989) and Santner, Williams, and Notz (2003).

In the behavioral sciences simulation studies have been used to study a wide array of statistical methods (e.g., Cribbie, Fiksenbaum, & Wilcox, 2012; Depaoli, 2012; Enders, Baraldi, & Cham, 2014; Hu & Bentler, 1999; Tomarken & Serlin, 1986). The general goal of these studies is to provide evidence of the behavior of statistical methods under a variety of data conditions that improves statistical practice and informs future statistical research. The goal here is to encourage methodological researchers to treat these studies as statistical sampling

Michael Harwell is a Professor in the Department of Educational Psychology. Email him at harwe001@umn.edu.

experiments subject to established principles of experimental design and data analysis.

An underappreciated facet of simulation studies in statistics is their role in enhancing the reproducibility of scientific findings. The importance of reproducibility has gained momentum in numerous scientific arenas because of growing evidence that many findings cannot be replicated (Stodden, 2015). Concerns over reproducibility and the role of statistics were captured in *Statistics and science: A report of the London workshop on the future of the statistical sciences* (2014) which noted: “The reproducibility problem goes far beyond statistics, of course, because it involves the entire reward structure of the scientific enterprise. Nevertheless, statistics is a very important ingredient in both the problem and the remedy.” (p. 27) Simulation studies in statistics can increase the likelihood that scientific findings can be reproduced by providing evidence of the impact of data that are perturbed on estimators, tests, bootstrapping methods, parameter estimation algorithms, model alterations, etc., and subsequent inferences (Stodden, 2015).

Computer simulation studies as statistical sampling experiments

Hoaglin and Andrews (1975) argued that simulation studies should be treated as statistical sampling experiments subject to established principles of research design and data analysis. Special attention is given to experimental design in simulation studies, because of its centrality in a research study and its ability to produce effects of interest, guide analyses of study outcomes, and enhance generalizability of study findings. Hoaglin and Andrews (1975) reviewed a sample of published studies using simulation methods and offered a harsh assessment of the state of the art: “Statisticians (who, of all people, should know better) often pay too little attention to their own principles of design, and they compound the error by rarely analyzing the results of experiments in statistical theory” (p. 124). Gentle (2003) reiterated this point: “A Monte Carlo study uses an experiment, and the principles of scientific experimentation should be observed.” (p. vii)

Hauck and Anderson (1984) surveyed studies in five statistics journals and reported that 216 (18%) studies used simulation methods and found little evidence that the recommendations of Hoaglin and Andrews (1975) were being adopted. Harwell, Kohli, and Peralta-Torres (2017) updated the Hauck and Anderson (1984) results by surveying studies in six statistics journals between 1985 and

2012 and found the use of simulation studies had basically doubled since 1984, but less than 5% of 371 simulation studies used an identifiable experimental design. Harwell, Kohli, and Peralta-Torres (2017) also reported that 99.9% of these studies relied exclusively on visual analysis of simulation findings (i.e., “eyeballing” the results).

It is important to emphasize simulation studies have made critical contributions to improving statistical practice; however, the recommendations of Hoaglin and Andrews (1975) imply that treating a simulation study as a statistical sampling experiment can further exploit the ability of these studies to inform statistical practice and a program of statistical research. The latter reflects the case in which a simulation study is part of a research program that includes previous studies whose results inform the conceptualization and execution of a proposed simulation study. The aim of the current study, therefore, is to encourage methodological researchers in the behavioral sciences to routinely treat computer simulation studies as statistical sampling experiments to fully exploit their strengths.

Experimental Design

Experimental design should play a crucial role in simulation studies because of its ability to produce effects of interest, guide analyses of study outcomes, and enhance generalizability of findings. The latter is particularly important because of concerns that generalizability of simulation study findings is frequently limited due to the way that values of simulation factors are selected (Paxton, Curran, Bollen, Kirby, & Chen, 2001; Skrondal, 2000). Modeling realistic conditions such as skewed data and small sample sizes is essential to generalizing simulation results in ways that improve statistical practice; our focus is designs that support generalizing results to simulation factor values beyond those explicitly modeled, which should further enhance generalizability and improve statistical practice.

Santner et al. (2003) defined inputs in a simulation as numerical values of simulation factors that collectively define the experimental region which in turn define the design. Thus experimental design is a specification of values of simulation factors in the experimental region at which we wish to compute an outcome. Input values are sampled from a defined pool of values using one of several sampling methods. The sampling methods are labeled space-filling, because they fill the experimental region in some fashion. More formally, an experimental design is defined by a matrix in which the columns correspond to simulation factors whose elements are researcher-specified numerical values for

EXPERIMENTAL DESIGN AND DATA ANALYSIS IN SIMULATION

the factors, and whose rows represent a combination of input values that define so-called design points. Consider the full factorial case in which all combinations of factor levels are examined. Let m^k represent k factors with m values (levels) which are being investigated using m^k input values; for two factors the experimental region is defined by m^{k_1} by m^{k_2} input values. For example, a binary factor (F_1) with researcher-specified values 10 and 20 crossed with a second binary factor (F_2) with values 18, 29, and 34 produces the values in Table 1. Factor levels are typically recoded for simplicity, for example, -1, 0, and +1 in Table 1, but this is not necessary (Sanchez, 2007).

The above design has six design points defined by the six rows in Table 1 with the coded values in a row representing inputs. In full factorials space-filling is the result of sampling the entire pool of researcher-specified simulation factor values. This practice generates a predictable pattern of space-filling that may answer specified research questions but can limit generalizations.

Table 1. Experimental Design for a 2×3 Full Factorial

Point	Original Values		Coded Values	
	F ₁	F ₂	F ₁	F ₂
1	10	18	-1	-1
2	20	18	+1	-1
3	10	29	-1	0
4	20	29	+1	0
5	10	34	-1	+1
6	20	34	+1	+1

An alternative to full factorials are incomplete fractional factorials. Skrondal (2000) described how these designs can be used in simulation studies in ways that enhance generalizability by employing more conditions than would typically be used in a full factorial because higher order interactions (reflected in combinations of factor conditions) are not modeled. These designs are especially appropriate for enhancing generalizability when there are many factors that take only a few values.

Space-filling by random sampling.

A related class of designs used to increase the generalizability of simulation findings relies on random sampling methods for space-filling (Santner et al., 2003). In some cases generalizability is increased by spreading points evenly over the experimental region, whereas in other instances points are concentrated on the boundaries of the experimental

region. One sampling method involves defining a pool of design points (with associated input values) assumed to follow a uniform distribution and taking a simple random sample.

Consider an exemplar simulation study investigating the impact of different numbers of clusters, within-cluster sample size, and distribution of cluster residuals when estimating fixed effects and the Type I error rate of tests of these effects for a two-level mixed (linear) model for continuous cross-sectional data. Suppose a pool of number of clusters (J) ($J = 10, 11, 12, \dots, 50$) was defined and a simple random sample taken; similarly, we could define design points as pairs of values of J and within-cluster sample size (n_j) that follow a uniform distribution (e.g., $J = 10, 11, 12, \dots, 50$; $n_j = 5, 6, 7, \dots, 100$) and take a simple random sample (assuming a normal distribution of cluster residuals for simplicity). This method should enhance generalizability relative to full factorials like that in [Table 1](#) but may not spread design points evenly across the experimental region. Stratified random sampling can potentially enhance generalizability by identifying a stratification variable and selecting a point at random from each stratum. For example, we could define strata using n_j (n_j strata defined as 5-10, 11-15, ..., 95-100) with a pool of values of J within each stratum (e.g., $J = 10, 11, 12, \dots, 50$) one of which is selected at random from each stratum. The resulting design points ensure space-filling as they include the entire range of values of n_j as captured by the strata.

Perhaps the most widely recommended sampling method for space-filling to increase generalizability of simulation findings is Latin hypercube sampling, which generates a Latin hypercube design (LHD) ([Santner et al., 2003](#)). Latin hypercube designs are a variation of traditional Latin squares and spread design points evenly across the range of an input. [Santner et al. \(2003\)](#), [Sanchez \(2007\)](#), and [Viana \(2013\)](#) illustrated the use of LHDs in simulation studies for relatively simple designs and pointed out their benefits generally increase with increases in k ; [Sanchez \(2007\)](#) noted the number of points (and potentially the generalizability) increases linearly with increases in k .

Let p denote the total number of design points and assume low and high levels (values) for a factor F_k are coded as 1 and p , and that the set of coded factor levels are $1, 2, \dots, p$. A $p \times k$ design matrix for a LHD can be written as $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_p]^T$ where each column represents a factor and each row $\mathbf{x}_i = (x_i^{(1)} x_i^{(2)} \dots x_i^{(k)})$ for $i = 1, \dots, p$ represents a design point. In a LHD each factor is divided into p equal levels and one point is sampled at each level using a random procedure. Different optimization algorithms for LHD have appeared such as genetic-type algorithms, simulated annealing, optimum Euclidian distance,

and column-pairwise optimization (Carnell, 2016; Viana, 2013), and specialized software like the lhs package in R (R Core team, 2016) is needed to implement even simple LHDs. This software is illustrated below.

Exemplar simulation study. The rationale for our two-level mixed model exemplar comes from a review of statistical theory and previous simulation results (Austin, 2010; Bell, Ferron, & Kromrey 2008; Clarke & Wheaton, 2007; Maas & Hox, 2004, 2005; Maeda, 2007; Moerbeek, van Breukelen, & Berger, 2000). This literature suggests the number of clusters needed to accurately estimate fixed effects and to have tests of these effects control Type I error rates at nominal levels is unresolved for non-normal cluster residuals. This prompted the research question: How many clusters are needed in a two-level model for continuous cross-sectional data with one predictor at each level for conditions of varying within-cluster sample sizes and non-normal cluster residuals to ensure: (a) accurate estimation of fixed effects and (b) statistical tests of these effects control Type I error rates at nominal levels?

For this simulation exemplar the statistical model with one predictor at each level was

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (\text{level 1}) \quad (1)$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + W_{1j}\gamma_{01} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + W_{1j}\gamma_{11} + u_{1j} \end{aligned} \quad (\text{level 2})$$

which implies the mixed model $Y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} + (\gamma_{10} + \gamma_{11}W_{1j} + u_{1j})X_{ij} + r_{ij}$. In equation (1), Y_{ij} represents the (continuous) outcome score of the i^{th} level 1 unit in the j^{th} level 2 unit (cluster), β_{0j} and β_{1j} are the intercept and linear slope for the j^{th} cluster, X_{ij} is a predictor value sampled from an $N(0,1)$ distribution), r_{ij} is that level 1 unit's residual ($r_{ij} \sim N(0, \sigma^2)$), γ_{00} is the average β_{0j} , γ_{10} is the average X_1, Y slope within clusters, γ_{01} is a slope capturing the effect of the level 2 predictor W_{1j} , γ_{11} is the slope capturing the cross-level interaction effect, and u_{0j} and u_{1j} are cluster residuals for the intercept and slope models (Raudenbush & Bryk, 2002, pp. 100-103). The fixed effects in equation (1) ($\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$) were set to zero to reflect the Type I error case meaning the mixed model underlying the data generation was simply $Y_{ij} = u_{0j} + u_{1j}X_{ij} + r_{ij}$.

To specify simulation conditions we relied on statistical theory (Raudenbush & Bryk, 2002, chpt. 3), previous simulation studies, and documented

characteristics of large multilevel datasets (Hedges & Hedberg, 2007). We assumed $\begin{bmatrix} u_{0t} \\ u_{1t} \end{bmatrix} \sim [\mathbf{0}, \mathbf{T}]$ followed a normal or chi-square distribution (see below), where $\mathbf{T} = \begin{bmatrix} 9 & 0 \\ 0 & .75 \end{bmatrix}$ was a 2×2 covariance matrix of random effects with diagonal entries τ_{00} (variance of u_{0j}) and τ_{11} (variance of u_{1j}), and covariance τ_{01} . We specified $\tau_{00} > \tau_{11}$ based on Lee and Bryk (1989) who reported a within-cluster variance for mathematics achievement data of 39.927 for their unconditional model, a between-cluster intercept variance of 9.335, and three between-cluster slope variances whose average was .75. Using values of 40 and 9 for σ^2 and τ_{00} in the unconditional model in our simulation produced an intra-class correlation (ICC) of .19, which is consistent with the results of Hedges and Hedberg (2007). The covariance component τ_{01} was set to 0 based on simulation evidence that this value typically has little impact on the number of clusters (Maas & Hox, 2004, 2005; Zhang, 2005). The resulting pool of inputs in our exemplar study was specified as $J = 10, 20, 30, 40, 50$ (number of clusters), $n_j = 18, 29, 34, 44, 60, 68$ (within-cluster sample sizes), and distribution = $\text{BVN} \sim \begin{pmatrix} 0 & 9 \\ 40 & 0 \end{pmatrix}, \chi_{10}^2$ (distribution of cluster residuals). n_j values were selected at random from a range of 5 to 100, because there was no empirical basis for specifying particular values. Data were simulated using the R software.

The estimated fixed effects in the exemplar study served as indicators of bias because the true values equaled zero, and were computed as an average across $R = 5,000$ replications. Type I error rates of tests of the fixed effects were estimated as the proportion of rejections of the associated statistical null hypothesis across R replications. $R = 5,000$, a number that generally provides accurate estimates of Type I error rates for general linear model-based statistical tests (Robey & Barcikowski, 1992) and should do the same for bias estimates. Next, the exemplar study is used to illustrate space-filling for a full factorial and LHD, and meta-analysis to analyze simulation results.

Results

The resulting design matrix for the exemplar had three columns and 60 rows (design points) and sampling all design points produced a 5×6×2 full factorial design with 60 cells. We conditioned the design on a particular distribution of cluster residuals (bivariate normal, chi-square); otherwise we must generate a

pool of input values representing distributions. If the focus was exclusively on the two distributions in the exemplar study these define the pool of inputs and the exemplar design matrix would have three columns and 60 rows. If instead the desire is to generalize findings to a family of skewed distributions such as chi-square a pool of input values defined by degrees of freedom could be specified, for example, $df = 1, 2, \dots, 20$, in which case the design matrix would have three columns and 600 rows. To simplify the graphical display we focus on J and n_j meaning the exemplar study design matrix has two columns and 60 rows. The `lhs` package in R was used to generate the experimental region for the 5×6 full factorial displayed in Figure 1, which is a grid composed of 30 points. Notice the lines of dots for J are equidistant from each other whereas those for n_j vary in distance because the latter vary in value. This figure highlights the non-random nature of space-filling for the 5×6 full factorial which limits generalizability to selected input values.

Employing a LHD signals we are interested in generalizing to design points not explicitly modeled in the simulation. This strategy supports generalizing findings to a pool of design points in ways not possible with a full factorial, and with less uncertainty compared to simple random sampling of points because space-filling throughout the experimental region is not assured.

To construct a LHD for the exemplar simulation study we used the `maximinLHS` function in the `lhs` package in R, which draws a Latin hypercube sample from a set of uniform distributions that can be rescaled to the range of interest (Carnell, 2016). The `maximinLHS` function optimizes the sample by maximizing the minimum distance between design points (Carnell, 2016). In order to create the LHD we drew a sample of 30 points considering two factors. The resulting design points were then rescaled to the ranges covered by factors one $J = (10, 11, \dots, 49, 50)$ and two ($n_j = 5, 6, 7, \dots, 100$) in our exemplar study. That is, F_1 (number of clusters) was rescaled to have values between 10 and 50 and F_2 (within-cluster sample size) to have values between 18 and 68. The number of sampled factor values (inputs) depends on the desired generalizability with more values expected to provide greater space-filling, although this may have to be weighed against available computing resources (Santner et al., 2003).

Figure 1. Experimental region for the exemplar simulation study with full factorial design conditioning on distribution of cluster residuals.

Shown in [Figure 2](#) are the design points associated with the LHD for the exemplar simulation study, which are spread evenly across the experimental region. The implication of the LHD in [Figure 2](#) is that findings of our exemplar simulation study are generalizable to the entire pool of researcher-specified values of J and n_j not just those explicitly modeled. [Figure 3](#) contrasts [Figures 1](#) and [2](#) and illustrates the systematic, non-random space-filling of a full factorial versus the random-sampling-based space-filling of a LHD. R code for generating the experimental regions illustrated in [Figures 1-3](#) appears in [Appendix A](#).

The enhanced generalizability linked to LHDs speaks to their potential to improve statistical practice and inform future statistical research. However, there are areas of statistical research employing simulation methods in which sampling all design points is appropriate because interest is limited to those inputs, perhaps because of theoretical or empirical reasons. For example, interest may be limited to a small number of distributions as was the case for the exemplar, where the space-filling illustrated in [Figure 1](#) is appropriate.

Figure 2. Experimental region with random selection of inputs for Latin Hypercube design conditioning on distribution of cluster residuals.

Figure 3. Contrasting the experimental region of the full factorial versus Latin Hypercube design.

Analysis of simulation results

Despite the recommendations of Hoaglin and Andrews (1975), Skrondal (2000), Boomsma (2013), Paxton et al. (2001) and others the analysis and reporting of results continues to rely heavily on visual analyses (Harwell et al., 2017). When there are exceptions they typically involve factorial ANOVA (e.g., Culpepper & Aguinis, 2011; MacCallum, Widaman, Preacher, & Hong, 2001), or less frequently logistic regression (e.g., Skrondal, 2000). Relying on visual analysis of simulation results is reasonable if key patterns and their magnitude are accurately captured such as interaction effects. On the other hand, reliance on tables and plots when summarizing information in dozens, hundreds, or even thousands of simulation results raises questions about how accurately important patterns can be detected and how precisely their magnitude can be estimated. We argue that visual analysis should typically be augmented by inferential analyses of results guided by the experimental design.

Visual analysis of simulation results. Methodological researchers have traditionally relied on visual analyses of simulation results which often appear in tables regardless of the number of simulation outcomes. For example, Wilcox (2009) reported three tables each containing 48 simulation results, Ramsey and Ramsey (2009) reported 1,750 values in five tables, and, as an extreme example, Aaron (2003) reported more than 7,000 values. The accuracy of visual analyses to summarize patterns and estimate the magnitude of effects in studies like Ramsey and Ramsey (2009) has not been tested experimentally, for example, by assembling a group of methodological researchers and assessing their ability to accurately detect patterns in simulation results using artificial sets of findings varying in known ways (e.g., entirely random pattern, only one effect). However, the ability to reliably and validly detect patterns using visual analysis has been studied in other research domains.

Single-case designs in psychology and education (Kratochwill et al., 2010; Kratochwill & Levin, 1992, 2010) involve collecting and plotting repeated measures data to assess the impact of one or more interventions (Smith, 2012). A good deal of research (Bailey, 1984; DeProspero & Cohen, 1979; Jones, Vaught, & Weinrott, 1977; Knapp, 1983; Matyas & Greenwood, 1990) assessing the ability of researchers, clinicians, and others to reliably and validly detect patterns using visual analysis highlighted the difficulties of doing so even for relatively small numbers of data points (e.g., 10-15), and the use of visual and inferential analyses has been recommended (Ferron, 2002; Kratochwill et al., 2010). Consider the estimated Type I error rates in Table 2 generated in the exemplar

EXPERIMENTAL DESIGN AND DATA ANALYSIS IN SIMULATION

study assuming a full factorial design. Values falling outside a 95% confidence interval are treated as sensitive to the conditions modeled. It's clear that a majority of $\hat{\alpha}$ values are inflated and that increases in the number of clusters seem to be associated with $\hat{\alpha}$ values closer to .05; within-cluster sample size and the distribution of cluster residuals do not seem to have much impact. Similarly, a visual analysis of average bias values in Table 3 suggests a chi-square distribution of cluster residuals produces somewhat more bias which generally shrinks as J increases. Careful visual analysis is important but performing inferential statistical analyses and estimating the magnitude of effects can provide additional insight into the impact of simulation factors on outcomes of interest.

Table 2. Estimated Type I error rates for tests of γ_{01} and γ_{11}

		<i>J</i>	10	20	30	40	50
u_{0j}, u_{1j} distribution	n_j	Type I Error Rate, γ_{01}					
$N(0,9)$ and $N(0,0.75)$	18	.085 *	.069 *	.055	.058 *	.057 *	
	29	.084 *	.069 *	.057 *	.060 *	.057 *	
	34	.093 *	.065 *	.059 *	.052	.058 *	
	44	.085 *	.062 *	.054	.055	.055	
	60	.084 *	.068 *	.057 *	.051	.047	
	68	.085 *	.065 *	.059 *	.053	.054	
χ^2_{10}	18	.090 *	.069 *	.063 *	.054	.059 *	
	29	.089 *	.064 *	.062 *	.053	.060 *	
	34	.086 *	.057 *	.061 *	.063 *	.052	
	44	.086 *	.063 *	.058 *	.058 *	.051	
	60	.085 *	.058 *	.063 *	.057 *	.060 *	
	68	.093 *	.069 *	.055	.055	.056	

u_{0j}, u_{1j} distribution	n_j	Type I Error Rate, γ_{11}					
$N(0,9)$ and $N(0,0.75)$	18	.051	.053	.049	.050	.052	
	29	.060 *	.055 *	.056	.055	.049	
	34	.063 *	.060 *	.062 *	.057 *	.050	
	44	.067 *	.065 *	.068 *	.063 *	.063 *	
	60	.071 *	.064 *	.059 *	.056	.058 *	
	68	.074 *	.072 *	.063 *	.063 *	.051	
χ^2_{10}	18	.086 *	.061 *	.067 *	.055	.055	
	29	.083 *	.062 *	.054	.059 *	.064 *	
	34	.086 *	.064 *	.055	.059 *	.059 *	
	44	.086 *	.068 *	.065 *	.058 *	.055	
	60	.091 *	.063 *	.057 *	.055	.054	
	68	.093 *	.063 *	.061 *	.054	.055	

Note: Tabled values represent estimated Type I error rate across $R = 5,000$ replications, * = an error rate falling outside the 95% confidence interval limits, u_{0j} and u_{1j} represent cluster residuals, J = number of clusters, n_j = within-cluster sample size.

Table 3. Average bias for γ_{01} and γ_{11}

		<i>J</i>	10	20	30	40	50
u_{0j}, u_{1j} distribution	n_j	Average bias for γ_{01}					
$N(0,9)$ and $N(0,0.75)$	18	-0.0138	-0.0249	-0.0202	-0.0148	-0.0141	
	29	-0.0126	0.0100	0.0045	-0.0054	0.0079	
	34	-0.0050	0.0104	0.0101	0.0059	0.0007	
	44	0.0276	-0.0141	0.0028	-0.0004	-0.0069	
	60	0.0467	-0.0199	-0.0067	0.0183	0.0042	
	68	-0.0180	0.0087	0.0059	0.0012	-0.0074	
χ^2_{10}	18	0.0226	-0.0120	-0.0231	-0.0154	0.0135	
	29	-0.0411	0.0154	-8.52E-05	-0.0230	-0.0055	
	34	-0.0432	-0.0450	0.0073	-0.0104	0.0055	
	44	0.0411	-0.0132	0.0126	-0.0035	0.0006	
	60	0.0571	0.0228	-0.0024	0.0141	0.0069	
	68	0.0076	0.0092	0.0247	0.0047	-0.0045	
u_{0j}, u_{1j} distribution	n_j	Average bias for γ_{11}					
$N(0,9)$ and $N(0,0.75)$	18	-0.0025	0.0077	-0.0035	-0.0029	-0.0036	
	29	0.0041	-0.0002	-0.0015	-0.0054	-0.0008	
	34	-0.0019	-0.0011	-0.0076	-0.0025	-0.0047	
	44	0.0111	-0.0058	-0.0055	-0.0049	0.0076	
	60	0.0023	-0.0034	-0.0061	0.0051	0.0005	
	68	0.0033	0.0008	0.0021	7.04E-05	0.0074	
χ^2_{10}	18	-0.0234	-0.0139	-0.0189	-0.0184	0.0066	
	29	-0.0088	-0.0078	0.0203	-0.0149	0.0063	
	34	0.0137	0.0236	-0.0260	0.0085	-0.0052	
	44	0.0032	-0.0336	-0.0018	-0.0231	0.0010	
	60	0.0146	0.0146	-0.0048	0.0135	0.0145	
	68	-0.0269	0.0249	0.0209	0.0216	0.0135	

Note: Tabled values represent average bias across $R = 5,000$ replications, and represent cluster residuals, J = number of clusters, n_j = within-cluster sample size.

Meta-analysis of simulation results.

Next, consider the use of meta-analysis to detect patterns in simulation results. Assume the typical case in which simulation outcomes are averaged across R replications in each cell of the design and a fixed effect full factorial design for our exemplar study. However, the method described below can be adapted to LHDs (see below). It is assumed model-checking will be performed to ensure underlying assumptions are plausible.

Meta-analytic methods permit the relationship between simulation factors and outcomes to be assessed and also provide a test of model misspecification. The averaged outcome for each cell serves as an effect size, for example, $\hat{\alpha}$ or

$\overline{bias} = \sum_{s=1}^{R_s} \frac{(\hat{\theta}_s - \theta)}{R_s}$, $\hat{\theta}_s = s^{\text{th}}$ estimated parameter, $\theta =$ parameter, and $R_s =$ number of replications $\hat{\theta}_s$ is based on. The mean and variance of outcomes must be available and for $\hat{\alpha}_s$ are well known. The expression $\sum_{s=1}^{R_s} \frac{(\hat{\theta}_s - \theta)^2}{R_s}$ provides a variance estimate for bias $\left[Var(\overline{bias})\right]$ that can serve as an effect size of the impact of simulation factors on the variability of bias estimates. To treat $\left[Var(\overline{bias})\right]$ as an effect size $\ln\left[Var(\overline{bias})\right]$ is computed under the assumption $\hat{\theta}_s$ values are normally-distributed (Raudenbush & Bryk, 1987). In this case $Var\left\{\ln\left[Var(\overline{bias})\right]\right\} = \frac{2}{S-H-1}$ ($S =$ total number of effect sizes) which allows inferential analyses of $\ln\left[Var(\overline{bias})\right]$ values. Similar expressions are available for other outcomes such as statistical power and model convergence rates.

Consider a meta-analytic regression model for Type I error rates:

$$\alpha_s = \beta_0 + \sum_{h=1}^H \beta_H X_{SH}, \hat{\alpha}_s = \alpha_s + \xi_s \quad (2)$$

In equation (2), α is the s^{th} effect size (population proportion, $s = 1, 2, \dots, S$) that depends on a set of H predictor variables X_{SH} which could include interactions, β_0 is a population intercept, β_H is a population regression coefficient that captures the linear relationship between a predictor and α_s , ξ_s is a population error term, and $\hat{\alpha}_s$ is an estimated Type I error rate (proportion) (Hedges & Olkin, 1985, p. 169). The fitted model has the form:

$$\hat{\alpha}'_s = \hat{\beta}_0 + \sum_h^H \hat{\beta}_H X_{SH} \quad (3)$$

In equation (3), $\hat{\beta}_H$ is an estimated slope and $\hat{\alpha}'_s$ is a model-predicted proportion. The relationship between a set of predictors and effect sizes can be tested using the Q_{Reg} statistic presented in Hedges and Olkin (1985, p. 169-171). Assume the distribution of errors is normal with a mean of zero and diagonal

covariance matrix $\Sigma_{\hat{\alpha}}$ with dimensions $S \times S$ and elements $\sigma_{\hat{\alpha}}^2$. The Q_{Reg} test statistic equals the weighted sum of squares due to regression for the model in equation (3) with weights $[\sigma_{\hat{\alpha}}^2]^{-1} = \frac{R_s}{\hat{\alpha}_s(1-\hat{\alpha}_s)}$, where R_s is the number of replications associated with $\hat{\alpha}_s$. Under the hypothesis $H_0: \beta = 0$, where β and 0 are $H \times 1$ vectors, Q_{Reg} follows a chi-square distribution with $df = H$. Because $\hat{\alpha}_s$ represents binomial data, a data-analytic alternative is to initially transform each $\hat{\alpha}_s$ using the arcsine transformation (Cox, 1970). The mean and variance of the transformed quantities ($\hat{\alpha}_s^{\text{arcsine}}$) are independent and the assumption of normality is typically plausible even for modest sample sizes. The transformed quantities follow $\hat{\alpha}_s^{\text{arcsine}} \sim N \left[E(\hat{\alpha}_s^{\text{arcsine}}) = \alpha_s^{\text{arcsine}}, \text{Var}(\hat{\alpha}_s^{\text{arcsine}}) = \frac{1}{S} \right]$ and serve as outcomes in equation (2).

A key feature of the meta-analytic approach is the ability to test model specification i.e., whether all predictors contributing to variation in effect sizes are in the model (Hedges & Olkin, 1985, p. 172). The test for misspecification relies on a weighted error sum of squares associated with the model in equation (2) that is computed using the test statistic $Q_{\text{Error}} = \hat{\alpha}' \sum_{\hat{\alpha}}^{-1} \hat{\alpha} - Q_{\text{Reg}}$, where $\hat{\alpha}$ is a $S \times 1$ vector of the $\hat{\alpha}_s$. If the model is correctly specified Q_{Error} it is distributed as a chi-square variable with $df = S - H - 1$. Rejection of the hypothesis that the model is correctly specified implies that the weighted error variance is larger than expected, results are subject to misspecification bias (Hedges & Olkin, 1985, p. 172), and adding additional predictors could reduce error and produce less biased estimates.

In all cases the Q tests assume normality and because of the large numbers of replications typically used in simulation the normality approximation for $\hat{\alpha}_s$ should be quite good. Alternatively weighted logistic regression could be used to estimate parameters and test hypotheses for $\hat{\alpha}_s$. The Hedges and Olkin (1985) Q tests were chosen because: (a) these tests can be applied to a variety of effect sizes, (b) this approach provides a widely adopted measure of explained variance (R^2) which is not always the case for weighted logistic regression although it is important to recall that R^2 in weighted least squares represents the variance in the weighted outcomes explained by the weighted prediction model (Willet & Singer, 1988), (c) existing data analysis software can be used to fit the models. Note the meta-analytic regression model in equation (2) assumes predictor values are fixed whereas for LHDs predictor values such as those for J and n_j are sampled at

EXPERIMENTAL DESIGN AND DATA ANALYSIS IN SIMULATION

random. In practice predictors whose values are fixed and those representing random variables produce the same statistical inferences since the former can be considered realizations of the latter (Sampson, 1974). Thus simulation results from LHDs can be analyzed using equation (2) by treating the sampled simulation factor values as realizations from a larger pool of such values.

To illustrate the Q tests consider the results in Table 2. The fixed effects ($\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$) could be treated as a within-subjects factor in the analyses but we chose to examine the γ_{01} and γ_{11} results separately (results for γ_{00} and γ_{10} were similar to those for γ_{01} and γ_{11}). The predictors were number of clusters, within-cluster sample size, and distribution of cluster residuals that were centered about their mean, and their two-way interactions. The resulting $Q_{Reg} = 428.2$ ($p < .05$) for the γ_{01} Type I error results signals a statistically significant relationship between Type I error rates and the set of predictors, and the associated R^2 of .66 indicates there is a strong predictive relationship almost all of which ($R^2 = .65$) is attributable to number of clusters. The model-predicted error rates for number of clusters were .077 ($J = 10$), .071 (20), .064 (30), .057 (40), and .051 (50). Post hoc analyses were performed testing each slope against zero (Hedges & Olkin, 1985, p. 174) and controlling for compounding Type I error rates using the method of Sidak (1967) such that the error rate for each test was $.05/6 = .0083$. Only the slope for the number of clusters predictor was significant ($-.001$), meaning that Type I error rates for the test of γ_{01} were on average insensitive to within-cluster sample size and cluster residual distribution as well as the three two-way interactions but were sensitive to number of clusters. Testing model misspecification produced a statistically significant test ($Q_{Error} = 223.4$, $p < .05$), implying that the regression findings should be interpreted cautiously and adding predictor variables could potentially reduce error variation and bias in parameter estimates.

The model in equation (3) was then fitted to $\hat{\alpha}_s$ for the test of γ_{11} and obtained $Q_{Reg} = 300.7$ ($p < .05$), meaning there was a statistically significant and, it turns out, strong ($R^2 = .67$) relationship between $\hat{\alpha}_s$ and the set of predictors. Post hoc analyses showed that cluster residual distribution, within-cluster sample size, and the interactions number of clusters $\times$ within-cluster sample size and number of clusters $\times$ cluster residual distribution were significant predictors. Approximately 18% ($R^2 = .18$) of the variance in was attributable to cluster residual distribution, followed by within-cluster sample size (11%), and the interactions number of clusters $\times$ level 2 residual distribution (8%) and within-cluster sample size $\times$ cluster residual distribution (6%).

Model-predicted error rates for cluster residual distribution were .059 (normal) and .064 (chi-square) and for number of clusters were .071 ($J = 10$), .066 (20), .062 (30), .057 (40), and .053 (50); for within-cluster sample size the average model-predicted error rates ranged from .059 to .065. The interaction plot for number of clusters $\times$ cluster residual distribution showed a discrepancy for $J = 10$ with an average error rate of .077 for a chi-square distribution and .065 for a normal distribution and .072, and .063 for $J = 20$; otherwise average error rates were similar for the remaining conditions. The interaction plot for within-cluster sample size $\times$ cluster residual distribution showed a modest difference for $J = 10$ with an average error rate of .061 for a chi-square distribution and .054 for a normal distribution, and .057 and .062 for $J = 20$; otherwise average error rates were quite similar. A test of model misspecification produced a significant result ($Q_{Error} = 149.8$, $p < .05$) meaning that the findings should be interpreted cautiously and adding predictor variables could reduce error variation and bias in parameter estimates.

Comparing a visual analysis of Table 2 with the inferential results reveals several important differences. For γ_{01} the tabular results showed a majority of $\hat{\alpha}_s$ values were inflated and that increases in the number of clusters seem to be associated with values closer to .05; within-cluster sample size and the distribution of cluster residuals did not seem to have much impact. The inferential analyses supported these inferences but quantified the predictive strength of number of clusters with 65% of the variance attributable to this factor. For γ_{11} a majority of Type I error rates were also inflated but also seemed to move toward .05 as J increased particularly for $J \geq 30$. The inferential analyses demonstrated that error rates were less sensitive to simulation factors than those for γ_{01} and more sensitive to cluster residual distribution than J . The results also showed that combinations of factors impacts Type I error rates although the strength of these effects was modest.

Conclusion

A substantial amount of simulation research is available that has unquestionably made important contributions to improving statistical practice and informing future statistical research, yet the potential of these studies has not yet been fully realized in large part because recommendations to treat them as statistical sampling experiments have not been widely adopted. Adopting the recommendations of Hoaglin and Andrews (1975) should enhance the contributions of simulation studies including their role in increasing the

EXPERIMENTAL DESIGN AND DATA ANALYSIS IN SIMULATION

reproducibility of findings of studies employing statistical analyses. Following Hoaglin and Andrews (1975), the focus was on two key facets of a simulation study: experimental design and analysis of results.

The presence of a literature focused on experimental designs in simulation studies that enhance generalizability, and the availability of software to construct these designs, provides an important resource for methodological researchers. It is argued it is first important to adopt some kind of identifiable experimental design. Of course, simulation studies in some statistical research areas are quite similar, so much so that this may explain why the design is not reported. For example, simulation studies such as Ramsey and Ramsey (2009) typically employed multiple categorical simulation factors and report results in a fashion consistent with a full factorial design but do not identify the design used. Reporting the experimental design used in the study (assuming there is one) and other relevant details is consistent with Hoaglin and Andrews (1975) recommendation “A published report of computation-based results must make it easy for the reader to make reasonable assessments of the numerical quality of the results.” (p. 124).

Describing the experimental design also allows readers to assess the generalizability of findings. Simulation studies by their nature offer strong internal validity but require special attention be given to generalizability. Designs in which simulation factor values are randomly sampled from a researcher-specified pool of values, such like Latin hypercube designs, speak to issues of generalizability. Of course, not every simulation study is focused on enhancing generalizability but there appear to be many instances in which adopting designs such as a Latin hypercube can increase their contribution. Construction of a Latin hypercube for our exemplar simulation study highlighted the enhanced generalizability this design offers.

A second facet was analysis of simulation results. Visual analysis of results as illustrated in our exemplar study was useful, but augmenting this approach with inferential methods should improve the accuracy with which patterns are detected and their magnitude estimated. Inferential analysis of simulation results is also consistent with the recommendations of Hoaglin and Andrews (1975). Meta-analytic methods treat simulation outcomes as effect sizes and simulation factors as predictors in a regression model. This approach provides a test of the relationship between the simulation factors and outcomes and an index of explained variance if this relationship is statistically significant. A test of model misspecification provides an important tool for properly modeling variation in outcomes as well as interpreting simulation findings.

What next?

Efforts to encourage methodological researchers to adopt recommendations to increase the impact of simulation studies by treating them as statistical sampling experiments have had limited success in the past four decades. Those who advocated recommendations of Hoaglin and Andrews (1975) be adopted seem to have assumed these recommendations possess a kind of face validity, i.e., their merit is obvious especially to individuals who subscribe to the importance of established principles of experimental design and data analysis. Clearly, this argument has not been sufficiently compelling and changing the conceptualization, execution, and reporting of computer simulation studies in ways consistent with Hoaglin and Andrews (1975) will require continued efforts to convince authors, reviewers, and editors of their merit.

References

- Aaron, L. A. (2003). *A comparative simulation of Type I error and power of four tests of homogeneity of effects for random- and fixed-effects models of meta-analysis*. Unpublished doctoral dissertation. University of South Florida.
- Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics*, 6(1), Article 16. doi: [10.2202/1557-4679.1195](https://doi.org/10.2202/1557-4679.1195)
- Bailey, D. B. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis*, 17(3), 359-365. doi: [10.1901/jaba.1984.17-359](https://doi.org/10.1901/jaba.1984.17-359)
- Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *Proceedings of the Joint Statistical Meetings, Survey Research Methods Section*, 1122-1129.
- Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3), 518-540. doi: [10.1080/10705511.2013.797839](https://doi.org/10.1080/10705511.2013.797839)
- Carnell, R. (2016). *lhs: Latin Hypercube Samples*. R package version 0.14. <https://CRAN.R-project.org/package=lhs>

- Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods. *Sociological Research & Methods*, 35, 311-351. doi: [10.1177/0049124106292362](https://doi.org/10.1177/0049124106292362)
- Cox, D. R. (1970). *Analysis of binary data*. London, U.K.: Chapman and Davis.
- Cribbie, R. A. A., Fiksenbaum, L., & Keselman, H. J. (2012). Effect of non-normality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology*, 65(1), 56–73. doi: [10.1111/j.2044-8317.2011.02014.x](https://doi.org/10.1111/j.2044-8317.2011.02014.x)
- Culpepper, S. A. & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, 16(2), 166-178. doi: [10.1037/a0023355](https://doi.org/10.1037/a0023355)
- Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(2), 178-203. doi: [10.1080/10705511.2012.659614](https://doi.org/10.1080/10705511.2012.659614)
- DeProspero A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, 12(4), 573–579. doi: [10.1901/jaba.1979.12-573](https://doi.org/10.1901/jaba.1979.12-573)
- Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, 19(1), 39–55. doi: [10.1037/a0035314](https://doi.org/10.1037/a0035314)
- Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments, & Computers*, 34(3), 324-331. doi: [10.3758/bf03195459](https://doi.org/10.3758/bf03195459)
- Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2nd ed.). New York, NY: Springer. doi: [10.1007/b97336](https://doi.org/10.1007/b97336)
- Harwell, M., Kohli, N., & Peralta-Torres, Y. (2017). A survey of reporting practices of computer simulation studies in statistical research. *The American Statistician*. Advance online publication. doi: [10.1080/00031305.2017.1342692](https://doi.org/10.1080/00031305.2017.1342692)
- Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*, 38(3), 214-216. doi: [10.1080/00031305.1984.10483206](https://doi.org/10.1080/00031305.1984.10483206)
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press: Orlando.

- Hedges, L.V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. doi: [10.3102/0162373707299706](https://doi.org/10.3102/0162373707299706)
- Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29(3), 122-126. doi: [10.1080/00031305.1975.10477393](https://doi.org/10.1080/00031305.1975.10477393)
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. doi: [10.1080/10705519909540118](https://doi.org/10.1080/10705519909540118)
- Jones, R. R., Vaught, R. S., & Weinrott, M. R. (1977). Time series analysis in operant research. *Journal of Applied Behavior Analysis*, 10(1), 151-166. doi: [10.1901/jaba.1977.10-151](https://doi.org/10.1901/jaba.1977.10-151)
- Knapp, T. J. (1983). Behavior analysts’ visual appraisal of behavior change in graphic display. *Behavioral Assessment*, 5(2), 155-164.
- Kratochwill, T. R., & Levin J. R. (1992). *Single-case research design and data analysis: New directions for psychology and education*. Howe, UK: Lawrence Erlbaum.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). *Single-case designs* [Technical documentation]. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf
- Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15(2), 124–144. doi: [10.1037/a0017736](https://doi.org/10.1037/a0017736)
- Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, 62(3), 172-192. doi: [10.2307/2112866](https://doi.org/10.2307/2112866)
- Lewis, P. A. W., & Orav, E. J. (1989). *Simulation methodology for statisticians, operations analysts, and engineers* (Vol. 1). Pacific Grove CA: Wadsworth and Brooks/Cole.
- Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2), 127-137. doi: [10.1046/j.0039-0402.2003.00252.x](https://doi.org/10.1046/j.0039-0402.2003.00252.x)
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 85-91. doi: [10.1027/1614-1881.1.3.86](https://doi.org/10.1027/1614-1881.1.3.86)

- MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, 36(4), 611–637. doi: [10.1207/s15327906mbr3604_06](https://doi.org/10.1207/s15327906mbr3604_06)
- Maeda, Y. (2007). *Monte Carlo evidence regarding the effects of violating assumed conditions of two-level hierarchical models for cross-sectional data*. Unpublished doctoral dissertation, University of Minnesota Twin Cities.
- Matyas T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, 23(3), 341–351. doi: [10.1901/jaba.1990.23-341](https://doi.org/10.1901/jaba.1990.23-341)
- Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, 25(3), 271–284. doi: [10.3102/10769986025003271](https://doi.org/10.3102/10769986025003271)
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287–312. doi: [10.1207/S15328007SEM0802_7](https://doi.org/10.1207/S15328007SEM0802_7)
- R Core Team. (2016). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from: <http://www.R-project.org/> (Version 3.3.1).
- Ramsey, P. H., & Ramsey, P. P. (2009). Power and Type I errors for pairwise comparisons of means in the unequal variances case. *British Journal of Mathematical and Statistical Psychology*, 62(2), 263–281. doi: [10.1348/000711008X291542](https://doi.org/10.1348/000711008X291542)
- Raudenbush, S. J., & Bryk A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, 12(3), 241–269. doi: [10.2307/1164686](https://doi.org/10.2307/1164686)
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). New York, NY: Sage Publications.
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45(2), 283–288. doi: [10.1111/j.2044-8317.1992.tb00993.x](https://doi.org/10.1111/j.2044-8317.1992.tb00993.x)
- Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, 69(347), 682–689. doi: [10.2307/2286002](https://doi.org/10.2307/2286002)
- Sanchez, S. M. (2007). Work smarter, not harder: Guidelines for designing simulation experiments. In Henderson, S. G., Biller, B., Hsieh, M. H., et al. (Eds.)

- Proceedings of the 2007 Winter Simulation Conference*, 9-12 December 2007, Washington DC. pp. 84-94. doi: [10.1109/WSC.2007.4419591](https://doi.org/10.1109/WSC.2007.4419591)
- Santner, T., Williams, B., & Notz, W. (2003). *The design and analysis of computer experiments*. New York, NY: Springer Verlag. doi: [10.1007/978-1-4757-3799-8](https://doi.org/10.1007/978-1-4757-3799-8)
- Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of American Statistical Association*, 62(318), 626–633. doi: [10.2307/2283989](https://doi.org/10.2307/2283989)
- Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35(2), 137–167. doi: [10.1207/S15327906MBR3502_1](https://doi.org/10.1207/S15327906MBR3502_1)
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17(4), 510–550. doi: [10.1037/a0029312](https://doi.org/10.1037/a0029312)
- Statistics and Science: A Report of the London Workshop on the Future of the Statistical Sciences* (2014). [Technical Report]. London, England. Retrieved from <http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf>.
- Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and its Applications*, 2, 1–19. doi: [10.1146/annurev-statistics-010814-020127](https://doi.org/10.1146/annurev-statistics-010814-020127)
- Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99(1), 90-99. doi: [10.1037/0033-2909.99.1.90](https://doi.org/10.1037/0033-2909.99.1.90)
- Viana, F. A. C. (2013). *Things you wanted to know about the Latin hypercube design and were afraid to ask*. 10th World Congress on Structural and Multidisciplinary Optimization, Orlando, FL.
- Wilcox, R. R. (2009). Robust ANCOVA using a smoother with bootstrap bagging. *British Journal of Mathematical and Statistical Psychology*, 62(2), 427–437. doi: [10.1348/000711008X325300](https://doi.org/10.1348/000711008X325300)
- Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics—Simulation and Computation*, 15(4), 933-943. doi: [10.1080/03610918608812553](https://doi.org/10.1080/03610918608812553)

EXPERIMENTAL DESIGN AND DATA ANALYSIS IN SIMULATION

Willett, J. B., & Singer, J. D. (1988). Another cautionary note about : It's use in weighted least squares regression analysis. *The American Statistician*, 42(3), 236-238. doi: [10.2307/2685031](https://doi.org/10.2307/2685031)

Zhang, D. (2005). *A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling*. Unpublished doctoral dissertation, Texas A & M University.

Appendix A: R code for Figures 1 - 3

Full Factorial Design (Figure 1)

```
# Libraries needed
library(ggplot2)
library(lhs)
library(scales)
grid.full <- expand.grid(f1 = c(10, 20, 30, 40, 50),
                        f2 = c(18, 29, 34, 44, 60, 68))
# Plot the full factorial design
ggplot(grid.full, aes(x = f1, y = f2)) +
  geom_point(size = 4) +
  xlab("Number of clusters") +
  ylab("Within-cluster sample size") +
  theme_bw()
```

Latin Hypercube Design (Figure 2)

```
# Set seed for reproducibility
set.seed(59832)
# Sample from a [0, 1] LHS design using lhs package
grid.lhd <- maximinLHS(n = 30, k = 2)
# Name columns of grid
colnames(grid.lhd) <- c("f1", "f2")
# Rescale grid to obtain the range of values factor 1 and factor 2 have
in the manuscript
grid.lhd[, 1] <- rescale(grid.lhd[, 1],
                        to = c(10, 50),
                        from = c(0, 1))
grid.lhd[, 2] <- rescale(grid.lhd[, 2],
                        to = c(18, 68),
                        from = c(0, 1))
# Convert the grid to a data frame
grid.lhd.data <- as.data.frame(grid.lhd)
# Plot the LHD
ggplot(grid.lhd.data, aes(x = f1, y = f2)) +
  geom_point(size = 4) +
  xlab("Number of clusters") +
```

```
ylab("Within-cluster sample size") +  
theme_bw()
```

Full Factorial Design versus Latin Hypercube Design (Figure 3)

```
# Create variable to identify the experimental design  
grid.full$factor_data <- c(1)  
grid.lhd.data$factor_data <- c(2)  
# Combine both data sets  
data.all <- rbind(grid.full, grid.lhd.data)  
# Create factor variable for experimental design  
data.all$factor_data <- factor(data.all$factor_data, levels = c(1,2),  
labels = c("Full factorial", "LHD"))  
# Plot both experimental designs  
ggplot(data.all, aes(x = f1, y = f2, shape = factor_data)) +  
  geom_point(size = 4) +  
  scale_shape_manual(values=c(1,17)) +  
  xlab("Number of clusters") +  
  ylab("Within-cluster sample size") +  
  theme_bw() +  
  theme(legend.position = "bottom", legend.title = element_blank())
```

The Regression Smoother LOWESS: A Confidence Band That Allows Heteroscedasticity And Has Some Specified Simultaneous Probability Coverage

Rand Wilcox

University of Southern California
Los Angeles, CA

Many nonparametric regression estimators (smoothers) have been proposed that provide a more flexible method for estimating the true regression line compared to using some of the more obvious parametric models. A basic goal when using any smoother is computing a confidence band for the true regression line. Let $M(Y|X)$ be some conditional measure of location associated with the random variable Y , given X and let x be some specific value of the covariate. When using the LOWESS estimator, an extant method that assumes homoscedasticity can be used to compute a confidence interval for $M(Y|X=x)$. A trivial way of computing a confidence band is to compute confidence intervals for K covariate values, each having probability coverage $1 - \alpha$. But an obvious concern is that the simultaneous probability coverage can be substantially smaller than $1 - \alpha$. A method is suggested for dealing with this issue that allows heteroscedasticity and simultaneously performs better than the Bonferroni method or the Studentized maximum modulus distribution.

Keywords: nonparametric regression, confidence band, heteroscedasticity

Introduction

Let $M(Y|X)$ be some conditional measure of location associated with the random variable Y , given X . Nonparametric regression estimators provide an approach to estimating $M(Y|X)$ that deal with curvature in a flexible manner beyond the more obvious parametric models that might be used. Numerous nonparametric regression estimators have been derived and their practical importance is well

Rand Wilcox is a professor at the University of Southern California. His main interests are robust methods. Email him at rwilcox@usc.edu.

established (e.g., Hastie & Tibshirani, 1990; Efromovich, 1999; Eubank, 1999; Fan & Gijbels, 1996; Fox, 2000; Green & Silverman, 1993; Györfi et al., 2002; Härdle, 1990; Wilcox, 2017).

The goal in this paper is to examine methods for computing a confidence interval for $M(Y|X)$ based on the smoother derived by Cleveland (1979), generally known as LOWESS, in manner that allows heteroscedasticity and provides simultaneous probability coverage $1 - \alpha$ for K values of the independent variable, where K is relatively large. From a robustness point of view, LOWESS is important because it includes a method that down-weights outliers among the dependent variable Y .

Let x be some specific value for the independent variable X . Assuming homoscedasticity, a method for computing a confidence interval for $M(Y|X=x)$ has been derived (Cleveland et al., 1992), which has been implemented via the R function `lowess`. So it is a trivial matter to compute a $1 - \alpha$ confidence interval for a collection of values for the covariate, say $x_1, \dots, x_K$. It is evident, however, that the simultaneous probability coverage will, in general, be substantially smaller than $1 - \alpha$. And there is the added concern that when in fact there is heteroscedasticity, an incorrect estimate of the standard error is being used.

Here, heteroscedasticity is addressed with a bootstrap estimate of the standard error of $\hat{\theta}_k$, where $\hat{\theta}_k$ is the estimate of $M(Y|X=x_k)$ ($k = 1, \dots, K$) based on LOWESS.

A simple way of achieving simultaneous probability coverage greater than or equal to $1 - \alpha$ is to compute a $1 - \alpha/K$ confidence interval for each of K covariate values of interest. That is, use the Bonferroni method. Another strategy is to use the Studentized maximum modulus distribution. But both of these strategies are too conservative meaning that the actual probability coverage will be substantially larger than the nominal level, in which case the widths of the confidence intervals will be larger than necessary. The strategy here is to find an adjustment for the confidence intervals that achieves simultaneous probability coverage under normality and homoscedasticity, and then study how well the method performs when dealing with non-normality and heteroscedasticity. The method for adjusting the confidence intervals has certain similarities to using a Studentized maximum modulus distribution, but it differs in ways that will be fairly evident.

LOWESS

Consider the basic strategy used by LOWESS. Not all of the many computational details are provided here, which are summarized by Cleveland et al. (1992). The main goal is to provide some sense of how the span is used and determined. The choice for the span turns out to play an important role given the goal of computing confidence intervals having simultaneous probability coverage $1 - \alpha$.

Given the goal of estimating $M(Y|X = x_k)$, let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample and let $\delta_i = |X_i - x_k|$. Next, sort the δ_i values and retain the pn pairs of points that have the smallest δ_i values, where p is a number between 0 and 1. The value of p represents the proportion of points used to predict Y and is generally referred to as the span. Let δ_m be the maximum value of the δ_i values that are retained. Set

$$Q_i = \frac{\delta_i}{\delta_m}$$

If $0 \leq Q_i < 1$, set $w_i = (1 - Q_i^3)^3$, otherwise set $w_i = 0$. Finally, use weighted least squares to predict Y using $w_1, \dots, w_n$ as weights. Because the weights change with X , generally a different regression estimate of Y is used when the value of X is altered.

There are refinements beyond the computational steps just described, but for the brevity the many details are omitted. As previously noted, the method includes the ability of down weighting outliers among the independent variable Y . The main point here is that the choice for the span, p , will be found to play a crucial role.

Description of the Proposed Method

Let $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$ be a bootstrap sample, which is obtained by resampling with replacement n points from $(X_1, Y_1), \dots, (X_n, Y_n)$. Let $\hat{\theta}_k^*$ be the estimate of $M(Y|X = x_k)$ based on this bootstrap sample. Repeat this process B times yielding $\hat{\theta}_{kb}^*$ ($b = 1, \dots, B$). From basic principles (e.g., Efron & Tibshirani, 1993), an estimate of the squared standard error of $\hat{\theta}_k$ is

REGRESSION CONFIDENCE BAND

$$s_k^2 = \frac{1}{B-1} \sum \left(\hat{\theta}_{kb}^* - \bar{\theta}_k \right)^2,$$

where $\bar{\theta}_k = \sum \hat{\theta}_{kb}^* / B$.

Here, two strategies for choosing $x_1, \dots, x_K$ were considered. The first used $K = n$ values based on all of the observed values of the independent variable X .

The adjusted confidence intervals, based on the basic strategy described momentarily, performed well in simulations for some situations, but not others, so this approach was abandoned.

To describe the second strategy, let M be the usual sample median based on $X_1, \dots, X_n$, let $L = M - 1.5 \text{ } MADN$ and $U = M + 1.5 \text{ } MAD$, where MAD is the median of

$$|X_1 - M|, \dots, |X_n - M|$$

and $MADN$ is MAD divided by 0.6745. To add perspective, it is noted that under normality, $MADN$ estimates the standard deviation. Then $x_1, \dots, x_K$ are taken to be K values evenly spaced between L and U , inclusive. Here the focus is on $K = 25$.

Now focus on a single value of the independent variable, x_k , and note that for some specified constant θ_0 ,

$$H_0 : \theta_k = \theta_0 \tag{1}$$

can be tested using the test statistic

$$T_k = \frac{\hat{\theta}_k - \theta_0}{s_k}, \tag{2}$$

where the null distribution is taken to be a Student's T distribution with degrees of freedom as indicated by Cleveland et al. (1992), which is computed by the R function `loess`. Let p_k be the resulting p -value, let

$$p_m = \min(p_1, \dots, p_K)$$

and let p_α be the α quantile of the distribution of p_m . As is evident, if the null hypothesis given by (1) is rejected if and only if $p_k \leq p_\alpha$, the probability of one or

more Type I errors is α . In terms of confidence intervals, if a $1 - p_\alpha$ confidence interval is computed for each θ_k , the simultaneous probability coverage is $1 - \alpha$.

Simulations are used to estimate p_α when dealing with independent standard normal distributions. More precisely, generate n pairs of points from a bivariate normal distribution having correlation zero, perform the K tests as just described, and determine p_m , the minimum p -value among these K tests. This process is repeated N times yielding say $p_{m1}, \dots, p_{mN}$, in which case p_α can be estimated with some quantile estimator. Here, the Harrell and Davis (1982) estimator is used with $N = 4,000$. For convenience, this method for computing confidence intervals will be called method C henceforth.

Shown in Table 1 are some estimates of p_α when $\alpha = 0.05$ and the sample size n ranges between 50 and 2,000. Note that based on the Bonferroni method with $K = 25$, each of the K tests would be performed at the 0.002 level. If, for example, the Studentized maximum modulus distribution is used with fifty degrees of freedom, in effect p_α is taken to be 0.0022. Generally, using the method described here will result in shorter confidence intervals. Roughly, the reason is that the T_k values are highly correlated, which is taken into account when computing p_α . Also note that as n increases, initially the estimates of p_α decrease and then they increase. The reason for this is unclear.

Table 1. Estimates of p_α based on 4,000 replications

n	p_α
30	0.00360
50	0.00266
70	0.00240
100	0.00288
150	0.00300
200	0.00354
300	0.00387
500	0.00440
1000	0.00408
2000	0.00451

Simulation Results

Simulations were used to study the small-sample properties of method C. Data were generated based on the model

REGRESSION CONFIDENCE BAND

$$Y = X^a + \lambda(X)\epsilon$$

for $a = 0, 1$ and 2 . Both X and ϵ were generated from one of four types of distributions: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, both the error term and the distribution of the independent variable were taken to be one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution and $g > 0$, then

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2/2)$$

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

$$W = Z \exp(hZ^2/2).$$

The four distributions used here were the standard normal ($g = h = 0.0$), an asymmetric heavy-tailed distribution ($h = 0.2, g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0, g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 2. Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Table 3 summarizes the simulation results for method C based on $a = 0$, sample sizes 50 and 100, and when the default value for the span is used, namely $p = 2/3$. Shown are estimates of α when the goal is to achieve simultaneous probability coverage $1 - \alpha = 0.95$. Similar results were obtained with $a = 1$ and $a = 2$. Bradley (1978) has suggested that as a general guide, when computing a

RAND WILCOX

0.95 confidence interval, at a minimum the actual probability coverage should be between 0.925 and 0.975. All of the estimates satisfy this basic criterion.

Table 3. Estimates of α when the goal is to achieve simultaneous probability coverage $1 - \alpha = 0.95$.

g	h	n	VP1	VP2	VP3
0.0	0.0	50	0.050	0.061	0.047
0.0	0.0	100	0.050	0.051	0.043
0.0	0.2	50	0.032	0.035	0.026
0.0	0.2	100	0.029	0.034	0.026
0.2	0.0	50	0.061	0.056	0.049
0.2	0.0	100	0.053	0.067	0.045
0.2	0.2	50	0.055	0.050	0.037
0.2	0.2	100	0.033	0.065	0.046

Method C continues to perform well with $n = 200$ and $n = 300$. But with $n = 500$ and when sampling from a skewed distribution, it can be unsatisfactory when there is heteroscedasticity. That is, the estimates of α exceed 0.075. Increasing the number of bootstrap samples to 400 improved matters in some cases. But what was more effective was reducing the span. For $n = 500$, reducing the span to $p = 0.5$ yielded estimates less than 0.05 for all situations considered. But under normality and homoscedasticity, the estimate was 0.016.

Illustration

Method C is illustrated using data from the Well Elderly 2 study (Clark et al., 2012) that dealt with an intervention program aimed at improving the physical and emotional wellbeing of older adults. A portion of the study focused on the association between the cortisol awakening response (CAR) and a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). CAR refers to the change in cortisol concentration that occurs 30-60 minutes after waking from sleep. A CESD score greater than 15 is regarded as an indication of mild depression. A score greater than 21 indicates the possibility of major depression.

Figure 1 shows the estimate of the regression line as well as a collection of confidence intervals having simultaneous probability coverage approximately equal to 0.95. (Leverage points were removed.) The horizontal dotted line in Figure 1 corresponds to CESD = 15. So Figure 1 indicates that for CAR values between -0.2 and 1.5 , after intervention, a reasonable decision is that the typical

REGRESSION CONFIDENCE BAND

participant does not have any indication of mild depression. Outside this interval, it is unclear the extent to which this is the case.

Figure 1. Confidence intervals for the typical CESD score using the CAR as the independent variable. The horizontal dotted line corresponds to CESD = 15. (CESD values greater than 15 are considered an indication of mild depression.)

Conclusion

Method C offers a more satisfactory way of computing confidence intervals compared to the simple approach of computing $1 - \alpha$ confidence intervals for each value of the independent variable of interest. The method performed well in simulations, in terms of achieving estimates of α less than 0.075 for $n \leq 500$, provided the span is chosen appropriately. However, there is room for improvement because as the sample size increases, the actual probability coverage becomes increasingly unstable in terms of how the data are generated. Avoiding estimates of α greater than 0.075 can be achieved by choosing the span to be sufficiently small, but at the expense of estimates less than 0.025 when there is normality and homoscedasticity.

A broader issue is whether some variation of method C can be used in conjunction with other smoothers. One of the many smoothers of interest is the running interval smoother (e.g., Wilcox, 2017) because it provides a simple and effective method for dealing with situations where $M(Y|X)$ is any robust measure of location of interest. Preliminary results indicate that an adjustment of the confidence intervals, similar to what was used here, is not straightforward. The details of how best to proceed are under investigation.

Finally, the R function `lp1otCI` applies method C and has been added to the library of R functions stored at Dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm in the file `Rallfun-v32`.

References

- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)
- Clark, F., Jackson, J., Carlson, M., et al. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, 66, 782-790. doi: [10.1136/jech.2009.099754](https://doi.org/10.1136/jech.2009.099754)
- Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829-836. doi: [10.2307/2286407](https://doi.org/10.2307/2286407)
- Cleveland, W. S., Grosse, E. & Shyu, W. M. (1992) Local regression models. In J. M. Chambers & T. J. Hastie, Eds. *Statistical Models in S*. New York: Wadsworth & Brooks/Cole (Springer-Verlag), Chapter 8.
- Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer-Verlag. doi: [10.1007/b97679](https://doi.org/10.1007/b97679)
- Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.
- Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Boca Raton, FL: CRC Press.
- Fox, J. (2000). *Multiple and Generalized Nonparametric Regression*. Thousands Oaks, CA: Sage. doi: [10.4135/9781412985154](https://doi.org/10.4135/9781412985154)

REGRESSION CONFIDENCE BAND

- Green, P. J. & Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Boca Raton, FL: CRC Press.
- Györfi, L., Kohler, M., Krzyżk, & A. Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer Verlag. doi: [10.1007/b97848](https://doi.org/10.1007/b97848)
- Hall, P. & Sheather, S. J. (1988). On the distribution of a Studentized quantile. *Journal of the Royal Statistical Society, B*, 50, 380-391.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.
- Härdle, W. (1990). *Applied Nonparametric Regression (Econometric Society Monographs No. 19)*. Cambridge, UK: Cambridge University Press.
- Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3), 635-640. doi: [10.1093/biomet/69.3.635](https://doi.org/10.1093/biomet/69.3.635)
- Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
- Hettmansperger, T. P. & McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.
- Hettmansperger, T. P., & Sheather, S. J. (1986). Confidence intervals based on interpolated order statistics. *Statistics & Probability Letters*, 4(2), 75-79. doi: [10.1016/0167-7152\(86\)90021-0](https://doi.org/10.1016/0167-7152(86)90021-0)
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. New York: Wiley, pp. 461-514. doi: [10.1002/9781118150702.ch11](https://doi.org/10.1002/9781118150702.ch11)
- Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics* (2nd Ed.). New York: Wiley. doi: [10.1002/9780470434697](https://doi.org/10.1002/9780470434697)
- Sheather, S. J. & McKean, J. W. (1987). A comparison of testing and confidence intervals for the median. *Statistics & Probability Letters*, 6(1), 31-36. doi: [10.1016/0167-7152\(87\)90055-1](https://doi.org/10.1016/0167-7152(87)90055-1)
- Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. doi: [10.1002/9781118165485](https://doi.org/10.1002/9781118165485)
- Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing* (4th Ed.). San Diego, CA: Academic Press.

Regular Articles

The Impact of Predictor Variable(s) with Skewed Cell Probabilities on Wald Tests in Binary Logistic Regression

Arwa Alkhalaf
University of British Columbia
Vancouver, BC, Canada

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

A series of simulation studies are reported that investigated the impact of a skewed predictor(s) on the Type I error rate and power of the Wald test in a logistic regression model. Five simulations were conducted for three different regression models. A detailed description of the impact of skewed cell predictor probabilities and sample size provide guidelines for practitioners wherein to expect the greatest problems.

Keywords: logistic regression, skewed cell probability, simulation study, categorical predictor, skewed predictor

Introduction

Logistic regression modeling is growing in popularity in psychological and educational research (Cohen, Cohen, West, & Aiken, 2013; Tabachnick & Fidell, 2013). In these disciplines, data analysts commonly encounter skewed predictor variables: either categorical predictor variables that reflect skewed cell probabilities or skewed continuous predictors. The purposes of this paper are to describe the issues surrounding skewed predictors and to document their consequences on parameter estimation, as well as on the Types I and II error (and statistical power) of their Wald tests.

The skewness of predictors is rarely discussed in statistical treatments of logistic regression for educational and psychological researchers. Moreover, while the mathematical statistics literature does mention skewed variables, as will be seen below, they are typically used as a motivation for employing alternative estimators, test statistics, and analysis strategies—which is quite reasonable given

Dr. Alkhalaf is a graduate of the University of British Columbia. Email her at arwaalkhalaf@hotmail.com. Dr. Zumbo is the Paragon UBC Professor of Psychometrics and Measurement and an editor of this journal. Email him at bruno.zumbo@ubc.ca.

the purpose of those studies. What is not found in either the methodological or the mathematical/statistical literature is a detailed documentation of the impact of predictor skewness on the convergence of estimators, and on the Types I and II error and statistical power of the hypothesis tests. There is no detailed information to guide researchers on the impact of skewed predictors in logistic regression. With the aim of filling this gap in the literature, consider the results of five simulation studies that aim to provide a comprehensive investigation of the convergence in maximum likelihood estimation (MLE) of the regression parameters (b-weights) and the operating characteristics of the Wald statistic for predictors in logistic regression with skewed cell probabilities. Note operating characteristics is used to here refer to the Type I and II error rates and statistical power of a hypothesis test (Ferris, Grubbs, & Weaver, 1946).

Problematic Data Structures: Sparse Tables, Skewness, and Separation in Logistic Regression, and Statement of the Problem

There are very few discussions of the issue of skewed or unequal cell probabilities in the logistic regression literature (Jennings, 1986; Larntz, 1978). A review of the literature on broad categorical data analysis reveals three types of data patterns that provide a context for issues potentially related to the impact of skewed cell probabilities and hence may offer insights on the problem.

Three types of data patterns.

To understand problematic data patterns, it is necessary to be able to visualize the data. In addition to the conventional data matrix (in which rows are participants and columns are variables), categorical data may be displayed as a multi-way table in which the cells are counts of occurrences of the corresponding row and column elements. The former display allows insight on the variety of covariate patterns, whereas the latter allows learning about potential small sample sizes in the cells of the table that result in sparse data. The statistical literature on categorical data analysis uses both of these data visualization tools, though it focuses more on cross-tabulation and the language of cell counts, and provides a few descriptions of problematic data structures and an extensive number of remedies (i.e. smoothing techniques and robust estimation procedures).

Sparse tables are a common concern in categorical data analysis. From the perspective of the cross-tabulation of the data, one is fitting a logistic regression

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

model with data in this table. In his discussion of empty cells and sparse tables, Agresti (2002) described them as contingency tables having small or zero cell counts. Sparse tables may occur when the sample size is small, when a variable contains a large number of categories, or when a model has many predictor variables and hence a high dimensional multi-way table. A sparse table in a logistic regression with a dichotomous predictor can be thought of as a two-way (row-by-column) table that has a similar format to Table 1. In Table 1, even though the outcome variable (Y) is symmetrically distributed and the predictor variable has a small skew in the marginal cell counts, there is a cell with zero occurrences—an empty cell. As such, it is clear that the marginal distributions are not necessarily indicative of the covariate pattern in the data.

Table 1. Two-way table with a zero count cell, an example of a sparse table or quasi-complete separation.

		<i>X</i>		
		<i>0</i>	<i>1</i>	<i>Total</i>
<i>Y</i>	<i>0</i>	40	10	50
	<i>1</i>	0	50	50
<i>Total</i>		40	60	100

The issue of separation was first introduced by Day and Kerridge (1967) to describe a problematic data configuration between the categorical outcome and predictor variables that negatively affects the MLE. Refining these earlier findings, Albert and Anderson (1984) identified three types of data configurations that may affect estimation: complete separation, quasi-complete separation, and overlap. They mathematically proved that although overlap yields a finite and unique solution, MLEs do not exist for the other two data patterns, although it was left to future researchers to develop new techniques to overcome this obstacle (e.g., Barreto, Russo, Brasil, & Simon, 2014; Gordóvil-Merino, Guàrdia-Olmos, & Perú-Cebollero, 2012; Heinze & Puhr, 2010; Míndrilã, 2010; Rousseeuw & Christmann, 2003).

Although skewness is a term rarely used in categorical data analysis, following Larntz's (1978) classic study, skewed probabilities is adopted to describe the row (or column) marginal distribution of the categorical predictor variables. This phrase has two uses in the statistical literature of interest. Larntz considered a case where the binary or multinomial predictor variables have an implicit order (help grade, or otherwise an ordered categorical variable of help), and where the marginal probabilities of the predictor variable are therefore

distributed in a skewed manner. Jennings (1986) did not use the phrase skewed probabilities, but instead described the marginal probabilities of the outcome variable as equal or unequal. The characterization Larntz described is more in line with the one adopted here, in good part because it corresponds more closely to how data analysts in education and psychology conceptualize such distributions.

Skewness here is used in lieu of sparse tables or separation for a few reasons. Sparse tables and separation are descriptive of a relationship between two variables (i.e., outcome and predictor), whereas skewness in the probability of occurrence for the categories in a predictor is a unique descriptor of one variable. Also, the term skewness can be generalized to the continuous cases. Because the interest is in the skewness of the predictor variable, the methods in conducting this study reflect this concept. In severe cases of skewed probabilities, however, sparse tables or separation may occur. Nevertheless, similar to the example shown in Table 1, a sparse table or separation does not indicate skewed marginal probabilities in a variable.

Relatively little is known about the impact of skewed probabilities on later statistical decisions of a logistic regression model. Therefore, if the skewness in the probabilities of a predictor is not severe enough to disrupt the MLEs in terms of convergence, to what extent could a researcher trust the test results and make valid decisions? The information in Table 2 represents an example of this problem, wherein the predictor X 's probability of obtaining category 0 is nine times more likely to occur than category 1, all while the probability of obtaining both categories in the outcome variable Y is approximately 0.5. The cell counts for the adjacent cells of $(X = 0, Y = 1)$ and $(X = 1, Y = 1)$ are very different. The question is, although estimation will yield a finite solution, to what extent are these estimates to be trusted? Is there bias? How large or small are the standard errors? And ultimately, how much can we trust the results of test statistics such as the Wald test?

Table 2. An example of the data structure examined in this study.

		<i>X</i>		<i>Total</i>
		<i>0</i>	<i>1</i>	
<i>Y</i>	<i>0</i>	89	18	107
	<i>1</i>	91	2	93
<i>Total</i>		180	20	200

What Is Known To Date

Skewed probabilities of a categorical variable

Although we know of no studies that have investigated the skewness of predictor variables in logistic regression, there are three others on related issues.

- 1) Jennings (1986) examined the impact of skewed probabilities (in the outcome variable) in a dichotomous logistic regression, where one category in the outcome variable was more likely to occur than the other. The MLEs of the parameter coefficients are upwardly biased as the cell with the lowest count in the row-by-column table becomes smaller. As a result, Jennings introduced a measure that detects inadequacies in estimation.
- 2) Larntz (1978) focused on the case of goodness of fit of binary and multinomial variables with two- and three-way tables and compared the performance of three multinomial goodness-of-fit statistics with varying sample sizes and degrees of skewness of cell probabilities. Working particularly with small samples because they often generate sparse tables, Larntz used a Monte Carlo simulation to induce skewness in the probabilities in the binary and multinomial variable. It was found that the fit statistics generally performed well.
- 3) In a study aimed to find a solution to the separation problem, Anderson and Richardson (1979) conducted a simulation study to investigate the effectiveness of a bias reduction method within MLEs. The recognition of potential skewness in the data set was interesting. They stated, “the distribution of the maximum likelihood estimators would be skew, particularly when the number of sample points from at least one population was disproportionately small” (p. 72). Because simulating complete separation or a cell with zero frequency would result in estimates that are extremely large (characterized as $\pm\infty$), these were eliminated, while only those data sets that were “acceptable” were included (p. 74).

Separation and MLE

Viewing the impact of skewed predictor cell probabilities from the different but potentially related lenses of separation and sparse tables resulting from particular data configurations (Anderson & Richardson, 1979; Jennings, 1986; Larntz, 1978), we predict that when these probabilities are skewed, the Type I error rate will be deflated and effect sizes will, in some cases, be inflated and may be infinite, however, the extent and under what conditions are unknown. Therefore, given the lack of an analytic solution, computer simulation experiments are needed to more fully explore the impact of predictor skewness.

There will be cases in this simulation when separation is inevitable—that is, when the sample size is small and the predictor variable is highly skewed. More generally, separation is caused by a linear combination of continuous or dichotomous predictors that perfectly separates events from non-events (the 1 and 0 of the outcome variable). Complete separation occurs when one or more of a model's predictors perfectly predict the outcome variable, therefore, no variance is left to be explained in the outcome variable by the model's other predictors. More commonly, quasi-complete separation occurs when only one covariate pattern has a zero count—expressed differently, when, for example, only one cell of the implied 2×2 table of X and Y is empty (Zorn, 2005, p. 161). Under such conditions, the parameter estimate for the separating variable will also be infinite, but the model's other predictors may remain unaffected (Zorn, 2005). Both complete and quasi-complete separation may be present in our simulation experiment as a by-product of the data configuration.

The problem with small samples and separated data lies in the estimation process—that is, a finite and unique MLE in logistic regression may not exist. The resulting estimates of the log odds ratios are biased, and the bias increases as the ratio of the number of observations to the number of parameters decreases (Cordeiro & McCullagh, 1991). The astronomically large estimates produced indicate that a variable perfectly predicts the outcome, which is in essence very desirable, but is an artifact of the data configuration. However, in small data sets, we must assume that separation is not due to truly infinite estimates, but is instead caused by random variation or the nature of the data configuration.

What is even more interesting is the effect of separation on test statistics, specifically the Wald test. Hauck and Donner (1977) demonstrate that for any sample size, the Wald test statistics decrease to zero as the distance between the parameter estimates and null values increases. Consequently, in all tests for model validation, validation variables are biased and the confidence intervals of the parameter estimates and the odds ratio are not efficient. In cases of separation, the

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

distance between the parameter estimates and their null value is very large, resulting in a nonsignificant Wald statistic.

In a simulation study conducted by Peduzzi, Concato, Kemper, Holford, and Feinstein (1996), in which they examined the effects of the number of events per predictor variable in a logistic regression model, it was found the MLE did not converge with two and five events per predictor. Moreover, when the MLE did converge, the Type I error was deflated (i.e., became more conservative), the power decreased, and the empirical distribution of the Wald statistic was not normally distributed. These problems did not exist with 10 or more events per predictor. However, Barreto et al. (2014) found that the Wald test can detect which variables are individually significant, but fails to determine the significance of the variable that presents separation. The maximum likelihood estimates become inefficient, providing inflated variances.

The Wald test was criticized for its limitations under both ideal (Pawitan, 2000) and problematic circumstances (Fears, Benichou, & Gail, 1996; Gregory & Veall, 1986; Lütkepohl & Burda, 1997; Vaeth, 1985). However, it is still widely reported and used to this day. In a recent review (Alkhalaf, 2014) of 323 articles in higher education research that use logistic regression, it was found that all of them reported the significance of parameters via the Wald test or z-statistic. Moreover, widely used software packages such as R, SAS, Stata, and SPSS provide the Wald statistic as output. For these reasons, we focus on the Wald test in this study.

Simulation Studies

The results of five simulation studies are reported, organized around three logistic regression models.

- The first model examined simple logistic regression with skewed probabilities of a dichotomous predictor. The results of two studies are reported. The first focused on the quality of the parameter coefficient estimates, including the convergence rates of the MLEs, as well as Type I error. The second simulation study investigated statistical power.
- The second model considered skewness in simple logistic regression with a continuous predictor. Because this model was included to check the generalizability to a continuous predictor case (rather than

a categorical predictor), only the MLE convergence and Type I error rate were investigated.

- The final model included two simulations that explored multiple logistic regression with skewed cell probabilities of two dichotomous predictors. Like the first model, the first simulation study focused on the convergence rates of the MLEs and Type I error, and the second on statistical power.

General methods

In this series of studies Monte Carlo simulations were used to examine the skewness of a predictor at the population level, meaning what happens when skewness is not a sampling artifact, but is rather the result of a population imbalance of the marginal probabilities of the predictor(s). This is sometimes called naturally occurring skewness. Examples of variables that are naturally skewed in the population include (a) the number of visually impaired undergraduate students in a certain discipline; (b) in clinical, psychological, health, or medical research, the presence of a rare diagnostic ailment; (c) in the social sciences, a large gender imbalance of the participants in a study due to culturally sensitive issues; and (d) as is well known, binary predictors in models 1 and 3 can be interpreted as being a design matrix in an experiment or clinical trial – note that the imbalance in the experiment or clinical trial reflects population imbalance or what is sometimes called unequal cell sample sizes reflecting population characteristics and therefore not due to selection bias or attrition (e.g., [Christensen, 2016](#)).

To directly answer the research question of the effect of a skewed predictor on the eventual statistical conclusions of a logistic regression model, outcome and predictor(s) variables were simulated with varying degrees of skewness, sample size, and predictor type (i.e., dichotomous and continuous). In all cases, the same statistical model that generated the data was fitted to the simulated sample using conventional MLE and Wald tests—that is, all of the models are correctly specified. The focus is on the Type I error rates and statistical power of the Wald test for the predictor(s). As is common practice, the nominal Type I error rate (α) was set at 0.05.

Accordingly, the overall research question can be stated more formally as: What is the empirical Type I error rate and statistical power for the Wald test for a

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

binary logistic regression when the predictor variable(s) has a skewed cell probability?

Model 1: Single Binary Predictor

The first model of interest involves simple logistic regression with one dichotomous predictor:

$$g(y) = \beta_0 + \beta_1 x,$$

where x is a predictor variable with skewed cell probability, β_0 and β_1 are fixed, $g(y)$ is a logit function, and y is a balanced outcome variable. This model acts as a baseline for comparing the results of the forthcoming studies.

Study A: Type I error rates and parameter estimates

Purpose of the study. The purpose of this first simulation experiment is to document the impact of skewed cell probability in a dichotomous predictor variable on the MLE, parameter estimates, and Type I error rate of the Wald test of the β_1 parameter. The outcome variable of the regression model, throughout, is balanced or nearly balanced (i.e., not skewed). A secondary aim is to provide diagnostic information by documenting the situations where skewness may affect decisions and inferences.

Method

Simulation factors. For this simulation, two experimental factors were varied: sample size and skewness of the predictor variable. Sample sizes ranged across 13 levels from 10 to 5000. This large range represents a wide space of sample sizes starting from very small at 10 and 50. Then the range includes sample sizes that are seen more frequently in educational research of 100 to 1,000 in increments of 100. A sample size of 5,000 was included to verify the simulation experiment. The expected probability p of the predictor variable, described in more detail below, ranged from 0.01 to 0.45 across 17 levels. Expected probabilities are directly linked to the degree of skewness as can be seen later in [Model 2](#). The degrees of expected probability range from extremely skewed to non-skewed distributions. These levels of skewed probability were chosen to reflect an array of distributional characteristics. Similar to sample size,

the slight increments in skewness levels provide a wide range of distributional characteristics for variables. In addition, for comparison purposes, the case where the predictor variable is balanced was considered, when the probability of occurrence for both categories is 0.5. The resulting experiment is an 18×13 fully-crossed factorial design involving 234 cells. This large range of sample sizes and skewness levels is necessary to more fully document the impact of skewed cell probabilities.

Simulation procedure. The simulation and analyses were conducted using the R software package. There were 1,000 replications in each cell of the experimental design, resulting in an empirical probability (either a Type I error rate or statistical power) per cell, as well as an empirical representation of the sampling distribution of the parameter estimate. The results based on one thousand replications were compared to 10,000 replications and we found that both yielded the same results in terms of Type I errors, percentage of non-convergences, standard errors, and statistical power. Therefore, there was no marginal gain from the additional replications and we report the results based on one thousand replications herein. For each replication in each cell, the simulation algorithm consists of multiple loops that achieve different purposes. There are a few important steps in this process.

Step 1. The experiment is built upon data generated from a Bernoulli distribution.

$$f(p) = \begin{cases} p, & k=1 \\ 1-p=q, & k=0 \end{cases},$$

with the expected probability $E(x) = p$ and the variance $V(x) = p(1 - p)$. The predictor is randomly drawn from a Bernoulli distribution with a specified sample size and expected probability. Similarly, the outcome variable was randomly chosen from a Bernoulli distribution with the same sample size and an expected probability that is calculated from the model as follows:

- (1) The mean of the Bernoulli distribution is a function of β_0 and β_1 , which are fixed to zero. The intercept term is fixed to zero because the balanced outcome variable results in a natural log of one, which is zero.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

- (2) The logit was calculated where $Logit = \beta_0 + \beta_1 X$ for the simple case.
- (3) The predicted probability was then calculated as $Predicted\ Probability = P/(1 - P) = e^{Logit}/(1 + e^{Logit})$. The predicted probability serves as the expected value for the Bernoulli distribution from which the outcome variable is drawn.
- (4) This process is repeated until the number of replications is complete.

Step 2. All the variables are aggregated in a data frame in preparation for analysis. The generalized linear models (glm) function in R is used to run the logistic regression. The parameter estimates and hypothesis test statistics are stored for each replication. In replications where the estimation does not converge (which is likely in this case due to separation), an N/A is recorded and the simulation outcome (e.g., rejecting the null hypothesis using the Wald test) for that instance in the experimental design is computed from the remaining converging replications in that cell of the simulation experimental design.

Step 3. The final step is to vary the sample size and skewed probability. Each combination of conditions is stored and analyzed separately. The Type I error rates are computed as the number of rejections of the null hypothesis out of the converged 1,000 replications. (To highlight the matter of non-convergence for day-to-day researchers, non-convergence rates based on the 1,000 replications were reported. The resulting Type I error rates are therefore based on the number of convergences and the simulation results are unbalanced (i.e., every empirical Type I error rate is not based on the same denominator). The reported results were compared against the findings wherein the number of replications within a cell continued until 1,000 convergences. The findings did not change, therefore the reported Type I error rates and statistical power results would not change. In the worst cases it took up to one million replications to achieve the 1,000 convergences, so the marginal computational gain was minimal.)

The nominal significance level was 0.05 throughout this study. Therefore, the empirical Type I error is defined as the proportion of times that a true null hypothesis was falsely rejected at a critical value of 0.05 (Mood, Graybill, & Boes, 1974).

Analysis of Type I error. Type I error rate was calculated for each condition. Bradley's (1978) approach was used to compare the nominal and empirical Type I error rates for each condition. Bradley specifies two criteria of robustness, one stringent and one liberal. The stringent criterion is for a robust test, the empirical Type I error should fall within the range of $\alpha \pm 0.1\alpha$, whereas for the liberal criterion the empirical Type I error should lie in a range of $\alpha \pm 0.5\alpha$. Given that a nominal Type I error rate of 0.05 was specified, the interval for an accepted empirical Type I error rate lies between 0.025 and 0.075 for a liberal study and between 0.045 and 0.055 for a stringent one.

Results and conclusions

Number of MLEs that do not converge. An important issue that was encountered was the non-convergence of some replications, as indicated in Table 3. Table 3 depicts the simulation experimental design, wherein each element is the number of non-convergences out of 1,000 replications. For example, for a sample size of 200 and $p = 0.02$, 21 of the 1,000 replications in that cell of the experimental design did not converge using conventional MLE. As expected, in the case of small sample sizes and a high degree of skewness in cell probability (i.e., small values of p), most of the replications did not converge. When the sample size was 10, non-convergence was present even when the predictor was balanced (i.e., $p = 0.5$). With a sample size of 50, the issue of non-convergence diminished as the predictor became less skewed. As the sample size increased, all replications converged, even with high levels of skewness. From the table we can see that a sample size of 500 is sufficient to ensure that the skewness of the predictor variable does not affect estimation for the single predictor model.

The summary statistics reflecting the outcomes of the simulation (i.e., the empirical Type I error rates, odds ratios (ORs), parameter estimates, and standard errors) are computed based solely on the replications that converged. Non-convergent replicates are excluded, mimicking what would go on in daily research practice.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Table 3. Number of non-convergences from 1,000 replications for [Model 1](#).

Probability	Sample Size													
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000	
.01	893	610	388	148	50	21	2	0	0	0	0	0	0	
.02	790	380	154	21	6	1	0	0	0	0	0	0	0	
.03	710	230	58	7	1	0	0	0	0	0	0	0	0	
.04	647	143	19	1	1	0	0	0	0	0	0	0	0	
.05	575	83	7	1	1	0	0	0	0	0	0	0	0	
.06	508	44	5	0	0	0	0	0	0	0	0	0	0	
.07	481	29	2	0	0	0	0	0	0	0	0	0	0	
.08	422	12	1	0	0	0	0	0	0	0	0	0	0	
.09	392	7	0	0	0	0	0	0	0	0	0	0	0	
.10	346	3	0	0	0	0	0	0	0	0	0	0	0	
.15	213	0	0	0	0	0	0	0	0	0	0	0	0	
.20	114	0	0	0	0	0	0	0	0	0	0	0	0	
.25	62	0	0	0	0	0	0	0	0	0	0	0	0	
.30	40	0	0	0	0	0	0	0	0	0	0	0	0	
.35	24	0	0	0	0	0	0	0	0	0	0	0	0	
.40	10	0	0	0	0	0	0	0	0	0	0	0	0	
.45	6	0	0	0	0	0	0	0	0	0	0	0	0	
.50	7	0	0	0	0	0	0	0	0	0	0	0	0	

Type I error rate. [Table 4](#) is structured in the same way as [Table 3](#) and provides Type I error rates for each experimental condition. These Type I error rates are compared against Bradley's criteria, which are shown in [Table 5](#). [Table 4](#) is greyscale coded to highlight two important areas. The darkly shaded area falls below the liberal criterion, while the unshaded area falls within it. Given the interaction of the sample size and the skewness of the cell probability of the predictor, researchers and practitioners should be careful when interpreting results with variable characteristics that are included in the darkly shaded part of the table. As will be shown in the next study, statistical power is greatly affected for these values. Similarly, to consider the shaded area a safe zone, consider statistical power. The Type I error rate rarely met the stringent criterion. Most of the time, it ranged from 0 to 0.044, falling below the lower limit of the stringent threshold of 0.045.

Two baseline conditions were included to serve as a check on our simulation methodology. In the first case, the Type I error rate for different sample sizes was computed for a balanced predictor to establish baselines for comparison with the conditions wherein various levels of probability (i.e., skewness in probability) were manipulated. In the second case, the Type I error rates for various levels of probability were computed for a large sample of 5000. As expected, in both cases, the empirical Type I error rate did not exceed the liberal criterion, for the nominal

level of .05 and hence verifying that the algorithm works as expected. In the balanced case, as shown in the last row of Table 4, all Type I error rates ranged from 0.052 to 0.062, meeting the liberal criterion. Also, the Type I error rates for the sample of 5000 varied from 0.031 to 0.059.

Table 4. Type I error rate for Model 1.

Probability	Sample Size													
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000	
.01	0	0	0	0	0	0	0	0	0	.02	.02	.02	.05	
.02	0	0	0	0	0	.01	.02	.03	.03	.04	.04	.04	.05	
.03	0	0	0	0	.01	.03	.05	.04	.03	.04	.04	.05	.05	
.04	0	0	0	.01	.03	.03	.05	.03	.03	.04	.05	.04	.04	
.05	0	0	0	.02	.04	.04	.05	.04	.04	.03	.05	.05	.03	
.06	0	0	0	.03	.03	.03	.05	.04	.04	.03	.05	.05	.04	
.07	0	0	.01	.04	.04	.04	.05	.05	.04	.04	.05	.05	.05	
.08	0	.01*	.01	.04	.04	.04	.05	.04	.05	.05	.06	.05	.06	
.09	0	.01	.02	.04	.04	.04	.05	.04	.05	.05	.06	.04	.05	
.10	0	.01	.02	.04	.04	.04	.04	.04	.04	.05	.05	.06	.05	
.15	0	.01	.03	.06	.04	.03	.04	.04	.04	.06	.05	.05	.05	
.20	0	.03	.04	.05	.06	.04	.04	.05	.04	.05	.05	.05	.05	
.25	0	.05	.05	.06	.05	.05	.05	.05	.05	.06	.05	.06	.06	
.30	0	.05	.04	.06	.05	.05	.06	.04	.04	.06	.05	.05	.05	
.35	0	.05	.03	.06	.05	.06	.05	.04	.04	.05	.05	.05	.05	
.40	0	.05	.04	.05	.06	.05	.05	.04	.04	.06	.04	.05	.06	
.45	0	.05	.04	.05	.05	.05	.06	.05	.05	.05	.05	.06	.05	
.50	0	.06	.05	.06	.06	.06	.05	.05	.05	.05	.06	.05	.06	

* Rounded to decimal points. **Note:** Cells depicted in grey have deflated Type I error rates, whereas those with no shading meet the adequacy condition using Bradley's criteria (see Table 5).

Table 5. Bradley's criteria.

Bradley's (1978) Criterion	Type I Error Rate
Violates liberal criterion, therefore deflated	$\alpha < 0.025$
Meets the liberal criterion	$0.025 < \alpha < 0.075$
Meets the stringent criterion	$0.045 < \alpha < 0.055$

In general, the Type I error rates ranged from 0 to 0.062, meaning that all conditions met Bradley's liberal criterion. Regardless of the sample size, the rates were consistently deflated with lower probabilities and closer to nominal values as they became more balanced. Sample size plays an important role in MLE and therefore arriving at more precise parameter estimates. For example, a sample of 600 and a probability level of 0.02, results in a Type I error rate of 0.026. On the

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

other hand, as the sample size decreased, the level of skewed probability did not inflate the empirical Type I error rate greatly. For instance, sample sizes of 50 and 200 can tolerate skewed cell probabilities of 0.2 and 0.06, respectively. Of particular note is the tolerance of the skewed probability of the predictor in this model. Even in the most extreme case of skewness (i.e., a probability of 0.01), with the largest sample size (5000), the empirical Type I error rate is at the nominal value.

Effect Size. [Table 6](#) is structured similarly to the previous tables, each element being the average odds ratio (OR) over the replications that converged. The average OR values for small samples and a highly skewed cell probability of the predictor are astronomical with values in the millions—whereas their true value is 1. Clearly, the degree of bias caused by the skewed predictor is very high. In cases where there was bias in the OR estimate, the sampling distribution of the OR was skewed and occasionally contained large gaps in the distribution. Because of the statistical nature of the sampling distribution, it is also useful to examine its median OR in each cell, as shown in [Table 7](#) (Birnbaum, 1964). This is referred to as median-unbiasedness.

Table 6. Average odds ratio, reflecting the widely used “mean unbiasedness.”

Probability	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.05
.02	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.20	1.20	1.20	1.00	1.00
.03	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	≈∞	1.10	1.10	1.10	1.10	1.10	1.00
.04	≈∞	≈∞	≈∞	≈∞	≈∞	1.20	1.20	1.10	1.10	1.10	1.10	1.10	1.00
.05	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.10	1.00	1.10	1.00	1.00
.06	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.10	1.00	1.10	1.00	1.00
.07	≈∞	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00
.08	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00
.09	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	0.99
.10	≈∞	≈∞	≈∞	1.20	1.10	1.10	1.10	1.03	1.02	1.02	1.03	1.01	1.00
.15	≈∞	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
.20	≈∞	≈∞	1.20	1.08	1.05	1.03	1.03	1.01	1.00	1.00	1.00	1.00	1.00
.25	≈∞	≈∞	1.10	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.30	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.35	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	≈∞	1.20	1.10	1.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	≈∞	1.20	1.10	1.10	1.03	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	≈∞	1.20	1.10	1.10	1.00	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00

Note that by ≈∞, we are indicating ORs in the millions.

Table 7. Median odds ratios, reflecting “median unbiasedness” for skewed sampling distributions.

Probability	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	0.00	1.08	0.95	0.98	1.00	0.98	1.00	1.00	1.00	0.99	1.03	0.99	1.01
.02	$\approx\infty$	1.00	0.96	0.95	0.98	0.97	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.03	$\approx\infty$	1.00	1.00	0.97	0.97	0.99	1.00	1.00	0.99	0.98	1.00	1.00	0.99
.04	1.80	1.00	1.00	1.00	0.97	1.01	1.00	1.00	1.00	0.97	1.00	0.99	0.99
.05	1.70	1.00	1.00	1.00	1.00	0.98	1.00	0.98	1.00	0.98	1.00	1.00	0.99
.06	1.70	1.00	1.00	0.97	1.01	0.99	1.00	0.99	1.00	0.98	0.99	1.00	0.99
.07	1.70	1.00	1.00	0.97	1.01	1.00	0.99	0.98	1.00	0.98	0.99	0.99	0.99
.08	1.70	0.92	1.00	1.00	1.00	0.99	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.09	1.50	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00	0.97	1.00	0.99	0.99
.10	1.40	1.00	1.00	1.00	1.02	1.00	1.00	0.99	1.00	0.99	0.99	0.99	0.99
.15	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.98	1.00	0.99	0.99
.20	1.00	1.00	1.03	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
.25	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	1.00	0.99	1.00	1.00	1.00
.30	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.99	0.99	0.99
.35	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99
.40	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99	1.00	1.00
.45	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
.50	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99

Note that by $\approx\infty$, we are indicating ORs in the millions.

Given the skewed nature of the sampling distribution of the OR, the OR medians are closer to the expected value of one. The median is biased upwards when the sample size is 10. The ORs displayed in Table 6 follow the trend in Table 4, wherein as the sample size and probability (i.e., skewness in probability) increase, the estimated ORs are closer to the simulated population value of one. For example, sample sizes of at least 400 perform very well and provide OR estimates closer to the simulated value when the skewed probability is at least 0.04.

Why is the Type I error rate consistently conservative?

The results in Tables 4 through 7 are based on the converged replications. Overall, the simulation agrees with the previous findings on parameter estimates (Peduzzi et al., 1996), that is, with a small sample size and few events per predictor, the average standard error and average slope estimates are highly biased. Figure 1 is a line graph that shows the slope and standard error, where the y-axis is the slope or standard error and the x-axis is the skewed probability of the

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

predictor. The line shades represent different sample sizes. Tables 8 and 9 contain the values from which these graphs were derived.

a. Estimated slope means.

b. Estimated standard error means.

Figure 1. Slope and standard error averages.

Figure 2. Distribution functions for the experimental condition: Sample size = 50, skewed probability = 0.01, 0.25, and 0.5.

As seen in Figure 1 and Table 8, the bias for the slope is both positive and negative when the sample size and highly skewed cell probability. The bias in slope is not as great as the bias in the ORs (as seen in Table 6). Consider a few examples to understand the distribution of the slope parameter estimate, and why it averages out to a small bias. Contrast a small sample size of 50 and a large one of 500 at three levels of skewed cell probability, 0.01, 0.25, and 0.50. The first of these levels represents a highly skewed predictor, the second is moderate, and the last is a balanced probability of both categories in the predictor. Shown in Figure 2 is a stacked density plot for a sample size of 50 and the three levels of skewed probability. For the first level of probability of 0.01, the slope estimate's range is $[-17.58, 18.04]$ with a mean of 0.71, as shown in Table 8. The 25th, 50th and 75th quantiles are $-15.52, 0.083,$ and $15.52,$ respectively. As indicated in Figure 2, the distribution of the slope estimates from this simulation is fragmented into three parts, such that there are no slope estimates that lie between them. Most of the slope estimates were in the range of $[-17.58, -14.75]$; the least number were in the range of $[-1.17, 0.78]$; and the rest, which comprised the last part, ranged from $[14.93, 18.04]$. For the same sample size and a skewed cell probability of 0.25, the shape of the distribution of the slope estimates mostly varies around zero, with a few outliers in the tails. The range is $[-18.42, 18.62]$ and the mean is

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

−0.012. The 25th, 50th, and 75th quantiles are −0.51, $-1.0 \times e^{-16}$, and 0.43, respectively.

Table 8. Average slope in each cell of the simulation design for Model 1.

Probability	Sample Size												
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	-0.36	0.71	-0.44	-0.54	-0.18	-0.30	0.05	0.08	-0.10	0.03	0.10	-0.01	0.00
.02	1.20	0.23	-0.27	-0.35	-0.10	-0.13	0.12	-0.03	-0.01	-0.01	0.04	-0.02	-0.01
.03	1.70	-0.08	-0.24	-0.14	-0.13	-0.04	0.06	-0.01	-0.02	-0.02	0.03	-0.01	0.00
.04	1.30	-0.57	-0.01	0.06	-0.01	0.01	0.01	0.00	-0.01	-0.01	0.02	-0.01	0.00
.05	1.40	-0.43	0.19	0.06	0.01	0.00	0.01	-0.01	0.00	-0.02	0.02	-0.01	0.00
.06	1.10	-0.36	0.15	-0.02	0.01	0.00	0.01	-0.01	0.00	-0.02	0.01	-0.01	-0.01
.07	1.00	-0.46	-0.02	-0.04	0.02	0.01	0.01	-0.01	0.00	-0.02	0.01	-0.01	0.00
.08	1.00	-0.38	-0.09	-0.02	0.02	0.01	0.00	-0.01	0.00	-0.02	0.01	-0.01	0.00
.09	0.90	-0.22	-0.02	0.00	0.02	0.01	0.00	-0.01	-0.01	-0.02	0.01	-0.01	-0.01
.10	0.80	-0.23	-0.07	0.01	0.01	0.01	0.00	-0.01	-0.01	-0.02	0.01	-0.01	0.00
.15	0.30	-0.15	0.02	0.01	0.00	0.01	0.00	0.00	-0.01	-0.01	0.01	0.00	0.00
.20	0.02	-0.11	0.00	0.01	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
.25	-0.04	-0.01	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
.30	0.12	-0.03	0.00	0.01	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
.35	0.24	-0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
.40	0.27	-0.02	-0.01	0.01	0.01	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.00
.45	-0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
.50	-0.13	0.01	0.00	0.01	0.00	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00

For comparison purposes, the distribution of the slope estimates was examined when the probability of the predictor is balanced (i.e., $p = 0.5$). As shown in Table 8, the estimated slopes are close to the simulated values of zero. In Figure 2, the distribution in this experimental condition is nearly symmetrical, with a range of $[-2.59, 2.58]$ and a mean of 0.014. This suggests that the distribution of the simulated slope estimates is disrupted by the skewness in the probability of the predictor.

Consider an example where the sample size is large, in this case 500, and examine the extent to which the distribution of the slope parameter changes with the aforementioned three levels of probability. Figure 3 demonstrates the stacked density plots for the three experimental conditions. For a probability of 0.01, the distribution is fragmented into three parts that cluster around zero. The distribution range is $[-15.7, 15.77]$ and the mean is 0.052. The 25th, 50th, and 75th quantiles are −0.66, 0.008, and 0.696, respectively. For the same sample size and a moderate probability of 0.25, we find that the distribution is symmetrical and nearly resembles a normal distribution. The slope estimates vary close to zero

(the actual value), with a mean of 0.004 and a range of $[-0.63, 0.62]$; the 25th, 50th, and 75th quantiles are -0.15 , -0.005 , and 0.14 , respectively. Finally, when the sample is 500 and the predictor is balanced, the distribution is tighter and varies closer to zero. It has a range of $[-0.57, 0.55]$ with three outliers equal to 1.57, 4.81, and 7.87. The mean, as seen in Table 8, is -0.004 and the 25th, 50th, and 75th quantiles are -0.123 , 0.0003 , and 0.12 , respectively.

Figure 3. Distribution functions for the experimental condition: Sample size = 500, skewed probability = 0.01, 0.25, and 0.5.

This wide range of the slope estimate when the skewed probability of a predictor is small clarifies a few things about the aforementioned small bias, and the largely upward bias of the ORs. Since the ORs are the exponentiation of the slope estimate, slopes with large positive values can create ORs that are in the order of magnitude of tens of millions, whereas negative slopes can result in ORs that tend toward zero. Therefore, the upwardly tending slopes will result in very large bias in the ORs.

Depicted in Table 9 are the average standard errors over replicates of the simulation. These standard errors range from highly biased to unbiased, with the concentration of high bias for small sample sizes and highly skewed cell probabilities (i.e., the top left corner of the table). For example, for sample sizes

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

of 50 to 300 and probability levels of 0.01 through 0.2, the average standard errors are in the thousands and range from [0.24, 4500], as shown in Table 9. As learned from examining the distributions of the slope estimates above, sample sizes of at least 400 perform very well and provide estimates closer to the simulated values when the skewed probability is at least 0.04. This supports the claim that the maximum likelihood estimation is affected by the skewed probabilities of the predictor. That is, even in replications where the MLE produced finite estimates, there was bias in the parameter estimates and standard error. However, as the sample size increases, the estimation becomes less influenced by the skewness.

Table 9. Average standard error of the slope in each cell of the simulation design for Model 1.

Probability	Sample Size													
	10	50	100	200	300	400	500	600	700	800	900	1,000	5,000	
.01	3956	1295	1015	503	302	135	95	56	18	16	8	4	0.28	
.02	4522	1108	695	217	70	23	8	3	1.10	0.54	0.50	0.48	0.20	
.03	4221	991	470	89	21	4	2	0.50	0.46	0.43	0.40	0.38	0.16	
.04	4360	837	303	39	6	0.54	0.48	0.43	0.39	0.37	0.35	0.33	0.15	
.05	414	715	193	13	0.57	0.48	0.43	0.38	0.36	0.33	0.31	0.29	0.13	
.06	4053	628	114	7	0.51	0.44	0.38	0.35	0.32	0.30	0.28	0.27	0.12	
.07	3914	487	76	4	0.47	0.40	0.36	0.33	0.30	0.28	0.27	0.25	0.11	
.08	3836	403	49	2	0.44	0.38	0.34	0.31	0.28	0.26	0.25	0.23	0.10	
.09	3830	339	31	0.52	0.42	0.36	0.32	0.29	0.27	0.25	0.24	0.22	0.10	
.10	3525	257	20	0.49	0.40	0.34	0.30	0.28	0.26	0.24	0.22	0.21	0.09	
.15	3617	90	0.59	0.41	0.33	0.28	0.25	0.23	0.21	0.20	0.18	0.18	0.08	
.20	3076	35	0.52	0.36	0.29	0.25	0.22	0.21	0.19	0.18	0.17	0.16	0.07	
.25	2434	7	0.47	0.33	0.27	0.23	0.21	0.19	0.17	0.16	0.15	0.15	0.07	
.30	2067	0.65	0.44	0.32	0.25	0.22	0.19	0.18	0.17	0.15	0.14	0.14	0.06	
.35	1744	0.62	0.43	0.30	0.24	0.20	0.18	0.17	0.16	0.15	0.14	0.13	0.06	
.40	1608	0.60	0.42	0.29	0.24	0.21	0.18	0.17	0.15	0.14	0.14	0.13	0.06	
.45	1535	0.59	0.41	0.28	0.23	0.20	0.18	0.16	0.15	0.14	0.13	0.13	0.06	
.50	1483	0.59	0.41	0.28	0.23	0.20	0.18	0.16	0.15	0.14	0.13	0.12	0.06	

Regardless of the low bias in the slope estimates, when the Wald statistic is calculated, the bias of the denominator is very high and outweighs the negligible bias of the numerator. This results in a Wald statistic that will likely not reject the null hypothesis, resulting in a conservative test. For instance, for a sample size of 100 and a skewed probability of .04 (which is quite a skewed predictor), the numerator of the Wald statistic is not highly biased, but the denominator is, resulting in a Type I error rate of zero.

Even though some modest bias exists in the parameter estimates, the conservative Type I error rates are clearly driven by the large standard errors. The apparent contradiction between a conservative Type I error rate and a highly inflated OR is best understood by examining the shape of the sampling distribution of the slope, wherein the large values of some of the replications with a cell of the experimental design influence the average value of the ORs. This is best seen by contrasting Tables 6 and 7 with the mean and median ORs, respectively.

In the extreme case of a sample size of 10, the average slope deviates far from the simulated value, even when the predictor is balanced. Likewise, the standard errors are always upwardly biased in the order of magnitude of the thousands. Because of these obvious biases and the impracticality of such a small sample size, it was removed from further analyses.

Study B: Power

Purpose of the study. Usually, a low probability of Type I errors is accompanied by low statistical power. Therefore, our next step is to examine the statistical power of the Wald test of the slope parameter for this model. Although there is no agreement on what magnitude of effect (i.e., effect size) is necessary to establish practical significance, Ferguson (2009) suggests three values related to risk estimates, i.e., measures comparing relative risk for a particular outcome between two or more groups. According to Ferguson (2009), ORs of 2, 3, and 4 represent small, moderate, and large effect sizes, respectively. As Cohen (1988) clearly stated, all effect size guidelines are research context dependent and should only be used in the research settings from which they were derived.

Method

Simulation factors and methodology. In addition to skewness and sample size, a third factor was manipulated in this study. As in the previous simulation, the sample size was varied across 13 levels ranging from 50 to 5000, and the probability of the occurrence of a category in the predictor from 0.01 to 0.45. For comparison purposes, the investigation pertained to what happens when the predictor variable is balanced. The third factor added is effect size, which was varied from small, moderate, and large. The resulting experiment is an 18×12×3 completely crossed factorial design.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

In this simulation, the estimation is built on the assumption that this model has an effect. Hence, it was assumed β_0 and β_1 are fixed to a number different from zero. The intercept parameter was fixed to -2 . For the slope parameter, three levels of effect size were considered: small effect of 0.683 (equivalent to OR = 2), moderate effect of 1.1 (equivalent to OR = 3), and large effect of 1.38 (equivalent to OR = 4). As in Study A, each cell includes 1,000 replications of the same model.

Analysis of simulation results. To assess power estimates, a framework was adopted similar to Bradley's for Type I error rates, to determine at what level of skewness 10% and 50% of expected statistical power was lost. The expected statistical power was identified as the power in the case of the balanced cell probability of the predictor. Hence, the estimated statistical power for each cell in the experiment is compared to the power for the same sample and effect size but with no skewness in the predictor's probability.

Results and conclusions

Tables 10, 11, and 12 follow the structure of previous tables and show the statistical power for each effect size level. The last row in each table is the power estimate for the balanced predictor variable. The tables are greyscale coded: no shading reflects losing 10% of power or less, light shading reflects losing 10% to 50% of power, and dark shading reflects losing over 50% of power.

Because sample size and effect size both significantly influence power, it is not surprising that as these two factors increase, power also increases. From the tables, it can be seen for a balanced predictor, an effect size of OR = 2, and a sample size of 200 or less, the power of the Wald test is less than 50%. It exceeds 50% after a sample size of 300, exceeds 75% after a sample size of 500, and reaches one after a sample size of 1,000. Moreover, the tables indicate that as the effect size grows, there is less of a need for larger sample sizes to detect the effects. For example, with a balanced predictor and a sample of 100, the statistical power is nearly 75% to detect an OR of 4, while it is 52% and 21% for ORs of 3 and 2, respectively.

For a low effect size, samples from 100 to 1,000, and a skewed cell probability less than or equal to 0.2, over 10% of power is lost compared to the balanced cases. As the effect size increases, the level of skewed probability that is tolerated slightly increases. For example, to retain 10% of power for sample sizes of 100-300, the level of probability tolerated for a low effect size ranges between

0.2 and 0.3. However, the level of skewed cell probability needed to retain 10% of power for a high effect is 0.15 to 0.25 for the same sample sizes. Power is highly influenced by skewed probabilities in small sample sizes, even when the effect size is moderate to large and hence highly detectable.

Table 10. Power with low effect size (OR = 2).

Probability	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	0.00	0.01	0.04	0.06	0.08	0.10	0.13	0.13	0.15	0.17	0.17	0.50
.02	0.01	0.04	0.09	0.12	0.13	0.17	0.18	0.19	0.22	0.25	0.27	0.75
.03	0.03	0.05	0.11	0.16	0.16	0.21	0.25	0.27	0.29	0.32	0.32	0.87
.04	0.03	0.07	0.12	0.18	0.21	0.24	0.30	0.32	0.34	0.38	0.41	0.94
.05	0.05	0.09	0.14	0.21	0.24	0.28	0.35	0.39	0.40	0.45	0.47	0.98
.06	0.06	0.11	0.16	0.24	0.26	0.33	0.38	0.43	0.45	0.49	0.52	0.99
.07	0.08	0.13	0.17	0.26	0.28	0.36	0.42	0.45	0.49	0.55	0.59	0.99
.08	0.08	0.14	0.20	0.28	0.32	0.40	0.46	0.51	0.54	0.59	0.63	0.99
.09	0.09	0.15	0.21	0.30	0.31	0.41	0.49	0.54	0.58	0.64	0.67	1.00
.10	0.09	0.17	0.22	0.32	0.36	0.45	0.52	0.57	0.62	0.67	0.72	1.00
.15	0.12	0.18	0.30	0.39	0.49	0.59	0.65	0.70	0.74	0.81	0.85	1.00
.20	0.13	0.22	0.34	0.46	0.56	0.67	0.73	0.80	0.85	0.87	0.91	1.00
.25	0.13	0.21	0.37	0.52	0.64	0.72	0.78	0.86	0.90	0.91	0.94	1.00
.30	0.13	0.22	0.39	0.55	0.68	0.76	0.83	0.89	0.92	0.94	0.96	1.00
.35	0.11	0.23	0.42	0.58	0.69	0.78	0.84	0.90	0.93	0.94	0.98	1.00
.40	0.10	0.22	0.43	0.60	0.72	0.80	0.85	0.91	0.95	0.96	0.97	1.00
.45	0.09	0.22	0.44	0.60	0.73	0.81	0.87	0.93	0.96	0.96	0.98	1.00
.50	0.06	0.21	0.42	0.61	0.70	0.80	0.87	0.92	0.96	0.96	0.98	1.00

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

However, sample sizes of 400 and over can retain 10% of power with low levels of skewness. For example, a sample of 500 with an OR of 2 retains 10% of power at a probability of 0.3. As the effect size increases to an OR of 3, 10% of power is retained at level 0.15; at an OR of 4, 10% is retained at level 0.07. A sample size of 1,000 with an OR of 2 retains 10% of power at level 0.2. The probability level that retains the same percentage of power quickly jumps to 0.06 and 0.04 for ORs of 3 and 4, respectively.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Table 11. Power with moderate effect size (OR = 3).

Probability	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	0.01	0.02	0.07	0.11	0.15	0.19	0.26	0.26	0.29	0.35	0.36	0.88
.02	0.01	0.07	0.15	0.23	0.29	0.35	0.41	0.43	0.45	0.54	0.54	0.99
.03	0.04	0.11	0.22	0.31	0.38	0.44	0.51	0.55	0.60	0.67	0.69	1.00
.04	0.05	0.14	0.27	0.37	0.45	0.53	0.63	0.66	0.71	0.79	0.79	1.00
.05	0.07	0.18	0.34	0.44	0.53	0.61	0.68	0.76	0.78	0.85	0.85	1.00
.06	0.09	0.21	0.36	0.50	0.58	0.69	0.74	0.82	0.84	0.89	0.90	1.00
.07	0.12	0.24	0.41	0.56	0.64	0.73	0.78	0.86	0.88	0.93	0.93	1.00
.08	0.14	0.27	0.44	0.60	0.69	0.77	0.83	0.90	0.92	0.95	0.95	1.00
.09	0.15	0.29	0.48	0.62	0.74	0.82	0.85	0.92	0.93	0.97	0.97	1.00
.10	0.16	0.32	0.52	0.66	0.78	0.85	0.89	0.94	0.96	0.98	0.98	1.00
.15	0.22	0.40	0.65	0.78	0.88	0.94	0.97	0.98	0.99	1.00	1.00	1.00
.20	0.25	0.46	0.70	0.86	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00
.25	0.26	0.47	0.77	0.90	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00
.30	0.28	0.51	0.80	0.94	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
.35	0.28	0.53	0.83	0.94	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	0.27	0.52	0.85	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	0.27	0.52	0.87	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	0.24	0.52	0.87	0.96	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

Table 12. Power with large effect size (OR = 4).

Probability	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1,000	5,000
.01	0.00	0.02	0.07	0.15	0.23	0.29	0.36	0.40	0.42	0.49	0.51	0.98
.02	0.01	0.08	0.21	0.34	0.42	0.52	0.58	0.61	0.66	0.73	0.76	1.00
.03	0.05	0.15	0.31	0.44	0.55	0.65	0.71	0.76	0.81	0.87	0.87	1.00
.04	0.07	0.20	0.40	0.53	0.65	0.75	0.81	0.85	0.89	0.94	0.94	1.00
.05	0.10	0.25	0.49	0.62	0.74	0.82	0.86	0.90	0.95	0.97	0.98	1.00
.06	0.13	0.30	0.54	0.69	0.80	0.87	0.90	0.94	0.97	0.99	0.98	1.00
.07	0.16	0.34	0.59	0.74	0.84	0.91	0.93	0.96	0.98	1.00	0.99	1.00
.08	0.19	0.40	0.65	0.70	0.88	0.93	0.95	0.98	0.99	1.00	1.00	1.00
.09	0.21	0.41	0.69	0.83	0.91	0.95	0.97	0.98	1.00	1.00	1.00	1.00
.10	0.24	0.46	0.72	0.85	0.93	0.97	0.98	0.99	1.00	1.00	1.00	1.00
.15	0.32	0.58	0.84	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.20	0.38	0.66	0.90	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.25	0.39	0.69	0.93	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.30	0.44	0.73	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.35	0.45	0.74	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.40	0.44	0.75	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.45	0.46	0.75	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.50	0.43	0.75	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

Power is largely affected by the skewed probability of the predictor variable. High levels of skewness [$p = 0.01, 0.1$] result in the loss of over 50% of power in most sample sizes, with the exception of 1,000 and 5,000. Although it was shown in the previous study Type I error is acceptable for the majority of factors examined, it has implications on the empirical power of the Wald test. This is evidence statistical power is biased downwards with the combination of smaller samples and higher degrees of skewed probability. Clearly, the skewed probability of a predictor diminishes the statistical power of the Wald test.

Model 2: Single Continuous Predictor

The second model of interest involves simple logistic regression with one continuous predictor.

$$g(y) = \beta_0 + \beta_1 x,$$

where x is a skewed continuous variable, β_0 and β_1 are fixed, $g(y)$ is a logit function, and y is a balanced outcome variable. The purpose of this model was to investigate whether findings from [Model 1](#) (which used one binary predictor) would generalize to a skewed continuous predictor. The aim is to determine whether issues with the skewness of the predictor are related to the categorical versus the numeric aspect of the variable—whether it is the skew or the binary nature that is causing the effect on the Type I error. To confirm this, focus only on the Type I error rate, because its reduction is accompanied by a corresponding reduction in statistical power. Therefore, a decreased Type I error rate is diagnostic of a problem with decreased power.

Study A: Type I error rates

Purpose of the study. Similarly to the previous study, the aim was to document the impact of a skewed continuous predictor variable on the estimation, parameter estimates, and Type I error rate of the Wald test.

Method

The simulation factors, methodology, and analysis of the Type I error rate are exactly the same as in the previous study. The only difference is in the nature of the predictor. This variable was generated from a Gamma distribution with the

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

rate and scale parameters fixed to 1 and varying the shape parameter across 17 levels. Skewness in the gamma distribution is a function of the shape parameter. To enable comparison, the skewness levels for this model were matched to the expected probabilities in [Model 1](#). Shown in [Table 13](#) are the shape parameter values used and the equivalent skewness levels. To create a baseline, the case investigated was where the predictor variable is drawn from a standard normal distribution with a mean of zero and a standard deviation of one. The resulting simulation experiment is a 12 (sample size) by 18 (skewness) completely crossed factorial design.

Table 13. Shape parameter and equivalent skewness level.

Shape Parameter	Skewness	Probability
0.047	9.250	0.010
0.086	6.870	0.020
0.130	5.550	0.030
0.200	4.750	0.040
0.250	4.040	0.050
0.300	3.700	0.060
0.370	3.250	0.070
0.400	3.190	0.080
0.500	2.830	0.090
0.600	2.720	0.100
1.000	1.960	0.150
1.750	1.500	0.200
3.000	1.150	0.250
5.500	0.873	0.300
10.000	0.630	0.350
25.000	0.410	0.400
50.000	0.200	0.450
Standard Normal	0.000	0.500

Results and conclusions

It is not surprising that with a continuous predictor, all of the replications converged for the 216 conditions of the simulation experiment. [Table 14](#), which is formatted similarly to the previous tables, shows that the Type I error rates ranged from 0.001 to 0.066, with an average of 0.043. The majority of the conditions met the liberal criterion, but not the stringent one. As can be seen from [Table 14](#), in only a few cases did the Type I error rate fall below 0.025, as dictated by the liberal criterion. These instances are with sample size 50 with skewness ≥ 2.6 , sample size 100 with skewness ≥ 5.54 , and sample sizes 200 and 300 with the highest skewness level (9.23).

Table 14. Liberal Type I error rate model.

Skewness	Sample Size											
	50	100	200	300	400	500	600	700	800	900	1,000	5,000
9.25	0.00	0.00	0.01	0.01	0.02	0.03	0.03	0.05	0.02	0.03	0.04	0.04
6.87	0.00	0.01	0.02	0.03	0.03	0.05	0.03	0.04	0.03	0.04	0.04	0.06
5.55	0.01	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.04	0.05
4.75	0.01	0.03	0.04	0.04	0.03	0.05	0.05	0.04	0.05	0.06	0.05	0.04
4.04	0.01	0.03	0.04	0.04	0.04	0.05	0.04	0.06	0.04	0.04	0.05	0.05
3.70	0.02	0.03	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.06	0.05	0.05
3.25	0.02	0.03	0.04	0.04	0.06	0.04	0.05	0.06	0.05	0.03	0.06	0.06
3.19	0.02	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.05	0.04	0.06	0.05
2.83	0.03	0.03	0.05	0.04	0.05	0.04	0.05	0.05	0.04	0.06	0.05	0.05
2.72	0.02	0.03	0.04	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.04	0.05
1.96	0.04	0.03	0.04	0.04	0.04	0.06	0.06	0.05	0.06	0.04	0.05	0.05
1.50	0.05	0.03	0.05	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.05
1.15	0.04	0.05	0.05	0.04	0.05	0.06	0.06	0.05	0.06	0.07	0.06	0.04
.873	0.04	0.06	0.04	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.06
.63	0.04	0.04	0.04	0.05	0.04	0.05	0.05	0.05	0.06	0.05	0.05	0.06
.41	0.04	0.04	0.04	0.05	0.06	0.04	0.04	0.07	0.06	0.06	0.06	0.05
.20	0.03	0.04	0.04	0.06	0.05	0.04	0.06	0.05	0.05	0.05	0.04	0.05
.00	0.05	0.06	0.06	0.04	0.05	0.06	0.05	0.05	0.06	0.06	0.05	0.06

Note: Cells depicted in grey have deflated Type I error rates, whereas those with no shading meet the adequacy condition using Bradley's criteria (see Table 5).

The estimation tolerated a skewed continuous predictor a great deal better than a dichotomous one. The same conclusions from the previous study can be drawn here in that as the sample size increases and the skewness becomes smaller, the Type I error rate gets closer to the nominal value. Hence, as with a dichotomous predictor, a highly skewed continuous predictor affects the estimation and inferences in the extreme case of a small sample size.

Model 3: Multiple Logistic Regression with Two Independent Binary Predictors

The last model investigated is a multiple logistic regression with two dichotomous predictors,

$$g(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where x_1 and x_2 are independent dichotomous variables with skewed probabilities; β_0 , β_1 , and β_2 are fixed; $g(y)$ is the logit function, and y is a balanced outcome

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

variable. The goal in including this model was to discover whether the skewed probability of one predictor could alter the parameter estimates of other variables in the model when the two predictors are independent. This model reflects, for example, a 2×2 (two-factor) randomized experiment or randomized clinical trial.

Study A: Type I error rates and non-convergences

Method

The simulation methodology and analysis of the type I error rate were the same as in [Models 1](#) and [2](#). Three factors were manipulated in this experiment. The first two are sample size and the probability of x_1 , while the additional factor is the probability of x_2 . The probability for each predictor varied from 0.01 to 0.045. As in previous studies, the case where the variables were balanced was also examined for comparison purposes. The resulting experiment is a $12 \times 18 \times 18$ completely crossed factorial design. Again, as in earlier models, the empirical and nominal Type I error rates were compared using Bradley's criteria.

Results and conclusions

There were 4,212 experimental conditions. Many results were identical or only a couple of decimals apart. Because of the sheer volume and the lack of variation, only a few sample sizes are presented.

Number of non-convergences. Shown in [Tables 15](#) and [16](#) are the number of non-converging replications with varying degrees of skewed probability on both predictors for samples of 100 and 400, respectively. As in previous studies, for a sample of 100 and a probability of 0.01, most of the replications did not converge. All replications converged when the probability of x_1 is 0.09 or higher and the probability of x_2 is equal or higher than 0.07. For sample sizes 100 and 400, the number of non-converging replications decreases as the probability of both variables becomes more balanced. This issue ceases to be important for samples of 900 and 5,000.

Table 15. Number of non-convergences from 1,000 replications for Model 3 when sample size is 100.

x_1 Probability	x_2 Probability								
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	...	0.5
.01	610	463	412	397	390	389	388	...	388
.02	474	260	195	165	157	155	154	...	154
.03	407	176	102	70	62	60	58	...	58
.04	384	143	68	33	23	21	19	...	19
.05	374	131	56	21	11	9	7	...	7
.06	373	129	54	19	9	7	5	...	5
.07	371	127	52	16	6	4	2	...	2
.08	371	127	52	16	6	3	1	...	1
.09	370	126	51	15	5	2	0	...	0
...	...	...	...	...	...	...	...	...	0
.50	370	126	51	15	5	2	0	0	0

Table 16. Number of non-convergences from 1,000 replications for Model 3 when sample size is 400.

x_1 Probability	x_2 Probability				
	.01	.02	.03	...	.50
.01	41	21	21	...	21
.02	21	1	1	...	1
.03	20	0	0	...	0
...	...	...	...	...	0
.50	20	0	0	0	0

Type I error rate. The effect of the skewness of x_2 on the Type I error rate was compared for the Wald test of x_1 . The results in Table 17 are formatted somewhat differently from those in other tables in this paper. Following Conover, Johnson, and Johnson (1981), average Type I error rates were used. Presented in the table are the Type I error rate of the Wald test for x_1 averaged across all levels of skewness of x_2 for sample size 100, 400, 900, and 5000 to represent the small, medium, and large sample sizes found in the literature. Like the tables in Models 1 and 2, Table 17 is greyscale coded, the darkly shaded areas falling below Bradley's liberal criterion and the unshaded ones falling within it. The Type I error was consistently deflated, with lower levels of probability for x_1 and x_2 , growing closer to the nominal value as the skewed probability for both predictors became more balanced. It satisfies the liberal criterion with ranges of [0, 0.06] for a sample size of 100, [0.002, 0.068] for a sample size of 400, [0.017, 0.065] for a sample size of 900, and [0.033, 0.057] for a sample size of 5000.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Table 17. Type I error rate for x_1 averaged across all levels of x_2 , and the range of Type I errors across all levels of the skewed probability of x_2 .

x_1 Probability	Sample Size			
	100	400	900	5000
.01	0	0.002	0.019	0.053
	(0,0)	(.001,.003)	(.017,.021)	(.051,.057)
.02	0	0.015	0.039	0.041
	(0,0)	(.013,.018)	(.035,.043)	(.035,.042)
.03	0	0.023	0.043	0.043
	(0,0)	(.021,.026)	(.040,.044)	(.041,.045)
.04	0	0.036	0.045	0.036
	(0,0)	(.033,.039)	(.042,.046)	(.033,.052)
.05	0.003	0.041	0.058	0.040
	(.001,.005)	(.038,.043)	(.056,.060)	(.039,.041)
.06	0.01	0.039	0.047	0.047
	(.005,.011)	(.036,.046)	(.045,.049)	(.046,.049)
.07	0.01	0.046	0.056	0.045
	(.010,.020)	(.043,.049)	(.053,.059)	(.043,.047)
.08	0.019	0.05	0.063	0.049
	(.016,.022)	(.048,.054)	(.061,.065)	(.049,.051)
.09	0.019	0.05	0.062	0.038
	(.017,.023)	(.048,.055)	(.060,.064)	(.038,.040)
.10	0.023	0.049	0.052	0.047
	(0,.030)	(.047,.052)	(.050,.055)	(.046,.049)
.15	0.037	0.062	0.055	0.049
	(.034,.040)	(.060,.065)	(.052,.061)	(.049,.051)
.20	0.045	0.049	0.045	0.042
	(.034,.048)	(.047,.057)	(.042,.047)	(.041,.043)
.25	0.041	0.048	0.047	0.052
	(.035,.044)	(.046,.049)	(.045,.050)	(.051,.053)
.30	0.047	0.054	0.038	0.055
	(.039,.053)	(.052,.056)	(.036,.041)	(.054,.056)
.35	0.056	0.053	0.048	0.047
	(.039,.060)	(.052,.055)	(.046,.050)	(.046,.048)
.40	0.049	0.066	0.053	0.050
	(.047,.052)	(.062,.069)	(.050,.054)	(.035,.055)
.45	0.049	0.058	0.040	0.049
	(.037,.052)	(.055,.061)	(.038,.045)	(.047,.051)
.50	0.056	0.057	0.047	0.042
	(.052,.059)	(.052,.060)	(.046,.049)	(.042,.044)

Note: Cells depicted in grey have deflated Type I error rates, whereas those with no shading meet the adequacy condition using Bradley's criteria (see Table 5).

The range of average Type I error rates for each cell in Table 17 does not vary greatly. Therefore, it is clear that the degree of skewness of the cell probability of x_2 has little to no impact on the Type I error rate of x_1 . In other words, the Type I error rate of x_1 with low probability on x_2 does not differ from

the Type I error rate of x_1 when x_2 is balanced. For example, for a sample of 400, the Type I error rate for x_1 is 0.035 when the probability of $x_1 = 0.04$ and the probability of $x_2 = 0.03$ or 0.45. Comparing the Type I error rate of x_1 in [Model 1](#) with [Model 3](#), we can see that factoring in the skewness of an additional variable that is completely independent from other variables in this model has minimal impact.

Study B: Power.

Purpose of the study. Pursuant to Study A, this simulation was designed to investigate the impact of two independent dichotomous predictors with skewed probabilities on the power of the Wald test. Ferguson's (2009) suggestion for a small, moderate, and large effect size was chosen, applying it to ORs.

Method

In addition to the three factors mentioned in Study A, a fourth factor was added: sample size, the probability of both predictors, and effect size, which was either small, moderate, or large. The resulting experiment is a $12 \times 18 \times 18 \times 3$ completely crossed factorial design.

The simulation procedure is the same, with the added assumption of model effect. Assume β_0 , β_1 , and β_2 are fixed to a number different from zero. The intercept parameter was fixed to -2 . Examined three levels of effect size for β_1 : small effect: 0.683 (equivalent to $OR = 2$), moderate effect: 1.1 (equivalent to $OR = 3$), and large effect: 1.38 (equivalent to $OR = 4$). However, we fixed the effect size for β_2 to a moderate value of 1.1. The simulation methodology is similar to that in Study A, with the addition of an extra loop to account for effect size.

Results and conclusions

To analyze power estimates, the best achievable power in the case of a balanced design was compared with other combinations of probabilities for each sample size. Both 10% and 50% loss of power were considered. The resulting number of conditions was 12636. Similarly to Study A, four typical sample sizes: 100, 400, 900, and 5,000 are presented. Shown in [Tables 18 through 20](#) are power estimates for x_1 averaged over all levels of probability of x_2 . As in [Table 17](#), [Tables 18 through 20](#) include the range of statistical power for each condition.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Table 18. Statistical power at an OR = 2 for x_1 averaged across all levels of x_2 , and the range of statistical power errors across all levels of the skewed probability of x_2 .

x_1 Probability	Sample Size			
	100	400	900	5,000
.01	0.01	0.10	0.17	0.53
	(.0,.013)	(.074,.110)	(.160,.180)	(.500,.600)
.02	0.04	0.14	0.26	0.80
	(.020,.210)	(.140,.150)	(.240,.280)	(.770,.850)
.03	0.05	0.166	0.33	0.92
	(.040,.055)	(.150,.180)	(.310,.360)	(.890,.950)
.04	0.08	0.213	0.41	0.97
	(.065,.085)	(.200,.230)	(.340,.450)	(.950,1)
.05	0.09	0.26	0.47	0.99
	(.011,.110)	(.230,.280)	(.440,.530)	(.980,1)
.06	0.10	0.30	0.53	0.99
	(.096,.011)	(.270,.397)	(.490,.590)	(.990,1)
.07	0.12	0.32	0.57	1
	(.110,.130)	(.300,.350)	(.530,.660)	(.990,1)
.08	0.13	0.36	0.62	1
	(.120,.140)	(.340,.390)	(.580,.700)	(.990,1)
.09	0.15	0.38	0.66	1
	(.130,.160)	(.360,.410)	(.620,.730)	(1,1)
.10	0.15	0.40	0.70	1
	(.140,.170)	(.380,.440)	(.660,.780)	(1,1)
.15	0.18	0.52	0.84	1
	(.160,.200)	(.480,.580)	(.810,.890)	(1,1)
.20	0.197	0.60	0.91	1
	(.170,.230)	(.560,.670)	(.890,.950)	(1,1)
.25	0.218	0.67	0.95	1
	(.190,.260)	(.620,.750)	(.930,.980)	(1,1)
.30	0.23	0.72	0.96	1
	(.190,.290)	(.680,.790)	(.950,.980)	(1,1)
.35	0.23	0.74	0.97	1
	(.200,.310)	(.690,.820)	(.960,.980)	(1,1)
.40	0.25	0.76	0.98	1
	(.220,.310)	(.720,.830)	(.970,.990)	(1,1)
.45	0.25	0.76	0.98	1
	(.200,.300)	(.710,.840)	(.960,.990)	(1,1)
.50	0.246	0.76	0.98	1
	(0.210,.300)	(.710,.850)	(.970,.990)	(1,1)

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

Table 19. Statistical power at an OR = 3 for x_1 averaged across all levels of x_2 , and the range of statistical power errors across all levels of the skewed probability of x_2 .

x_1 Probability	Sample Size			
	100	400	900	5,000
.01	0.02	0.15	0.33	0.90
	(.010,.300)	(.120,.170)	(.320,.305)	(.880,.930)
.02	0.05	0.28	0.54	0.99
	(.040,.060)	(.270,.290)	(.520,.560)	(.980,1)
.03	0.09	0.38	0.68	0.99
	(.080,.100)	(.370,.400)	(.660,.730)	(.990,1)
.04	0.14	0.46	0.79	0.99
	(.120,.150)	(.440,.480)	(.770,.830)	(.990,1)
.05	0.18	0.53	0.86	0.99
	(.180,.200)	(.510,.570)	(.840,.900)	(.980,1)
.06	0.22	0.6	0.90	0.99
	(.210,.230)	(.570,.630)	(.880,.940)	(.990,1)
.07	0.25	0.66	0.94	1
	(.240,.270)	(.640,.710)	(.930,.950)	(.990,1)
.08	0.27	0.71	0.96	1
	(.260,.290)	(.690,.750)	(.950,.970)	(.990,1)
.09	0.3	0.76	0.97	1
	(.290,.330)	(.740,.780)	(.960,.980)	(1,1)
.10	0.32	0.80	0.98	1
	(.300,.360)	(.750,.920)	(.970,.990)	(1,1)
.15	0.40	0.90	0.99	1
	(.320,.440)	(.800,.930)	(.980,1)	(1,1)
.20	0.46	0.96	0.99	1
	(.430,.520)	(.930,.980)	(.990,1)	(1,1)
.25	0.49	0.98	1	1
	(.450,.560)	(.940,.990)	(1,1)	(1,1)
.30	0.55	0.99	1	1
	(.500,.620)	(.980,.990)	(1,1)	(1,1)
.35	0.56	0.99	1	1
	(.520,.630)	(.980,.990)	(1,1)	(1,1)
.40	0.58	0.99	1	1
	(.530,.670)	(.990,.990)	(1,1)	(1,1)
.45	0.59	0.99	1	1
	(.540,.670)	(.990,.990)	(1,1)	(1,1)
.50	0.59	0.99	1	1
	(.440,.680)	(.990,.990)	(1,1)	(1,1)

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Table 20. Statistical power at an OR = 4 for x_1 averaged across all levels of x_2 , and the range of statistical power errors across all levels of the skewed probability of x_2 .

x_1 Probability	Sample Size			
	100	400	900	5,000
.01	0.02	0.20	0.50	0.98
	(.010,.040)	(.170,.230)	(.450,.990)	(.970,1)
.02	0.07	0.39	0.74	0.99
	(.050,.090)	(.380,.410)	(.720,.770)	(.990,1)
.03	0.13	0.55	0.87	1
	(.110,.160)	(.530,.570)	(.850,.900)	(1,1)
.04	0.19	0.65	0.95	1
	(.170,.220)	(.630,.680)	(.940,.970)	(1,1)
.05	0.26	0.74	0.97	1
	(.240,.270)	(.700,.780)	(.970,.990)	(1,1)
.06	0.31	0.80	0.98	1
	(.290,.330)	(.770,.830)	(.970,.990)	(1,1)
.07	0.36	0.86	0.99	1
	(.340,.370)	(.830,.890)	(.990,1)	(1,1)
.08	0.39	0.89	0.99	1
	(.380,.410)	(.860,.910)	(.990,1)	(1,1)
.09	0.44	0.92	0.99	1
	(.420,.460)	(.910,.940)	(.990,1)	(1,1)
.10	0.47	0.94	0.97	1
	(.450,.500)	(.930,.960)	(.980,1)	(1,1)
.15	0.59	0.99	1	1
	(.570,.630)	(.980,.990)	(1,1)	(1,1)
.20	0.67	0.99	1	1
	(.630,.720)	(.990,1)	(1,1)	(1,1)
.25	0.72	0.99	1	1
	(.600,.780)	(.990,1)	(1,1)	(1,1)
.30	0.77	0.99	1	1
	(.720,.830)	(.990,1)	(1,1)	(1,1)
.35	0.79	1	1	1
	(.730,.850)	(1,1)	(1,1)	(1,1)
.40	0.81	1	1	1
	(.760,.870)	(1,1)	(1,1)	(1,1)
.45	0.82	1	1	1
	(.770,.880)	(1,1)	(1,1)	(1,1)
.50	0.82	1	1	1
	(.770,.880)	(1,1)	(1,1)	(1,1)

Note: No shading reflects losing 10% of power or less, light shading reflects losing 10%-50% of power, and dark shading reflects losing over 50% of power.

Statistical power of x_1 is not affected by changes in the probability of x_2 , but rather, is affected by its own skewed probability regardless of the sample and effect sizes. Considering the range of statistical power for each condition, changes

in the power of x_1 when the probability of x_2 is at its extreme are within [0.01, 0.03] of the power when the probability of x_2 is balanced. However, it is evident that the power of x_1 is highly influenced by its own skewed probability. For example, the highest achievable power for this model under the circumstances identified in this simulation for x_1 with a sample of 100 and a small effect size is 0.3, as seen in Table 18. Power is dramatically affected by the deflation in the Type I error rate. It is shown that with a skewed probability of 0.01, the power of the Wald test for the same predictor plummets to a range of [0, 0.5] for all sample and effect sizes. Similar to the findings in Model 1, as the sample and effect sizes increase, the skewed probability tolerance accelerates significantly.

Conclusion

It is not uncommon to encounter data from skewed populations. In these cases, the skewness of the sample and predictor variables reflects the true character of the population rather than a sampling bias. Hence, the skewness in the predictor(s) may influence estimation if separation occurs or decrease the reliability of parameter estimates. Detecting separation through data configurations, infinite parameter estimates, and the non-convergence of the MLE is straightforward. However, with a skewed predictor, these clear indicators are not present. This leaves the question of the impact of skewed predictors on the eventual statistical results of a logistic regression. To answer this general question, five inter-related simulation studies were conducted, which to our knowledge are the first of their kind to be done for skewed dichotomous predictors.

A broad picture of the effects of skewed cell probabilities in dichotomous predictors on the logistic regression model is provided, specifically regarding how a categorical predictor's skewness in probabilities affect estimation, parameter estimates, and the Wald test. In many cases, the estimator came to a convergence and results were produced, but there is no warning that a potential problem may exist. Data analysts can carry on without being aware that the standard errors are greatly inflated, resulting in low to no statistical power and (at times) greatly enlarged ORs.

Skewed probabilities can induce separation, which automatically affects estimation and results in non-convergence (Albert & Anderson, 1984). When separation does not occur—even in severe cases of skewed probability—ML converges and estimates are produced.

MLEs are biased upwards in severe conditions of small samples and highly skewed probability. Lastly, when skewness is less severe, with a range of [0.25,

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

0.5], or the sample size is sufficiently large, Type I error rates reach a nominal value and power is high. Overall, these findings demonstrate why it is important to consider the descriptive characteristics of the predictor(s) before conducting a logistic regression analysis. Researchers may encounter situations wherein the Type I error rate of their hypothesis test is highly deflated, ostensibly declaring a strong test when this may not be the case. Also, the power of the hypothesis test performs in a complementary manner to the Type I error rates. That is, the power is deflated when the Type I error is, and reaches full power when the rate achieves a nominal value.

The skewed binary predictors can depending on the research design as (i) observed groups (e.g., gender differences or rare diagnostic or disease states) that have skewed probabilities of occurrence, or (ii) in other research settings these binary predictor(s) can be considered elements of a design matrix for experiments or clinical trials wherein the imbalance in group sizes is not due to selection bias or attrition.

Therefore, with data analysts and consumers of research materials in mind, guidelines are suggested on how to think about and handle skewed predictor(s) in a logistic regression analysis based on whether the skewness is severe or not.

- 1) If skewness is severe (i.e. the shaded areas in the tables included in this paper), there are two cases to consider. The first case under severe skewness is when separation (by inspecting the data structure) or zero cell counts (by looking at multi-way table) occur in the data. In this case, the conventional maximum likelihood estimator will not converge and estimates are not produced. Data analysts are encouraged to use alternative estimation procedures for example the ones found in Bull, Mak, and Greenwood (2002) and Heinze (2006). The second case under severe skewness occurs when estimates are produced by the conventional maximum likelihood estimator. This is due to the fact that one does not have zero cell counts nor separation. In this case of severe skewness, however, the Wald test is not reliable and an alternative test, such as the likelihood ratio test, is recommended.
- 2) If skewness is not severe (i.e. the non-shaded areas in the tables mentioned in this paper), the Wald statistic is reliable and the interpretation of the test statistic should follow general statistical analysis recommendations (i.e. do not rely on the test statistic

exclusively but also examine effect size, parameter estimates, standard errors, fit, etc.). In addition, because skew of the predictor may impact on the operating characteristics of the statistical test, when planning a study a researcher/reader should take this in to account when computing the level of power and assumed Type I error rate.

Acknowledgments

The first author would like to acknowledge the supporting and funding from King Abdulaziz University. This project is funded by the UBC-Paragon Research Agreement.

References

- Agresti, A. (2002). *Categorical data analysis* (3rd Ed). Hoboken, NJ: John Wiley & Sons, Inc. doi: [10.1002/0471249688](https://doi.org/10.1002/0471249688)
- Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1), 1–10. doi: [10.1093/biomet/71.1.1](https://doi.org/10.1093/biomet/71.1.1)
- Alkhalaf, A. A. (2014). *Application of Logistic Regression in Higher Education Research: A sample from 2000-2013*. Unpublished Manuscript, Measurement, Evaluation, & Research Methodology Program, University of British Columbia, Vancouver, Canada.
- Anderson, J. A., & Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1), 71–78. doi: [10.2307/1268582](https://doi.org/10.2307/1268582)
- Barreto, I. D. de C., Russo, S. L., Brasil, G. H., & Simon, V. H. (2014). Separation phenomena logistic regression. *Revista GEINTEC*, 4(1), 716–728. doi: [10.7198/s2237-0722201400010024](https://doi.org/10.7198/s2237-0722201400010024)
- Birnbaum, A. (1964). Median-unbiased estimators. *Bulletin of Mathematical Statistics (Tokyo)*, 11, 25–34.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)
- Bull, S. B., Mak, C., & Greenwood, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples.

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Computational Statistics and Data Analysis, 39, 57–74. doi: [10.1016/s0167-9473\(01\)00048-2](https://doi.org/10.1016/s0167-9473(01)00048-2)

Christensen, R. (2016). *Analysis of variance, design, and regression: linear modeling for unbalanced data* (2nd Ed.). Boca Raton, Florida: CRC Press, Taylor & Francis Group.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.

Conover, A. 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. doi: [10.2307/1268225](https://doi.org/10.2307/1268225)

Cordeiro, G. M., & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3), 629–643.

Day, N. E., & Kerridge, D. F. (1967). A general likelihood discriminant. *Biometrics*, 23(2), 313–323. doi: [10.2307/2528164](https://doi.org/10.2307/2528164)

Fears, T. R., Benichou, J., & Gail, M. H. (1996). A reminder of the fallibility of the Wald statistic. *The American Statistician*, 50(3), 226–227. doi: [10.1080/00031305.1996.10474384](https://doi.org/10.1080/00031305.1996.10474384)

Ferguson, C. J. (2009). An effect size primer : a guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532–538. doi: [10.1037/a0015808](https://doi.org/10.1037/a0015808)

Ferris, C. D., Grubbs, F. E., & Weaver, C. L. (1946). Operating characteristics for the common statistical tests of significance. *The Annals of Mathematical Statistics*, 17(2), 178–197. doi: [10.1214/aoms/1177730979](https://doi.org/10.1214/aoms/1177730979)

Gordóvil-Merino, A., Guàrdia-Olmos, J., & Però-Cebollero, M. (2012). Estimation of logistic regression models in small samples: a simulation study using a weakly informative default prior distribution. *Psicologica*, 33, 345–361.

Gregory, W., & Veall, R. (1986). Wald tests of common factor restrictions. *Economica Letters*, 22, 203–208. doi: [10.1016/0165-1765\(86\)90232-6](https://doi.org/10.1016/0165-1765(86)90232-6)

Hauck, Jr., W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360), 851–853. doi: [10.2307/2286473](https://doi.org/10.2307/2286473)

- Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, 25, 4216–4226. doi: [10.1002/sim.2687](https://doi.org/10.1002/sim.2687)
- Heinze, G., & Puh, R. (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in Medicine*, 29, 770–777. doi: [10.1002/sim.3794](https://doi.org/10.1002/sim.3794)
- Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81(394), 471–476. doi: [10.2307/2289237](https://doi.org/10.2307/2289237)
- Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics, *Journal of the American Statistical Association*, 73(362), 253–263. doi: [10.2307/2286650](https://doi.org/10.2307/2286650)
- Lütkepohl, H., & Burda, M. M. (1997). Modified Wald tests under nonregular conditions. *Journal of Econometrics*, 78, 315–332. doi: [10.1016/S0304-4076\(97\)80015-2](https://doi.org/10.1016/S0304-4076(97)80015-2)
- Mîndrilă, D. (2010). Maximum likelihood (ML) and diagonally wighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, 1(1), 60–66. doi: [10.20533/ijds.2040.2570.2010.0010](https://doi.org/10.20533/ijds.2040.2570.2010.0010)
- Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3rd Ed.). McGraw Hill Inc.
- Pawitan, Y. (2000). A reminder of the fallibility of the Wald statistic : likelihood explanation. *The American Statistician*, 54(1), 54–56. doi: [10.2307/2685612](https://doi.org/10.2307/2685612)
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373–1379. doi: [10.1016/s0895-4356\(96\)00236-3](https://doi.org/10.1016/s0895-4356(96)00236-3)
- Rousseeuw, P. J., & Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, 43, 315–332. doi: [10.1016/s0167-9473\(02\)00304-3](https://doi.org/10.1016/s0167-9473(02)00304-3)
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th Ed.). Pearson Education.
- Vaeth, M. (1985). On the use of Wald’s test in exponential families. *International Statistical Review*, 53(2), 199–214. doi: [10.2307/1402935](https://doi.org/10.2307/1402935)

SKEWED PREDICTOR VARIABLE(S) LOGISTIC REGRESSION

Zorn, C. (2005). A solution to separation in binary response models.
Political Analysis, 13(2), 157–170. doi: [10.1093/pan/mpi009](https://doi.org/10.1093/pan/mpi009)

Using Pratt's Importance Measures in Confirmatory Factor Analyses

Amrey D. Wu

University of British Columbia
Vancouver, BC

Bruno D. Zumbo

University of British Columbia
Vancouver, BC

When running a confirmatory factor analysis (CFA), users specify and interpret the pattern (loading) matrix. It has been recommended that the structure coefficients, indicating the factors' correlation with the observed indicators, should also be reported when the factors are correlated (Graham, Guthrie, & Thompson, 2003; Thompson, 1997). The aims of this article are: (1) to note the structure coefficient should be interpreted with caution if the factors are specified to correlate. Because the structure coefficient is a zero-order correlation, it may be partially or entirely a reflection of factor correlations. This is elucidated by the matrix algebra of the structure coefficients based on the example in Graham et al. (2003). (2) The second aim is to introduce the method of Pratt's (1987) importance measures to be used in a CFA. The method uses the information in the structure coefficients, along with the pattern coefficients, into unique measures that are not confounded by the factor correlations. These importance measures indicate the proportions of the variation in an observed indicator that are attributable to the factors – an interpretation analogous to the effect size measure of eta-squared. The importance measures can further be transformed to eta correlations, a measure of unique directional correlation of a factor with an observed indicator. This is illustrated with a real data example.

Keywords: Variable importance ordering, Pratt's importance measures, pattern coefficient, structure coefficient, **D** matrix, eta correlation, coefficient of determination, confirmatory factor analysis, factor interpretation, multidimensional factor analysis, oblique factors

Introduction

When running a confirmatory factor analysis (CFA), users specify the pattern (loading) matrix and interpret the results. It has been recommended that the

Dr. Amery D. Wu is an Assistant Professor in the department of Educational and Counselling Psychology, and Special Education. Email her at: amery.wu@ubc.ca. Dr. Bruno D. Zumbo is the Paragon UBC Professor of Psychometrics and Measurement. Email him at: bruno.zumbo@ubc.ca.

structure coefficients, which represent the factors' correlation with the observed indicators, should be reported in addition to the loading matrix when the factors are allowed to be correlated (Graham et al., 2003; Thompson, 1997). This recommendation was made based on the argument that ignoring the structure coefficient is an omission of important information and leads to misinterpretation of CFA results.

It is important to attend to the information in the structure coefficient. In addition, the first aim of this study is to show the structure coefficient can be partly or entirely a reflection of the inter-factor correlations depending on the loading specification and the extent to which the factors are correlated. Therefore, structure coefficients should be interpreted with caution. In particular, there is a higher chance of misinterpretation when the two conditions coexist, namely; a model that has moderate or high correlations between factors and also few observed indicators cross-load on factors.

The second aim is to show how the directional and unique relationship, unconfounded by factor correlation, can be revealed by adapting Pratt's importance measures for factor analysis (Pratt, 1987). In doing so, we show that the structure coefficients, along with their corresponding pattern coefficients, can be transformed to importance measures in terms of variance explained. Thus, structure coefficients can be used to order the importance of the factors. The importance measures can further be transformed into unique, directional correlation coefficients (i.e., eta correlation) to aid in interpreting a CFA with correlated factors.

CFA Pattern and Structure Coefficients of Correlated Factors

In using a CFA, there should be a firm expectation of the underlying factor structure based on theoretical and/or empirical grounds (Church & Burke, 1994; Floyd & Widaman, 1995; Henson & Roberts, 2006). CFA requires a priori model specification regarding four elements of the factor structure: the number of factors, factor correlations, the pattern coefficients (referred to as "loadings" when the factor solution is unidimensional or orthogonal), and if necessary the residual correlations; see Jöreskog and Sörbom (1999) for the single group case and Wu, Li, and Zumbo (2007) for the multi-group case. In statistical terms, a CFA constrains a subset of the model parameters to some fixed values (typically zeros or ones) according to the investigator's hypothesis. For this reason, CFA is also referred to as restricted factor analysis in contrast to unrestricted factor analysis for exploratory

factor analysis (EFA) (Ferrando & Lorenzo-Seva, 2000; Jöreskog & Sörbom, 1999). Typically, the interest is in specifying the factor correlation(s) and the pattern coefficients.

The pattern coefficients are the slope coefficients, i.e., the partial regression weights estimated for the factors to yield the prediction of the observed indicators. These slope coefficients reflect the unique directional effect, that is, the amount of change in the observed score per unit change in the factor score taking into account the overlapping relationships among the factors when the factor solution allows for the factors to be correlated. In addition to the pattern coefficients, the structure coefficients may provide useful information that aid in interpreting the factor solution. The structure coefficients are the zero-order correlations between the observed indicators and the factors representing the non-directional relationship. The structure coefficients are analogous to the zero-order bivariate Pearson correlations without isolating the overlapping relationships among the factors (Graham et al., 2003; Thompson, 1997).

The matrix of the pattern coefficients is often denoted as $\mathbf{P}$, the matrix of the structure coefficients as $\mathbf{S}$, and matrix of the factor correlations as $\mathbf{R}$. Both $\mathbf{P}$ and $\mathbf{S}$ are of size $q \times p$ and $\mathbf{R}$ is of size $p \times p$, where q is the number of observed indicators and p is the number of factors. The relationship between $\mathbf{P}$, $\mathbf{S}$, and $\mathbf{R}$ is given as

$$\mathbf{S}_{q \times p} = \mathbf{P}_{q \times p} \mathbf{R}_{p \times p} \quad (1)$$

Note when factors are uncorrelated the $\mathbf{R}$ is an identity matrix and in this case $\mathbf{S} = \mathbf{P}$. When the factors are uncorrelated, the zero-order bivariate correlation also represents the unique directional effect. This is because the factors contain no overlapping information to be isolated. In this case, the pattern coefficients are equal to the structure coefficients and are indistinctly referred to as factor loadings.

The structure correlation is by definition non-directional. It is inconsequential whether the factor or the observed indicator comes first in the pair when estimating the correlation. Also, the structure coefficient is a zero-order correlation representing a relationship without controlling for the confounding relationships with other variable(s). This is necessary to understand why the structure coefficient can be a reflection of confounding relationships with the factor correlation. The level of confounding depends on the loading specification and the extent to which the factors are correlated. Equation (1) will be used to demonstrate that a structure coefficient can be misleading in a CFA due to these specifications.

PRATT'S MEASURES IN CFA

Table 1. CFA results for Graham et al.'s (2003) example and the Pratt's measures

	P		S		L		PS		$\sqrt{\mathbf{PS}}$		h²	D	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2		F1	F2
A	0.849(g)	0.000(h)	0.849(i)	0.580(j)	0.836	0.000	0.721	0.000	0.849	0.000	0.721	1.000	0.000
B	0.726	0.000	0.726	0.495	0.721	0.000	0.527	0.000	0.726	0.000	0.527	1.000	0.000
C	0.817	0.000	0.817	0.557	0.836	0.000	0.667	0.000	0.817	0.000	0.667	1.000	0.000
D	0.000	0.875	0.597	0.875	0.000	0.855	0.000	0.766	0.000	0.875	0.766	0.000	1.000
E	0.000	0.774	0.528	0.774	0.000	0.777	0.000	0.599	0.000	0.774	0.599	0.000	1.000
F	0.000	0.808	0.552	0.808	0.000	0.794	0.000	0.653	0.000	0.808	0.653	0.000	1.000
R													
F1	1.000(k)	0.680(l)											
F2	0.680(m)	1.000(n)											

Note: **P**: pattern matrix; **S**: structure matrix; **L**: loading matrix; **PS**: a matrix of which the elements are the products of a given pattern coefficient and its corresponding structure coefficients, i.e., the unstandardized Pratt's measures, they are analogous to the coefficient of determination η^2 (eta-squared); $\sqrt{\mathbf{PS}}$: the square root of **PS**, i.e., η (eta) correlation; **h²**: Communality; **D**: a matrix of communality-standardized Pratt's measure matrix; **R**: factor correlation matrix.

Caveats to Interpreting Structure Coefficients in CFA

Graham et al. (2003) and Thompson (1997) called for the reporting of the structure coefficients in CFA. Their recommendation was based on the argument that constraining a factor's pattern coefficient to be zero does not automatically constrain its corresponding structure coefficient to be zero. Hence, the structure coefficients should not be ignored. These earlier works suggested that, to properly interpret CFA results, the structure coefficients should be juxtaposed and interpreted along with the pattern coefficients; otherwise, the interpretation may be problematic.

The first data set generated by Graham et al. (2003) was re-analyzed; based on which they highlighted that the structure coefficients were not zero when the pattern coefficients were specified to be zeros. For that data set, two factors that correlate at 0.68 were hypothesized to be underlying six observed indicators, as shown in Table 1, the **P** and **S** matrices reported by Graham et al. The second and third columns (under the heading **P**) show that the first factor (F1) only had partial effects on the first three observed indicators; the second factor (F2) only had partial effects on the last three observed indicators; all the other pattern effects were fixed to be zeros indicating no factorial complexities (e.g., no cross-loadings). This example is an ideal representation of simple structure, which is often a common and preferred configure for CFA specification. Graham et al.'s point was: despite the zero constraint on the pattern coefficients, the corresponding structure coefficients still yielded substantial values as highlighted in bold face in Table 1 under the heading of **S**. For example, although the pattern coefficient of F2 on indicator A was constrained to be zero, its corresponding structure coefficient of 0.58 was salient enough and should not be ignored. Using the examples in Table 1, Graham et al. raised the concern of missing out on important information if the structure coefficient is not interpreted.

Ignoring the structural relationship may omit important information; however, they should be interpreted with careful consideration. Below is an explanation for why a structure coefficient can be misleading in a CFA with correlated factors when accompanied by a zero pattern coefficient, as in the case for all six indicators in Table 1.

Numerical Calculations to Demonstrate How Structure Coefficients Can Be Misleading

The structure coefficient in Table 1 can be misleading because the estimated correlation of 0.58 between F2 and indicator A is a result of factor correlation between F1 and F2. The correlation between F2 and indicator A is due to indicator A's correlation with F1, which in turn correlates with F2. That is, both F2 and indicator A are correlated with F1. The substantial zero-order bivariate correlation between F2 and indicator A would turn to zero once the factor correlation between F1 and F2 is controlled for. Hence, the substantial correlation between F2 and indicator A, as indicated by the structure coefficient, is simply a result of factor correlation. Interpreting the structure relationship while neglecting the factor correlation can mislead the conclusions.

The matrix algebra multiplication in equation (1) demonstrates the above account. Plugging the information in Table 1 into equation (1), the resulting structure coefficient of 0.58 between F2 and indicator A, denoted as (j) in Table 1, is the sum of two product terms calculated by the values in cells denoted as (g), (l), (h), and (n) such that

$$\begin{aligned}
 (j) &= 0.58 \\
 &= (g) \times (l) + (h) \times (n) \\
 &= (\text{pattern of F1 on A}) \times (\text{corr. between F1 \& F2}) \\
 &\quad + (\text{pattern of F2 on A}) \times (\text{corr. between F1 \& F2}) \\
 &= 0.849 \times 0.68 + 0 \times 1
 \end{aligned}$$

Because the second product term is equal to zero due to the zero constraint on the pattern coefficient of F2 on indicator A, the structure coefficient (j) between F2 and indicator A is entirely attributable to the first product term. The first product term is the partial effect of F1 on indicator A (0.849) times the correlation between F1 and F2 (0.68). This product term, however, has nothing to do with any relationships between F2 and indicator A. Demonstrating the calculation of the structure coefficient (j) clearly shows that, when the corresponding pattern coefficient is constrained to be zero, the moderately high structural relationship of 0.58 between F2 and indicator A is simply a result of the correlation between F1 and F2.

The above account was not meant to negate the information in the structure coefficients. It has been shown that ice cream sales and drowning rate are highly

correlated. Indeed, there is some useful information embedded in this correlation and should not be simply ignored. However, once controlling for both ice cream sales and drowning rates are also highly correlated to temperature, there may be little to no relationship between ice cream sale and drowning rate. The intention is to bring the users of CFA to this realization when interpreting the structure coefficients. This caveat for interpreting the structure coefficient is heightened in the case of dealing with latent variables (the example of ice cream sales, drowning rates, and temperature consists of only observed variables). The latent variables are mathematical creations that do not have inherent meaning. This makes interpretation even more prone to confounding factors than the already confounded case of the observed variables for ice cream sales and drowning rate. The substantive meanings of the factors are inferred from the indicators. In turn, the indicators' relationships with the factors, through the structure coefficients, are being estimated and interpreted at the same time. This circularity makes the interpretation the zero-order structure coefficients even more subtle.

Historical Method to Sidestep the Problem

Because of factor correlation, the structure coefficients can be confounded and can sometimes point to different conclusions from those of the pattern coefficients. This can lead to difficulty in drawing conclusions. Conventionally, the interpretation difficulties arising from factor correlation are often avoided by constraining the factors to be uncorrelated. This is because, as shown above, when factor correlations are zero, estimates of the pattern and the structure coefficients will be identical and synonymously called loadings. They represent both a factor's correlations with as well as a factor's partial effects on an observed indicator. In this case, the structure coefficient does indeed represent the unique relationship with an observed variable because it is not confounded by that factor's correlation with other factors.

In addition, resorting to factor orthogonality lends to the additive property in terms of variance explained by the factors. When factors are uncorrelated, the square of a loading represents the amount of variance in an observed indicator that is accounted for by a factor. Hence, the sum of the squared loadings across the factors will add up to the communality of an observed indicator – in CFA terms, this is the R-squared of a regression equation for an observed indicator variable. This additive property makes the interpretation very straightforward. Unfortunately, due to factor correlation, correlated factor models do not hold this additive property for straightforward interpretation.

PRATT'S MEASURES IN CFA

Although resorting to orthogonal factors avoids the interpretational difficulties arising from factor correlation, it may lead to the problem of an incorrect model because the factors may indeed be correlated in the population. We fit the orthogonal model to the data for Table 1, which were generated by an oblique model with a correlation of 0.68. The problem of model misspecification was evidenced by the poor fit indices: the $\chi^2_{(df=9)} = 130.519$, $p < 0.001$, CFI = 0.865, and RMSEA = 0.190 (90% CI: 0.172 – 0.206) as a result of fitting the incorrect orthogonal model. In contrast, fitting a correlated factor model dramatically improved the fit with only one degree of freedom difference. The fit indices were almost perfect when an oblique model with a simple structure shown in Table 1 was specified; viz., $\chi^2_{(df=8)} = 2.792$, $p = 0.904$, CFI = 1.00, and RMSEA = 0.000 (90% CI: 0.000 – 0.000). This almost perfect fit was a consequence of recovering the model that generated the data. The problem of model misspecification due to fixing the inter-factor correlation to zero also led to biased estimates of loadings. These biases can be seen in Table 1 by comparing the estimated loadings reported as in the **L** matrix to the corresponding pattern coefficients in the **P** matrix (i.e., loadings estimated by the model that generated the data. Biases are the evident differences in comparing the orthogonal loading estimates to those of corresponding oblique loading estimates, rather than to the population parameters).

Resorting to uncorrelated factors to avoid the interpretation difficulties due to factor correlation often contradicts the rationale for using a CFA if the factors are a priori hypothesized to be correlated. Many constructs in the social, behavioural, and health sciences are by their very nature assumed to be not entirely distinct. Frequently, allowing factor correlation for better theoretical and statistical fit occurs, leading to potential difficulties in interpreting the results (inconsistent conclusions based between the pattern and the structure coefficients). Still, an orthogonal model may be chosen over a correlated factor model for its interpretational simplicity, even when the factors are theoretically or empirically shown to be otherwise (Conway & Huffcutt, 2003; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Henson & Robert, 2006; Kieffer, 1998; Preacher & MacCallum, 2003).

Pratt's Importance Measures in CFA

Pratt's relative importance measures transform the information in the structure and pattern coefficients into unique measures that are readily attributable to the factors despite factor correlation. Pratt's relative importance measures were initially developed for use in multiple regression (Pratt, 1987; Thomas, Hughes, & Zumbo,

1998). This method was adapted to EFA (Wu, 2008; Wu, Zumbo, & Marshall, 2014) by considering factor analysis as a form of multiple regression such that a factor analysis simultaneously regresses the q observed indicators (i.e., dependent variables) onto the p common factors (i.e., predictor variables) (Gorsuch, 1983; Wu et al., 2014). In this paper, we will explain the use of Pratt's importance measures in CFA.

Pratt's Importance Measures

It is sometimes recommended the importance of a set of p independent variables can be order by the absolute value of $\hat{\beta}_p$, the standardized partial regression coefficient. It is believed that $\hat{\beta}_p$ is a standardized measure that circumvents the issues of incomparability; namely, the incomparability due to the unstandardized regression coefficients being estimated for the independent variables and having different units of measurement. This suggestion is problematic because it ignores the fact that the partial regression coefficient, whether it be standardized or not, is a measure of relationship between a specific predictor variable with the outcome variable controlling for the relationships with the rest of the $(p - 1)$ predictor variables. However, for different predictor variables, the set of $(p - 1)$ controlled relationships will be different, and hence their importance is not directly comparable. This problem was resolved by Pratt (1987).

Pratt (1987) showed that this unique measure of the importance of an predictor variable could be expressed as the product of $\hat{\beta}_p \hat{\rho}_p$ where $\hat{\rho}_p$ denotes the estimate of Pearson's product moment correlation between the predictor and the dependent variable, and $\hat{\beta}_p$ denotes the standardized regression coefficient. The standardized Pratt's measure, d_p for the relative importance of the p^{th} predictor variable is given by

$$d_p = \frac{\hat{\beta}_p \hat{\rho}_p}{R^2} \quad (2)$$

Because

$$\sum_{p=1}^w \hat{\beta}_p \hat{\rho}_p = R^2$$

PRATT'S MEASURES IN CFA

it follows that

$$\sum_{p=1}^w \frac{\hat{\beta}_p \hat{r}_p}{R^2} = 1$$

hence

$$\sum_{p=1}^w d_p = 1$$

a result that was shown by Thomas et al.'s (1998) geometric derivation. Accordingly, the importance of the predictor variables then can be ordered by d_p . The essential feature of Pratt's importance measures is the *additive property* such that the sum of the unstandardized Pratt's measures is equal to the R^2 and the sum of the standardized Pratt's measures is equal to one. See Table 1 in Wu et al. (2014, p. 99) for an example of multiple regression.

Pratt's Importance Measures in CFA

Consider factor analysis as a form of q simultaneous regression analyses wherein one regresses each of the q observed indicators onto the p common factors. From this framework, Pratt's importance measures can be easily applied to multidimensional factor analysis. The outcome of applying Pratt's measures in a factor analysis is the Pratt's measure matrix, referred to as the **D** matrix. The elements of the **D** matrix are the Pratt's measure of the p^{th} factor for the q^{th} observed indicator. The three building blocks for producing the **D** matrix in factor analysis are the pattern matrix **P**, the structure matrix **S**, as well as the vector of the communalities $\mathbf{h}^2$, in which the elements are the equivalent to the R -squared values in a multiple regression. Using matrix algebra, the **D** matrix is expressed as

$$\mathbf{D} = \mathbf{P} \otimes \mathbf{S} \quad (3)$$

where **P** and **S** are defined above, and $\otimes$ denotes the Hadamard product of matrices of the same order. The Hadamard product expresses the elementwise product of matrices (Rao & Rao, 1998; Styán, 1973). Because it is seldom used, the Hadamard product is not available in widely used statistical software. However, it can be easily handled in a spreadsheet such as Excel. To obtain the unstandardized Pratt's

measure of the p^{th} factor (predictor variable) for the q^{th} indicator (dependent variable), simply multiply the pattern coefficient by its corresponding structure coefficient. One can complete the computation of the **D** matrix by repeating the same procedures for all q indicators. The corresponding standardized Pratt measure can then be obtained by dividing the unstandardized value by the communality of the q^{th} indicator. See Wu (2008) and Wu et al. (2014) for a full explanation the **D** matrix and more examples for its calculation.

Real Data Illustration

The application of Pratt's measures will be illustrated in a CFA with real data. The data consists of 314 participants' responses to 13 items measuring the two dimensions (knowledge and action) of health self-care reported on a 4-point Likert-type scale. Accordingly, a two correlated factor model was fit to the data. Based on the previous results from EFA, items 1 to 7 were specified to indicate only the first factor, items 10 to 13 to indicate only the second factor; however, items 8 and 9 were specified to indicate both factors. This is an example of two factorial complexities (i.e., factor cross-loading on item 8 and 9). The estimates of the pattern and structure coefficients are shown in Table 2 under the headings of **P** and **S**.

Pratt's Importance Measures with Cross-Loadings:

In Table 2, the observed items of V8 and V9 were in bold face to highlight the cross-loading specification as shown by the pattern coefficients. The products of pattern coefficients and their corresponding structure coefficients are under the heading of "**PS**." These are the unstandardized Pratt's measures indicating the amount of variance of an item explained by each of the two factors. Take item 8 (V8) for example, the unstandardized Pratt's measure of 0.388 for Factor 1 (F1) was obtained by 0.544×0.714 , and the unstandardized Pratt's measure of 0.136 for Factor 2 (F2) was obtained by 0.208×0.653 . These unstandardized Pratt's importance measures are equivalently to the concepts of coefficient of determination or eta-squared (η^2), and can be interpreted as the unique contribution of a factor to an item's observed variation.

Each value under the heading of " $\sqrt{\text{PS}}$ " in Table 2 (i.e., the square roots of PS) is the unique directional correlation between a given factor and an item by taking into account the factor correlation. Their interpretation is analogous to the η (eta) correlation in ANOVA except that, in this case, the factors in this application are continuous latent variables instead of observed grouping variables in an

PRATT'S MEASURES IN CFA

Table 2. Real data demonstration of the use of Pratt's measures in an oblique CFA

	P		S		PS		$\sqrt{\mathbf{PS}}$			D	
	F1	F2	F1	F2	F1	F2	F1	F2	h²	F1	F2
V1	0.773	0.000	0.773	0.632	0.598	0.000	0.773	0.000	0.598	1.000	0.000
V2	0.801	0.000	0.801	0.655	0.642	0.000	0.801	0.000	0.641	1.000	0.000
V3	0.771	0.000	0.771	0.631	0.594	0.000	0.771	0.000	0.595	1.000	0.000
V4	0.752	0.000	0.752	0.615	0.566	0.000	0.752	0.000	0.566	1.000	0.000
V5	0.785	0.000	0.785	0.642	0.616	0.000	0.785	0.000	0.617	1.000	0.000
V6	0.837	0.000	0.837	0.685	0.701	0.000	0.837	0.000	0.701	1.000	0.000
V7	0.842	0.000	0.842	0.689	0.709	0.000	0.842	0.000	0.710	1.000	0.000
V8	0.544	0.208	0.714	0.653	0.388	0.136	0.623	0.369	0.524	0.740	0.260
V9	0.273	0.551	0.724	0.774	0.198	0.427	0.444	0.653	0.624	0.320	0.680
V10	0.000	0.843	0.690	0.843	0.000	0.711	0.000	0.843	0.710	0.000	1.000
V11	0.000	0.883	0.722	0.883	0.000	0.780	0.000	0.883	0.781	0.000	1.000
V12	0.000	0.800	0.654	0.800	0.000	0.640	0.000	0.800	0.639	0.000	1.000
V13	0.000	0.661	0.541	0.661	0.000	0.437	0.000	0.661	0.437	0.000	1.000
R											
F1	1.000	0.818									
F2	0.818	1.000									

Note: **P**: pattern matrix; **S**: structure matrix; **PS**: a matrix of which the elements are the products of a given pattern coefficient and its corresponding structure coefficients, i.e., the unstandardized Pratt's measures, they are analogous to the coefficient of determination η^2 (eta-squared); $\sqrt{\mathbf{PS}}$: the square root of **PS**, i.e., η (eta) correlation; **h²**: Communality; **D**: a matrix of communality-standardized Pratt's measure matrix; **R**: factor correlation matrix

ANOVA. This eta correlation was also referred to as correlation ratio by Pearson (1905); refer to Huberty (2002) for a historical review. For V8, the unique eta correlation with F1 of .623 and with F2 of .369, as indicated by the $\sqrt{\mathbf{PS}}$, were noticeably smaller than the structural correlation of 0.714 and 0.653 as shown in Table 2. This is because the structural correlations did not take into account the factor correlation of 0.818. That is, the structure coefficient is a confounded measure of unique correlation between a factor and an observed item.

To compute the standardized Pratt measures, the values in the **PS** matrix were further divided by the item's communality value. This communality-standardization resulted in proportions that add up to one across the factors for each item. For V8, dividing the two values 0.388 and 0.136 by the communality 0.524 yields the standardized Pratt's measures of 0.74 and 0.26. They are reported in the matrix under the heading of **D**. The interpretations are as follows: The two standardized Pratt's measures indicate the proportion of communality of V8 that was accounted for by each of the two factors, respectively. Namely, F1 accounted for 74% and F2 accounted for 26% of the communality of V8. In other words, Pratt's measures partitioned the communality (*R*-squared) of an item into two additive parts that could be readily attributable to the two factors despite the high factor correlation of 0.818 between the two factors. The standardized Pratt's measures are particularly useful in ordering the relative importance of a greater number of factors because of their additive property. This is the case despite the complex correlation pattern among factors.

Pratt's Measures Without Cross-Loading

V8 was chosen as an example because of the cross-loading specification, i.e., neither the pattern coefficient for F1 nor for F2 was constrained to be zero. One of the key points of this paper is to demonstrate what happens to the values in **PS** (η^2), $\sqrt{\mathbf{PS}}$ (η), and **D** when the factors do not cross-load, i.e., the pattern coefficient of one of the two factors was constrained to be zero. Note the no-cross-loading specification is the same as that for all the six indicators in Table 1 taken from Graham et al.'s (2003) paper.

For example, if the focus is on item one (V1) in Table 2, the value in the matrix **PS** shows that F1 accounted for 59.8% (that is all) of the observed variance (i.e., 100% of the communality), whereas F2 accounted for 0.0% of the observed variance (0% of the communality)! The reason that F2 explained none of the observed variance in V1 was that the structure coefficient of .632 shown in Table

2 was multiplied by the pattern coefficient of zero which had been a priori specified by the authors. Once the factor correlation was accounted for by the application of Pratt's method, the unique correlation, $\sqrt{\mathbf{PS}}$, between V1 and F2 turned out to be zero. This indicates that F2 was not uniquely related with V1 and was unable to explain any variance of V1. For the same reason, the values of $\mathbf{PS}$, $\sqrt{\mathbf{PS}}$, and $\mathbf{D}$ were all zero in Table 2 where the pattern coefficients were constrained to be zero (V1 to V7 for F2 and V10 to V13 for F1); Conclusions of CFA based on these fairly large face values of the structure coefficients (at least 0.615 among the no-cross-loadings items in Table 2) without realizing that they are merely a reflection of the factor correlation. This is the same as the example in Table 1 wherein all the values for $\mathbf{PS}$, $\sqrt{\mathbf{PS}}$, and $\mathbf{D}$ are also equal to zero where the pattern coefficients were constrained to zero after the factor correlation was taken into account by Pratt's measures.

Conclusion

In CFA, be warned that the size of structure coefficients is confounded with the level of the factor correlations and should be interpreted with caution. The manner and the extent of the confounding depend on the following conditions. When factor cross-loading is allowed, a structure coefficient over-represents a factor's unique correlation with an observed indicator to the extent that the factors inter-correlate. When factor cross-loading is restricted, for the factor of which the pattern coefficient is specified to be zero, the structure coefficient of that factor is merely a reflection of the factor correlation.

The interpretation difficulties arising from factor correlation were traditionally avoided by constraining the factors to be orthogonal – i.e., uncorrelated. Factor orthogonality holds the additive property in terms of unique variance explained by the factors that is not confounded by factor correlation. Such a property makes the interpretation straightforward. Nonetheless, this approach raises many concerns with respect to theory and model misfit. Pratt's measures applied to CFA restore the additive property distorted by factor correlation; hence it resolves the interpretational complexities arising from factor correlation without having to constrain factors to be uncorrelated.

Pratt's measures integrate the information in a pattern and a structure coefficient by transforming them into one single unique measure that is grounded on Pratt's axioms and Thomas et al.'s geometry. The transformed measure

represents the proportion of variance that is uniquely attributable to a given factor despite its correlations with other factors. The interpretation is analogous to that of effect size measure, eta-squared. The communality-standardized version of a Pratt's measure indicates the proportion of communality (R -squared) accounted for by each of the factors. They can be used to order the importance of the factors and help to enhance the interpretation of the results, in particular, when the solution allows for cross-loadings, is highly dimensional and correlated. In so doing, Pratt's measures applied to factor analysis resolves a longstanding problem in the interpretation of factor analysis solutions with correlated factors.

By taking the square root of an unstandardized Pratt's measure, one can obtain a measure of which the meaning is analogous to the eta correlation. The eta correlation can be understood as a directional, unique, simple correlation between an observed indicator and a factor even in the case when the factors are correlated. When a factor cross-loads, the eta correlation downward adjusts the relationship of a factor with an observed indicator by removing the confounding with factor correlation. When a factor does not cross-load (the pattern coefficient being constrained to be zero), the Pratt's measures method will yield an importance measure of zero, hence an eta correlation equals zero. In this case, even though a factor may have a notable zero-order relationship with an observed indicator as shown by the structure coefficient, it actually accounts for zero variation in the observed indicator. Interpretation of the structure coefficient should take this fact into account.

Pratt's measures can also be useful for Exploratory Factor Analysis (EFA). This is because EFA can be seen as a particular type of CFA specification where the all factors' pattern relationships are estimated for all observed indicators (with no zero constraints at all; see [Wu, 2008](#); [Wu et al., 2014](#)).

It was shown how the unique directional correlation between factors and observed indicator, un-confounded by factor correlation, can be revealed by synthesizing the information in the structure and pattern coefficients via the method of Pratt's importance measures. Following the simultaneous regression logic one may ask about the use of partial and semi-partial (part) correlations to handle the confounding effect arising from factor correlation. These n^{th} order-controlled correlations (n = number of controlled variables) reflect non-directional relationship between two variables, as is the structure coefficient. They are the correlation between two scores that are residualized by the n variables. Indeed, they can be computed to indicate the un-confounded correlations. However, unlike Pratt's measures, these measures are not comparable across the factors. The incomparability issue is the same as that of the standardized partial regression

PRATT'S MEASURES IN CFA

coefficients in multiple regression and the pattern coefficients in factor analysis; such that the set of $(p - 1)$ variables being controlled for are not the same. The ultimate advantages of Pratt's measures over the partial and semi-partial measures are: (1) their intuitive meaning as proportion variance explained makes the interpretation very straightforward and (2) their additive property makes the comparison across the factors meaningful.

There is no intent to negate the importance of the structure coefficients in CFA. In fact, recent recommendations that the information in the structure coefficients should not be ignored. Nonetheless, CFA users should note the structure coefficient should be interpreted cautiously knowing that they may be partially or entirely a reflection of factor correlation. Better still, consider applying Pratt's easily computed measures.

References

- Church, J. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the big five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology*, 66(1), 93-114. doi: [10.1037/0022-3514.66.1.93](https://doi.org/10.1037/0022-3514.66.1.93)
- Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, 6(2), 147-168. doi: [10.1177/1094428103251541](https://doi.org/10.1177/1094428103251541)
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299. doi: [10.1037/1082-989x.4.3.272](https://doi.org/10.1037/1082-989x.4.3.272)
- Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspect of the problem and some suggestions. *Psicológica*, 21(2), 301-323. Retrieved from <http://www.redalyc.org/articulo.oa?id=16921206>
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286-299. doi: [10.1037/1040-3590.7.3.286](https://doi.org/10.1037/1040-3590.7.3.286)
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder.

Structural Equation Modeling: A Multidisciplinary Journal, 10(1), 142-152. doi: 10.1207/s15328007sem1001_7

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393-416. doi: 10.1177/0013164405282485

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, 62(2), 227-240. doi: 10.1177/0013164402062002002

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 user's reference guide*. Chicago, IL: Scientific Software International.

Kieffer, K. M. (1998). *Orthogonal versus oblique factor rotation: A review of the literature regarding the pros and cons*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA.

Pearson, K. (1905). *Mathematical contributions to the theory of evolution, XIV: On the general theory of skew correlation and non-linear regression*. Drapers' Company research memoirs, biometric series II. London, UK: Dulau.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In T. Pukilla & S. Duntaneu (Eds.), *Proceedings of Second Tampere Conference in Statistics* (pp. 245-260). Tampere, Finland: University of Tampere.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2(1), 13-43. doi: 10.1207/s15328031us0201_02

Rao, C. R., & Rao, M. B. (1998). *Matrix algebra and its applications to statistics and econometrics*. Singapore: World Scientific.

Styan, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra and its Applications*, 6, 217-240. doi: 10.1016/0024-3795(73)90023-2

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275. doi: 10.1023/a:1006954016433

Thompson, B. (1997). The importance of structure coefficients in structural equation modeling confirmatory factor analysis. *Educational and Psychological Measurement*, 57(1), 5-19. doi: 10.1177/0013164497057001001

PRATT'S MEASURES IN CFA

Wu, A. D. (2008). *Pratt's importance measures in factor analysis: A new technique for interpreting oblique factor model* (Unpublished doctoral dissertation). University of British Columbia, Vancouver, BC.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3). Retrieved from <http://pareonline.net/genpare.asp?wh=0&abt=12>

Wu, A. D., Zumbo, B. D., & Marshall, S. K. (2014). A method to aid in the interpretation of EFA results: An application of Pratt's measures. *International Journal of Behavioral Development, 38*(1), 98-110. doi: 10.1177/0165025413506143

Robust Measures of Variable Importance for Multivariate Group Designs

Tolulope T. Sajobi
University of Calgary
Calgary, AB

Lisa M. Lix
University of Manitoba
Winnipeg, MB

Variable importance measures based on discriminant analysis and multivariate analysis of variance are useful for identifying variables that discriminate between two groups in multivariate group designs. Variable importance measures are developed based on trimmed and Winsorized estimators for describing group differences in multivariate non-normal populations.

Keywords: Variable importance, discriminant analysis, multivariate analysis of variance, trimmed estimators, assumption violations

Introduction

In educational and behavioral research studies where two groups (e.g., treatment versus control) are compared on a battery of outcome variables, testing for significant differences between the groups and identifying the relative importance of the variables that may discriminate between groups may be of interest. For example, Shih (2012) examined how Taiwanese junior high school students' perfectionistic tendencies and achievement goals were related to their academic burnout versus work engagement. The differences between indicators of burnout versus engagement among students with different subtypes of perfectionism was also investigated. The contributions of perfectionistic tendencies to academic burnout and engagement were found to be far greater than those of achievement goals.

Bogler (2002) also investigated differences between teachers with a high level of job satisfaction and those with low job satisfaction on occupational, leadership, and demographic characteristics. Teachers' perception of their occupation and their principal's transformational style were the most important characteristics that

Dr. Tolulope Sajobi is an Assistant Professor of Biostatistics in the Cumming School of Medicine. Email him at: ttsajobi@ucalgary.ca. Dr. Lisa Lix is a Professor of Biostatistics and Manitoba Research Chair in the Department of Community Health Sciences.

ROBUST MEASURES OF VARIABLE IMPORTANCE

discriminated between the two types of teachers. There are several reasons why it may be interesting to identifying the outcomes on which group differences exist.

There may be limited research knowledge about which outcome(s) will be responsive to treatment, or little consensus about which outcome(s) is (are) relevant for behavioral or educational intervention. The intervention might be intended to have a multi-faceted effect, necessitating a research question that focuses on more than one outcome.

Measures to identify the relative importance of outcomes that discriminate between two independent groups were developed based on descriptive discriminant analysis (DDA) and multivariate analysis of variance (MANOVA). They quantify the relative importance of a variable (or outcome) based on its contribution to grouping effects and discriminant function scores (Huberty & Wisenbaker, 1992; Thomas, 1997). DDA measures of variable importance identify one or more linear combinations of the variables that maximize group separation; they are based on functions of the discriminant function coefficients and include standardized discriminant function coefficients, discriminant ratio coefficients, and total discriminant ratio coefficients (Huberty & Wisenbaker, 1992; Thomas, 1992). The *F*-to-remove statistic, a stepwise procedure based on MANOVA, has also been recommended as a variable importance measure; it quantifies the importance of a variable based on its unique contribution to group separation beyond that contributed by the remaining study variables (Huberty, 1984; Huberty & Wisenbaker, 1992). Applications of these measures have appeared in several disciplines including behavioral psychology (Sperling, Schilling, Glosser, Tracy, & Asadi-Pooya, 2008), criminology (Eastman & Bunch, 2009), development of questionnaires (Richardson, 2007), and educational research (Holder, 2007; Curenton, McWey, & Bolen, 2009).

However, DDA and MANOVA rest on the assumption of multivariate normality and covariance homogeneity, two assumptions which may not always be satisfied in practice. Normality of the outcome variables may not be a tenable assumption in behavioral or educational research investigations, which frequently exhibit multi-modal, skewed, or heavy-tailed distributions (Cressie & Whitford, 1986; Micceri, 1989). As well, the treatment group may exhibit greater variability than the control group (Blair & Sawilowsky, 1993; Troendle, Blair, Rumsey, & Moke, 1997). Previous research has shown that measures of variable importance may result in incorrect rank ordering of a set of correlated variables when applied to non-normal data with heterogeneous covariances (McLachlan, 1992). Thus, departures from the assumption of multivariate normality may have serious

consequences for investigators interested in identifying the outcomes that explains the differences between two or more groups.

Linear discriminant analysis procedures that are robust (i.e., insensitive) to departures from the assumption of multivariate normality have been proposed (Todorov, Neykov, & Neytchev, 1994) by replacing the conventional least-squares estimators of means and covariances with robust estimators, such as M-estimators (Croux & Dehon, 2001), minimum covariance determinant (MCD) estimators (Hubert & Van Driessen, 2004; Rousseeuw, 1984), minimum volume ellipsoid (MVE) estimators (Rousseeuw, 1984), and trimmed estimators (Ahmed & Lachenbruch, 1977; Gnanadesikan & Kettenring, 1972; Srivastava & Mudholkar, 2001). However, their emphasis has been primarily on prediction and not on describing the variables that contribute to group separation in non-normal data. There has been limited investigation of robust variable importance measures for evaluating the relative importance of variables in multivariate non-normal data.

Robust measures of variable importance are developed here in which conventional least squares (LS) estimates of the means and covariances are replaced by trimmed means and Winsorized covariance parameters based on coordinatewise trimming (CT) of the multivariate data. Trimmed means and Winsorized covariances possess good theoretical properties for heavy-tailed and skewed distributions, are computationally efficient, and straightforward to implement (Wilcox, 1994; Srivastava & Mudholkar, 2001). The sensitivity of the robust variable importance measures to departures from derivational assumptions and other data-analytic characteristics are investigated using Monte Carlo techniques.

Relative Importance Measures based on Descriptive Discriminant Analysis

Consider the two-group problem, although all of the procedures can be generalized to multi-group designs. Let $\mathbf{y}_{ij}$ be the $p \times 1$ vector of observed measurements for the i^{th} study participant ($i = 1, \dots, n_j$) in the j^{th} group ($j = 1, \dots, J$). Initially, assume $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the population mean and covariance for the j^{th} group and are estimated by $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$, respectively. For the linear DA procedure, the discriminant function coefficient vector is estimated by

$$\hat{\mathbf{a}} = \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \quad (1)$$

where

ROBUST MEASURES OF VARIABLE IMPORTANCE

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2}{n_1 + n_2 - 2} \quad (2)$$

The number of uncorrelated discriminant functions that separates G groups is equal to $\min(J - 1, p)$.

In DDA, variable importance measures based on discriminant function coefficients can be used to rank variables according to their contributions to group separation (Huberty & Wisenbaker, 1992). The computation and implementation of these measures have been described in detail elsewhere (Sajobi, Lix, Clara, et al., 2012), and are only briefly reviewed here.

The standardized discriminant function coefficient (SDFC) is one commonly adopted variable importance measure. It quantifies the importance of a variable by taking into account the presence of other variables in the study. The SDFC for the k^{th} variable, denoted by $\hat{a}_k^*$, is

$$\hat{a}_k^* = \hat{a}_k s_k \quad (3)$$

where $\hat{a}_k$ and s_k are the corresponding estimated k^{th} discriminant function coefficient and standard deviation, respectively. By placing a constraint on the discriminant function coefficients such that $\hat{\mathbf{a}}^T \mathbf{S}_E \hat{\mathbf{a}} = 1$, where T is the transpose operator, $\mathbf{S}_E = \hat{\Sigma}$, and the coefficients range in value from -1 to +1. SDFCs can be positive or negative and the absolute magnitude determines relative importance. Although there have been arguments in favor of using SDFCs to measure variable importance (Rencher, 2002), they are known to be sensitive to variable correlations (Rencher, 2002).

Discriminant ratio coefficients (DRCs) are sometimes recommended instead of SDFCs (Thomas, 1992; Thomas & Zumbo, 1996). DRCs measure the importance of a variable as a proportion of the group differences explained by the variable. The k^{th} DRC is given by

$$q_k = a_k^* f_k \quad (4)$$

where f_k is the k^{th} structure coefficient, the correlation between the k^{th} outcome variable and the discriminant function. DRCs generally range in value from zero to one, with larger values indicating greater importance. However, they can have

negative values; a large negative value may be indicative of collinearity or suppression (Thomas & Zumbo, 1996). In MANOVA, suppression occurs when a variable makes little or no direct contribution to group separation on its own but contributes indirectly through another variable.

The F -to-remove statistic (Huberty & Wisenbaker, 1992) is obtained by conducting p MANOVA tests, each time removing one variable from the analysis. For the k^{th} domain, the statistic is

$$F_{(k)} = \frac{k_2 \left(\hat{a}_k / s_{(kk)} \right)^2}{\left(\bar{z}_1 - \bar{z}_2 \right) + k_3 - \left(\hat{a}_k / s_{(kk)} \right)^2} \quad (5)$$

where $k_2 = (n_1 + n_2 - 2 - q)$, $k_3 = (n_1 + n_2)(n_1 + n_2) / n_1 n_2$, $\hat{a}_k$ is the discriminant function coefficient for the k^{th} domain, $\bar{z}_1$ and $\bar{z}_2$ are the group means for the discriminant function score corresponding to $\hat{\mathbf{a}}$, and $s_{(kk)}$ is the positive square root of the k^{th} diagonal element of the inverse of $\mathbf{E}$, the error sums of square and cross product matrix. F -to-remove statistics have a lower bound of zero, but no upper bound. Variable importance is assessed by the magnitude of F -to-remove-statistic, with the most important domain yielding the largest statistic.

The total discriminant ratio coefficient (TDRC) for the k^{th} variable is

$$t_k = \frac{|\hat{a}_k| \mathbf{S}_{\mathbf{T}kk}}{\sqrt{(\hat{\mathbf{a}} \mathbf{S}_{\mathbf{T}} \hat{\mathbf{a}})}} \quad (6)$$

where

$$\mathbf{E} = \sum_{j=1}^2 \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)^{\text{T}} \quad (7)$$

is the error sum of squares and cross product matrix, $\mathbf{S}_{\mathbf{T}kk}$ is the $(k, k)^{\text{th}}$ element of $\mathbf{S}_{\mathbf{T}}$, $\mathbf{S}_{\mathbf{T}} = \mathbf{T} / (N - 1)$, $\mathbf{T} = \mathbf{H} + \mathbf{E}$,

$$\mathbf{H} = \sum_{j=1}^2 n_j (\mathbf{y}_j - \bar{\mathbf{y}})(\mathbf{y}_j - \bar{\mathbf{y}})^{\text{T}} \quad (8)$$

ROBUST MEASURES OF VARIABLE IMPORTANCE

is the hypothesis sum of squares and cross product matrix, and T is the transpose operator. Like DRCs, the TDRCs can range in value from zero to one, with a larger value indicating greater relative importance.

Robust Estimation

In CT of multivariate data, each outcome variable is independently trimmed by removing a pre-specified proportion of the data at both of tails of the variable's distribution. Let $y_{(1)jm} \leq y_{(2)jm} \leq \dots \leq y_{(n_j)jm}$ denote the order statistics of the j^{th} group for the m^{th} ($m = 1, \dots, p$) outcome variable (Srivastava & Mudholkar, 2001). Define $b_j = [\delta n_j]$, where δ represents the proportion of the observations to be trimmed, or censored, from each tail of the distribution and $[x]$ denotes the integer less than or equal to x . When symmetric trimming is adopted, so that the same number of observations are removed from each tail of the distribution, the effective sample size for the j^{th} group is $f_j = n_j - 2b_j$. The trimmed mean for the j^{th} group on the k^{th} outcome variable is

$$\bar{y}_{ijk} = \frac{1}{f_j} \sum_{i=b_j+1}^{n_j-b_j} y_{(i)jk} \quad (9)$$

Define z_{ijk} as

$$z_{ijk} = \begin{cases} y_{(b_j+1)jk} & \text{if } y_{ijk} \leq y_{(b_j+1)jk} \\ y_{ijk} & \text{if } y_{(b_j+1)jk} < y_{ijk} < y_{(n_j-b_j)jk} \\ y_{(n_j-b_j)jk} & \text{if } y_{ijk} \geq y_{(n_j-b_j)jk} \end{cases}$$

Then

$$\bar{y}_{wjk} = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{ijk} \quad (10)$$

is the j^{th} group Winsorized mean for the k^{th} outcome variable. The Winsorized sum of squared deviations for the k^{th} and l^{th} outcome variable in the j^{th} group is

$$SS_j = \sum_{i=1}^{n_j} (z_{ijk} - \bar{z}_{jk})(z_{ijl} - \bar{z}_{jl}) \quad (11)$$

and $S_{jw} = (ss_{jk})$ is the estimated Winsorized sum of squares and cross product matrix. Given that CT estimators are derived by independently trimming each outcome variable (Wilcox, 1994; Maronna, Martin, & Yohai, 2006), they share similar robustness properties to univariate trimmed estimators. Robust variable importance measures are derived by replacing the least squares means and covariances by trimmed means and Winsorized covariances of the data.

Simulation Study

A Monte Carlo study was used to evaluate the performance of measures of relative importance in rank ordering a set of correlated variables under a variety of data analytic conditions. Measures of relative importance investigated include (a) SDFC, (b) DRC, (c) FTR, and (d) TDRC. These were developed based on least squares and CT means and Winsorized covariances. All variable importance measures were investigated for the case of two independent groups.

The conditions investigated were (a) number of outcome variables, (b) total sample size and equality/in equality of group sizes, (c) magnitude and pattern of variable correlation, (d) mean configuration, and (e) shape of the population distribution. The number of outcome variables was set at $p = 4, 6$, and 8 . Similarly investigated numbers of outcome variables ranging from 4 to 10 were previously considered (LeBreton, Polyhart, & Ladd, 2004). Total sample sizes of $N = 60, 90, 140$, and 200 were investigated. Although previous simulation studies for relative importance measures have primarily focused on equal group sizes (Finch & Laking, 2008), unequal group sizes have also been shown to influence the size of discriminant function coefficients (Barön, 1991; He & Fung, 2000), and may influence the consistency of the measures in accurately rank ordering the variables. Therefore, both equal and unequal group size conditions were investigated. For $n = 60$, the group sizes were $(n_1, n_2) = (30, 30)$ and $(n_1, n_2) = (24, 36)$. For $n = 90$, they were $(n_1, n_2) = (45, 45)$ and $(n_1, n_2) = (36, 52)$. For $n = 140$, the group sizes were $(n_1, n_2) = (70, 70)$ and $(n_1, n_2) = (56, 84)$. For $n = 2000$, the group sizes were $(n_1, n_2) = (100, 100)$ and $(n_1, n_2) = (80, 120)$. These group sizes were chosen based on previous research to represent small to large degrees of group size imbalance (Barön, 1991; Sajobi, Lix, Laverty, & Li, 2011).

ROBUST MEASURES OF VARIABLE IMPORTANCE

Previous investigations have shown that variable importance measures are sensitive to the magnitude of correlation among the outcome variables (Huberty & Wisenbaker, 1992; Thomas & Zumbo, 1996). Therefore, investigated here are different correlation structures and sizes of correlations: (a) $\mathbf{Q}_1$: an independent correlation structure with $\rho = 0.0$, (b) $\mathbf{Q}_2$: compound symmetric structure with $\rho = 0.3$, (c) $\mathbf{Q}_3$: compound symmetric structure with $\rho = 0.6$, (d) $\mathbf{Q}_4$: unstructured with average correlation among the off-diagonal elements of 0.3, and (e) $\mathbf{Q}_5$: unstructured with average correlation among the off-diagonal elements of 0.6. Previous research studies have investigated variable correlations ranging between 0.1 and 0.8 (Tonidandel, LeBreton, & Johnson, 2009).

The measures of relative importance were investigated when the data were both multivariate normal and non-normal. For the former, in which skewness (γ_1) and kurtosis (γ_2) values were equal to $\gamma_1 = 0$ and $\gamma_2 = 0$, respectively, pseudo-random observation vectors $\mathbf{X}_{ij}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ were generated using the RANNOR function in SAS (SAS Institute, 2008). A vector of p standard normal deviates, $\mathbf{R}_{ij}$, was transformed to a vector of multivariate observations via $\mathbf{X}_{ij} = \boldsymbol{\mu}_j + \mathbf{L}\mathbf{R}_{ij}$. The Cholesky decomposition was used to obtain $\mathbf{L}$, an upper triangular matrix of dimension m satisfying the equality $\mathbf{L}^T\mathbf{L} = \boldsymbol{\Sigma}_j$. The measures were also considered when the data were sampled from populations with multivariate skewed and heavy-tailed distributions.

For non-normal distributions, two skewed distributions and two heavy-tailed distributions were investigated. In the former, a moderately skewed non-normal distribution with $\gamma_1 = 1.8$ and $\gamma_2 = 5.9$ (SK-I) and a largely skewed non-normal distribution with $\gamma_1 = 13.2$ and $\gamma_2 = 42892.9$ (SK-II) were investigated. Two heavy-tailed non-normal distributions were also considered; the first is a moderately heavy-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 33$ (HT-I), while the second distribution is a heavy-tailed non-normal distribution similar to a heavy-tailed Cauchy distribution for which γ_1 and γ_2 are undefined (HT-II). Field and Genton (2006) described a flexible family of multivariate non-normal distributions obtained by modifying their quantiles. The variables g and h , which control the magnitude of γ_1 and γ_2 , are used to transform a standard normal random variate R_{ijk} via

$$Y_{ijk} = \left(\frac{\exp(gR_{ijk}) - 1}{g} \right) \left(\exp\left(\frac{h}{2} R_{ijk}^2\right) \right) \quad (12)$$

Table 1. Values of μ_1 selected for the Monte Carlo study

# Variables	Mean Pattern	Mean Vector	D^2
4	I	(2.5,2,1.5,1)	13.50
	II	(1,0.75,0.5,0.25)	1.88
	III	(1.5,1,0.5,2)	7.50
	IV	(0.75,0.5,0.25,1)	1.88
6	I	(3.5,3,2.5,2,1.5,1)	34.75
	II	(1.25,1,0.75,0.5,0.5,0.25)	3.69
	III	(2,1.5,1,0.5,2.5,3)	22.75
	IV	(1,0.75,0.5,0.25,1.25,1.5)	5.69
8	I	(4.5,4,3.5,3,2.5,2,1.5,1)	71.00
	II	(2.5,2.25,2,1.75,1.5,1.25,1,0.75,0.5)	17.50
	III	(2.5,2,1.5,1,0.5,3,3.5,4)	51.00
	IV	(1.25,1,0.75,0.5,0.25,1.5,1.75,2)	12.75

when $g = 0$, this equation reduces to

$$Y_{ijk} = \left(\exp \frac{h}{2} R_{ijk}^2 \right) R_{ijk} \quad (13)$$

The g and h parameters provide a convenient approach to generate data from a wide range of multivariate non-normal distributions. When $g = h = 0$, the distribution is multivariate normal. The parameter h determines the heaviness of the tails of a distribution, while the parameter g controls the magnitude of skewness of the distribution.

Four configurations of the variable means were selected for group 1 (Table 1). Configurations I and II had a linearly decreasing mean trend across the variables for group 1, while configuration III and IV had a quadratic mean trend across the variables for group 1. In all cases, the mean vector for group 2 was the null vector. For robust variable importance measures, a 20% symmetric trimming rule was adopted as recommended by Wilcox (1994). All combinations of conditions were investigated for each procedure and each method of estimation, resulting in a total of 1200 combinations of simulation conditions with 10,000 replications for each combination. The Monte Carlo study was conducted using SAS/IML version 9.2 (SAS Institute, 2008).

To evaluate the ranking accuracy of each variable importance measure, the estimated rank ordering of variables in the data was compared to the population variable rankings, which were derived from the rank ordering the variables based on the magnitude differences in the population group means. Other methods for

ROBUST MEASURES OF VARIABLE IMPORTANCE

generating the population variable ranking are described in the Appendix. Three indices were used to compare the ranking accuracy of these measures: (a) any-variable correct ranking (ACR) percentage, (b) average all-variable correct ranking (ALLCR) percentage, and (c) average per-variable correct ranking (PCR) percentage. The average ACR rate is the percent of simulations in which at least one variable in the sample was ranked the same as the corresponding variable in the population. The average ALLCR is the percentage of simulations in which all variable in the sample was ranked the same as the population rank ordering of the variables. The average PCR percentage is the overall average percentage of simulations in which a variable in the sample had the same rank as the variable in the population.

Results

Described in [Table 2](#) are the average ACR percentages for variable importance measures based on least squares and CT estimators by the type of distribution and mean configuration. For normally distributed data, the ACR percentage values for all importance measures increased as the magnitude of group separation increased. However, under large departures from a multivariate normal distribution, the magnitude of change in ACR percentage values were substantial for variable importance measures based on least squares estimation as the magnitude of group separation increased.

For example when the data for group 1 were sampled from a population with a linear decreasing mean configuration (configuration I), the average ACR percentage values for DRCs based on least squares estimation under mean configuration I and II were 99.1% and 96.1%, respectively. When the data were sampled from a skewed distribution (SK-II), the average ACR percentage values for DRCs based on least squares estimation under mean configuration I and II were 88.5% and 77.9%, respectively. The magnitude of change in ACR percentage values for the robust variable importance measures decreased as the amount of group separation increased. However, it was not substantially different irrespective of the population distribution. For example, the average ACR percentage values for the robust FTR under mean configurations III and IV were 97.15% and 95.22%, respectively when the data were sampled from normal distribution.

Table 2. Average percentage of any-variable correct ranking (ACR) for variable importance measures by population distribution and mean configuration

Distribution	Mean Configuration	Least Squares Estimators				Robust Estimators			
		SDFC	DRC	FTR	TDRC	SDFC	DRC	FTR	TDRC
Normal	I	89.81	99.08	95.62	97.18	93.98	97.80	93.22	98.52
	II	85.45	96.10	91.11	89.32	89.99	95.18	89.44	94.00
	III	87.75	98.55	91.44	95.29	88.58	97.15	90.78	95.71
	IV	84.47	96.70	87.49	88.61	85.31	95.22	87.26	89.99
HT-I	I	89.15	97.29	91.68	92.88	94.69	97.70	94.88	98.37
	II	83.01	91.35	86.20	84.95	90.16	94.06	90.38	92.34
	III	83.14	96.62	86.72	89.85	80.14	95.76	82.67	91.07
	IV	81.22	92.52	83.27	83.79	78.97	92.08	80.32	85.18
SK-I	I	91.07	97.94	93.40	95.49	94.29	97.53	94.26	98.19
	II	84.99	93.79	88.68	87.72	90.02	94.50	90.17	93.00
	III	84.34	97.81	88.23	91.20	83.70	96.94	86.49	93.02
	IV	82.22	94.38	84.84	85.68	81.47	93.49	83.05	87.14
HT-II	I	86.28	94.19	88.58	90.99	95.04	97.58	95.46	98.12
	II	79.45	87.26	82.43	80.24	89.84	93.57	90.54	91.36
	III	81.34	94.32	83.80	86.34	76.00	94.74	77.09	88.14
	IV	78.11	87.67	79.46	79.38	74.80	88.87	75.57	81.89
SK-II	I	80.99	88.50	83.56	83.00	95.51	97.59	96.05	97.88
	II	72.92	77.91	75.32	70.43	89.57	92.45	90.42	89.31
	III	78.50	87.95	79.44	79.02	70.04	91.05	68.96	82.92
	IV	74.17	79.54	74.82	72.85	71.63	85.52	71.36	78.88

Note: See Table 1 for a description of mean configurations. SDFC = standardized discriminant function coefficient; DRC = discriminant ratio coefficient; FTR = *F*-to-remove statistic; TDRC = total discriminant ratio coefficient; HT-I = heavy-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 33$; HT-II = a heavy-tailed distribution with $\gamma_1 = \infty$ and $\gamma_2 = \infty$; SK-I = skewed distribution with $\gamma_1 = 1.8$ and $\gamma_2 = 5.9$; SK-II = skewed distribution with $\gamma_1 = 13.2$ and $\gamma_2 = 42892.9$

ROBUST MEASURES OF VARIABLE IMPORTANCE

Table 3. Average percentage of all-variable correct ranking (ALLCR) by population distribution and correlation structure

Distribution	Mean Configuration	Least Squares Estimators				Robust Estimators			
		SDFC	DRC	FTR	TDRC	SDFC	DRC	FTR	TDRC
Normal	Independent	26.69	34.93	26.85	37.92	24.46	29.75	24.60	37.87
	CS(0.3)	14.14	35.81	13.94	17.27	15.50	30.94	15.29	21.21
	CS(0.6)	1.59	38.24	1.52	2.44	2.66	32.77	2.57	5.21
	UN(0.3)	9.69	24.84	9.27	12.80	11.02	24.46	10.76	16.53
	UN(0.6)	2.26	7.03	3.84	5.61	4.98	8.80	6.05	14.01
HT-II	Independent	14.74	16.38	14.76	20.65	23.78	27.53	23.86	33.41
	CS(0.3)	10.02	18.24	9.99	12.04	11.61	28.05	11.56	15.62
	CS(0.6)	2.91	23.09	2.86	3.34	3.23	27.38	3.18	4.91
	UN(0.3)	9.40	16.52	9.20	11.01	11.27	19.77	11.13	14.99
	UN(0.6)	3.78	6.46	4.91	5.71	5.10	6.18	5.45	10.38
SK-II	Independent	8.28	8.81	8.33	10.14	22.17	25.48	22.26	30.07
	CS(0.3)	6.04	9.59	6.11	6.61	9.13	23.58	9.10	11.75
	CS(0.6)	3.06	12.31	3.25	3.13	3.37	21.20	3.38	4.51
	UN(0.3)	6.38	9.48	6.39	6.72	10.75	15.50	10.80	13.12
	UN(0.6)	3.30	4.63	3.93	3.61	4.56	4.75	4.54	8.40

Note: See Table 1 for a description of mean configurations. SDFC = standardized discriminant function coefficient; DRC = discriminant ratio coefficient; FTR = *F*-to-remove statistic; TDRC = total discriminant ratio coefficient; HT-II = a heavy-tailed distribution with $\gamma_1 = \infty$ and $\gamma_2 = \infty$; SK-II = skewed distribution with $\gamma_1 = 13.2$ and $\gamma_2 = 42892.9$; CS(ρ) = compound symmetric correlation structure with $\rho = 0.3$; UN(ρ) = unstructured correlation with average off-diagonal correlation of ρ

Compiled in Table 3 are the results for average ALLCR percentage values when the data were sampled from multivariate normal and non-normal distributions for independent, compound symmetric, and unstructured correlations. The average ALLCR percentage values for SDFC, TDRC, and FTR decreased as the magnitude of correlation among the variables increased, regardless of the population distributions, number of variables, or method of estimation. In contrast, the change in ALLCR percentage values for the DRCs based on least squares and robust estimators as the magnitude of correlation increased varied across correlation structures. For example, when the data were sampled from a population with compound symmetric correlation structure, the average ACR values for DRCs based on least squares estimators increased but there was negligible change in the average values of DRC based on robust estimators, as the magnitude of correlation increased from 0.0 to 0.6, when the data were sampled from a normal distribution.

However, when the data were sampled from a population with an unstructured variable correlation, the decrease in average ALLCR values were more than 10%. Moreover, the magnitude of change in average ALLCR percentage values for the variable importance measures varied across population distribution and correlation

structure. The average ALLCR percentage values for each variable importance measure were largest when the data were sampled from a multivariate normal distribution but smallest under a multivariate skewed distribution, as the magnitude of correlation increased. For example, when the data were sampled from a normally distributed population, the average ALLCR percentage values for the FTR measure based on the least-squares method under independent and unstructured variable correlations (i.e. UN[0.6]) were 26.9% and 3.8%, respectively. However, under a SK-II distribution, average ALLCR percentage values were 8.3% and 3.9% when the data were sample from a population with independent and unstructured variable correlations, respectively.

Moreover, when the data were sampled from a multivariate normal distribution, differences between the average ALLCR percentage values for variable importance measures based on least squares estimators and those based on robust estimators were not more than 6%, except for the TDRC measure. However, when the data were sampled from non-normal distributions, the robust variable importance measures resulted in substantially higher average ALLCR percentage values than the variable importance measures based on least squares estimators. For instance, when the data were sampled from a normal distribution with an independent correlation structure, the average ALLCR values for DRC based on least squares and robust estimators were 34.9% and 289.8%, respectively. But when the data were sampled from a population with a multivariate heavy-tailed non-normal distribution (HT-II), the corresponding average PCR values were 16.4% and 27.5%, respectively.

Depicted in [Figure 1](#) are the average PCR for all the investigated measures by method of estimation, and total sample size. The analyses reveal the average PCR values for each variable importance measure were smallest when $N = 50$ and largest when $N = 200$, regardless of the population distribution and the method of estimation. The average PCR value for each variable importance measure increased as N increased, regardless of the method of estimation. Specifically, the DRC was most sensitive to changes in sample size when the data were sampled from a multivariate normal distribution. For example, for DRC based on least squares estimators and robust estimators, the change in average ACR percentage values were 20.6% and 22.70%, respectively, as N increased from 50 to 200. For other variable importance measures, the increase in average PCR percentage values were not more than 15% as N increased from 50 to 200. When the data were sampled from a highly skewed non-normal distribution, the average PCR values increased as N increased from 50 to 200, but the magnitude of change were not as large as under multivariate normal distribution. For example, as N increased from 50 to 200,

ROBUST MEASURES OF VARIABLE IMPORTANCE

the change in average PCR values for the TDRC based on least squares estimators were 9.7% and 4.9% when the data were sampled from a multivariate normal and highly skewed distributions, respectively. For example, the change in average ACR values for the FTR measure based on robust and least squares estimators as n increased from 50 to 200 were 8.9% and 7.2%, respectively, when the data were sampled from a heavy-tailed distribution (Figure 1).

The average PCR values for all the variable importance measures increased as p increased from 4 to 8 (Figure 2). There were negligible differences among the average PCR values of SDFC, DRC, and FTR procedures based on least squares and robust estimators when the data were sampled from a normal distribution. But the robust variable importance measures resulted in higher PCR values than measures based on least-squares estimators as p increased from 4 to 8, when the data were sampled from a highly skewed or heavy-tailed non-normal distribution.

Figure 1. Average percentage of per-variable correct ranking (ACR) for variable importance measures by population distribution and total sample size

Figure 2. Average percentage of per-variable correct ranking (PCR) for variable importance measures by population distribution and number of variables

Conclusion

Robust measures of relative importance were developed based on trimmed means and Winsorized covariances for non-normal two-group multivariate data. The performance of these measures were investigated in a variety of data analytic conditions including multivariate non-normality, number of variables, sample size, magnitude of correlation, and mean configuration. Our findings suggest that variable importance measures based on coordinatewise trimmed means and Winsorized covariance estimators result in higher percentages of correct variable ranking than least-squares measures under moderate to strong departures from multivariate normality. Robust DRCs, which achieved the highest proportion of correctly ranked variables, were the least sensitive to group separation, variable correlation, and sample size, regardless of the distribution of the data.

One criticism of trimmed estimators is that they may not perform well when sample size is small and the data are normally distributed (Wilcox, 1994). This study results suggest robust measures of relative importance performed equally well as measures based on least squares estimators under small sample size

ROBUST MEASURES OF VARIABLE IMPORTANCE

conditions and when data are normally distributed. Hence, they can be adopted for rank ordering the variables regardless of the shape of the population distribution.

The ability of a variable importance measure to correctly rank order a set of correlated variables in a simulation study may depend on the method for generating the population rank order of the variables. Although the variable importance measures described in this study may not be equally sensitive to choice of methods for generating population variable rankings (See [Tables A1 and A2](#)), one finding that is consistent across these methods is that the robust variable importance measures developed in this study resulted in higher correct ranking percentage values than variable importance measures based on least squares estimators regardless of the population ranking method.

Although previous research investigations have shown that variable importance measures based on DDA and MANOVA are sensitive to non-normal data ([Sajobi, Dansu, & Lix, 2013](#)), adopting these measures for evaluating the relative importance of variables in non-normal data is likely to lead to incorrect conclusions about the study variables. Based on the findings, it is recommended robust variable importance measures should be adopted for rank ordering variables in studies that are likely to be characterized by non-normal data distributions.

The limitations of this study should be noted. The simulation study focused on conditions in which group covariances were homogeneous. This may not be a reasonable assumption in all clinical studies; the treatment group may exhibit greater variability than the control group ([Blair & Sawilowsky, 1993](#)). The DRC is the only measure among the investigated measures that is designed to detect the presence of suppression effects among the variables ([Thomas & Zumbo, 1996](#)). Although, the presence of suppression effects among the variables can reduce the percentage of correctly rank variables, previous studies have suggested the exclusion of the suppressor variable from the relative importance analysis since it makes little direct contribution to group discrimination ([Thomas, Hughes, & Zumbo, 1998](#); [Ochieng & Zumbo, 2001](#)). However, the exclusion of collinear or suppressor effects may also reduce statistical power and can affect the rank ordering of the study variables. In our simulation study, we ignored the presence of the suppression among the variables by ranking all the variables according to the signed DRC values. We must also reiterate that other importance measures are also may not always be robust to suppression effects among the variables. Moreover, the robust procedures described in this study are sensitive to strong variable correlations. This might be attributed to the coordinate-wise approach for trimming multivariate data. An alternative trimmed estimator is the maximum trimmed likelihood estimators (MTLE), which trims the log-likelihood function of the

multivariate data (Bednarski & Clarke, 1993). Future investigations might examine the development of variable importance measures based on MTLE for describing group separation in multivariate data.

Another criticism of variable importance measures is that the absolute importance of a set of correlated outcome variables is derived based on estimates of variable importance indices obtained from sample data and do not account for sampling variations around these estimates. There are currently limited recommendations to guide researchers about contextual relevance/meaning of the rank ordering of outcome variables from. Rules of thumb and descriptive analyses of relative importance are being used. For example, Dalglish (1994) suggested that variables with structure coefficients above 0.4 should be considered as important. But this approach fails to account for sampling error in the observed data. More formal parametric methods that assess the statistical significance of the estimated variable importance index have been proposed (Rao, 1970; Huberty & Wisenbaker, 1992).

Although the variable importance measures developed in this study are robust to non-normal data, other variable importance measures that do not assume multivariate normality have been developed based on logistic regression model. This includes relative weights (Azen & Traxel, 2009; Tonidandel & LeBreton, 2010), dominance analysis (Azen & Budescu, 2003), standardized logistic regression coefficients, and Pratt's index for logistic regression (Thomas, Zumbo, Zhu, & Dutta, 2008). The relative weights are derived from an orthogonal transformation of the predictor variables, whereas the dominance analysis method relies on regression model's pseudo coefficient of determination (i.e., model R^2) to evaluate variable importance. Although some of these measures (e.g., rescaled relative weights) may perform better under some conditions (e.g., strong variable correlations), there has not been a comparison of the performance of these logistic regression-based variable importance measures under a broad range of data analytic conditions.

There are additional considerations for conducting relative importance analysis. The conclusion that one outcome variable is more important than another outcome variable depends on the set of variables under investigation. Hence, changing the mix of variables included in a study could change a researcher's conclusions about variable importance. The choice of study variables and the grouping variables are assumed to have been determined *a priori*. However, conclusions about a variable's importance as estimated from a sample may not be generalizable to other populations because it does not account for sampling variability in the data. An internal or external validation of the ranks should also be

ROBUST MEASURES OF VARIABLE IMPORTANCE

considered, in order to assess the generalizability of the results. For example, a split-sample validation or resampling based methods such as bootstrap may be adopted to examine statistical significance of a variable's importance.

In conclusion, robust measures for describing the relative importance of investigated set of correlated variables for group discrimination in normal and non-normal data were developed. Measures of relative importance have a number of applications (Baek et al., 2008; Johnson & LeBreton, 2004). They can be used to build a parsimonious statistical model for use in further research or to choose a subset of variables on which to focus in clinical investigations.

References

- Ahmed, S. W., & Lachenbruch, P. A. (1977). Discriminant analysis when scale contamination is present in the initial sample. In: J. Van Ryzin (Ed.), *Classification and clustering* (pp. 351-353). New York, NY: Academic Press. doi: [10.1016/b978-0-12-714250-0.50016-4](https://doi.org/10.1016/b978-0-12-714250-0.50016-4)
- Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8(2), 129-148. doi: [10.1037/1082-989x.8.2.129](https://doi.org/10.1037/1082-989x.8.2.129)
- Azen, R., & Traxel, N. M. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, 34(3), 319-347. doi: [10.3102/1076998609332754](https://doi.org/10.3102/1076998609332754)
- Baek, S., Moon, H., Ahn, H., Kodell, R. L., Lin, C., & Chen, J. J. (2008). Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking. *Journal of Biopharmaceutical Statistics*, 18(5), 853-868. doi: [10.1080/10543400802278023](https://doi.org/10.1080/10543400802278023)
- Barön, A. E. (1991). Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in Medicine*, 10(5), 757-766. doi: [10.1002/sim.4780100511](https://doi.org/10.1002/sim.4780100511)
- Bednarski, T., & Clarke, B. R. (1993). Trimmed likelihood estimation of location and scale of the normal distribution. *Australian & New Zealand Journal of Statistics*, 35(2), 141-153. doi: [10.1111/j.1467-842x.1993.tb01321.x](https://doi.org/10.1111/j.1467-842x.1993.tb01321.x)
- Blair, R. C., & Sawilowsky, S. (1993). Comparison of two tests useful in situations where a treatment is expected to increase variability relative to controls. *Statistics in Medicine*, 12(23), 2233-2243. doi: [10.1002/sim.4780122308](https://doi.org/10.1002/sim.4780122308)

- Bogler, R. (2002). Two profiles of school teachers: A discriminant analysis of job satisfaction. *Teaching and Teacher Education*, 18(6), 665-673. doi: [10.1016/s0742-051x\(02\)00026-4](https://doi.org/10.1016/s0742-051x(02)00026-4)
- Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal*, 28(2), 131-148. doi: [10.1002/bimj.4710280202](https://doi.org/10.1002/bimj.4710280202)
- Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics*, 29(3), 473-493. doi: [10.2307/3316042](https://doi.org/10.2307/3316042)
- Curenton, S. M., McWey, L. M., & Bolen, M. G. (2009). Distinguishing maltreating versus non-maltreating at-risk families: Implications for foster care and early childhood education interventions. *Families in Society: The Journal of Contemporary Social Services*, 90(2), 176-182. doi: [10.1606/1044-3894.3871](https://doi.org/10.1606/1044-3894.3871)
- Dalgleish, L. I. (1994). Discriminant analysis: Statistical inference using the jackknife and bootstrap procedures. *Psychological Bulletin*, 116(3), 498-508. doi: [10.1037/0033-2909.116.3.498](https://doi.org/10.1037/0033-2909.116.3.498)
- Eastman, B. J., & Brunch, S. G. (2009). Variables associated with treatment outcome in a sample of female offenders. *The Open Criminology Journal*, 2, 18-23. doi: [10.2174/1874917800902010018](https://doi.org/10.2174/1874917800902010018)
- Field, C., & Genton, M. G. (2006). The multivariate **g**-and-**h** distribution. *Technometrics*, 48(1), 104-111. doi: [10.1198/004017005000000562](https://doi.org/10.1198/004017005000000562)
- Finch, W. H., & Laking, T. (2008). Evaluation of the use of standardized weights for interpreting results from a descriptive discriminant analysis. *General Linear Model Journal*, 34(1), 19-34.
- Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and other outlier detection with multiresponse data. *Biometrics*, 28(1), 81-124. doi: [10.2307/2528963](https://doi.org/10.2307/2528963)
- He, X., & Fung, W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72(2), 151-162. doi: [10.1006/jmva.1999.1857](https://doi.org/10.1006/jmva.1999.1857)
- Holder, B. (2007). An investigation of hope, academics, environment, and motivation as predictors of persistence in higher education online programs. *The Internet and Higher Education*, 10(4), 245-260. doi: [10.1016/j.iheduc.2007.08.002](https://doi.org/10.1016/j.iheduc.2007.08.002)
- Hubert, M., & Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45(2), 301-320. doi: [10.1016/s0167-9473\(02\)00299-2](https://doi.org/10.1016/s0167-9473(02)00299-2)

ROBUST MEASURES OF VARIABLE IMPORTANCE

- Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin*, 95(1), 156-171. doi: [10.1037/0033-2909.95.1.156](https://doi.org/10.1037/0033-2909.95.1.156)
- Huberty, C. J., & Wisenbaker, J. M. (1992). Variable importance in multivariate group comparisons. *Journal of Educational and Behavioral Statistics*, 17(1), 75-91. doi: [10.3102/10769986017001075](https://doi.org/10.3102/10769986017001075)
- Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3), 238-257. doi: [10.1177/1094428104266510](https://doi.org/10.1177/1094428104266510)
- LeBreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods*, 7(3), 258-282. doi: [10.1177/1094428104266017](https://doi.org/10.1177/1094428104266017)
- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Hoboken, NJ: Wiley. doi: [10.1002/0470010940](https://doi.org/10.1002/0470010940)
- McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York, NY: Wiley. doi: [10.1002/0471725293](https://doi.org/10.1002/0471725293)
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166. doi: [10.1037/0033-2909.105.1.156](https://doi.org/10.1037/0033-2909.105.1.156)
- Ochieng, C. O., & Zumbo, B. D. (2001). *Examination of a variable ordering index in linear regression models: An assessment of the relative Pratt index in Likert data*. Paper presented at the Bob Conry Conference on Measurement Evaluation Research and Methodology, Vancouver, Canada.
- Rao, C. R. (1970). Inference on discriminant coefficients. In R. C. Bose (Ed.), *Essays in probability and statistics* (pp. 587-602). Chapel Hill, NC: University of North Carolina Press.
- Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). Hoboken, NJ: Wiley. doi: [10.1002/0471271357](https://doi.org/10.1002/0471271357)
- Richardson, J. T. E. (2007). Mental models of learning in distance education. *British Journal of Educational Psychology*, 77(2), 253-270. doi: [10.1348/000709906x110557](https://doi.org/10.1348/000709906x110557)
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 851-857. doi: [10.1080/01621459.1984.10477105](https://doi.org/10.1080/01621459.1984.10477105)
- Sajobi, T. T., Dansu, B. M., & Lix, L. M. (2013). Variable importance measures for non-normal data: An application to patient-reported outcomes.

Chronic Diseases and Injuries in Canada, 33(2). Retrieved from <https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-33-no-2-2013/cseb-student-conference-2012-abstract-winners.html>

Sajobi, T. T., Lix, L. M., Clara, I., Walker, J., Graff, L. L., Rawsthorne, P.,... Bernstein, C. N. (2012). Measures of relative importance for health-related quality of life. *Quality of Life Research*, 21(1), 1-11. doi: 10.1007/s11136-011-9914-7

Sajobi, T. T., Lix, L. M., Laverty, W. H., & Li, L. (2011). Discriminant analysis for repeated measures data: Effects of covariance misspecification on bias and error in discriminant function coefficients. *Journal of Modern Applied Statistical Methods*, 10(2), 571-582. doi: 10.22237/jmasm/1320120840

SAS Institute. (2008). *SAS/IML user's guide* (Version 9.2). Cary, NC: SAS Institute.

Shih, S. (2012). An examination of academic burnout versus work engagement among Taiwanese adolescents. *The Journal of Educational Research*, 105(4), 286-298. doi: 10.1080/00220671.2011.629695

Sperling, M. R., Schilling, C. A., Glosser, D., Tracy, J. I., & Asadi-Pooya, A. A. (2008). Self-perception of seizure precipitants and their relation to anxiety level, depression, and health locus control in epilepsy. *Seizure*, 17(4), 302-307. doi: 10.1016/j.seizure.2007.09.003

Srivastava, D. K., & Mudholkar, G. S. (2001). Trimmed $\tilde{T}^2$: A robust analog of Hottelling's T^2 . *Journal of Statistical Planning and Inference*, 97(2), 343-358. doi: 10.1016/s0378-3758(00)00239-1

Thomas, D. R. (1992). Interpreting discriminant functions: a data analytic approach. *Multivariate Behavioral Research*, 27(3), 335-362. doi: 10.1207/s15327906mbr2703_3

Thomas, D. R. (1997). A note on Huberty and Wisenbaker's "views of variable importance". *Journal of Educational and Behavioral Statistics*, 22(3), 309-322. doi: 10.3102/10769986022003309

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275. doi: 10.1023/A:10069540

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal*

ROBUST MEASURES OF VARIABLE IMPORTANCE

of Educational and Behavioral Statistics, 21(2), 110-130. doi:
[10.3102/10769986021002110](https://doi.org/10.3102/10769986021002110)

Thomas, D. R., Zumbo, B. D., Zhu, P., & Dutta, S. (2008). On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7(1), 21-38. doi:
[10.22237/jmasm/1209614580](https://doi.org/10.22237/jmasm/1209614580)

Todorov, V., Neykov, N., & Neytchev, P. (1994). Robust two-group discrimination by bounded influence regression. A Monte Carlo simulation. *Computational Statistics & Data Analysis*, 17(3), 289-302. doi: [10.1016/0167-9473\(94\)90122-8](https://doi.org/10.1016/0167-9473(94)90122-8)

Tonidandel, S., & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weight analysis. *Organizational Research Methods*, 13(4), 767-781. doi:
[10.1177/1094428109341993](https://doi.org/10.1177/1094428109341993)

Tonidandel, S., LeBreton, J. M., & Johnson, J. W. (2009). Determining the statistical significance of relative weights. *Psychological Methods*, 14(4), 387-399. doi: [10.1037/a0017735](https://doi.org/10.1037/a0017735)

Troendle, J. F., Blair, R. C., Rumsey, D., & Moke, P. (1997). Parametric and non-parametric tests for the overall comparison of several treatments to a control when treatment is expected to increase variability. *Statistics in Medicine*, 16(23), 2729-2739. doi: [10.1002/\(sici\)1097-0258\(19971215\)16:23<2729::aid-sim694>3.3.co;2-y](https://doi.org/10.1002/(sici)1097-0258(19971215)16:23<2729::aid-sim694>3.3.co;2-y)

Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal*, 36(3), 259-273. doi: [10.1002/bimj.4710360302](https://doi.org/10.1002/bimj.4710360302)

Appendix

Although this simulation study defined the rank ordering of variables based on the differences amongst population means, there are other methods for defining the population rank ordering, including the population discriminant function coefficients and the population Mahalanobis distance (M-distance). In the former, the rank ordering of variables is based on the population discriminant function coefficients derived from the population covariance matrix and group means. Consider two population groups (A and B) measured on p outcome variables, where μ_1 and μ_2 denote the $p \times 1$ vector of means for the population groups A and group B, respectively.

Population Discriminant Function Coefficients

The population vector of discriminant function coefficients is

$$\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$$

where Σ is the $p \times p$ pooled covariance matrix for both population groups. The population variable ranking is derived by rank ordering the variables based on the magnitude of the elements of α .

Population M-Distance

Given the population means and pooled covariance matrix, Σ , the M-distance for the two populations is

$$\Delta = (\mu_1 - \mu_2) \Sigma^{-1} (\mu_1 - \mu_2)^T$$

where T is the transpose operator. The contribution of the k^{th} ($k = 1, \dots, m$) variable to the group differences can be estimated as

$$\Delta_k = \Delta - \Delta_{-k}$$

Where Δ_{-k} is value of Δ when the k^{th} variable is excluded from the analyses. A variable's importance in the population can be derived by ranking according to the

ROBUST MEASURES OF VARIABLE IMPORTANCE

magnitude of each variable's Δ_k . We present some summary results from both approaches (See Table A1 and A2).

Table A1. Average percentage per-variable correct ranking (PCR) of variable importance measures by population distribution and correlation structure

Distribution	Mean Configuration	Least Squares Estimators				Robust Estimators			
		SDFC	DRC	FTR	TDRC	SDFC	DRC	FTR	TDRC
Normal	Independent	55.36	63.76	55.59	66.34	53.36	59.28	53.56	66.48
	CS(0.3)	51.80	43.52	51.70	54.61	47.75	40.65	47.61	49.01
	CS(0.6)	55.10	25.06	56.19	46.28	48.10	24.94	48.68	39.04
	UN(0.3)	54.27	39.14	51.13	52.79	47.08	36.16	44.94	46.54
	UN(0.6)	73.72	21.39	40.57	47.97	33.88	16.56	26.02	30.49
HT-II	Independent	42.59	45.31	42.60	49.54	52.10	56.46	52.23	61.59
	CS(0.3)	39.33	37.74	39.57	42.34	46.24	41.68	46.44	50.48
	CS(0.6)	42.61	28.01	44.07	43.12	45.01	28.53	45.32	43.48
	UN(0.3)	42.84	36.38	41.62	44.23	48.36	39.35	47.79	49.15
	UN(0.6)	56.08	22.02	41.24	50.35	37.11	20.38	32.83	34.66
SK-II	Independent	33.70	34.80	33.78	36.85	51.06	54.98	51.17	59.12
	CS(0.3)	31.14	32.14	31.51	31.72	42.70	40.48	42.97	47.31
	CS(0.6)	32.98	27.47	34.55	32.97	41.14	29.09	41.43	41.64
	UN(0.3)	33.43	32.19	33.11	33.44	45.43	39.61	45.62	47.24
	UN(0.6)	44.59	22.84	35.79	39.63	36.27	22.36	33.61	34.54

Note: See Table 1 for a description of mean configurations. SDFC = standardized discriminant function coefficient; DRC = discriminant ratio coefficient; FTR = F -to-remove statistic; TDRC = total discriminant ratio coefficient; HT-II = A heavy-tailed distribution with $\gamma_1 = \infty$ and $\gamma_2 = \infty$; SK-II = Skewed distribution with $\gamma_1 = 13.2$ and $\gamma_2 = 42892.9$; CS(ρ) = Compound Symmetric correlation structure with $\rho = 0.3$; UN(ρ) = Unstructured correlation with average off-diagonal correlation of ρ

Appendix II: Results Based on Population Variable Ranking Derived from Population Mahalanobis Distance

Table A2. Average percentage any-variable correct ranking (PCR) of variable importance measures by population distribution and correlation structure

Distribution	Mean Configuration	Least Squares Estimators				Robust Estimators			
		SDFC	DRC	FTR	TDRC	SDFC	DRC	FTR	TDRC
Normal	Independent	95.81	97.63	95.90	97.77	95.14	96.78	95.22	97.69
	CS(0.3)	95.01	93.65	94.89	96.35	92.97	91.18	92.85	94.52
	CS(0.6)	96.73	92.73	97.14	97.58	94.67	88.85	95.15	95.34
	UN(0.3)	95.95	90.56	95.86	97.14	93.51	89.15	92.60	95.10
	UN(0.6)	99.16	74.43	99.09	98.92	93.21	88.96	90.74	96.50
HT-II	Independent	87.58	89.02	87.59	90.45	92.08	93.60	92.17	94.54
	CS(0.3)	86.08	84.14	86.39	87.93	90.44	86.97	90.55	92.47
	CS(0.6)	89.65	82.57	90.95	90.96	92.71	86.53	93.62	94.41
	UN(0.3)	88.07	81.54	87.78	89.74	91.87	80.03	91.75	93.40
	UN(0.6)	94.66	73.33	94.34	93.67	87.86	83.87	87.97	93.82
SK-II	Independent	81.14	81.93	81.18	83.15	92.19	93.65	92.28	94.55
	CS(0.3)	78.58	78.79	79.23	78.50	88.63	84.30	88.83	90.96
	CS(0.6)	80.57	76.97	82.72	80.63	90.46	84.34	91.56	92.46
	UN(0.3)	81.24	77.14	81.24	81.22	90.55	76.69	90.54	92.21
	UN(0.6)	88.13	72.38	88.15	85.09	84.01	81.82	85.17	90.45

Note: See Table 1 for a description of mean configurations. SDFC = standardized discriminant function coefficient; DRC = discriminant ratio coefficient; FTR = F -to-remove statistic; TDRC = total discriminant ratio coefficient; HT-II = A heavy-tailed distribution with $\gamma_1 = \infty$ and $\gamma_2 = \infty$; SK-II = Skewed distribution with $\gamma_1 = 13.2$ and $\gamma_2 = 42892.9$; CS(ρ) = Compound Symmetric correlation structure with $\rho = 0.3$; UN(ρ) = Unstructured correlation with average off-diagonal correlation of ρ

On Variance Balanced Designs

Dilip Kumar Ghosh

Saurashtra University Rajkot
Gujarat, India

Sangeeta Ahuja

Indian Agricultural Statistics Research Institute
New Delhi, India

Balanced incomplete block designs are not always possible to construct because of their parametric relations. In such a situation another balanced design, the variance balanced design, is required. This construction of binary, equal replicated variance balanced designs are discussed using the half fraction of the $2n$ factorial designs with smaller block sizes. This method was also extended to construct another variance balanced design by deleting the last block of the resulting variance balanced designs. Its efficiency factor compared with randomized block designs was compared and found to be highly efficient.

Keywords: Binary, efficiency factor, balanced incomplete block designs, information matrix and incidence matrix

Introduction

The balanced incomplete block design (BIBD) is binary, proper, connected, equi-replicated, balanced, and non-orthogonal. It is simple to construct and analyze. However, it is not available for all parameters because of its following parametric relations: (a) $vr = bk$; (b) $\lambda(v - 1) = r(k - 1)$; and (c) $b \geq v$. Thus, an incomplete block design is needed that should be connected and balanced. This type of incomplete block design is called a variance balanced (VB) design.

Rao (1958) noted that, if the information matrix $\mathbf{C}$ of a block design satisfies

$$\mathbf{C} = \theta \left[\mathbf{I}_v - \frac{1}{v} \mathbf{E}_{vv} \right] \quad (1)$$

where θ is the non-zero eigenvalue of the matrix $\mathbf{C}$, $\mathbf{I}_v$ is an identity matrix of order v , and $\mathbf{E}_{vv}$ is a matrix of Unity with v rows and v columns, then such a design is a variance balanced design. BIBD satisfies this property and hence is a variance balanced design.

Dr. D. K. Ghosh is a UGC BSR Faculty Fellow in the Department of Statistics. Email him at: ghosh_dkg@yahoo.com. Dr. Sangeeta Ahuja is a Senior Scientist.

Chakrabarti (1963) gave the useful concept of a **C**-matrix of design. Pearce (1964) obtained VB design with varying block sizes. Mukerjee and Kageyama (1985) obtained a resolvable VB design with unequal replications. Das and Ghosh (1985) obtained unequal replicated, unequal/equal block size VB design from BIBD and partially balanced incomplete block design (PBIBD) using augmented blocks and treatments. Rao (1958), Hedayat and Federer (1974), Raghavarao (1962), and Puri and Nigam (1977) defined that a design is said to be variance balanced if every normalized estimable linear function of treatment effect can be estimated with same precision. Kageyama (1988) discussed the construction of VB design using BIBD, PBIBD, and some incidence matrices.

Khatri (1982) gave a formula to measure the overall A-efficiency of VB designs along with method of construction of VB designs. Gupta and Jones (1983) obtained VB designs using BIBD and PBIBD with two associate classes. Calvin and Sinha (1989) extended the technique of Calvin (1986) to construct VB designs with more than two distinct block sizes that permit fewer replications. Das and Ghosh (1985) defined generalized efficiency balanced (GEB) design which include both VB as well as efficiency balanced (EB) designs. Ghosh (1988), Ghosh and Karmoker (1988), Ghosh, Divecha, and Kageyama (1991), and Ghosh, Joshi, and Kageyama (1993) provided several methods for the construction of VB designs. Ghosh and Joshi (1995) constructed VB designs through a triangular design. Agarwal and Kumar (1985) constructed a VB design which is associated with group divisible design. Ghosh and Joshi (1991) constructed a VB design through a group divisible (GD) design.

Definition: A block design is said to be variance-balanced if it permits the estimation of all estimable normalized treatment contrasts with the same variance,

$$V(\hat{t}_i - \hat{t}_j) = k\sigma^2, \quad \forall i \neq j = 1, 2, \dots, v$$

Variance balance designs can also be further defined as in the following way: A connected block design is said to be variance balanced if and only if all the nonzero eigenvalues of the matrix **C** of the block design are equal.

Corollary: A connected block design is variance balanced if and only if its **C**-matrix has all its diagonal elements equal and all its off-diagonal element equal, i.e.

$$\mathbf{C} = (a - b)\mathbf{I} + b\mathbf{j}\mathbf{j}'$$

Remark: For an equi-replicated, proper, binary, variance balanced design, the concurrence matrix satisfies

$$\mathbf{NN}' = (r - \lambda)\mathbf{I}_v + \lambda \mathbf{jj}' \quad (2)$$

Here, the variance balanced design is constructed using fractional factorial experiments, and by developing some incidence matrices.

Method of Construction of Variance Balanced Designs

Variance Balanced Design from Half Fraction of 2^n Factorial Experiment

Here, use the half fraction of 2^n factorial designs to construct binary, equi-replicated variance balanced designs with smaller block sizes.

Theorem 1: The half fraction of 2^n factorial designs always gives binary, equi-replicated variance balanced designs of unequal block sizes with smaller block sizes having parameters $v = n$, $b = 2^{n-1} - 1$, $r = 2^{n-2}$, $k = \{2, \dots, 2; 4, \dots, 4; \dots, p\}$, where $p = n$ if n is an even number; otherwise, $p = n - 1$.

Proof: Construct a half fraction of 2^n factorial design in blocks of sizes 2^{n-1} . That is, n factors or columns and 2^{n-1} rows. Consider the n factors as n treatments and the 2^{n-1} rows as blocks. Delete the first row having all the factors at zero levels. Consider this as an incidence matrix of an incomplete block design having n rows as treatments and $2^{n-1} - 1$ columns as blocks. Thus, $v = n$, $b = 2^{n-1} - 1$. Here, all the elements are either zero or one so the design is binary. Again, each row contains '1' 2^{n-2} times so $r = 2^{n-2}$. Similarly, each block contains '1' either 2, 4, ..., n times or 2, 4, ..., $n - 1$ times depending on n being even or odd. So $k = \{2, \dots, 2; 4, \dots, 4; \dots; n \text{ or } n - 1\}$.

The incidence matrix of the incomplete block design is

$$\begin{array}{cccccc}
 0 & 0 & 0 & 1 & \dots & 0 \\
 0 & 0 & 1 & 1 & \dots & 1 \\
 1 & 1 & 0 & 1 & \dots & 0 \\
 \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & 1 & 0 & 1 & \dots & 1
 \end{array}$$

If it can be proven the incidence matrix satisfies (1) for

$$\mathbf{C} = \text{diag}(r_1, r_2, \dots, r_v) - \mathbf{N}\mathbf{K}^{-1}\mathbf{N}' \quad (3)$$

then the design will be variance balanced. Because each row of the incidence matrix has (half – 1) ‘0’ and half ‘1’, the diagonal and off-diagonal elements of $\mathbf{N}\mathbf{K}^{-1}\mathbf{N}'$ are given as follows:

(a) Diagonal elements:

$$(i) \frac{\binom{n-2}{0}}{2} + \frac{\binom{n-2}{1}}{2} + \frac{\binom{n-2}{2}}{4} + \frac{\binom{n-2}{3}}{4} + \dots + \frac{1}{n} \text{ provided } n \text{ is even,}$$

$$(ii) \frac{\binom{n-2}{0}}{2} + \frac{\binom{n-2}{1}}{2} + \frac{\binom{n-2}{2}}{4} + \frac{\binom{n-2}{3}}{4} + \dots + \frac{1}{n-1} \text{ provided } n \text{ is odd,}$$

(b) Off-diagonal elements:

$$(i) \frac{\binom{n-3}{0}}{2} + \frac{\binom{n-3}{1}}{4} + \frac{\binom{n-3}{2}}{4} + \frac{\binom{n-3}{3}}{6} + \dots + \frac{1}{n} \text{ provided } n \text{ is even,}$$

$$(ii) \frac{\binom{n-3}{0}}{2} + \frac{\binom{n-3}{1}}{4} + \frac{\binom{n-3}{2}}{4} + \frac{\binom{n-3}{3}}{6} + \dots + \frac{1}{n-1} \text{ provided } n \text{ is odd.}$$

Using (3), we have the $\mathbf{C}$ -matrix of the incomplete block design as

ON VARIANCE BALANCED DESIGNS

$$\mathbf{C} = \begin{bmatrix} 2^{n-2} & 0 & \dots & 0 \\ 0 & 2^{n-2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2^{n-2} \end{bmatrix} - \begin{bmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \dots & A \end{bmatrix} \quad (4)$$

where

$$A = \frac{\binom{n-2}{0}}{2} + \frac{\binom{n-2}{1}}{2} + \frac{\binom{n-2}{2}}{4} + \frac{\binom{n-2}{3}}{4} + \dots + \frac{1}{p}$$

$$B = \frac{\binom{n-3}{0}}{2} + \frac{\binom{n-3}{1}}{4} + \frac{\binom{n-3}{2}}{4} + \frac{\binom{n-3}{3}}{6} + \dots + \frac{1}{p}$$

$p = n$ if n is even; otherwise, $p = n - 1$.

After solving (4),

$$\mathbf{C} = \begin{bmatrix} 2^{n-2} - A & -B & \dots & -B \\ -B & 2^{n-2} - A & \dots & -B \\ \vdots & \vdots & \ddots & \vdots \\ -B & -B & \dots & 2^{n-2} - A \end{bmatrix}$$

$\mathbf{C}$ can be rewritten in the form

$$\mathbf{C} = \theta \left[\mathbf{I}_v - \frac{1}{v} \mathbf{E}_{vv} \right] \quad (5)$$

where θ is the non-zero eigenvalue of the $\mathbf{C}$ -matrix of incomplete block design given by $\theta = (2^{n-2} - A + B)$. This satisfies the criteria of variance balanced design. Hence the design is variance balanced design with the required parameters.

Example 1: Let $n = 5$. Using method 1, the incidence matrix of the variance balanced design is given as

$$\mathbf{N} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Now, from the above incidence matrix and Theorem 1, it can be verified that the parameters of the VB design are $v = n = 5$, $b = 2^{n-1} - 1 = 15$, $r = 2^{n-2} = 8$, and $k = \{2, \dots, 2; 4, \dots, 4, \dots, 4\}$, as n is an odd number.

Here,

$$A = \frac{\binom{3}{0}}{2} + \frac{\binom{3}{1}}{2} + \frac{\binom{3}{2}}{4} + \frac{1}{4} = 1 + \frac{3}{2} + \frac{3}{4} + \frac{1}{4} = \frac{12}{4}$$

$$B = \frac{1}{2} + \frac{2}{4} + \frac{1}{4} = \frac{5}{4}$$

so

$$\mathbf{N}\mathbf{K}^{-1}\mathbf{N}' = \left(\frac{1}{4}\right) \begin{bmatrix} 12 & 5 & 5 & 5 & 5 \\ 5 & 12 & 5 & 5 & 5 \\ 5 & 5 & 12 & 5 & 5 \\ 5 & 5 & 5 & 12 & 5 \\ 5 & 5 & 5 & 5 & 12 \end{bmatrix}$$

Using (3), the $\mathbf{C}$ -matrix of the variance balanced design is computed as

$$\mathbf{C} = \left(\frac{1}{4}\right) \begin{bmatrix} 12 & -5 & -5 & -5 & -5 \\ -5 & 12 & -5 & -5 & -5 \\ -5 & -5 & 12 & -5 & -5 \\ -5 & -5 & -5 & 12 & -5 \\ -5 & -5 & -5 & -5 & 12 \end{bmatrix}$$

$$\mathbf{C} = \frac{25}{4} \left[\mathbf{I}_5 - \frac{1}{4} \mathbf{E}_{55} \right]$$

Therefore, the design illustrated in Example 1 is a variance balanced design with nonzero eigenvalue $\theta = 25/4$ with multiplicity four. The variance of the treatment effects t_i and t_j is computed as $V(\hat{t}_i - \hat{t}_j) = (8/25)\sigma^2$.

Construction of Variance Balanced Designs by Deleting the Last Row of the Half Fraction of the 2^n Factorial Design

It was shown that the variance balanced design can be constructed by deleting the last row of the half fraction of the 2^n factorial designs discussed in the previous section, provided n is even and greater than 4. This is because, with $n = 4$, it gives a balanced incomplete block design with parameters $v = 4$, $b = 6$, $r = 3$, $k = 2$, and $\lambda = 1$. This design is currently known in the literature.

Theorem 2: Deleting the last row of the half fraction of the 2^n factorial designs, a variance balanced design can always be constructed provided that n is an even number and greater than four, with parameters $v = n$, $b = 2^{n-1} - 2$, $r = 2^{n-2} - 1$, $k = \{2, \dots, 2; 4, \dots, 4; \dots, n - 2\}$.

Proof: Construct a half fraction of the 2^n factorial designs when n is an even number and greater than four. This design contains n factors or columns and 2^{n-1} runs or rows. Delete the control treatment and last row which has '1' for all its elements. After deletion, we have $2^{n-1} - 2$ rows. Consider the n columns as treatments and $2^{n-1} - 2$ rows as blocks which form an incidence matrix. Call this an incidence matrix N of an incomplete block design d with n treatments arranged in $2^{n-1} - 2$ blocks. Using the previous section, we can easily see that, for this incomplete block design, $v = n$, $b = 2^{n-1} - 2$, $r = 2^{n-2} - 1$, $k = \{2, \dots, 2; 4, \dots, 4; \dots, n - 2\}$.

The design will be variance balanced if it can be proven the incidence matrix satisfies

$$C = \theta \left[I_v - \frac{1}{v} E_{vv} \right] \quad (6)$$

To prove this, derive the diagonal and off diagonal elements of $NK^{-1}N'$.

Diagonal elements:

$$\frac{\binom{n-2}{0}}{2} + \frac{\binom{n-2}{1}}{2} + \frac{\binom{n-2}{2}}{4} + \frac{\binom{n-2}{3}}{4} + \dots + \frac{\binom{n-2}{n-3}}{n-2}$$

Off-diagonal elements:

$$\frac{\binom{n-3}{0}}{2} + \frac{\binom{n-3}{1}}{4} + \frac{\binom{n-3}{2}}{4} + \frac{\binom{n-3}{3}}{6} + \dots + \frac{\binom{n-3}{n-4}}{n-2}$$

provided n is an even and greater than four. Using (3), we have that the $\mathbf{C}$ -matrix of the incomplete block design is

$$\mathbf{C} = \begin{bmatrix} 2^{n-2}-1 & 0 & \dots & 0 \\ 0 & 2^{n-2}-1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2^{n-2}-1 \end{bmatrix} - \begin{bmatrix} A & B & \dots & B \\ B & A & \dots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \dots & A \end{bmatrix} \quad (7)$$

where

$$A = \frac{\binom{n-2}{0}}{2} + \frac{\binom{n-2}{1}}{2} + \frac{\binom{n-2}{2}}{4} + \frac{\binom{n-2}{3}}{4} + \dots + \frac{\binom{n-2}{n-3}}{n-2}$$

$$B = \frac{\binom{n-3}{0}}{2} + \frac{\binom{n-3}{1}}{4} + \frac{\binom{n-3}{2}}{4} + \frac{\binom{n-3}{3}}{6} + \dots + \frac{\binom{n-3}{n-4}}{n-2}$$

provided n is an even and greater than four. After solving (7),

$$\mathbf{C} = \begin{bmatrix} 2^{n-2}-1-A & -B & \dots & -B \\ -B & 2^{n-2}-1-A & \dots & -B \\ \vdots & \vdots & \ddots & \vdots \\ -B & -B & \dots & 2^{n-2}-1-A \end{bmatrix}$$

and hence $\mathbf{C}$ can be rewritten as

$$\mathbf{C} = \theta \left[\mathbf{I}_v - \frac{1}{v} \mathbf{E}_{vv} \right] \quad (8)$$

where θ is the non-zero eigenvalue of the $\mathbf{C}$ -matrix of the incomplete block design given by $\theta = (2^{n-2} - 1 - A + B)$. This satisfies the criteria of variance balanced design. Hence, the design is variance balanced design with the required parameters.

Example 2: Let $n = 6$. Using Theorem 2, the incidence matrix of the variance balanced design is given as

$$\mathbf{N} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From above incidence matrix and Theorem 2, it can be verified the parameters of the VB design are $v = n = 6$, $b = 2^{n-1} - 2 = 30$, $r = 2^{n-2} - 1 = 15$, $k = \{2, \dots, 2; 4, \dots, 4, \dots, 4\}$, as n is even and greater than 4.

$$A = \frac{\binom{4}{0}}{2} + \frac{\binom{4}{1}}{2} + \frac{\binom{4}{2}}{4} + \frac{\binom{4}{3}}{4} = \frac{20}{4}$$

$$B = \frac{\binom{3}{0}}{2} + \frac{\binom{3}{1}}{4} + \frac{\binom{3}{2}}{4} = \frac{8}{4}$$

so

$$\mathbf{N}\mathbf{K}^{-1}\mathbf{N}' = \left(\frac{1}{4}\right) \begin{bmatrix} 20 & 8 & 8 & 8 & 8 & 8 \\ 8 & 20 & 8 & 8 & 8 & 8 \\ 8 & 8 & 20 & 8 & 8 & 8 \\ 8 & 8 & 8 & 20 & 8 & 8 \\ 8 & 8 & 8 & 8 & 20 & 8 \\ 8 & 8 & 8 & 8 & 8 & 20 \end{bmatrix}$$

Using (3), the $\mathbf{C}$ -matrix of the variance balanced design is computed as

$$\mathbf{C} = \left(\frac{1}{4}\right) \begin{bmatrix} 40 & -8 & -8 & -8 & -8 & -8 \\ -8 & 40 & -8 & -8 & -8 & -8 \\ -8 & -8 & 40 & -8 & -8 & -8 \\ -8 & -8 & -8 & 40 & -8 & -8 \\ -8 & -8 & -8 & -8 & 40 & -8 \\ -8 & -8 & -8 & -8 & -8 & 40 \end{bmatrix}$$

$$\mathbf{C} = \frac{48}{4} \left[\mathbf{I}_6 - \frac{1}{6} \mathbf{E}_{66} \right]$$

Thus, the design illustrated in Example 2 is a variance balanced design with non-zero eigenvalue $\theta = 48/4$ with multiplicity five. The variance of the treatment effects t_i and t_j is computed as $V(\hat{t}_i - \hat{t}_j) = (8/48)\sigma^2$.

Efficiency Factor

The efficiency factor of variance-balanced designs compare to completely randomized design is given by

$$E = \frac{V(\hat{t}_i - \hat{t}_j)_{\text{RBD}}}{V(\hat{t}_i - \hat{t}_j)_{\text{VB}}}$$

The efficiency factor will be derived separately from equations (5) and (8):

From equation (5), $\theta = (2^{n-2} - A + B)$ so $\hat{t}_i = (1/\theta)Q_i$ and $V(\hat{t}_i - \hat{t}_j) = (2/\theta)\sigma^2$. That is

ON VARIANCE BALANCED DESIGNS

$$V(\hat{t}_i - \hat{t}_j)_{\text{VB}} = \frac{2}{(2^{n-2} - A + B)} \sigma^2$$

$$V(\hat{t}_i - \hat{t}_j)_{\text{RBD}} = \frac{2}{r} \sigma^2$$

where r is the replication size of the randomized block designs. Hence the efficiency factor E is given by

$$E = \frac{2^{n-2} - A + B}{r}$$

The efficiency factor for Example 1 can be computed as

$$E = \frac{\frac{2}{8} \sigma^2}{\frac{8}{25} \sigma^2} = \frac{25}{32} = 0.78$$

The efficiency of the variance balanced design compared with the randomized block design is 78 percent.

From equation (8) we have $\theta = (2^{n-2} - 1 - A + B)$, so $\hat{t}_i = (1/\theta)Q_i$ and $V(\hat{t}_i - \hat{t}_j) = (2/\theta) \sigma^2$. That is

$$V(\hat{t}_i - \hat{t}_j)_{\text{VB}} = \frac{2}{(2^{n-2} - 1 - A + B)} \sigma^2$$

$$V(\hat{t}_i - \hat{t}_j)_{\text{RBD}} = \frac{2}{r} \sigma^2$$

where r is the replication size of the randomized block designs. The efficiency factor E is given by

$$E = \frac{2^{n-2} - 1 - A + B}{r}$$

Similarly, the efficiency factor for Example 2 can be computed as

$$E = \frac{\frac{2}{15} \sigma^2}{\frac{8}{48} \sigma^2} = \frac{4}{5} = 0.8$$

The efficiency of the variance balanced design compared to the randomized block design is 80 percent.

Consider the parameters and efficiency factor of the variance balanced designs with the usual blocks and deletion of last block respectively in Table 1 and Table 2. The range of n from 3 to 10 was selected for convenience.

Remarks:

- (i). From Table 1 and Table 2, it is obvious that efficiency factor increases as number of treatment increases.
- (ii). If the efficiency factor of VB designs shown in Table 1 and Table 2 are compared, note the efficiency factor decreases when a block is deleted from a given VB design with the same number of treatment.

Table 1. The parameters and efficiency factor of the VB designs

SN	$n = v$	b	r	k	Efficiency Factor
1	4	7	4	{2,..., 2; 4}	0.7500
2	5	15	8	{2,..., 2; 4,..., 4; 4}	0.7800
3	6	31	16	{2,..., 2; 4,..., 4; 4,..., 4; 6}	0.8125
4	7	63	32	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6; 6}	0.8300
5	8	127	64	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6; 8}	0.8600
6	9	256	128	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6; 8,..., 8; 8}	0.8814
7	10	512	256	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6; 8,..., 8; 10}	0.9023

Table 2. The parameters and efficiency factor of the VB designs with deletion of last block

SN	$n = v$	b	r	k	Efficiency Factor
1	4	6	3	{2,..., 2}	0.6600
2	6	30	15	{2,..., 2; 4,..., 4}	0.8000
3	8	126	63	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6}	0.8441
4	10	511	255	{2,..., 2; 4,..., 4; 4,..., 4; 6,..., 6; 8,..., 8}	0.9022

Acknowledgements

The first author is thankful to UGC for providing him UGC BSR Faculty Fellow. We are thankful to the referee for his valuable suggestions so as to bring this articles in better form.

References

- Agarwal, G. G., & Kumar, S. (1985). A note on construction of variance balanced designs. *Journal of the Indian Society of Agricultural Statistics*, 37(2), 181-183.
- Calvin, J. A. (1986). A new class of variance balanced designs. *Journal of Statistical Planning and Inference*, 14(2-3), 251-254. doi: [10.1016/0378-3758\(86\)90162-x](https://doi.org/10.1016/0378-3758(86)90162-x)
- Calvin, J. A., & Sinha, K. (1989). A method of constructing variance balanced designs. *Journal of Statistical Planning and Inference*, 23(1), 127-131.
- Chakrabarti, M. C. (1963). On the C-matrix in design of experiments. *Journal of the Indian Statistical Association*, 1, 8-23.
- Das, M. N., & Ghosh, D. K. (1985). Balancing incomplete block designs. *Sankhyā: The Indian Journal of Statistics, Series B*, 47(1), 67-77. Available from <http://www.jstor.org/stable/25052375>
- Ghosh, D. K. (1988). A series of generalized efficiency balanced designs. *Gujarat Statistical Review*, 15(1), 33-38.
- Ghosh, D. K., Divecha, J., & Kageyama, S. (1991). Equireplicate variance balanced designs from group divisible designs. *Journal of the Japan Statistical Society*, 21(1), 205-209.
- Ghosh, D.K., & Joshi, K. (1991). Note on construction of variance balanced designs through group divisible designs. *Utilitas Mathematica*, 39, 249-253.
- Ghosh, D. K., & Joshi, K. (1995). Construction of variance balanced designs through triangular PBIB designs. *Calcutta Statistical Association Bulletin*, 45(1-2), 111-118. doi: [10.1177/0008068319950107](https://doi.org/10.1177/0008068319950107)
- Ghosh, D. K., Joshi, K., & Kageyama, S. (1993). Ternary variance balanced designs through BIB and GD designs. *Journal of the Japan Statistical Society*, 23(1), 75-81.

- Ghosh, D. K., & Karmoker, P. K. (1988). Some series of efficiency balanced designs. *Australian & New Zealand Journal of Statistics*, 30(1), 47-51. doi: [10.1111/j.1467-842x.1988.tb00610.x](https://doi.org/10.1111/j.1467-842x.1988.tb00610.x)
- Gupta, S. C., & Jones, B. (1983). Equireplicate balanced block designs with unequal block sizes. *Biometrika*, 70(2), 443-440. doi: [10.2307/2335557](https://doi.org/10.2307/2335557)
- Hedayat, A., & Federer, W. T. (1974). Pairwise and variance balanced incomplete block design. *Annals of the Institute of Statistical Mathematics*, 26(1), 331-338. doi: [10.1007/bf02479828](https://doi.org/10.1007/bf02479828)
- Kageyama, S. (1988). Existence of variance-balanced binary designs with fewer experimental units. *Statistics & Probability Letters*, 7(1), 27-28. doi: [10.1016/0167-7152\(88\)90083-1](https://doi.org/10.1016/0167-7152(88)90083-1)
- Khatri, C. G. (1982). A note on variance balanced designs. *Journal of Statistical Planning and Inference*, 6(2), 173-177. doi: [10.1016/0378-3758\(82\)90086-6](https://doi.org/10.1016/0378-3758(82)90086-6)
- Mukerjee, R., & Kageyama, S. (1985). On resolvable and affine resolvable variance-balanced designs. *Biometrika*, 72(1), 165-172. doi: [10.2307/2336346](https://doi.org/10.2307/2336346)
- Pearce, S. C. (1964). Experimenting with blocks of natural size. *Biometrics*, 20(4), 699-706. doi: [10.2307/2528123](https://doi.org/10.2307/2528123)
- Puri, P. D., & Nigam, A. K. (1977). Balanced block designs. *Communications in Statistics – Theory and Methods*, 6(12), 1171-1179. doi: [10.1080/03610927708827560](https://doi.org/10.1080/03610927708827560)
- Raghavarao, D. (1962). On balanced unequal block designs. *Biometrika*, 49(3/4), 561-562. doi: [10.2307/2333990](https://doi.org/10.2307/2333990)
- Rao, V. R. (1958). A note on balanced designs. *The Annals of Mathematical Statistics*, 29(1), 290-294. doi: [10.1214/aoms/1177706729](https://doi.org/10.1214/aoms/1177706729)

Unit Root Test for Panel Data AR(1) Time Series Model With Linear Time Trend and Augmentation Term: A Bayesian Approach

Jitendra Kumar

Central University of Rajasthan
Rajasthan, India

Anoop Chaturvedi

University of Allahabad
Allahabad, India

Umme Afifa

Central University of Rajasthan
Rajasthan, India

Shafat Yousuf

Central University of Rajasthan
Rajasthan, India

Saurabh Kumar

Central University of Rajasthan
Rajasthan, India

The univariate time series models, in the case of unit root hypothesis, are more biased towards the acceptance of the Unit Root Hypothesis especially in a short time span. However, the panel data time series model is more appropriate in such situation. The Bayesian analysis of unit root testing for a panel data time series model is considered. An autoregressive panel data AR(1) model with linear time trend and augmentation term has been considered and derived the posterior odds ratio for testing the presence of unit root hypothesis under appropriate prior assumptions. A simulation study and real data analysis are carried out for the derived theorem.

Keywords: Panel data, stationarity, autoregressive time series, unit root, prior and posterior, posterior odds ratio

Introduction

Analysis of chronologically recorded data became more popular because of its usability, which combines cross-section information like a series collected from multiple locations, a group of people who are surveyed periodically over a given period of time, etc. This is called a panel data time series, which is also equally popular in the social sciences, having been used in economics to study the behavior of firms and wages of people over a time, as well as in marketing to study market share changes across different market structures (Yaffee, 2003; Hsiao, 2007). Panel data are more appropriate in comparison to the univariate model because of its

*Dr. Jitendra Kumar is an Associate Professor and Head of the Department of Statistics.
Email him at: jitendra_20932@rediffmail.com.*

applicability in controlling individual's heterogeneity, which is not controlled alone through univariate time series model. There are lots of works appearing in literature dealing with Bayesian unit root tests for univariate time series (DeJong & Whiteman, 1991; Schotman & van Dijk, 1991; Uhlig, 1994; Lubrano, 1995). The panel data time series model is able to study the dynamics of adjustment, which is capable of identifying and measures the effects that are simply not detectable in pure cross-section or pure time-series data. See also Harris and Tzavalis (1999), Maddala and Wu (1999), Kruiniger and Tzavalis (2002), Moon and Perron (2004), De Wachter, Harris, and Tzavalis (2007), Moon and Perron (2008), Madsen (2010), and De Blander and Dhaene (2012).

In an analysis of financial time series, testing the stationarity of series is very important. A series may be non-stationary due to time trend or unit root. The order of time trend may be reduced by one polynomial degree in the case of unit root. If there is a linear time trend then, under the unit root hypothesis, the model becomes difference stationary (Dickey & Fuller, 1979, 1981). The traditional theory of unit root associated with univariate approach has low power if it is close to unity, particularly in short term time span; see e.g. Shiller and Perron (1985). This makes the test more favorable to the acceptance of unit root hypothesis in various economic time series.

The literature on the unit root test has been extended to include the situations where panel data are available. Levin and Lin (1992, 1993) proposed to apply the unit root test on a pooled cross-sectional data set instead of an equation unit root test. The main inspiration behind the panel data unit root tests, as conferred by Maddala and Kim (1998), was to increase the power of the test by increasing the sample size through the panel data approach. These panel data unit root tests take advantage of cross-sectional information and escort to boost in power of the test. Asymptotic normality of the Dickey-Fuller test statistic for panel data with arbitrarily large cross-sectional dimensions and small fixed time series dimensions was obtained by Breitung and Meyer (1994), and Papell (1997) extended the panel data unit root test in case of purchase power parity to test the stationarity. Im, Pesaran, and Shin (1997) observed that the t-bar statistic has higher power than the Levin-Lin test by allowing a greater heterogeneity across individuals. Under the alternative hypothesis, they assumed the same long-run multiplier across countries and proposed a new test based on the mean group approach which was applied by Wu (2000) in panel data unit root tests to obtain support for the mean-reverting property of the current account series. Levin, Lin, and Chu (2002) analyzed the asymptotic and finite sample properties of the panel data unit root test when intercept and trend are allowed to vary across individuals.

BAYESIAN UNIT ROOT TEST FOR AR(1)

Westerlund (2014) proposed a panel data unit root test where errors may be not only serial and/or cross-correlated, but also unconditionally heteroscedastic. Westerlund also shows evidence through Monte Carlo and suggested that the new test performs well in small samples in the case of panel data. The classical testing is predominantly based on the assumption that parameters are fixed and the population is finite; however, Bayesian analysis is free from such assumptions. The Bayesian approach for testing the unit root hypothesis was proposed by Sims (1988) and Sims and Uhlig (1991).

Kuma, Chaturvedi, and Afifa (2016) tested the stationarity of NAV series of NPS (new pension scheme) using unit root hypothesis and found that series are trend stationary. Karavias and Tzavalis (2016) discussed asymptotic local power properties in reference to the panel data unit root test for various fixed T and serially correlated error. They also studied the case considering the instrumental variable and found that variables are dominant in the case of the test based on the within-groups estimator. Schotman and van Dijk (1991) considered a panel data time series model with linear time trend incorporating augmentation term.

Here, a posterior odds ratio is derived for testing the unit root hypothesis considering prior assumption. A simulation study is carried out to explore the panel data unit root test. Considered here are three series in AR(1) panel data time series. First, take fixed values of intercept terms and generate the series for all possible combinations of fixed coefficients of time trend. Similarly, the simulation is repeated for taking a constant coefficient of time trend and generating the panel data for all possible combinations of fixed intercept terms. For numerical simplification, we have taken the prior odds ratio to be one and then we have obtained posterior odds ratios by applying the results of the theorem. The simulation results demonstrate that, for all combinations of selected values of the parameters, the derived posterior odds ratio correctly identifies the true hypothesis.

Model and Hypothesis

Consider the panel data time series model with linear time trend

$$y_{it} = \mu_i + \delta_i t + u_{it} \quad (1)$$

with $\{y_{it} : i = 1, 2, \dots, n; t = 1, 2, \dots, T\}$. This is a time series of observations on each of n cross sections, and its error is from an AR(1) process

$$u_{it} = \rho u_{it-1} + \varepsilon_{it} \quad (2)$$

Here, ρ is an autoregressive time series model and the ε_{it} are iid random variables, each following normal distribution with mean zero and variance τ^{-1} .

Write the model (1), incorporating an augmentation term and error equation (2) as:

$$y_{it} = \rho y_{it-1} + [(1-\rho)\mu_i + \rho\delta_i] + (1-\rho)\delta_i t + \sum_{j=1}^{k_i} \theta_{ij} \Delta y_{it-j} + \varepsilon_{it} \quad (3)$$

Considering $\alpha_i = [(1-\rho)\mu_i + \rho\delta_i]$ and $\beta_i = (1-\rho)\delta_i$, and $\Delta y_{it} = y_{it} - y_{it-1}$. Rewrite the above model (3) as

$$y_{it} = \rho y_{it-1} + \alpha_i + \beta_i t + \sum_{j=1}^{k_i} \theta_{ij} \Delta y_{it-j} + \varepsilon_{it} \quad (4)$$

The interest is in testing the unit root hypothesis $H_0: \rho = 1$ against the alternative $H_1: \rho \in S$ with $S = \{a < \rho < 1; a > -1\}$. Under the null hypothesis of unit root, the model reduces to

$$\Delta y_{it} = \delta_i + \sum_{j=1}^{k_i} \theta_{ij} \Delta y_{it-j} + \varepsilon_{it} \quad (5)$$

Let $\mathbf{l}_t$ be a $T \times 1$ vector with all elements 1 and let $\xi_T = (1, 2, \dots, T)'$. Further, define

$$\begin{aligned} \Delta \mathbf{y}_i &= (\Delta y_{i1}, \Delta y_{i2}, \dots, \Delta y_{iT})', \quad \Delta \mathbf{y} = (\Delta \mathbf{y}'_1, \Delta \mathbf{y}'_2, \dots, \Delta \mathbf{y}'_n)', \\ \mathbf{y} &= (\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_n)', \quad \mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})', \quad \boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \dots, \boldsymbol{\varepsilon}'_n)' \\ \mathbf{Z} &= [(\mathbf{I}_n \otimes \mathbf{l}_t)(\mathbf{I}_n \otimes \xi_T)], \quad \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)', \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)', \\ \boldsymbol{\gamma} &= (\boldsymbol{\alpha}, \boldsymbol{\beta})', \quad \boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_n)' \end{aligned} \quad (6)$$

Notice the i^{th} equation includes an augmentation term of order k_i . For writing the models (5) and (6) in matrix notations, let us write

BAYESIAN UNIT ROOT TEST FOR AR(1)

$$k = \sum_{i=1}^n k_i, \quad \boldsymbol{\theta}_i = \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \\ \vdots \\ \theta_{ik_i} \end{pmatrix}, \quad \mathbf{X}_i = \begin{pmatrix} \Delta y_{i0} & \Delta y_{i-1} & \cdots & \Delta y_{i1-k_i} \\ \Delta y_{i-1} & \Delta y_{i0} & \cdots & \Delta y_{i2-k_i} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta y_{iT-1} & \Delta y_{iT-2} & \cdots & \Delta y_{iT-k_i} \end{pmatrix}, \quad (7)$$

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}'_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}'_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}'_n \end{pmatrix}$$

Utilizing the above notations, along with the notations defined in (6), write the model under the null and alternative hypothesis as

$$\text{Under } H_0 : \Delta \mathbf{y} = (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta} + \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (8)$$

$$\text{Under } H_1 : \mathbf{y} = \rho \mathbf{y}_{-1} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (9)$$

where $\mathbf{y}_{-1} = (\mathbf{y}'_{-1,1}, \mathbf{y}'_{-1,2}, \dots, \mathbf{y}'_{-1,n})$ and $\mathbf{y}_{-1,i} = (y_{i0}, y_{i1}, \dots, y_{iT-1})'$.

Posterior Odds Ratio

The posterior odds ratio is now derived for the unit root hypothesis. Assume prior distributions as in Schotman and van Dijk (1991) for the following parameters of the model:

$$\begin{aligned} \boldsymbol{\delta} &\sim N\left(0, \frac{1}{\mathfrak{g}\tau} \mathbf{I}_n\right), \quad \boldsymbol{\alpha} \sim N\left((1-\rho)y_0, \frac{(1+\rho)}{(1-\rho)\tau} \mathbf{I}_n\right), \\ \boldsymbol{\beta} &\sim N\left(0, \frac{(1-\rho)^2}{\mathfrak{g}\tau} \mathbf{I}_n\right), \\ p(\tau) &\propto \frac{1}{\tau}, \quad 0 < \tau < \infty, \\ p(\rho) &= \frac{1}{1-a}, \quad a < \rho < 1; a > -1 \\ p(\theta) &\propto 1 \end{aligned} \quad (10)$$

where $\mathfrak{P}$ is considered to be a hyper parameter. Then prior distribution for γ is given by

$$\gamma \sim N\left((1-\rho)\phi_0, \frac{1}{\tau}V(\rho)^{-1}\right) \quad (11)$$

where, if y_0 is the vector of initial observations,

$$\phi_0 = \begin{pmatrix} y_0 \\ 0 \end{pmatrix}, \quad V(\rho) = \begin{pmatrix} \frac{(1+\rho)}{(1-\rho)} \mathbf{I}_n & 0 \\ 0 & \frac{\mathfrak{P}}{(1-\rho)^2} \mathbf{I}_n \end{pmatrix} \quad (12)$$

Assume that the prior probability in favor of H_0 is $P(\rho = 1) = p_0$ and prior probability in favor of H_1 is $P(\rho \in S) = 1 - p_0$.

Define

$$\begin{aligned} \Sigma &= \mathbf{I}_{nT} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{A} = (\mathbf{I}_n \otimes \mathbf{I}_T)' \Sigma (\mathbf{I}_n \otimes \mathbf{I}_T) + \mathfrak{P} \mathbf{I}_n, \\ \varsigma &= \Delta \mathbf{y}' \left(\Sigma - \Sigma (\mathbf{I}_n \otimes \mathbf{I}_T) \mathbf{A}^{-1} (\mathbf{I}_n \otimes \mathbf{I}_T)' \Sigma \right) \Delta \mathbf{y}, \quad \mathbf{B}(\rho) = \mathbf{Z}' \Sigma \mathbf{Z} + V(\rho), \\ \varsigma(\rho) &= (\mathbf{y} - \rho \mathbf{y}_{-1})' \Sigma (\mathbf{y} - \rho \mathbf{y}_{-1}) \\ &\quad - \left(\mathbf{Z}' (\mathbf{y} - \rho \mathbf{y}_{-1}) + (1 - \rho^2) \phi_0 \right)' \mathbf{B}(\rho) \left(\mathbf{Z}' (\mathbf{y} - \rho \mathbf{y}_{-1}) + (1 - \rho^2) \phi_0 \right) \end{aligned} \quad (13)$$

Theorem: An AR(1) panel data time series model with linear time trend and augmentation term is difference stationary or trend stationary equivalent to $H_0: \rho = 1$ against the alternative $H_1: \rho \in S$ with $S = \{a < \rho < 1; a > -1\}$ with prior odds ratio $p_0 / (1 - p_0)$ can be tested by the posterior odds ratio is given by

$$\beta_{01} = \frac{p_0}{1 - p_0} \frac{1 - a}{|\mathbf{A}|^{\frac{1}{2}} \varsigma^{\frac{nT-k}{2}}} \left[\int_a^1 \frac{(1+\rho)^{\frac{n}{2}}}{(1-\rho)^{\frac{3n}{2}} |\mathbf{B}(\rho)|^{\frac{1}{2}} (\varsigma(\rho))^{\frac{nT-k}{2}}} d\rho \right]^{-1} \quad (14)$$

Proof: The likelihood function under the unit root hypothesis is given by

BAYESIAN UNIT ROOT TEST FOR AR(1)

$$p(\mathbf{y} | \boldsymbol{\delta}, \tau) = \frac{\tau^{\frac{nT}{2}}}{(2\pi)^{\frac{nT}{2}}} \exp \left[-\frac{\tau}{2} \left\{ (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta} - \mathbf{X} \boldsymbol{\theta})' (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta} - \mathbf{X} \boldsymbol{\theta}) \right\} \right] \quad (15)$$

Combining the likelihood function with the prior distributions (10) of parameters, we get

$$p(\mathbf{y} | H_0) = \int_0^\infty \int_{\mathbb{R}^n} \int_{\mathbb{R}^{nk}} \frac{\tau^{\frac{nT+n-1}{2}} \boldsymbol{\vartheta}^{\frac{n}{2}}}{(2\pi)^{\frac{nT+n}{2}}} \exp \left[-\frac{\tau}{2} \left\{ (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta} - \mathbf{X} \boldsymbol{\theta})' (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta} - \mathbf{X} \boldsymbol{\theta}) + \boldsymbol{\vartheta} \boldsymbol{\delta}' \boldsymbol{\delta} \right\} \right] d\boldsymbol{\theta} d\boldsymbol{\delta} d\tau \quad (16)$$

Let $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta})$. Then, write (16) as

$$\begin{aligned} p(\mathbf{y} | H_0) &= \int_0^\infty \int_{\mathbb{R}^n} \int_{\mathbb{R}^k} \frac{\tau^{\frac{nT+n-1}{2}} \boldsymbol{\vartheta}^{\frac{n}{2}}}{(2\pi)^{\frac{nT+n}{2}}} \exp \left[-\frac{\tau}{2} \left\{ (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta})' (\mathbf{I}_{nk} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') (\Delta \mathbf{y} - (\mathbf{I}_n \otimes \mathbf{I}_T) \boldsymbol{\delta}) \right. \right. \\ &\quad \left. \left. + \boldsymbol{\vartheta} \boldsymbol{\delta}' \boldsymbol{\delta} + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\} \right] d\boldsymbol{\theta} d\boldsymbol{\delta} d\tau \\ &= \frac{\boldsymbol{\vartheta}^{\frac{n}{2}} |\mathbf{X}\mathbf{X}|^{\frac{1}{2}} \Gamma\left(\frac{nT-k}{2}\right)}{\pi^{\frac{nT-k}{2}} |\mathbf{A}|^{\frac{1}{2}} \boldsymbol{\varsigma}^{\frac{nT-k}{2}}} \end{aligned} \quad (17)$$

Further, under H_1 , the likelihood function is given by

$$p(\mathbf{y} | \boldsymbol{\gamma}, \rho, \tau) = \frac{\tau^{\frac{nT}{2}}}{(2\pi)^{\frac{nT}{2}}} \exp \left[-\frac{\tau}{2} \left\{ (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\theta})' (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\theta}) \right\} \right] \quad (18)$$

Combining the likelihood function with the prior distribution, (10), (11), and (12) leads to

$$p(\mathbf{y} | H_1) = \int_a^1 \int_0^\infty \int_{\mathbb{R}^{2n}} \int_{\mathbb{R}^k} \frac{\tau^{\frac{nT}{2}+n-1} \boldsymbol{\Sigma}^{\frac{n}{2}} |\mathbf{V}(\rho)|^{\frac{1}{2}}}{(2\pi)^{\frac{nT}{2}+n} (1-a)} \exp \left[-\frac{\tau}{2} \left\{ (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\theta})' (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\theta}) \right. \right. \\ \left. \left. + (\boldsymbol{\gamma} - (1-\rho)\boldsymbol{\phi}_0)' \mathbf{V}(\rho) (\boldsymbol{\gamma} - (1-\rho)\boldsymbol{\phi}_0) \right\} \right] d\boldsymbol{\theta} d\boldsymbol{\gamma} d\tau d\rho$$

writing

$$\tilde{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\gamma}} = \mathbf{B}(\rho)^{-1} \left[\mathbf{Z}'\boldsymbol{\Sigma}(\mathbf{y} - \rho \mathbf{y}_{-1}) + (1-\rho^2)\boldsymbol{\phi}_0 \right]$$

which obtains

BAYESIAN UNIT ROOT TEST FOR AR(1)

$$\begin{aligned}
p(\mathbf{y} | H_1) &= \int_a^1 \int_0^\infty \int_{\mathbb{R}^{2n}} \int_{\mathbb{R}^k} \frac{\tau^{\frac{nT}{2}+n-1} \mathfrak{G}^{\frac{n}{2}} |\mathbf{V}(\rho)|^{\frac{1}{2}}}{(2\pi)^{\frac{nT}{2}+n} (1-a)} \exp \left[-\frac{\tau}{2} \left\{ (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\gamma)' \boldsymbol{\Sigma} (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\gamma) \right. \right. \\
&\quad \left. \left. + (\gamma - (1-\rho)\phi_0)' \mathbf{V}(\rho) (\gamma - (1-\rho)\phi_0) \right\} + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right] d\boldsymbol{\theta} d\gamma d\tau d\rho \\
&= \int_a^1 \int_0^\infty \int_{\mathbb{R}^{2n}} \frac{\tau^{\frac{nT-k}{2}+n-1} \mathfrak{G}^{\frac{n}{2}} |\mathbf{V}(\rho)|^{\frac{1}{2}} |\mathbf{X}' \mathbf{X}|^{-\frac{1}{2}}}{(2\pi)^{\frac{nT}{2}+n} (1-a)} \exp \left[-\frac{\tau}{2} \left\{ (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\gamma)' \boldsymbol{\Sigma} (\mathbf{y} - \rho \mathbf{y}_{-1} - \mathbf{Z}\gamma) \right. \right. \\
&\quad \left. \left. + (\gamma - (1-\rho)\phi_0)' \mathbf{V}(\rho) (\gamma - (1-\rho)\phi_0) \right\} \right] d\gamma d\tau d\rho \\
&= \int_a^1 \int_0^\infty \int_{\mathbb{R}^{2n}} \frac{\tau^{\frac{nT-k}{2}+n-1} \mathfrak{G}^{\frac{n}{2}} |\mathbf{V}(\rho)|^{\frac{1}{2}} |\mathbf{X}' \mathbf{X}|^{-\frac{1}{2}}}{(2\pi)^{\frac{nT}{2}+n} (1-a)} \exp \left[-\frac{\tau}{2} \left\{ \varsigma(\rho) + (\gamma - \tilde{\gamma})' \mathbf{B}(\rho) (\gamma - \tilde{\gamma}) \right\} \right] d\gamma d\tau d\rho \\
&= \int_a^1 \int_0^\infty \frac{\tau^{\frac{nT-k}{2}-1} \mathfrak{G}^{\frac{n}{2}} (1+\rho)^{\frac{n}{2}} |\mathbf{X}' \mathbf{X}|^{-\frac{1}{2}}}{(2\pi)^{\frac{nT-k}{2}} (1-\rho)^{\frac{3n}{2}} |\mathbf{B}(\rho)|^{\frac{1}{2}} (1-a)} \exp \left[-\frac{\tau}{2} \varsigma(\rho) \right] d\tau d\rho \\
&= \frac{\mathfrak{G}^{\frac{n}{2}} |\mathbf{X}' \mathbf{X}|^{-\frac{1}{2}} \Gamma\left(\frac{nT-k}{2}\right)}{\pi^{\frac{nT-k}{2}} (1-a)} \int_a^1 \frac{(1+\rho)^{\frac{n}{2}}}{(1-\rho)^{\frac{3n}{2}} |\mathbf{B}(\rho)|^{\frac{1}{2}} \varsigma(\rho)^{\frac{nT-k}{2}}} d\rho
\end{aligned} \tag{19}$$

Using (17) and (19), obtain the expression (14) required of the theorem.

Table 1. Posterior odds ratio and autoregressive coefficient value

β_{01}	$\hat{\rho}$	$SE(\hat{\rho})$	$\hat{\sigma}^2$
1.25E-214	0.674	1.23E+03	7.83E+06

Numerical Illustration

To understand the need and worthiness of proposed study, empirical as well as simulation studies are frequently used methodology, therefore an empirical analysis for the model under study is explored and then a simulation study is also applied for the same.

Empirical Analysis

The objective is to develop the testing procedure in order to test the unit root hypothesis for the panel data time series model with linear time trend incorporating an augmentation term of order one. Consider the Bayesian procedure for testing the unit root hypothesis and obtain the posterior odds ratio and empirical analysis to justify the Bayesian testing procedure for proposed model. To fulfill the objective, data were taken from *Statistics at a Glance 2014* (Ministry of Agriculture, India, 2015). The import series of fertilizers were analyzed, namely Nitrogen (N), Phosphate (P), and Potash (K), in a panel dataset covering the period from 1980-81 to 2013-14.

Consider the series of fertilizers as a time series in three panels where panels are taken in respect to different fertilizers. Test the hypothesis whether the observed fertilizer series are difference stationary or trend stationary or, equivalently, the unit root hypothesis $H_0: \rho = 1$ against the alternative $H_1: \rho \in S$ with $S = \{a < \rho < 1; a > -1\}$. Let $\{y_{it} : i = 1, 2, \dots, n; t = 1, 2, \dots, T\}$ be the recorded import series of fertilizers are assumed by the model (5) for analysis purpose.

The posterior odds ratio, estimated value of $\hat{\rho}$ with $SE(\hat{\rho})$, and error variance ($\hat{\sigma}^2$) are recorded on Table 1, and the maximum likelihood estimates of autoregressive regression coefficients: intercept, trend, and $\hat{\rho}$ are given in Table 2 with variance covariance matrix of regression coefficients.

As the posterior odds ratio works on the comparison of probabilities, it is clear from Table 1 that observed posterior odds ratio under study is less than one which indicates that the unit root hypothesis $H_0: \rho = 1$ is rejected and we may accept the alternative hypothesis $H_1: \rho \in S$ with $S = \{a < \rho < 1; a > -1\}$. Therefore, series of

BAYESIAN UNIT ROOT TEST FOR AR(1)

fertilizers under study are trend stationary. For justification of real data analysis, a simulation study has also conducted.

Simulation Study

Data were generated from the panel data time series model:

$$y_{it} = \rho y_{it-1} + [(1-\rho)\mu_i + \rho\delta_i] + (1-\rho)\delta_i t + \sum_{j=1}^{k_i} \theta_{ij} \Delta y_{it-j} + \varepsilon_{it}$$

with $i = 1, 2, \dots, n$; $t = 1, 2, \dots, T$. For generating a panel data time series of size 25, consider three panels with $\varepsilon_{it} \sim N(0, 1)$, where the initial observations are $y_{10} = 1000$, $y_{20} = 1500$, and $y_{30} = 2000$, the coefficient of augmentation term with order 1; $\theta_{11} = \theta_{21} = \theta_{31} = 1$. We have generated the data for two situations. In the first situation, take the fixed values of the intercept term, which are $\{\mu_1 = 750, \mu_2 = 1000, \mu_3 = 1250\}$ and generate the series for all the possible 27 combinations of values $\{\delta_1, \delta_2, \delta_3\} = \{1, 1.25, 1.5\}$. Similarly, in the second situation, take the fixed values of the coefficient of time trend $\{\delta_1 = 1, \delta_2 = 1.25, \delta_3 = 1.5\}$ and generate the series for all the 27 possible combinations of $\{\mu_1, \mu_2, \mu_3\} = \{750, 1000, 1250\}$.

Using the derived theorem for the model (3), the presence of unit root in autoregressive panel data is tested with linear time trend and augmentation term for the values of $\rho = \{0.90, 0.92, 0.94, 0.96, 0.98\}$. The posterior odds ratio and estimated value of $\hat{\rho}$ with $SE(\hat{\rho})$ for testing unit root hypothesis, i.e. series is difference stationary (equivalent to: $H_0: \rho = 1$) against the alternative, i.e. the series is trend stationary (equivalent to $H_1: \rho \in S$ with $S = \{a < \rho < 1; a > -1\}$) are calculated. Because the posterior odds ratio is decreasing with the increasing value of ρ , therefore we have reported the results only for $\rho = \{0.90, 0.94, 0.98\}$ in [Tables 3 to 5](#), considering equal prior probability for the null and alternative hypothesis.

AR(1) panel data time series model is generated considering the trend stationary model and we got all posterior odds ratios less than one. In a Bayesian testing procedure using the posterior odds ratio, the concept of accepting the hypothesis is directly decided with respect to the hypothesis which has more chance in comparison to other. Here, if the probability of the alternate is more than the probability of null that means the results are confirming that all the generated series under the setup of trend stationary are concluded trend stationary. Therefore, it may be concluded the results support the derived theorem for identifying the unit root hypothesis correctly.

Conclusion

The posterior odds ratio was derived for testing the unit root hypothesis in panel data time series models with a linear trend and augmentation term. The simulation and empirical study correctly tested the hypothesis. This may be extended for the cases of non-linear time trends, a model with non-normal errors, as well as other multivariate time series models.

Acknowledgements

All authors gratefully acknowledge the anonymous reviewers for their careful work and thoughtful suggestions that have helped improve this paper substantially, and the first author gratefully acknowledges the support from UGC under MRP Scheme (Grant No.42-43/2013).

References

- Breitung, J., & Meyer, W. (1994). Testing for unit roots in panel data: Are wages on different bargaining levels cointegrated? *Applied Economics*, 26(4), 353-361. doi: [10.1080/000368494000000081](https://doi.org/10.1080/000368494000000081)
- De Blander, R., & Dhaene, G. (2012). Unit root tests for panel data with AR(1) errors and small T. *The Econometrics Journal*, 15(1), 101-124. doi: [10.1111/j.1368-423x.2012.00363.x](https://doi.org/10.1111/j.1368-423x.2012.00363.x)
- De Wachter, S., Harris, R. D., & Tzavalis, E. (2007). Panel data unit roots tests: The role of serial correlation and the time dimension. *Journal of Statistical Planning and Inference*, 137(1), 230-244. doi: [10.1016/j.jspi.2005.11.004](https://doi.org/10.1016/j.jspi.2005.11.004)
- DeJong, D. N., & Whiteman, C. H. (1991). Reconsidering 'trends and random walks in macroeconomic time series'. *Journal of Monetary Economics*, 28(2), 221-254. doi: [10.1016/0304-3932\(91\)90051-o](https://doi.org/10.1016/0304-3932(91)90051-o)
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427-431. doi: [10.2307/2286348](https://doi.org/10.2307/2286348)
- Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4), 1057-1072. doi: [10.2307/1912517](https://doi.org/10.2307/1912517)

BAYESIAN UNIT ROOT TEST FOR AR(1)

Harris, R. D., & Tzavalis, E. (1999). Inference for unit roots in dynamic panels where the time dimension is fixed. *Journal of Econometrics*, 91(2), 201-226. doi: [10.1016/s0304-4076\(98\)00076-1](https://doi.org/10.1016/s0304-4076(98)00076-1)

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *TEST*, 16(1), 1-22. doi: [10.1007/s11749-007-0046-x](https://doi.org/10.1007/s11749-007-0046-x)

Im, K. S., Pesaran, M. H., & Shin, Y. (1997). *Testing for unit roots in heterogeneous panels* [Mimeo]. Cambridge, UK: University of Cambridge Department of Applied Economics.

Karavias, Y., & Tzavalis, E. (2016). Local power of fixed- T panel unit root tests with serially correlated errors and incidental trends. *Journal of Time Series Analysis*, 37(2), 222-239. doi: [10.1111/jtsa.12144](https://doi.org/10.1111/jtsa.12144)

Kruiniger, H., & Tzavalis, E. (2002). *Testing for unit roots in short dynamic panels with serially correlated and heteroscedastic disturbance terms* (University of London Queen Mary Economics Working Paper No. 459). London, UK: University of London Queen Mary Department of Economics. doi: [10.2139/ssrn.317162](https://doi.org/10.2139/ssrn.317162)

Kuma, J., Chaturvedi, A., & Afifa, U. (2016). Bayesian unit root test for panel data (Economics and Econometrics Research Institute Research Paper No. 14/2016). Brussels, Belgium: Economics and Econometrics Research Institute. Retrieved from http://www.eeri.eu/documents/wp/EERI_RP_2016_14.pdf

Levin, A., & Lin, C.-F. (1992). *Unit root tests in panel data: Asymptotic and finite sample properties* (Discussion Paper 93-23). San Diego, CA: Department of Economics, University of California, San Diego.

Levin, A., & Lin, C.-F. (1993). *Unit root tests in panel data: New results* (Economics Working Paper Series 1993/56). San Diego, CA: University of California, San Diego.

Levin, A., Lin, C.-F., & Chu, C.-S. J. (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. *Journal of Econometrics*, 108(1), 1-24. doi: [10.1016/s0304-4076\(01\)00098-7](https://doi.org/10.1016/s0304-4076(01)00098-7)

Lubrano, M. (1995). Testing for unit roots in a Bayesian framework. *Journal of Econometrics*, 69(1), 81-109. doi: [10.1016/0304-4076\(94\)01663-k](https://doi.org/10.1016/0304-4076(94)01663-k)

Maddala, G. S., & Kim, I.-M. (1998). *Unit roots, cointegration, and structural change*. Cambridge, UK: Cambridge University Press. doi: [10.1017/cbo9780511751974](https://doi.org/10.1017/cbo9780511751974)

- Maddala, G. S., & Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and Statistics*, 61(S1), 631-652. doi: [10.1111/1468-0084.0610s1631](https://doi.org/10.1111/1468-0084.0610s1631)
- Madsen, E. (2010). Unit root inference in panel data models where the time-series dimension is fixed: A comparison of different tests. *The Econometrics Journal*, 13(1), 63-94. doi: [10.1111/j.1368-423x.2009.00302.x](https://doi.org/10.1111/j.1368-423x.2009.00302.x)
- Ministry of Agriculture, India. (2015). *Agricultural statistics at a glance 2014*. New Delhi, India: Oxford University Press. Retrieved from <http://eands.dacnet.nic.in/PDF/Agricultural-Statistics-At-Glance2014.pdf>
- Moon, H. R., & Perron, B. (2004). Testing for a unit root in panels with dynamic factors. *Journal of Econometrics*, 122(1), 81-126. doi: [10.1016/j.jeconom.2003.10.020](https://doi.org/10.1016/j.jeconom.2003.10.020)
- Moon, H. R., & Perron, B. (2008). Asymptotic local power of pooled t-ratio tests for unit roots in panels with fixed effects. *The Econometrics Journal*, 11(1), 80-104. doi: [10.1111/j.1368-423x.2008.00236.x](https://doi.org/10.1111/j.1368-423x.2008.00236.x)
- Papell, D. H. (1997). Searching for stationarity: Purchasing power parity under the current float. *Journal of International Economics*, 43(3), 313-332. doi: [10.1016/s0022-1996\(96\)01467-5](https://doi.org/10.1016/s0022-1996(96)01467-5)
- Schotman, P., & van Dijk, H. K. (1991). A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics*, 49(1-2), 195-238. doi: [10.1016/0304-4076\(91\)90014-5](https://doi.org/10.1016/0304-4076(91)90014-5)
- Shiller, R. J., & Perron, P. (1985). Testing the random walk hypothesis: Power versus frequency of observation. *Economics Letters*, 18(4), 381-386. doi: [10.1016/0165-1765\(85\)90058-8](https://doi.org/10.1016/0165-1765(85)90058-8)
- Sims, C. A. (1988). Bayesian skepticism on unit root econometrics. *Journal of Economic Dynamics and Control*, 12(2), 463-474. doi: [10.1016/0165-1889\(88\)90050-4](https://doi.org/10.1016/0165-1889(88)90050-4)
- Sims, C. A., & Uhlig, H. (1991). Understanding unit rooters: A helicopter tour. *Econometrica*, 59(6), 1591-1599. doi: [10.2307/2938280](https://doi.org/10.2307/2938280)
- Uhlig, H. (1994). What macroeconomists should know about unit roots: A Bayesian perspective. *Econometric Theory*, 10(3-4), 645-671. doi: [10.1017/s0266466600008719](https://doi.org/10.1017/s0266466600008719)
- Westerlund, J. (2014). Heteroscedasticity robust panel unit root tests. *Journal of Business & Economic Statistics*, 32(1), 112-135. doi: [10.1080/07350015.2013.857612](https://doi.org/10.1080/07350015.2013.857612)

BAYESIAN UNIT ROOT TEST FOR AR(1)

Wu, J.-L. (2000). Mean reversion of the current account: Evidence from the panel data unit-root test. *Economics Letters*, 66(2), 215-222. doi: [10.1016/S0165-1765\(99\)00198-6](https://doi.org/10.1016/S0165-1765(99)00198-6)

Yaffee, R. (2003). A primer for panel data analysis. *Connect: Information Technology at NYU, 2003(Fall)*. Retrieved from <https://wp.nyu.edu/connect/print-edition-2/>

Appendix A: Tables 2-5

Table 2. Variance and covariance matrix for fertilizers panel data

Coeff	Variance Covariance Matrix ($\hat{\Sigma}$)											
$\hat{\rho}$	6.75E-01	1.50E+06	5.90E+07	4.10E+08	-8.80E+08	-2.60E+08	-2.20E+08	-1.90E+08	-8.40E+05	-6.50E+05	-1.10E+06	
$\hat{\mu}_1$	-1.74E+01	5.90E+07	1.30E+13	1.60E+10	-3.50E+10	-9.00E+11	-8.60E+09	-7.40E+09	2.70E+08	-2.60E+07	-4.20E+07	
$\hat{\mu}_2$	2.20E+01	4.10E+08	1.60E+10	1.30E+13	-2.40E+11	-6.90E+10	-9.50E+11	-5.10E+10	-2.20E+08	3.40E+08	-2.90E+08	
$\hat{\mu}_3$	4.49E+02	-8.80E+08	-3.50E+10	-2.40E+11	1.40E+13	1.50E+11	1.30E+11	-8.00E+11	4.90E+08	3.80E+08	-9.20E+08	
$\hat{\beta}_1$	6.36E+01	-2.60E+08	-9.00E+11	-6.90E+10	1.50E+11	1.30E+11	3.80E+10	3.30E+10	-3.60E+07	1.10E+08	1.90E+08	
$\hat{\beta}_2$	4.08E+01	-2.20E+08	-8.60E+09	-9.50E+11	1.30E+11	3.80E+10	1.10E+11	2.70E+10	1.20E+08	-2.20E+06	1.60E+08	
$\hat{\beta}_3$	1.82E+01	-1.90E+08	-7.40E+09	-5.10E+10	-8.00E+11	3.30E+10	2.70E+10	1.10E+11	1.10E+08	8.20E+07	2.80E+08	
$\hat{\theta}_{11}$	4.04E-01	-8.40E+05	2.70E+08	-2.20E+08	4.90E+08	-3.60E+07	1.20E+08	1.10E+08	9.00E+06	3.60E+05	6.00E+05	
$\hat{\theta}_{12}$	1.52E-01	-6.50E+05	-2.60E+07	3.40E+08	3.80E+08	1.10E+08	-2.20E+06	8.20E+07	3.60E+05	6.70E+06	4.70E+05	
$\hat{\theta}_{13}$	1.38E-01	-1.10E+06	-4.20E+07	-2.90E+08	-9.20E+08	1.90E+08	1.60E+08	2.80E+08	6.00E+05	4.70E+05	7.90E+06	

BAYESIAN UNIT ROOT TEST FOR AR(1)

Table 3. For $\rho = 0.90$, posterior odds ratio and estimated autoregressive coefficient for different combinations of intercept (μ) and coefficient of time trend value (δ)

For fixed intercept term: $\mu_1 = 750; \mu_2 = 1000; \mu_3 = 1250$						For fixed coefficient of time trend: $\delta_1 = 1.0; \delta_2 = 1.25; \delta_3 = 1.5$					
δ_1	δ_2	δ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}	μ_1	μ_2	μ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}
1.00	1.00	1.00	0.9000	8.39E-04	3.13E-160	750	750	750	0.90002	8.34E-04	1.13E-140
1.00	1.00	1.25	0.9000	6.75E-04	2.86E-170	750	750	1000	0.90000	6.02E-04	3.06E-150
1.00	1.00	1.50	0.9000	8.43E-04	1.55E-170	750	750	1250	0.90002	8.39E-04	1.61E-160
1.00	1.25	1.00	0.9000	6.05E-04	2.80E-170	750	1000	750	0.90003	8.38E-04	1.24E-145
1.00	1.25	1.25	0.9000	9.45E-04	1.67E-170	750	1000	1000	0.90003	8.41E-04	3.02E-157
1.00	1.25	1.50	0.9000	8.43E-04	1.26E-170	750	1000	1250	0.90003	8.43E-04	1.26E-170
1.00	1.50	1.00	0.9000	8.43E-04	1.92E-170	750	1250	750	0.90000	6.01E-04	6.30E-149
1.00	1.50	1.25	0.9000	6.05E-04	1.66E-170	750	1250	1000	0.90003	8.43E-04	3.34E-162
1.00	1.50	1.50	0.9000	9.45E-04	9.94E-171	750	1250	1250	0.90003	8.16E-04	1.51E-140
1.25	1.00	1.00	0.8999	8.96E-04	3.59E-170	1000	750	750	0.90003	8.24E-04	1.49E-141
1.25	1.00	1.25	0.9000	8.43E-04	1.92E-170	1000	750	1000	0.90000	8.27E-04	1.81E-151
1.25	1.00	1.50	0.9001	5.83E-04	1.67E-170	1000	750	1250	0.90003	8.30E-04	3.80E-162
1.25	1.25	1.00	0.9000	8.43E-04	2.14E-170	1000	1000	750	0.90003	8.28E-04	1.17E-146
1.25	1.25	1.25	0.9000	6.05E-04	1.85E-170	1000	1000	1000	0.90000	6.04E-04	1.23E-158
1.25	1.25	1.50	0.9000	8.43E-04	1.14E-170	1000	1000	1250	0.90003	8.33E-04	7.28E-173
1.25	1.50	1.00	0.9000	8.43E-04	1.74E-170	1000	1250	750	0.90003	8.30E-04	4.00E-150
1.25	1.50	1.25	0.9000	8.43E-04	1.27E-170	1000	1250	1000	0.90000	6.05E-04	7.83E-164
1.25	1.50	1.50	0.9000	8.43E-04	9.27E-171	1000	1250	1250	0.90000	9.46E-04	2.28E-181
1.50	1.00	1.00	0.9000	6.05E-04	2.83E-170	1250	750	750	0.90003	8.45E-04	1.65E-178
1.50	1.00	1.25	0.9000	8.43E-04	1.74E-170	1250	750	1000	0.90003	8.19E-04	4.68E-150
1.50	1.00	1.50	0.9000	8.43E-04	1.27E-170	1250	750	1250	0.90003	8.21E-04	3.82E-160
1.50	1.25	1.00	0.9000	8.43E-04	1.94E-170	1250	1000	750	0.90000	6.02E-04	2.10E-145
1.50	1.25	1.25	0.9000	8.43E-04	1.42E-170	1250	1000	1000	0.90003	8.22E-04	6.34E-157
1.50	1.25	1.50	0.9000	8.43E-04	1.03E-170	1250	1000	1250	0.90000	6.06E-04	5.76E-170
1.50	1.50	1.00	0.9000	8.43E-04	1.58E-170	1250	1250	750	0.90000	6.03E-04	1.00E-148
1.50	1.50	1.25	0.8999	1.21E-03	1.81E-170	1250	1250	1000	0.90003	8.24E-04	8.17E-162
1.50	1.50	1.50	0.9000	8.43E-04	8.40E-171	1250	1250	1250	0.90003	8.27E-04	9.48E-178

Table 4. For $\rho = 0.94$, posterior odds ratio and estimated autoregressive coefficient for different combinations of intercept (μ) and coefficient of time trend value (δ)

For fixed intercept term: $\mu_1 = 750; \mu_2 = 1000; \mu_3 = 1250$						For fixed coefficient of time trend: $\delta_1 = 1.0; \delta_2 = 1.25; \delta_3 = 1.5$					
δ_1	δ_2	δ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}	μ_1	μ_2	μ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}
1.00	1.00	1.00	0.9400	8.10E-04	1.05E-181	750	750	750	0.94000	8.13E-04	4.33E-160
1.00	1.00	1.25	0.9400	9.88E-04	2.00E-192	750	750	1000	0.93998	6.06E-04	2.73E-170
1.00	1.00	1.50	0.9400	8.09E-04	1.64E-193	750	750	1250	0.94000	8.10E-04	2.49E-182
1.00	1.25	1.00	0.9400	6.03E-04	1.65E-192	750	1000	750	0.94000	8.12E-04	2.66E-165
1.00	1.25	1.25	0.9400	7.17E-04	9.46E-193	750	1000	1000	0.94000	8.11E-04	2.01E-178
1.00	1.25	1.50	0.9400	8.09E-04	1.06E-193	750	1000	1250	0.94000	8.09E-04	1.06E-193
1.00	1.50	1.00	0.9400	8.09E-04	2.69E-193	750	1250	750	0.93998	6.04E-04	4.84E-168
1.00	1.50	1.25	0.9400	6.03E-04	5.43E-193	750	1250	1000	0.94000	8.10E-04	3.06E-183
1.00	1.50	1.50	0.9400	7.17E-04	3.14E-193	750	1250	1250	0.94000	8.08E-04	5.31E-159
1.25	1.00	1.00	0.9400	5.95E-04	1.16E-191	1000	750	750	0.94000	8.11E-04	1.25E-160
1.25	1.00	1.25	0.9400	8.09E-04	2.66E-193	1000	750	1000	0.94000	8.10E-04	1.27E-171
1.25	1.00	1.50	0.9400	8.94E-04	9.88E-193	1000	750	1250	0.94000	8.08E-04	2.26E-183
1.25	1.25	1.00	0.9400	8.09E-04	3.40E-193	1000	1000	750	0.94000	8.10E-04	6.20E-166
1.25	1.25	1.25	0.9400	6.03E-04	6.93E-193	1000	1000	1000	0.93998	6.01E-04	8.86E-179
1.25	1.25	1.50	0.9400	8.09E-04	8.73E-194	1000	1000	1250	0.94000	8.07E-04	3.46E-195
1.25	1.50	1.00	0.9400	8.09E-04	2.22E-193	1000	1250	750	0.94000	8.09E-04	3.91E-169
1.25	1.50	1.25	0.9400	8.09E-04	1.12E-193	1000	1250	1000	0.93998	5.99E-04	7.29E-184
1.25	1.50	1.50	0.9400	8.09E-04	5.65E-194	1000	1250	1250	0.93996	7.13E-04	3.78E-203
1.50	1.00	1.00	0.9400	6.03E-04	1.76E-192	1250	750	750	0.94000	8.08E-04	8.02E-202
1.50	1.00	1.25	0.9400	8.09E-04	2.21E-193	1250	750	1000	0.94000	8.07E-04	2.50E-169
1.50	1.00	1.50	0.9400	8.09E-04	1.12E-193	1250	750	1250	0.94000	8.05E-04	4.35E-180
1.50	1.25	1.00	0.9400	8.09E-04	2.83E-193	1250	1000	750	0.94000	5.97E-04	1.51E-163
1.50	1.25	1.25	0.9400	8.09E-04	1.43E-193	1250	1000	1000	0.94000	8.06E-04	1.81E-176
1.50	1.25	1.50	0.9400	8.09E-04	7.23E-194	1250	1000	1250	0.93998	5.95E-04	8.22E-190
1.50	1.50	1.00	0.9400	8.09E-04	1.84E-193	1250	1250	750	0.91998	7.14E-04	6.71E-154
1.50	1.50	1.25	0.9400	9.83E-04	5.91E-194	1250	1250	1000	0.94000	8.05E-04	5.40E-181
1.50	1.50	1.50	0.9400	8.09E-04	4.67E-194	1250	1250	1250	0.94000	8.03E-04	7.19E-198

BAYESIAN UNIT ROOT TEST FOR AR(1)

Table 5. For $\rho = 0.98$, posterior odds ratio and estimated autoregressive coefficient for different combinations of intercept (μ) and coefficient of time trend value (δ)

For fixed intercept term: $\mu_1 = 750; \mu_2 = 1000; \mu_3 = 1250$						For fixed coefficient of time trend: $\delta_1 = 1.0; \delta_2 = 1.25; \delta_3 = 1.5$					
δ_1	δ_2	δ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}	μ_1	μ_2	μ_3	$\hat{\rho}$	$SE(\hat{\rho})$	β_{01}
1.00	1.00	1.00	0.9800	7.17E-04	4.57E-176	750	750	750	0.97999	7.29E-04	2.65E-154
1.00	1.00	1.25	0.9800	6.08E-04	1.47E-188	750	750	1000	0.98000	5.17E-04	3.63E-166
1.00	1.00	1.50	0.9800	7.11E-04	3.52E-189	750	750	1250	0.97999	7.16E-04	3.84E-178
1.00	1.25	1.00	0.9800	5.09E-04	5.04E-188	750	1000	750	0.97999	7.24E-04	3.12E-159
1.00	1.25	1.25	0.9800	6.03E-04	2.79E-187	750	1000	1000	0.97999	7.17E-04	5.46E-174
1.00	1.25	1.50	0.9800	7.11E-04	8.96E-190	750	1000	1250	0.97999	7.11E-04	8.96E-190
1.00	1.50	1.00	0.9800	7.11E-04	2.40E-188	750	1250	750	0.98000	5.15E-04	1.66E-160
1.00	1.50	1.25	0.9800	5.09E-04	1.46E-189	750	1250	1000	0.97999	7.12E-04	1.62E-176
1.00	1.50	1.50	0.9800	6.03E-04	4.16E-189	750	1250	1250	0.98000	7.26E-04	1.14E-149
1.25	1.00	1.00	0.9800	5.18E-04	3.22E-187	1000	750	750	0.97999	7.27E-04	2.41E-153
1.25	1.00	1.25	0.9800	7.11E-04	2.36E-188	1000	750	1000	0.98000	7.20E-04	4.02E-165
1.25	1.00	1.50	0.9799	4.65E-04	4.56E-190	1000	750	1250	0.97999	7.14E-04	2.23E-176
1.25	1.25	1.00	0.9800	7.11E-04	6.05E-188	1000	1000	750	0.97999	7.22E-04	3.90E-158
1.25	1.25	1.25	0.9800	5.09E-04	3.81E-189	1000	1000	1000	0.98000	5.16E-04	2.91E-172
1.25	1.25	1.50	0.9800	7.11E-04	6.46E-190	1000	1000	1250	0.97999	7.09E-04	6.89E-188
1.25	1.50	1.00	0.9800	7.11E-04	1.77E-188	1000	1250	750	0.97999	7.17E-04	1.16E-159
1.25	1.50	1.25	0.9800	7.11E-04	1.74E-189	1000	1250	1000	0.98000	5.13E-04	8.90E-175
1.25	1.50	1.50	0.9800	7.10E-04	1.71E-190	1000	1250	1250	0.98000	5.98E-04	4.82E-194
1.50	1.00	1.00	0.9800	5.10E-04	1.14E-187	1250	750	750	0.97999	7.06E-04	1.91E-195
1.50	1.00	1.25	0.9800	7.11E-04	1.89E-188	1250	750	1000	0.97999	7.20E-04	5.67E-160
1.50	1.00	1.50	0.9800	7.11E-04	2.03E-189	1250	750	1250	0.98000	7.13E-04	1.35E-169
1.50	1.25	1.00	0.9800	7.11E-04	4.87E-188	1250	1000	750	0.98000	5.24E-04	6.50E-154
1.50	1.25	1.25	0.9800	7.11E-04	4.96E-189	1250	1000	1000	0.97999	7.14E-04	3.69E-166
1.50	1.25	1.50	0.9800	7.11E-04	5.03E-190	1250	1000	1250	0.98000	5.15E-04	6.13E-179
1.50	1.50	1.00	0.9800	7.11E-04	1.41E-188	1250	1250	750	0.98000	5.21E-04	3.39E-155
1.50	1.50	1.25	0.9800	6.74E-04	4.30E-190	1250	1250	1000	0.97999	7.10E-04	1.07E-168
1.50	1.50	1.50	0.9800	7.10E-04	1.32E-190	1250	1250	1250	0.97999	7.04E-04	2.86E-184

Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha

Heather J. Turner
Dallas ISD
Dallas, TX

Prathiba Natesan
University of North Texas
Denton, TX

Robin K. Henson
University of North Texas
Denton, TX

The aim of this study was to investigate the performance of the Fisher, Feldt, Bonner, and Hakstian and Whalen (HW) confidence intervals methods for the non-parametric reliability estimate, ordinal alpha. All methods yielded unacceptably low coverage rates and potentially increased Type-I error rates.

Keywords: Ordinal alpha, confidence interval, coverage, Monte Carlo, simulation

Introduction

Reliability is an estimate of the consistency of results from a measurement (Crocker & Algina, 2008; Cronbach, 1951) and an essential component to establish validity of a scale (Allen & Yen, 1979, 2002). Social scientists often measure attitudes and opinions with ordinal Likert-type ratings. The individual options on the scale are assumed to be discrete realizations of an underlying continuously-scaled construct (Flora & Curran, 2004). Nevertheless, researchers often treat ordinally-scaled data as continuous by using statistical methods that assume continuity of data. This causes an empirical mismatch with the data analyzed (Gadermann, Guhn, & Zumbo, 2012; Schmitt, 1996; Sijtsma, 2009; Streiner, 2003), underestimation of sample coefficient alpha, and may lead researchers to incorrect conclusions (Duhachek & Iacobucci, 2004; Flora & Curran, 2004; Gadermann et al., 2012; Maydeu-Olivares, Coffman, & Hartmann, 2007; Zumbo, Gadermann, & Zeisser, 2007).

Dr. Turner is an evaluation analyst in the Dallas Independent School District. Email her at heatherjeanturner@gmail.com. Dr. Natesan is an Associate Professor of Research Methods, Measurement, and Statistics. Email her at prathibachaj@gmail.com. The first and second authors contributed equally to this manuscript and are listed alphabetically by first name. Dr. Henson is a Professor of Educational Psychology. Email him at robin.henson@unt.edu.

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

One strategy to overcome the violation of continuity is to use ordinal coefficient alpha, which uses polychoric correlations instead of Pearson correlations (Gadermann et al., 2012; Zumbo et al., 2007). Although ordinal coefficient alpha has been shown to be a more appropriate measure of reliability for ordinal data, it is still just a point estimate. Fan and Thompson (2001) emphasized the need to report confidence intervals for coefficient alpha as a method for assessing the relative stability of the statistic as opposed to invoking rigid cutoff decisions about whether a value is large enough or not. For instance, an alpha coefficient value of 0.8 is generally considered acceptable (Cronbach & Shavelson, 2004). A 95% confidence interval is expected to contain the true value of the statistical estimate 95% of the time when resampled. This resampling is often hypothetical. Confidence intervals are a function of the standard error of the statistic and their coverage rates indicate Type-I error rate. The probability of Type-I error increases when the coverage rate of confidence intervals is less than expected. Therefore, it is important to examine the performance of confidence intervals for statistical estimates. One such diagnostic is coverage rate. Coverage rate is defined as the percentage of confidence intervals that contain the true value. In empirical research we do not know what the true value is. However, in simulation the true value is known. Comparing the coverage rate to confidence level through simulation helps us verify if the confidence interval given by the theoretical formulas are accurate.

There are several confidence interval approaches for conventional coefficient alpha (e.g. Bonett, 2002; Feldt, 1965; Fisher, 1950; Hakstian & Whalen, 1976). Various confidence interval methods were investigated for conventional coefficient alpha (Bonett, 2002; Cui & Li, 2012; Duhachek & Iacobucci, 2004; Feldt, 1965; Fisher, 1950; Hakstian & Whalen, 1976; Maydeu-Olivares et al., 2007; Padilla, Divers, & Newton, 2012; van Zyl, Neudecker, & Nel, 2000; Yuan & Bentler, 2002). However, the confidence intervals have not been investigated for ordinal coefficient alpha. Given the widespread use of Likert-type data in educational and behavioral research and the increased emphasis on reporting and interpreting confidence intervals of estimates (Cohen, 1994; Cumming, 2012; Cumming & Fidler, 2009; Finch, Cumming, & Thomason, 2001; Thompson, 2006a, 2006b; Wilkinson & APA Task Force on Statistical Inference, 1999), there is a need to evaluate the performance of currently available confidence interval methods for the ordinal coefficient alpha.

The purpose of this study is to investigate the coverage rates, widths, and biases of the four types of confidence intervals (Bonnet, Feldt, Fisher, and Hakstian Whalen), and the accuracy of the ordinal alpha point estimates under

varying data conditions. Sample size, the number of items on an instrument, skewness of the responses, and population alpha were chosen as the data conditions because these have known to have an impact on the confidence intervals of alpha (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Romano, Kromrey, Owens, & Scott, 2011).

Literature Review

The reliability of a composite score may be estimated in a factor model as the ratio of item variances to total variances. The factor analytic representation of classical test theory is expressed as

$$X_i = \lambda_i \xi + u_i \quad i = 1, 2, \dots, p, \quad (1)$$

where X_i denotes the observed scores on the i^{th} item, λ_i denotes the factor pattern coefficient of the i th item, ξ is the true score common factor, and u_i is the uniqueness or random error up to p number of items. Novick and Lewis (1967) derived coefficient alpha as an unbiased estimate when the factor coefficients of each variable are equal to the common factor. Coefficient alpha can be calculated in a factor model as

$$\rho_{xx} = \frac{\left(\sum_{i=1}^p \lambda_i\right)^2}{\left(\sum_{i=1}^p \lambda_i\right)^2 + \sum_{i=1}^p \text{var}(e)_{ii}} \quad (2)$$

where $\text{var}(e)_{ii}$ is the error variance of the i^{th} item in a factor analytic model. Generally score reliability increases as coefficient alpha approaches a value of one.

Ordinal Coefficient Alpha

Gadermann et al. (2012) recommended using a non-parametric estimate of reliability coefficients for ordinal data, especially when there are few item response categories and skewed response distributions. Ordinal alpha is analogous to coefficient alpha, only differing by the type of correlation matrix used for computation. While coefficient alpha uses the Pearson correlation matrix and

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

assumes data are continuously scaled, ordinal alpha uses the polychoric correlation matrix.

Polychoric correlation estimates the linear relationship between two ordinally scaled variables as the correlation between their respective underlying latent variable distributions (Jöreskog, 1990). By not assuming normality for the underlying distribution, the thresholds are allowed to be unequally spaced. The observed ordinal response y_j for item j with C response categories, where the response option $c = 0, 1, 2, \dots, C - 1$ is defined as

$$y_j = c, \text{ if } \tau_c < y_j^* < \tau_{c+1}, \quad (3)$$

where τ_c, τ_{c+1} are the thresholds on the underlying continuum y_j^* and satisfy the constraint

$$-\infty = \tau_0 < \tau_1 < \dots < \tau_{C-1} < \tau_C = \infty. \quad (4)$$

The polychoric correlation, Φ , between two ordinal items y_i and y_j is given by the Pearson product-moment correlation between their corresponding underlying latent variables y_i^* and y_j^* , respectively. By treating the observed item's responses in this manner, ordinal alpha is a nonparametric reliability estimate. The formula for ordinal alpha is expressed as

$$\alpha_{ordinal} = \left[\frac{k}{k-1} \right] * \left[\frac{k(\lambda^2) - h^2}{k(\lambda^2) + \mu^2} \right], \quad (5)$$

where $\alpha_{ordinal}$ is ordinal alpha, k is the number of items, λ^2 is the squared factor pattern coefficient, h^2 is the communality where for a 1-factor model $h^2 = \lambda^2$, μ is uniqueness ($\mu^2 = 1 - h^2$). Theoretically, ordinal alpha equals the true reliability when the items are tau-equivalent and fit a one-factor model with equal factor pattern coefficients (Maydeu-Olivares et al., 2007; Zumbo et al., 2007).

Confidence Intervals

Authors should report a reliability coefficient even when the focus is not psychometric because it is a critical component to interpreting observed effects (Wilkinson and APA Task Force, 1999). Cronbach and Shavelson (2004)

suggested that researchers report a reliability coefficient for their specific study and not rely on published psychometrics, due to sampling and random errors. As with any point estimate, reliability coefficients are estimates of population parameters and tend to vary from sample to sample. This point is explicitly highlighted in reliability generalization studies that examine reliability fluctuation across studies (Vacha-Haase, Henson, & Caruso, 2002; Vacha-Haase & Thompson, 2011). Therefore, estimating the standard error of reliability coefficients with confidence intervals is critical.

A confidence interval (CI) provides information about the standard error of sample statistics and estimated range of values that most likely capture the true parameter (Cumming, 2012; Cumming & Finch, 2005). Larger standard errors and wider CIs are associated with low score reliability. The nominal width of a CI quantifies uncertainty and provides information regarding the precision of a point estimate (Cumming & Fidler, 2009). The standard error for the sample reliability coefficient is sensitive to sample size, the number of items, inter-item correlations, and homogeneity of variance (Duhachek & Iacobucci, 2004).

The sampling distribution for coefficient alpha follows a typical F distribution for large sample sizes (Feldt, 1965; Kristof, 1963). The sampling distribution of ordinal coefficient alpha can be conceived as having similar properties as coefficient alpha (B. Zumbo, personal communication, December 13, 2013). The reasoning is that ordinal alpha is simply coefficient alpha on the latent response distribution. The computation for ordinal alpha remains the same as that for coefficient alpha, except ordinal alpha is computed on the underlying latent continuous variable whereas coefficient alpha is computed on the measured continuous variable. From this point of view, the polychoric methods can be thought of as classes of complex transformations so that any property of coefficient alpha will likely hold for ordinal alpha. Therefore, the sampling distribution of ordinal alpha is highly likely to follow that of coefficient alpha. A natural question that follows is whether the methods for confidence intervals of coefficient alpha are appropriate to be used with ordinal alpha.

Consider the following four CI methods developed initially for coefficient alpha: Feldt (1965), Fisher (1950), Bonett (2002), and Hakstian and Whalen (1976). The differences between the methods are procedural transformations of sample coefficient alpha and the computation of variance. The formulae for the Feldt (1965) interval computations are

$$CI_{upper} = 1 - \left[(1 - \hat{\alpha}) \times F_{(\gamma/2), df_1, df_2} \right], \quad (6)$$

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

$$CI_{lower} = 1 - \left[(1 - \hat{\alpha}) \times F_{(1-\gamma/2), df_1, df_2} \right], \quad (7)$$

where $\hat{\alpha}$ is the sample reliability coefficient, γ is the specified level of significance, F represents the values at γ and $\gamma/2$ levels, n is the sample size with k items where $df_1 = (n - 1)$ and $df_2 = (n - 2)(k - 1)$.

Several methods transform sample coefficient alpha so confidence intervals may be computed within a normal theory framework. First, Fisher (1950) normalized a product moment correlation, such that

$$\text{Fisher's } z = \frac{1}{2} \ln \left(\frac{1 + |\hat{\alpha}|}{1 - |\hat{\alpha}|} \right), \quad (8)$$

where Fisher's z is the transformed parameter estimate and $\hat{\alpha}$ is the sample reliability coefficient. The z critical value (crit_z) is determined by the level of confidence where 95% has a value of 1.96. The standard error of z is estimated as

$$SE_z = \frac{1}{\sqrt{n-3}} \quad (9)$$

and applied to the lower and upper bounds of CI respectively as $\exp^{(2 \times (\text{Fisher}_z \pm \text{crit}_z SE_z)) - 1} / \exp^{(2 \times (\text{Fisher}_z \pm \text{crit}_z SE_z)) + 1}$. The transformed z statistic can be appropriately computed within a normal theory framework for a confidence interval and transformed back into the original units (Romano et al., 2011). Bonett (2002) extended Fisher's z (1950) as

$$\text{Bonett's } z = \ln(1 - |\hat{\alpha}|), \quad (10)$$

where Bonett's z statistic is closely normally distributed compared to that of sample coefficient alpha. The variances of Fisher's (1950) and Bonett (2002) z statistics are the same, yielding the lower and upper limits of CI respectively as $1 - \exp^{(\text{Bonett}_z \pm \text{crit}_z SE_z)}$. Hakstian and Whalen (1976) suggested another transformation of alpha, such that:

$$z = (1 - \hat{\alpha})^{\frac{1}{3}}, \quad (11)$$

where the resulting z statistic is normally distributed with a variance, σ^2 of

$$\sigma^2 = \frac{18k(n-1)(1-\hat{\alpha})^{\frac{2}{3}}}{(k-1)(9n-11)^2}. \quad (12)$$

The lower and upper limits of CI respectively are $\Pr\left[1 - c^{*3}(1-\hat{\alpha})^{\frac{1}{3}} \pm z_{1-(\alpha/2)}\sigma\right]^3$ where

$$c^{*3} = \frac{(9n-11)(k-1)}{9(n-1)(k-1)-2}. \quad (13)$$

Hakstian and Whalen (1976) argued their method is generally less biased than that of Fisher's (1950) z transformation, because they used the correction term $(1-\hat{\alpha})^{\frac{1}{3}}$ as an estimate of $(1-\alpha)^{\frac{1}{3}}$. There are notable performance differences among these CI methods as noted in the literature.

Recent Developments

Although several simulation studies have analyzed the performance of various methods for confidence intervals for coefficient alpha (Cui & Li, 2012; Iacobucci & Duhachek, 2003; Padilla et al., 2012; Romano et al., 2011), no known published study has analyzed confidence intervals for ordinal alpha. Romano et al. (2011) found negligible differences between the following eight confidence interval methods, with respect to bias, coverage, and precision for coefficient alpha computed for ordinal data: (a) Maydeu-Oliveres et al. (2007) asymptotic distribution free (ADF), (b) Bonett (2002), (c) Feldt (1965), (d) Fisher (1950), (e) Hakstian and Whalen (1976), (f) Duhachek and Iacobucci (2004), (g) Koning and Franses asymptotic (2003), and (h) Koning and Franses exact (2003) method. The findings suggest the ADF method was the least accurate for small sample sizes, and little was gained from departing from the Fisher approach. This finding is especially noteworthy because many other simulation studies suggested that ADF method outperformed other normal theory approaches, and that the Fisher approach yielded low coverage rates (Duhachek & Iacobucci, 2004; Hakstian & Whalen, 1976; Maydeu-Olivares et al., 2007). Romano et al. (2011) provided evidence that sophisticated CI methodology does not necessarily yield better

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

performance. However, Romano et al. (2011) computed alpha coefficient for ordinal data based on Pearson and not polychoric correlations.

Romano et al.'s (2011) findings are important because advancements of ADF methods were considered the most robust in skewed distributions and small sample sizes. van Zyl et al. (2000) derived an asymptotic (i.e. large sample) distribution for sample coefficient alpha, only assuming a multivariate normal distribution and positive-definite matrix (Maydeu-Olivares, et al., 2007). Although van Zyl et al.'s (2000) intervals have been shown to yield the most narrow intervals, they often have undercoverage (Cui & Li, 2012).

Duhachek and Iacobucci (2004) extended van Zyl et al.'s (2000) method and presented statistics for coefficient alpha's standard error and computed an ADF-based CI. They found ADF intervals repeatedly outperformed other normal theory based intervals, including Feldt (1965) and Hakstian and Whalen (1976). This finding was consistent across all study conditions, but their study was not generalizable to Likert-type data. Maydeu-Olivares et al. (2007) found that the empirical coverage rate of the ADF intervals for coefficient alpha outperforms that of normal theory intervals, regardless of observed skewness and kurtosis of item distributions (Cui & Li, 2012; Romano et al., 2011). These results are significant because researchers are no longer bound by normality assumptions (i.e. normal theory) that were often violated when analyzing Likert-type data. Padilla et al. (2012) found that the normal theory bootstrap method had the most acceptable coverage rate followed by Bonett, and normal theory for non-normal data. Fisher method yielded unacceptably high variability, except when the scale had more than 15 items.

In sum, there is a need to evaluate the performance of the confidence intervals of ordinal alpha because Likert-type data is very commonly used in educational and behavioral research. Therefore, the present study investigated the coverage rates, widths, and biases of the four types of CIs (Bonnet, Feldt, Fisher, and Hakstian Whalen), and the accuracy of the ordinal alpha point estimates. Sample size, the number of items, skewness of the responses, and population alpha were varied.

Method

The program code was written using R (Version 3.0.2) using the R Studio interface (Version 0.98.976). The code was executed in a Windows-based environment (Version 8). Based on Maydeu-Olivares et al. (2007) and Hakstian

and Whalen (1976) we generated the data from the factor analytic classical test theory model, assuming the parallel items model as follows:

- a) For a given condition, generate a population of 1 million subjects by k number of items, with p population alphas, c response categories, and s skewness.
- b) For a given sample size n , generate a $n \times k$ theoretical ability matrix θ^* such that $\theta^* \sim N(0,1)$.
- c) Generate a $n \times k$ random error matrix U such that $U \sim MVN(0,\sigma)$ where

$$\sigma = \begin{bmatrix} .25 & 0 & \dots & 0 \\ 0 & .25 & \dots & \dots \\ \dots & \dots & .25 & 0 \\ 0 & \dots & 0 & .25 \end{bmatrix}.$$

- d) Calculate the $n \times k$ matrix X^* such that $x_{ik} = \lambda\theta_{ik} + u_{ik}$ where λ values are specified below.
- e) Categorize the scores in the item response distributions in X^* by applying rigid thresholds, τ (Muthén & Kaplan, 1985; Zumbo et al., 2007) to generate a skewed and symmetric distribution. The exact threshold values are provided in Table 1.

Design Factors

Population alphas (a). Three population alphas were specified at .6, .8, and .9 as used in previous simulation studies (Cui & Li, 2012; Padilla et al., 2012; Romano et al., 2011; Zumbo et al., 2007). Factor pattern coefficients values (λ s) were based on Zumbo et al. (2007) with values of .311, .471, and .625 for population ordinal coefficient alphas of .6, .8, and .9, respectively.

Sample size (n). The design conditions included four levels of sample size (20, 50, 100, 200). The sample sizes were selected based on previous studies and represent sample sizes often noted in applied research (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Romano et al., 2011; Yuan & Bentler, 2002; Zumbo et al., 2007). While

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

large sample sizes are always desirable, they are not always realistic. Duhachek and Iacobucci (2004) indicated that sample sizes beyond 200 have diminishing returns for coefficient alpha, given a sufficient number of items and strong inter-item correlations. Therefore, they were not simulated in the present study. Similarly, we considered sample sizes as small as 20 because this is not an uncommon sample size in educational research and has therefore been included as a data condition in other similar simulation studies (e.g. Natesan & Thompson, 2007). Moreover, considering a sample size as low as 20 helps the researcher understand a possible lower bound of sample size necessary for estimating ordinal alpha.

Number of items (k). The number of items chosen were $k = 5, 10, 25,$ and 40. Previous studies have simulated between two and 40 items, which also reflects the test widths of interest to applied researchers (Cui & Li, 2012; Duhachek & Iacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Romano et al., 2011; Yuan & Bentler, 2002; Zumbo et al., 2007). It is not uncommon to consider 5 and 10 Likert-type items per factor in simulation studies (e.g. Ankermann & Stone, 1992; Kieftenbeld & Natesan, 2012; Reise & Yu, 1990). Forty items were considered as the upper bound of test length.

Skewness (s). Two types of observed item response distributions were selected: $s = 0, -1.217$. These values were selected to demonstrate the impact of symmetry on precision of confidence intervals for ordinal coefficient alpha (Zumbo et al., 2007). Threshold values are used to categorically score the individual item's value computed in steps (a) through (e) described above. The following thresholds for the two item response distributions and relative response categories are based on the works of Zumbo et al. (2007) and specified in Table 1.

Response categories (C). Two scales of response categories (C) were selected: the five-point and seven-point scales. Duhachek and Iacobucci (2004) demonstrated that confidence interval performance does not improve beyond seven response categories. Therefore, simulating more than seven response categories was not deemed necessary. A five-point Likert scale is commonly used in behavioral research. The resulting design is a fully-crossed $2(s) \times 2(C) \times 3(\alpha) \times 4(n) \times 4(k)$ factorial design with 192 conditions.

Table 1. Likert Scale Thresholds

<i>Skewness</i> y_j	Five-point Scale		Seven-point Scale	
	0	-1.217	0	-1.217
1	$y_j^* \leq -1.8$	$y_j^* > 1.8$	$y_j^* \leq -2.14$	$y_j^* > 2.4$
2	$-1.8 < y_j^* \leq -0.6$	$1.8 \geq y_j^* > 1.34$	$-2.14 < y_j^* \leq -1.29$	$2.4 \geq y_j^* > 1.95$
3	$-0.6 < y_j^* \leq 0.6$	$1.34 \geq y_j^* > 0.77$	$-1.29 < y_j^* \leq -0.43$	$1.95 \geq y_j^* > 1.42$
4	$0.6 < y_j^* \leq 1.8$	$0.77 \geq y_j^* > 0.05$	$-0.43 < y_j^* \leq 0.43$	$1.42 \geq y_j^* > 0.99$
5	$1.8 < y_j^*$	$0.05 \geq y_j^*$	$-0.43 < y_j^* \leq 0.43$	$0.99 \geq y_j^* > 0.47$
6	NA	NA	$0.43 < y_j^* \leq 2.14$	$0.47 \geq y_j^* > -0.2$
7	NA	NA	$2.14 < y_j^*$	$-0.2 \geq y_j^*$

Diagnostics

Coverage rates were computed as,

$$\text{coverage rate} = (A/B) \times 100\%, \quad (14)$$

where A is the frequency of intervals which contain the true population parameter, and B is the total number of intervals. Coverage rate should have a value close to the nominal level but are not a sufficient diagnostic, particularly when the skewness of the sampling distribution is not provided or unknown (Jennings, 1987; Schall, 2012; Zhang, Gutiérrez Rojas, & Cuervo, 2010). In addition to coverage rate, positive and negative bias of intervals that do not contain the true value must be reported. Among the $(B - A)$ intervals that did not contain the true value, the number of intervals which were below and above the true value when expressed as percentage of the total number of intervals indicate negative and positive bias of CI, respectively. An imbalance in these biases indicate possible systematic bias in the estimation (Natesan, 2015). An unbiased interval is equally likely to be above or below the true value. Therefore an unbiased CI estimate would have roughly equal number of negatively and positively biased intervals. CI width is the difference between the upper and lower limits of the CI. Following Zhang et al. (2010), mean and variance of CI widths were computed. A highly variable CI width indicates poor precision for the interval estimate method.

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Precision of point estimates. Bias and RMSE of ordinal alpha were computed. Bias is the difference between the true population parameter value and the sample estimate. RMSE across N replications is computed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}, \quad (15)$$

where μ is the true population parameter value, x_i is the estimate of ordinal alpha in the i^{th} replication. To determine whether a confidence interval is unacceptably wide, the empirical standard error (SE) of ordinal alpha was computed. The empirical SE is the standard deviation of all sample ordinal alpha estimates for a given condition.

Data Analysis

Following the simulation, η^2 effect sizes were examined for separate ANOVAs to understand the variance in the simulation diagnostics explained by the data conditions. The independent variables were population alphas, sample size, number of items, skewness, and response categories. The dependent variables were coverage rates, CI width, variance of CI width, RMSE, and bias of ordinal alpha. Both main effects and all higher order interactions were examined. Following Cohen (1988), 1%, 6%, and 14% were considered small, medium, and large effect sizes for η^2 . Only large main and higher order interaction effects are interpreted and discussed.

Results

Practical Computation Issues

To minimize the standard error of the simulation, 1,000 samples were drawn for each condition (Fan & SAS Institute, 2002; Wang & Thompson, 2007). The Appendix includes a list of 25 conditions (of the 192 conditions) that did not execute due to repeated crashing. The error message stated that a “not positive-definite matrix” was computed which caused computations to stop. This error occurred when there was a large number of items (e.g. 25 or 40 items) with a small sample size (e.g. $n = 20$). The rigid thresholds set for the 5-point and 7-point Likert scales removed important differences between the available response options (1, 2, ... 5 or 1, 2, ... 7). Ultimately, there simply was not enough

sampling variability generated across each repetition and the variables became constant (e.g. all responses were scored “3”). When no variability was generated either across items or subjects, the covariance and standard deviation are essentially zero. When this occurs, estimation stops because one cannot divide by a standard deviation of zero to compute polychoric correlation. The issues related to the lack of variance generated seem to be an artifact of restricted range with the ordinal data. The resulting dataset contained 167,000 replications (167 workable conditions $\times$ 1,000 samples each). The total time elapsed was approximately 691 computing hours. The simulations were executed on a Dell Precision T3600 Intel (R) Xeon (R) CPU E5-1620 3.60 GHz Windows 8 machine.

Coverage rate and CI bias.

Overall, coverage rates were much lower than the 95% nominal rate as seen in Figure 1, ranging from 46% (Feldt) to 62% (Fisher). Feldt method had the lowest coverage rate due to the confounding impact of several independent variables summarized in Table 2. ANOVA results show that skewness explained most of the variance in mean coverage rates (23.651% to 62.915%) except for the Feldt method. Interaction effects have a dominating presence, especially for the Feldt method where $\eta^2 = 64.968\%$.

Figure 1. Coverage rates by confidence interval method across $N = 167,000$ replications

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Table 2. η^2 (%) by confidence interval method for coverage rates

Independent variable	<i>Fisher</i>	<i>Feldt</i>	<i>Bonett</i>	<i>HW</i>
population alpha (α)	0.422	0.252	0.110	0.381
sample size (n)	4.122	7.453	4.692	7.085
items (k)	1.934	2.722	5.312	6.116
skewness (s)	62.915	23.651	55.108	46.769
response categories (C)	2.157	0.322	0.997	1.034
Interactions ^a	27.197	64.968	32.756	37.590
$\alpha \times s$	9.375	13.129	13.165	16.919
$\alpha \times n$	4.764	6.705	6.307	5.371
$\alpha \times n \times k$		6.430		
$\alpha \times n \times k \times C$		8.979		

Note. ^a Interactions includes all possible 2nd, 3rd, 4th, and 5th order interactions; only η^2 (%) values greater than 4% are reported for the interaction terms

The largest two-way interaction effect, population alpha by skewness, is shown in [Figure 2](#). Two-way mean interaction plots of population alpha by skewness effects. As population alpha increased from .6 to .9, the estimated marginal means (EMMs) of captured coverage rates increased for skewed distributions. Coverage rates increased due to the joint influence of population alpha levels and skewness. Coverage rates were higher in skewed data except for the Feldt method, when population alpha levels were .6. The CIs that did not contain the true value more often underestimated ordinal alpha for all methods. That is, negatively biased intervals (35-38%) occurred more frequently than positively biased intervals (4-17%) as shown in [Figure 2](#).

Figure 2. Two-way mean interaction plots of population alpha by skewness effects on coverage rates.

CI widths. The boxplot shown in Figure 3 depicts the interquartile ranges of the 95% CI width for the four confidence interval methods. The Fisher confidence intervals consistently yielded the narrowest intervals, while Bonett intervals were the widest across all conditions. Table 3 shows that all intervals became narrower with increase in sample size and population alpha with one exception. The exception occurred with the Feldt interval when $n = 200$ and the population alpha increased from .8 to .9.

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Figure 3. Boxplot of 95% CI widths for four estimation methods

Table 3. Mean confidence interval width (SD) at study condition level

Method	Population alpha	Sample Size			
		20	50	100	200
Fisher	0.6	.654 (.073)	.286 (.032)	.184 (.017)	.133 (.010)
	0.8	.573 (.082)	.223 (.029)	.140 (.015)	.091 (.008)
	0.9	.510 (.077)	.184 (.024)	.108 (.014)	.060 (.007)
Feldt	0.6	.673 (.125)	.254 (.041)	.153 (.019)	.100 (.001)
	0.8	.524 (.114)	.178 (.029)	.103 (.014)	.071 (.008)
	0.9	.499 (.107)	.162 (.026)	.094 (.014)	.073 (.008)
Bonett	0.6	.742 (.140)	.281 (.045)	.172 (.022)	.126 (.013)
	0.8	.610 (.135)	.202 (.033)	.126 (.017)	.080 (.009)
	0.9	.516 (.112)	.165 (.027)	.093 (.013)	.050 (.006)
HW	0.6	.665 (.125)	.270 (.043)	.168 (.021)	.125 (.013)
	0.8	.547 (.121)	.194 (.031)	.123 (.017)	.079 (.009)
	0.9	.463 (.100)	.158 (.026)	.091 (.013)	.050 (.006)

Shown in Figures 4 through 6, the CIs became narrower with increase in both sample size and the number of items simultaneously for all methods. The

intervals became quite narrow when sample size = 200 and the number of items = 40. There are no striking visual differences in the confidence interval widths between the Fisher, Feldt, Bonett, and HW methods across various levels of population alpha because the patterns are similar for all methods.

Figure 4. Mean confidence limits for population alpha = .6. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient alpha.

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Figure 5. Mean confidence limits for population alpha = .8. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient alpha.

Figure 6. Mean confidence limits for population $\alpha = .9$. Dashed line references population parameter. Bottom marker is the mean lower limit and top marker is the mean upper limit. Middle marker is the mean sample ordinal coefficient α .

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Figure 7. Interaction of sample size by skewness on CI width

As seen in Table 4, sample size explained the most variance in CI width across all methods. Specifically, the CI widths decreased with increase in sample size. Of the interaction effects, the largest amount of the interaction effects were explained by the sample size by skewness interaction. The sample size by skewness interactions ranged from 2.55% (Fisher) to 5.03% (Bonett). While the η^2 values for both interactions may be considered “small,” (i.e., $< 1\%$, Cohen, 1988), the implications are meaningful. The CI widths were consistently smaller across all methods when sample size increased to 200 and skewness = 0 as shown in Figure 7.

Table 4. η^2 (%) by confidence interval method for CI width

Independent variable	<i>Fisher</i>	<i>Feldt</i>	<i>Bonett</i>	<i>HW</i>
population alpha (α)	3.882	2.516	3.852	4.290
sample size (n)	67.254	55.737	57.940	55.062
items (k)	13.599	11.795	14.709	16.566
skewness (s)	6.053	7.914	6.527	7.068
Interactions ^a	5.280	15.487	10.148	10.010

Note. ^a Interactions includes all possible interactions. Response categories had η^2 less than 1%

Variance of CI widths. Effect sizes summaries for variance CI widths for all CI methods are shown in Table 5. Sample size by number of items had the largest effect across the four confidence interval methods. The η^2 values ranged from 2.140% (Fisher) to 10.394% (HW) with the mean plots provided below in Figure 8. All four methods followed the same pattern with variance of CI widths sharply decreasing as both sample size and the number of items increased. In summary, the joint influence of the number of items and sample size impacted the mean variance of CI width across all methods.

Table 5. η^2 (%) by confidence interval method for variance of CI width

Independent variable	<i>Fisher</i>	<i>Feldt</i>	<i>Bonett</i>	<i>HW</i>
population alpha (α)	0.144	0.500	0.738	0.829
sample size (n)	83.358	63.538	67.264	66.409
items (k)	2.821	5.989	7.285	7.915
skewness (s)	0.254	3.011	2.257	2.492
Interactions ^a	12.780	26.866	22.378	22.276

Note. ^a Interactions includes all possible 2nd, 3rd, 4th, and 5th order interactions; number of categories had $\eta^2 < 1\%$

Point estimates. Given the large η^2 values of sample size (43.377%) and items (34.102%), further post-hoc analyses were conducted to see which levels of the independent contributed the most to the variance of RMSE as seen in Table 6. The RMSEs decreased from .083 to .022 as sample size increased from 20 to 200. As the number of items increased from 5 to 40, the RMSE values decreased from .068 to .011.

Overall, sample ordinal alpha was negatively biased ($M = -.054$, $SD = .103$, $N = 167,000$) ranging from $-.69$ to $.328$. The distribution of bias was negatively

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

skewed (-1.10 , $SE = .006$) with a leptokurtic shape (1.940 , $SE = .012$). Skewness levels (0 , -1.217) had the largest impact on the bias of ordinal coefficient alpha. Sample size $\times$ skewness explained 5.36% of the variance in bias. The rest of the interactions explained less than 4% of the variance. Negative skewness resulted in a less biased estimate ($EMM = .013$, $SE < .001$) compared to no skewness ($EMM = -.122$, $SE < .001$). These results support the use of ordinal coefficient alpha when analyzing Likert-type or ordinal data because less bias is present when data are skewed. In summary, the precision of ordinal coefficient alpha, in terms of RMSE, is best explained by the main effects of sample size and the number of items. Bias is best explained by the main effect of skewness and a combination of small interaction effects.

Figure 8. Interaction of samples size by number of items on CI width

Table 6. η^2 (%) for RMSE and bias

Source	RMSE	Bias
population alpha (α)	5.295	2.811
sample size (n)	42.377	10.376
items (k)	34.102	1.600
skewness (s)	10.094	41.754
response categories (C)	0.036	2.166
Interactions ^a	8.084	19.085

Conclusion

The aim in this study was to evaluate the performance of Feldt, Fisher, Bonett, and HW confidence interval methods for ordinal coefficient alpha. The simulation findings are only applicable to study designs where the sample sizes range from 20 to 200, the number of items range from five to 40, scores are categorized into symmetric and skewed item response distributions, with five or seven response categories. None of the CI methods suggested for coefficient alpha have adequate coverage for ordinal alpha. Skewness had the largest impact on coverage rates. Mean coverage rates were 46% - 62%, low, and unacceptably low for all methods. This could lead to high type-I error rates. Moreover, for all methods the CIs that did not contain the true values were more negatively biased. Clearly these findings show the need for a new method that specifically formulates CI for ordinal alpha.

CI widths were statistically significantly different across Feldt, Fisher, Bonett, and HW methods ($p < .05$). CI widths became narrower as population alphas increased and sample size increased. There are small, but notable differences observed with CI width between methods. The Feldt method is the only CI method that did not use any transformation of sample ordinal coefficient alpha, and was therefore, impacted differently than Fisher, Bonett, and HW. The Feldt CI width is determined as a function of the degrees of freedom based on sample size and number of items, and was therefore heavily impacted by the interactions of these conditions. The Fisher, Bonett, and HW methods apply logarithmic transformations of sample ordinal alpha and were therefore more easily explained by varying sample size and the number of items.

Sample size and number of items best explained the precision of ordinal alpha. Interestingly, the number of response categories is a strong predictor of coefficient alpha, but not necessarily for ordinal coefficient alpha (Zumbo et al.,

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

2007). EMMs of RMSE were statistically significant across all levels of sample size and number of items. The practical implications suggest keeping an instrument, with Likert-type data, between 10-25 items, while striving for at least 50 participants. However, researchers should not use any of the CI formulae tested in the present study to compute confidence interval for ordinal alpha.

Bias was best explained by skewness, and sample size by skewness interaction effect. Overall, bias is persistently negative across all design levels except for skewed data. Bias approached zero when $n = 200$. This shows that regardless of the method, when the estimate is more biased the coverage rate will be lower. Again, very little confidence should be placed on confidence intervals methods for ordinal alpha.

As with any simulation study, the results are limited to the conditions specified. The conditions were justified with previous research to portray scenarios in applied research. The conclusions hold for the study conditions specified; therefore, a number of opportunities exist to extend the current research. First, a confidence interval method specifically for ordinal alpha which improves coverage rates closer to the nominal rate needs to be developed. Additionally, the contiguous points between 10 and 25 items may be explored to determine the optimal point of precision of ordinal alpha for both RMSE and bias.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Allen, M. J., & Yen, W. M. (2002). *Classical true-score theory*. Prospect Heights, IL: Waveland Press.
- Ankemann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, 27(4), 335-340. doi: [10.3102/10769986027004335](https://doi.org/10.3102/10769986027004335)
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997-1003. doi: [10.1037/0003-066x.49.12.997](https://doi.org/10.1037/0003-066x.49.12.997)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

- Crocker, L., & Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. doi: [10.1007/bf02310555](https://doi.org/10.1007/bf02310555)
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418. doi: [10.1177/0013164404266386](https://doi.org/10.1177/0013164404266386)
- Cui, Y., & Li, J. (2012). Evaluating the performance of different procedures for constructing confidence intervals for coefficient alpha: A simulation study. *British Journal of Mathematical and Statistical Psychology*, 65(3), 467-498. doi: [10.1111/j.2044-8317.2012.02038.x](https://doi.org/10.1111/j.2044-8317.2012.02038.x)
- Cumming, G. (2012). *Understanding the new statistics*. New York: NY: Taylor & Francis Group.
- Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology*, 217(1), 15-26. doi: [10.1027/0044-3409.217.1.15](https://doi.org/10.1027/0044-3409.217.1.15)
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170-180. doi: [10.1037/0003-066X.60.2.170](https://doi.org/10.1037/0003-066X.60.2.170)
- Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, 89(5), 792-808. doi: [10.1037/0021-9010.89.5.792](https://doi.org/10.1037/0021-9010.89.5.792)
- Fan, X., & SAS Institute. (2002). *SAS® for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute.
- Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes: Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, 61(4), 517-531. doi: [10.1177/0013164401614001](https://doi.org/10.1177/0013164401614001)
- Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson coefficient twenty. *Psychometrika*, 30(3), 357-370. doi: [10.1007/bf02289499](https://doi.org/10.1007/bf02289499)
- Finch, S., Cumming, G., & Thomason, N. (2001). Colloquium on effect Sizes: The roles of editors, textbook authors, and the publication manual: Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement*, 61(2), 181-210. doi: [10.1177/0013164401612001](https://doi.org/10.1177/0013164401612001)

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

- Fisher, R. A. (1950). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466-491. doi: [10.1037/1082-989x.9.4.466](https://doi.org/10.1037/1082-989x.9.4.466)
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, 17(3). Available online: <http://pareonline.net/getvn.asp?v=17&n=3>
- Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, 41(2), 219-231. doi: [10.1007/bf02291840](https://doi.org/10.1007/bf02291840)
- Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, 13(4), 478-487. doi: [10.1207/s15327663jcp1304_14](https://doi.org/10.1207/s15327663jcp1304_14)
- Jennings, D. E. (1987). How do we judge confidence-interval adequacy? *The American Statistician*, 41(4), 335-337. doi: [10.2307/2684768](https://doi.org/10.2307/2684768)
- Jöreskog, K.G., 1990. New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity*, 24(4), 387-404. doi: [10.1007/bf00152012](https://doi.org/10.1007/bf00152012)
- Kieftenbeld, V. & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, 36(5), 399-419. doi: [10.1177/0146621612446170](https://doi.org/10.1177/0146621612446170)
- Koning, A. J., & Franses, P. H. (2003). *Confidence intervals for Cronbach's coefficient alpha values*. (No. ERS-2003-041-MKT). *ERIM Report Series Research In Management*. Retrieved from <http://hdl.handle.net/1765/431>
- Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28(3), 221-238. doi: [10.1007/bf02289571](https://doi.org/10.1007/bf02289571)
- Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, 12(2), 157-176. doi: [10.1037/1082-989x.12.2.157](https://doi.org/10.1037/1082-989x.12.2.157)
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of*

Mathematical and Statistical Psychology, 38(2), 171-189. doi: [10.1111/j.2044-8317.1985.tb00832.x](https://doi.org/10.1111/j.2044-8317.1985.tb00832.x)

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology*, 6, 1599. doi: [10.3389/fpsyg.2015.01599](https://doi.org/10.3389/fpsyg.2015.01599)

Natesan, P. & Thompson, B. (2007). Extending improvement-over-chance I-index effect size simulation studies to cover some small-sample cases. *Educational and Psychological Measurement*, 67(1), 59-72. doi: [10.1177/0013164406292028](https://doi.org/10.1177/0013164406292028)

Novick, M. R., & Lewis, M. R. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1), 1-13. doi: [10.1007/bf02289400](https://doi.org/10.1007/bf02289400)

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement*, 36(5), 331-348. doi: [10.1177/0146621612445470](https://doi.org/10.1177/0146621612445470)

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27(2), 133-144. doi: [10.1111/j.1745-3984.1990.tb00738.x](https://doi.org/10.1111/j.1745-3984.1990.tb00738.x)

Romano, J. L., Kromrey, J. D. , Owens, C. M. , & Scott, H. M. (2011). Confidence interval methods for coefficient alpha on the basis of discrete, ordinal response items: Which one, if any, is the best? *Journal of Experimental Education*, 79(4), 382-403. doi: [10.1080/00220973.2010.510859](https://doi.org/10.1080/00220973.2010.510859)

Schall, R. (2012). The empirical coverage of confidence intervals: Point estimates and confidence intervals for confidence levels. *Biometrical Journal*, 54(4), 537-551. doi: [10.1002/bimj.201100134](https://doi.org/10.1002/bimj.201100134)

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353. doi: [10.1037/1040-3590.8.4.350](https://doi.org/10.1037/1040-3590.8.4.350)

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120. doi: [10.1007/s11336-008-9101-0](https://doi.org/10.1007/s11336-008-9101-0)

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103. doi: [10.1207/s15327752jpa8001_18](https://doi.org/10.1207/s15327752jpa8001_18)

Thompson, B. (2006a). *Foundations of behavioral statistics*. New York, NY: The Guilford Press.

EVALUATION OF CI FOR ORDINAL COEFFICIENT ALPHA

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583-603). New Jersey: Lawrence Erlbaum Associates.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, 62(4), 562-569. doi: [10.1177/0013164402062004002](https://doi.org/10.1177/0013164402062004002)

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, 44(3), 159-168. doi: [10.1177/0748175611409845](https://doi.org/10.1177/0748175611409845)

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, 65(3), 271-280. doi: [10.1007/bf02296146](https://doi.org/10.1007/bf02296146)

Wang, Z., & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *Journal of Experimental Education*, 75(2), 109-125. doi: [10.3200/jexe.75.2.109-125](https://doi.org/10.3200/jexe.75.2.109-125)

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594-604. doi: [10.1037/0003-066x.54.8.594](https://doi.org/10.1037/0003-066x.54.8.594)

Yuan, K., & Bentler, P. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, 67(2), 251-259. doi: [10.1007/bf02294845](https://doi.org/10.1007/bf02294845)

Zhang, H., Gutiérrez Rojas, H. A., & Cuervo, E. C. (2010). Confidence and credibility intervals for the difference of two proportions. *Revista Colombiana De Estadística*, 33(1), 63-88.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods* 6, 21-29. doi: [10.22237/jmasm/1177992180](https://doi.org/10.22237/jmasm/1177992180)

Appendix A: Non-Executable Conditions

<i>Obs num</i>	<i>SimID</i>	<i>Population ordinal coeff alpha</i>	<i>Factor loading</i>	<i>k items</i>	<i>c response categories</i>	<i>n sample size</i>	<i>Skewness</i>
1	3	0.9	0.625	25	5	20	0.000
2	4	0.6	0.311	40	5	20	0.000
3	7	0.6	0.311	25	7	20	0.000
4	8	0.8	0.471	40	7	20	0.000
5	12	0.9	0.625	40	5	50	0.000
6	35	0.8	0.471	25	5	20	-1.217
7	36	0.9	0.625	40	5	20	-1.217
8	39	0.9	0.625	25	7	20	-1.217
9	40	0.6	0.311	40	7	20	-1.217
10	67	0.6	0.311	25	5	20	0.000
11	68	0.8	0.471	40	5	20	0.000
12	71	0.8	0.471	25	7	20	0.000
13	72	0.9	0.625	40	7	20	0.000
14	99	0.9	0.625	25	5	20	0.000
15	100	0.6	0.311	40	5	20	-1.217
16	103	0.6	0.311	25	7	20	-1.217
17	104	0.8	0.471	40	7	20	-1.217
18	131	0.8	0.471	25	5	20	0.000
19	132	0.9	0.625	40	5	20	0.000
20	135	0.9	0.625	25	7	20	0.000
21	136	0.6	0.311	40	7	20	0.000
22	163	0.6	0.311	25	5	20	0.000
23	164	0.8	0.471	40	5	20	-1.217
24	167	0.8	0.471	25	7	20	-1.217
25	168	0.9	0.625	40	7	20	-1.217

Effectively Comparing Differences in Proportions

Lonnie Turpin
McNeese State University
Lake Charles, LA

A single framework of developing and implementing tests about proportions is outlined. It avoids some of the pitfalls of methods commonly put forward in an introductory data analysis course.

Keywords: Binary data analysis, proportions, significance, inference

Introduction

Proportions derived from binary variables are simple predominantly due to the nature of the variables involved. Because of this, the logic behind the methods are able to be grasped, as opposed to resorting to memorizing formulas. However, confusion arises when making a connection within the equations between a Bernoulli random variable X and the associated estimator $\hat{p}$ of the sample proportion p . A way to mitigate this confusion will be shown where only a basic knowledge of descriptive/inferential statistics and linear combinations are required. With all necessary formulations included, this study is essentially self-contained and aimed at analysts (practitioners and teachers focusing on applications who need a quick guide for analyzing proportions).

A subject can represent any object of analysis (people, products, etc.) and the method refers to the two examples used to illustrate proportions, not the statistical technique used in analysis. A single framework of developing and implementing tests about proportions will be outlined, which avoids some of the pitfalls of methods commonly put forward in introductory classes. The methods will entail using the simple two-way probability table to easily calculate confidence intervals and Z-scores while accounting for the correlation between the proportions. The key contributions are: 1) to make use of this simple two-way table as an easy way to

Dr. Lonnie Turpin is an Assistant Professor of Operations Management and Business Statistics. Email him at: lturpin@mcneese.edu.

compute the correlation (which is often ignored in intro level treatments); and 2) to motivate the methods in a way that makes use of the student's intuition about how sample means relate to population means. This study was motivated by experience teaching graduate-level applied business statistics courses. Common points of confusion are highlighted as remarks. For theoretical expositions on the topics presented, Brownlee (1965) and Bickel and Doksum (2015) are recommended.

Comparing Proportions

Begin with an example that involves comparing the effectiveness of two different methods of testing a product in order to provide a single method of comparing two proportions. An illustrative example concerns comparing the effectiveness of two methods. The methods are denoted as $M1$ (a chosen method 1) and $M2$ (a chosen method 2 that is different from method 1). It is a loose assumption that there exists an initial method of comparison ($M0$) with a known statistical measure of effectiveness. However, it is not necessary for the comparison of the test methods ($M1$ and $M2$), and this is therefore omitted for the remainder of the paper. Both of these methods will be given to a group of n test subjects. For a classic example of comparing two proportions from an identical survey, see Scott and Seber (1983) and Wild and Seber (1993).

Procedure

Assumption 1: To control for the potential individual-specific factors, each subject in the test group is given both methods. Practical outlets for this assumption are in clinical trials where each subject receives both the treatment and the control (Senn, 2002).

For each subject i , define $X_i^M = 1$ if method M is successful and 0 otherwise. Prior to beginning the experiment, it is important to note that X_i^M is a random variable, since we do not know in advance whether the method will work for any given subject.

Assumption 2: Outcomes for different subjects are independent and identically distributed (i.i.d.).

Given this assumption, and based on the nature of the variables, we let the distribution of each variable expressed in terms of p as

EFFECTIVELY COMPARING DIFFERENCES IN PROPORTIONS

$$X_i^M \sim \text{Bernoulli}(p^M) \text{ i.i.d.} \quad (1)$$

where p^M is the familiar probability $\Pr(X_i^M) = 1$ representing the success of a given method for the i^{th} subject. Let $\hat{p}^M$ be the fraction of subjects for which method M is successful. For dummy variables, let

$$\hat{p}^M = \frac{1}{n} \sum_{i=1}^n X_i^M \quad (2)$$

be the estimator for p^M , representing the sample proportion of ones contained in the data set.

Assumption 3: The sample size n is large enough to follow the Central Limit Theorem, thus define the sampling distribution of $\hat{p}^M$ as

$$\hat{p}^M \sim N\left(p^M, \frac{p^M(1-p^M)}{n}\right) \quad (3)$$

For effective techniques in working with smaller data sets (especially in teaching the logic behind these techniques), see Agresti and Caffo (2000).

Notice that in Assumption 3, the sampling distribution of $\hat{p}^M$ depends on the unknown parameter p^M . With a large n , by equation (2), rewrite equation (3) in the form

$$\hat{p}^M \approx N\left(\hat{p}^M, \frac{\hat{p}^M(1-\hat{p}^M)}{n}\right) \quad (4)$$

Replace p with $\hat{p}$ in the equations to follow with the understanding that $\hat{p} \approx p$ by Assumption 3 and equation (4). Then, define the mean and variance of X_i^M in terms of p^M and $\hat{p}^M$ as

$$\begin{aligned} E[X_i^M] &= p^M \\ &\approx \hat{p}^M \end{aligned} \quad (5)$$

and

$$\begin{aligned} V[X_i^M] &= p^M (1 - p^M) \\ &\approx \hat{p}^M (1 - \hat{p}^M) \end{aligned} \quad (6)$$

Now, figure out which of the two methods is more effective. That is, the goal is to find and explain the difference $p^{M1} - p^{M2}$. To do this, think about the two methods together. Considering the potential unique nature of the subjects highlighted by Assumption 1, the following assumption is introduced:

Assumption 4: Although it was assumed the outcomes are i.i.d. across different subjects, do not assume the two outcomes X_i^{M1} and X_i^{M2} are independent for the same subject. Because there is no assumption that the two methods are independent, the question arises how strongly they are related. Use the familiar covariance formula

$$\begin{aligned} \text{Cov}[X_i^{M1}, X_i^{M2}] \\ = \sum_{X_i^{M1}, X_i^{M2} \in \{0,1\}} \text{Pr}(X_i^{M1}, X_i^{M2}) (X_i^{M1} - E[X_i^{M1}]) (X_i^{M2} - E[X_i^{M2}]) \end{aligned} \quad (7)$$

where the weights $\text{Pr}(X_i^{M1}, X_i^{M2})$ are the joint probabilities easily derived using the two-way Table 1. In Appendix B, the numerical example is used to show a connection between covariance and the assumption of independence via Bayes' Rule.

Table 1. Two-way joint probability table

		X_i^{M1}		$\text{Pr}(X_i^{M2})$
		0	1	
X_i^{M2}	0	$\text{Pr}(0, 0)$	$\text{Pr}(1, 0)$	$1 - p^{M2}$
	1	$\text{Pr}(0, 1)$	$\text{Pr}(1, 1)$	p^{M2}
$\text{Pr}(X_i^{M1})$		$1 - p^{M1}$	p^{M1}	

EFFECTIVELY COMPARING DIFFERENCES IN PROPORTIONS

Table 2. Two-way joint probability table with $\hat{p}^M$ estimators

		X_i^{M1}		
		0	1	$\Pr(X_i^{M2})$
X_i^{M2}	0	$\Pr(0,0)$	$\Pr(1,0)$	$1 - \hat{p}^{M2}$
	1	$\Pr(0,1)$	$\Pr(1,1)$	$\hat{p}^{M2}$
$\Pr(X_i^{M1})$		$1 - \hat{p}^{M1}$	$\hat{p}^{M1}$	

Replace the unknown p^{M1} , p^{M2} , and $\Pr(X_i^{M1}, X_i^{M2})$ in [Table 1](#) with the estimators $\hat{p}^{M1}$, $\hat{p}^{M2}$, and $\Pr(X_i^{M1}, X_i^{M2})$ in the modified [Table 2](#).

As with equation (5) and equation (6), rewrite equation (7) in terms of p as

$$\begin{aligned}
 & \text{Cov}[X_i^{M1}, X_i^{M2}] \\
 &= \sum_{X_i^{M1}, X_i^{M2} \in \{0,1\}} \Pr(X_i^{M1}, X_i^{M2}) (X_i^{M1} - E[X_i^{M1}]) (X_i^{M2} - E[X_i^{M2}]) \\
 &= \sum_{X_i^{M1}, X_i^{M2} \in \{0,1\}} \Pr(X_i^{M1}, X_i^{M2}) (X_i^{M1} - p^{M1}) (X_i^{M2} - p^{M2}) \\
 &\approx \sum_{X_i^{M1}, X_i^{M2} \in \{0,1\}} \Pr(X_i^{M1}, X_i^{M2}) (X_i^{M1} - \hat{p}^{M1}) (X_i^{M2} - \hat{p}^{M2})
 \end{aligned} \tag{8}$$

where $\Pr(X_i^{M1}, X_i^{M2})$ and $\text{Cov}[X_i^{M1}, X_i^{M2}]$ are estimates for $\Pr(X_i^{M1}, X_i^{M2})$ and $\text{Cov}[X_i^{M1}, X_i^{M2}]$, respectively.

Remark 1: The covariance of the methods are represented in the familiar probabilistic form as $\text{Cov}[M1, M2] = \sigma_{M1, M2}$. Let $\text{Cov}[X_i^{M1}, X_i^{M2}] = \sigma_{M1, M2}$. Like p^{M1} and p^{M2} , treat $\sigma_{M1, M2}$ as an unknown parameter.

Following Remark 1, for a particular subject i , define the mean and variance of X_i^{M1} and X_i^{M2} in two steps just as $\text{Cov}[X_i^{M1}, X_i^{M2}]$ in equation (7) and equation (8). First, use the linear formulas to write $E[X_i^{M1} - X_i^{M2}]$ and

$V[X_i^{M1} - X_i^{M2}]$ in terms of the means, variances, and the covariance of X_i^{M1} and X_i^{M2} as

$$E[X_i^{M1} - X_i^{M2}] = E[X_i^{M1}] - E[X_i^{M2}] \quad (9)$$

and

$$V[X_i^{M1} - X_i^{M2}] = V[X_i^{M1}] + V[X_i^{M2}] - 2(\text{Cov}[X_i^{M1}, X_i^{M2}]) \quad (10)$$

Then, figure out what each mean, variance, and covariance is in terms of p^{M1} , p^{M2} , and $\sigma_{M1,M2}$ as

$$\begin{aligned} E[X_i^{M1} - X_i^{M2}] &= E[X_i^{M1}] - E[X_i^{M2}] \\ &= p^{M1} - p^{M2} \\ &\approx \hat{p}^{M1} - \hat{p}^{M2} \end{aligned} \quad (11)$$

and

$$\begin{aligned} V[X_i^{M1} - X_i^{M2}] &= V[X_i^{M1}] + V[X_i^{M2}] - 2(\text{Cov}[X_i^{M1}, X_i^{M2}]) \\ &= p^{M1}(1 - p^{M1}) + p^{M2}(1 - p^{M2}) - 2\sigma_{M1,M2} \\ &\approx \hat{p}^{M1}(1 - \hat{p}^{M1}) + \hat{p}^{M2}(1 - \hat{p}^{M2}) - 2\sigma_{M1,M2} \end{aligned} \quad (12)$$

This leads to estimating the mean and variance of $\hat{p}^{M1} - \hat{p}^{M2}$. The mean tends to cause less confusion than the variance since a very simple connection can be made in the form of the average difference

$$\hat{p}^{M1} - \hat{p}^{M2} = \frac{1}{n} \sum_{i=1}^n (X_i^{M1} - X_i^{M2}) \quad (13)$$

Remark 2: The benefit of equation (13) is that it follows from a bit of basic algebra, so a formal proof is not needed. The analyst only needs to recall that the expected difference in two variables is the difference in expected values. For the

EFFECTIVELY COMPARING DIFFERENCES IN PROPORTIONS

variance, it is important to recall the connection between the variance of averages and individual values.

To achieve Remark 2, recognize that the average simply equals $1 / n$ times the sum. Therefore, $1 / n$ acts as a constant. When a random variable is multiplied by a constant, the mean gets multiplied by the same constant and the variance gets multiplied by that constant squared.

$$\begin{aligned} V[\hat{p}^{M1} - \hat{p}^{M2}] &= V\left[\frac{1}{n} \sum_{i=1}^n (X_i^{M1} - X_i^{M2})\right] \\ &= \left(\frac{1}{n}\right)^2 n V[X_i^{M1} - X_i^{M2}] \\ &= \frac{V[X_i^{M1} - X_i^{M2}]}{n} \end{aligned} \quad (14)$$

Then, the difference in means can be estimated as

$$\begin{aligned} E[\hat{p}^{M1} - \hat{p}^{M2}] &= E[\hat{p}^{M1}] - E[\hat{p}^{M2}] \\ &= p^{M1} - p^{M2} \\ &\approx \hat{p}^{M1} - \hat{p}^{M2} \end{aligned} \quad (15)$$

and the variance as

$$\begin{aligned} V[\hat{p}^{M1} - \hat{p}^{M2}] &= \frac{V[X_i^{M1} - X_i^{M2}]}{n} \\ &\approx \frac{\hat{p}^{M1}(1 - \hat{p}^{M1}) + \hat{p}^{M2}(1 - \hat{p}^{M2}) - 2\sigma_{M1,M2}}{n} \end{aligned} \quad (16)$$

Remark 3: p^{M1}, p^{M2} , or $\text{Cov}[X_i^{M1}, X_i^{M2}]$ is not known, and the values of $\hat{p}^{M1}$, $\hat{p}^{M2}$, and $\text{Cov}[X_i^{M1}, X_i^{M2}]$ are just approximations even though the sample size n is such that $\hat{p}^{M1} \approx p^{M1}$ and $\hat{p}^{M2} \approx p^{M2}$ by Assumption 3.

Testing

Suppose the intent is test the null hypothesis $H_0: p^{M1} = p^{M2}$, which says the two methods are equally effective. To clarify this point, another way to write the null hypothesis is $H_0: p^{M1} - p^{M2} = 0$. However, the consequence of Remark 3 is there is no guarantee of the equivalency $\hat{p}^{M1} - \hat{p}^{M2} = p^{M1} - p^{M2}$. By all preceding logic, at best $\hat{p}^{M1} - \hat{p}^{M2} \approx p^{M1} - p^{M2}$, which still leaves unknown values for p^{M1} and p^{M2} . Therefore, construct a 95% confidence interval for the difference $p^{M1} - p^{M2}$ as

$$(\hat{p}^{M1} - \hat{p}^{M2}) \pm 1.96 \sqrt{V[\hat{p}^{M1} - \hat{p}^{M2}]} \quad (17)$$

which follows from considerations about the weight in the tails of the standard normal distribution. Two may be used instead of the usual 1.96 in equation (17). If the interval does not contain 0, the null hypothesis is rejected at the usual 5% level (Bickel & Doksum, 2015).

The hypothesis could also be tested using the Z-score

$$Z = \frac{\hat{p}^{M1} - \hat{p}^{M2}}{\sqrt{V[\hat{p}^{M1} - \hat{p}^{M2}]}} \quad (18)$$

where Z represents the number of standard errors, $\sqrt{V[\hat{p}^{M1} - \hat{p}^{M2}]}$, the estimate, $\hat{p}^{M1} - \hat{p}^{M2}$, and the null, 0, are from each other. In the event that $|Z| > 1.96$, then based on the data, there is evidence that one method is more effective than the other. For an example of the power of the traditional Z-test in comparing Bernoulli proportions, see Suissa and Shuster (1984). An advantage of these tests is that the confidence interval is always consistent with the hypothesis test decision for a two-tailed test. A formal connection is provided in Appendix A.

Remark 4: Although the method of examining the overlap between two confidence intervals is a recognized technique, avoid using this method in formal significance testing; for justification, see Schenker and Gentleman (2001). This is highlighted the following example:

Example

Suppose the success of two types of nail polish removal methods are tested on $n = 100$ people. Each participant has the exact type of nail polish on each testable finger. For each subject i , define $X_i^{M1} = 1$ if method M is successful (that is, reducing the target amount of polish on the applied nail within a given timeframe) and 0 otherwise. As the aim of this article is for statistical clarity more than method content, for brevity assume this can be nail polish remover of any type (acetone-based vs. acetone-free, remover strips vs. soaking, etc.). The uniqueness of people helps make clear Assumptions 1 and 4. To satisfy Assumption 1, we can apply Method 1 ($M1$) to any right-hand finger and Method 2 ($M2$) to the corresponding finger on the left hand simultaneously. The data on each of the 100 participants, including Excel functions, is shown in the supplementary material. Initially, use the logic of Table 2 to complete Table 3.

The mean values given in equation (5) for each variable are shown in the table and, by equation (1), note $X_i^{M1} \sim \text{Bernoulli}(0.65)$ and $X_i^{M2} \sim \text{Bernoulli}(0.49)$. The corresponding variances given in equation (6) are derived as $\hat{p}^{M1}(1 - \hat{p}^{M1}) = 0.2275$ and $\hat{p}^{M2}(1 - \hat{p}^{M2}) = 0.2499$. Notice this is multiplying the two marginal probabilities $\Pr(X_i^M = 1) \times \Pr(X_i^M = 0)$ for each method M . By equation (8), calculate the covariance as

$$\begin{aligned} \text{Cov}[X_i^{M1}, X_i^{M2}] &= \Pr(0,0)(0-0.65)(0-0.49) + \Pr(1,0)(1-0.65)(0-0.49) \\ &\quad + \Pr(0,1)(0-0.65)(1-0.49) + \Pr(1,1)(1-0.65)(1-0.49) \\ &= 0.0615 \end{aligned} \quad (19)$$

Table 3. Two-way joint probability table for the persuasion example

		X_i^{M1}		$\Pr(X_i^{M2})$
		0	1	
X_i^{M2}	0	$\Pr(0,0) = 0.24$	$\Pr(1,0) = 0.27$	$1 - \hat{p}^{M2} = 0.51$
	1	$\Pr(0,1) = 0.11$	$\Pr(1,1) = 0.38$	$\hat{p}^{M2} = 0.49$
$\Pr(X_i^{M1})$		$1 - \hat{p}^{M1} = 0.35$	$\hat{p}^{M1} = 0.65$	

Then, calculate the variance of the difference by equation (15) as

$$\begin{aligned} V[\hat{p}^{M1} - \hat{p}^{M2}] &= \frac{0.65(0.35) + 0.49(0.51) - 2(0.0615)}{100} \\ &= 0.0035 \end{aligned} \quad (20)$$

Using equation (20), the interval (17) for the difference $\hat{p}^{M1} - \hat{p}^{M2}$ is calculated to be (0.0409, 0.2791). Notice it does not contain 0, reject the null, and conclude the proportions are significantly different. In this case, $M1$ is more effective.

To highlight Remark 4, consider the interval

$$\hat{p}^M \pm 1.96 \sqrt{\frac{\hat{p}^M (1 - \hat{p}^M)}{n}} \quad (21)$$

representing the 95% confidence interval for a single sample $\hat{p}^M$. The interval in (21) is the familiar nominal 95% confidence interval shown in Brownlee (1965). Should Assumption 4 be violated, the covariance would be 0, thereby eliminating the covariance term in equation (16). This condition (via the assumption of independence) mirrors the calculation for $V[\hat{p}^{M1} - \hat{p}^{M2}]$ in Brownlee, and is used to compare proportions discussed in Schenker and Gentleman (2001). Applying the interval (21) to the proportions $\hat{p}^{M1}$ and $\hat{p}^{M2}$ results in (0.5546, 0.7454) and (0.3900, 0.5899), respectively. By the overlap method, conclude that the proportions are not significantly different resulting in a contradiction with our preceding analysis.

Conclusion

Proportions are a key part of applied statistics and merit the attention of useful guides in clarifying common techniques. A classic technique was discussed for comparing proportions with a target audience of students and practitioners (and to a degree, teachers) dealing with statistics of everyday life. A step-by-step procedure was presented for analysis to mitigate the confusion of the audience when making a connection within the equations between a Bernoulli random variable X and the associated proportion p (as well as its respective estimator $\hat{p}$). This procedure culminated with a basic example where we showed the connection between

EFFECTIVELY COMPARING DIFFERENCES IN PROPORTIONS

confidence intervals and hypothesis testing and highlighted the deficiencies of relying on overlapping confidence intervals as a means of inference. This article should serve as a good secondary reference analysts who needs to not only apply statistical procedures to their research, but also to appreciate the basic connections within them.

Acknowledgements

The authors would like to thank the anonymous reviewers for their detailed and constructive comments. We would also like to thank Dr. William Galose and Morgan Turpin for their discussions involving the logical framework and examples presented in this research.

References

- Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54(4), 280-288. doi: [10.2307/2685779](https://doi.org/10.2307/2685779)
- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed.) (Vol. 2). Boca Raton, FL: CRC Press.
- Brownlee, K. A. (1965). *Statistical theory and methodology: In science and engineering* (2nd ed.). New York: John Wiley & Sons.
- Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3), 182-186. doi: [10.1198/000313001317097960](https://doi.org/10.1198/000313001317097960)
- Scott, A. J., & Seber, G. A. (1983). Difference of proportions from the same survey. *The American Statistician*, 37(4), 319-320. doi: [10.2307/2682774](https://doi.org/10.2307/2682774)
- Senn, S. (2002). *Cross-over trials in clinical research* (2nd ed.). Chichester, UK: John Wiley & Sons. doi: [10.1002/0470854596](https://doi.org/10.1002/0470854596)
- Suissa, S., & Shuster, J. J. (1984). Are uniformly most powerful unbiased tests really best? *The American Statistician*, 38(3), 204-206. doi: [10.2307/2683654](https://doi.org/10.2307/2683654)
- Wild, C. J., & Seber, G. A. F. (1993). Comparing two proportions from the same survey. *The American Statistician*, 47(3), 178-181. doi: [10.2307/2684972](https://doi.org/10.2307/2684972)

Appendices

In the following appendices, we first present a formal connection of confidence intervals and hypothesis tests; then, via Bayes' Rule, we show a connection between covariance and the assumption of independence.

Appendix A

The confidence interval and hypothesis test are two ways of saying what we think about the true value of the unknown difference $p^{M1} - p^{M2}$. To make clear this notion, recall the formula for the 95% confidence interval in (17) expressed in its equivalency as

$$(\hat{p}^{M1} - \hat{p}^{M2}) \pm 1.969 \text{ SE} \quad (\text{A22})$$

Now consider the formula for the test statistic in equation (18) in its comparative form

$$Z = \frac{\hat{p}^{M1} - \hat{p}^{M2}}{\text{SE}} \quad (\text{A23})$$

allowing us to compare two values for $p^{M1} - p^{M2}$. These two values are the value we guessed, 0 (from $H_0: p^{M1} - p^{M2} = 0$), and the value we actually estimated from our data, $\hat{p}^{M1} - \hat{p}^{M2}$. The difference between the two values gets divided by the standard error $\sqrt{V[\hat{p}^{M1} - \hat{p}^{M2}]}$, labeled simply as SE.

Just like with the single sample hypothesis test, we want to calculate the number of standard errors away from the null hypothesis value our estimate actually is. So if $\hat{p}^{M1} - \hat{p}^{M2}$ and 0 are more than 1.96 standard errors apart, we will get a Z-score greater than 1.96 and will reject the null at the 5% level. Now, recall that the 95% confidence interval contains all the values within 1.96 standard errors of $\hat{p}^{M1} - \hat{p}^{M2}$. If our guess, 0, lies outside the 95% confidence interval, we will reject the null.

Appendix B

There is a positive relationship between the two methods by the sign of the covariance 0.0615 calculated in equation (19). We can verify this by Bayes' Rule

$$\Pr(X_i^{M2} | X_i^{M1}) = \frac{\Pr(X_i^{M2}, X_i^{M1})}{\Pr(X_i^{M1})} \quad (\text{B24})$$

where $\Pr(X_i^{M2} | X_i^{M1})$ represents the conditional distribution and $\Pr(X_i^{M2}, X_i^{M1})$ and $\Pr(X_i^{M1})$ each represent the joint and marginal distributions discussed previously. We could also structure equation (B1) with respect to X_i^{M2} by solving for $\Pr(X_i^{M2} | X_i^{M1})$. Using the data from the supplemental material, we estimate

$$\begin{aligned} \Pr(X_i^{M2} = 1 | X_i^{M1} = 1) &= \frac{\Pr(X_i^{M2} = 1, X_i^{M1} = 1)}{\Pr(X_i^{M1} = 1)} \\ &= \frac{0.38}{0.65} \\ &= 0.5846 \end{aligned} \quad (\text{B25})$$

With equation (B2) yielding $0.5846 \approx 0.58$, we can now compare the marginal distribution $X_i^{M2} \sim \text{Bernoulli}(0.49)$ to the conditional distribution $X_i^{M2} | X_i^{M1} \sim \text{Bernoulli}(0.58)$. Notice that we didn't really need to calculate $\Pr(X_i^{M2} = 0 | X_i^{M1} = 1)$ since

$$\Pr(X_i^{M2} = 1 | X_i^{M1} = 1) + \Pr(X_i^{M2} = 0 | X_i^{M1} = 1) = 1$$

by the definition of a distribution. Thus, the marginal distribution $\Pr(X_i^{M2})$ and the conditional distribution $\Pr(X_i^{M2} | X_i^{M1} = 1)$ are not the same. The distribution of X_i^{M2} depends on what we observe for X_i^{M1} . Therefore, they are not independent, validating Assumption 4.

Given

$$\Pr(X_i^{M2}) \neq \Pr(X_i^{M2} | X_i^{M1} = 1)$$

it is easily inferred that

$$\Pr(X_i^{M2}) \neq \Pr(X_i^{M2} | X_i^{M1} = 0)$$

Thus, the conditional distributions $\Pr(X_i^{M2} | X_i^{M1} = 1)$ and $\Pr(X_i^{M2} | X_i^{M1} = 0)$ are not the same. They each depend on what we observe for X_i^{M1} .

To help see the how X_i^{M1} and X_i^{M2} are positively related, we compare the conditional distributions $\Pr(X_i^{M2} = 1 | X_i^{M1} = 1)$ and $\Pr(X_i^{M2} = 1 | X_i^{M1} = 0)$ to $\Pr(X_i^{M2} = 0 | X_i^{M1} = 1)$ and $\Pr(X_i^{M2} = 0 | X_i^{M1} = 0)$. Applying equation (A2), we get the following calculations:

$$\Pr(X_i^{M2} = 0 | X_i^{M1} = 0) = 0.6857 \approx 0.69$$

$$\Pr(X_i^{M2} = 1 | X_i^{M1} = 0) = 0.3143 \approx 0.31$$

$$\Pr(X_i^{M2} = 0 | X_i^{M1} = 1) = 0.4154 \approx 0.42$$

$$\Pr(X_i^{M2} = 1 | X_i^{M1} = 1) = 0.5846 \approx 0.58$$

Notice that conditional on X_i^{M1} being small (large), the probabilities get larger for X_i^{M2} when it is also small (large). Therefore, they are positively related.

'Parallel Universe' or 'Proven Future'? The Language of Dependent Means t -Test Interpretations

Anthony M. Gould

Laval University
Quebec City, QC

Jean-Etienne Joullié

Gulf University for Science and Technology
Mubarak Al-Abdullah, Kuwait

Of the three kinds of two-mean comparisons which judge a test statistic against a critical value taken from a Student t -distribution, one – the repeated measures or dependent-means application – is distinctive because it is meant to assess the value of a parameter which is not part of the natural order. This absence forces a choice between two interpretations of a significant test result and the meaning of the test hypothesis. The parallel universe view advances a conditional, backward-looking conclusion. The more practical proven future interpretation is a non-conditional proposition about what will happen if an intervention is (now) applied to each population element. Proven future conclusions are subject to the corrupting influence of time-displacement, which include the effects of learning, development, and history. These two interpretations are explored, and a proposal for new conceptual categories and nomenclature is given to distinguish them, applicable to other repeated measures procedures derived from the general linear model including ANOVA.

Keywords: t -test, parameter, dependent-means, language

Introduction

In the social sciences, whether knowledge is a socially constructed discourse or refers to a stable and given reality that is objectively accessible at least in principle is a question that has attracted contributions for decades. Statistical analysis, traditionally advanced as a means to measure reality, has not been spared criticism. It has been compared to storytelling and thus viewed as a form of conventional discourse (Ainsworth & Hardy, 2012). Although not taking a side in this debate, this article illustrates the difficulty resulting from the belief the notion of a natural order refers to a given something that can be objectively known. Indeed, when

Dr. Anthony M. Gould is a Full Professor of Employment Relations. Email him at: Anthony.Gould@rld.ulaval.ca.

making statistical inferences it is sometimes difficult to specify a parameter and to formulate an accurate linguistic description of what a result actually means.

There are two types of reasons why this is so: The first arises from contextual elements concerning the problem or research question itself (Lewin & Somekh, 2011; Tajfel & Fraser, 1986). In practice, this kind of concern manifests when a population is difficult to discern. The second – that which is the focus of this article – occurs because in some circumstances parameters literally do not exist because they are merely conceptual a-priori and post-hoc to an analysis. For example, assuming no control group (elements acting as their own controls) in early drug trial situations, a sample of rats may show tumor reduction following treatment with a putative anti-cancer agent. In such a case, because the treatment has been applied only to the sample, the parameter does not exist in the population at the moment the statistic is calculated. Other examples exist in diverse research paradigms, including: counselling intervention research, organizational development research, and where economic interventions are being assessed. The absence of parameter seems paradoxical, because parameter estimation is the *raison d'être* of parametric statistics.

There is a distinction between dependent means *t*-tests and other mean-related *t*-test applications. Parameters for dependent means *t*-tests literally do not exist. The problem is not that certain parameters are theoretical in the sense that the sampling distribution of means, Student *t*-distributions and the central limit theorem are abstract natural phenomena that conveniently support the logic of an analysis and can as such be simulated. Rather, the point is that dependent means *t*-test parameters are simply not out there to be discovered.

There are two possible interpretations of this non-existence and therefore two ways to interpret a significant dependent means *t*-test result. One of these is conceptually and technically sound but somewhat impractical; it is referred to here as the ‘parallel universe’ view, because it invites whoever reads about research results to imagine an alternative reality in which an entire population has been subjected to an intervention, rather than just a subset of elements. The other interpretation is less theoretically defensible but more practical; it is called here the ‘proven future’ view, because it corrals consumers of research to accept the proposition that the future is solely and exclusively determined by the past such that, if the intervention is applied to all subsequent cases, it would yield an outcome comparable to the sample-based result.

One-sample, dependent means, and independent means *t*-tests are intended to identify an actual state of reality, albeit one that it is not easily discoverable and therefore must be inferred from observations of samples (or more precisely sample

statistics). However, of the various two-mean comparisons available, one in particular – the repeated measures procedure – requires that a distinction be made about which of two possible interpretations should accompany a decision to reject the null hypothesis. The literature addressing the mechanics of two-mean comparisons as well as that discussing advanced repeated measures procedures which use the general linear model (e.g. ANOVA-based analyses) mostly either overlooks or does not well elucidate this point. This is unfortunate, because there are methodological and conceptual advantages that flow from giving a more nuanced understanding of the consequence of a missing parameter. Such benefits concern, from an applied perspective, interpretation; and, from a teaching/explanatory perspective, a deepening of understanding. Whatever the case, the existence of two possible interpretations of a dependent means *t*-test result has implications for experimental control which mostly have not received enough attention. They are sufficiently important to necessitate the creation of a new nomenclature and a new way of distinguishing between population frequency distributions.

Three Research Designs Necessitating a Two-Mean Comparison: The Dependent Means Case as Special

There are so-called parametric data analysis situations where population parameters literally do not exist. This is not to say that their values are impractical to calculate, nor does it imply that an understanding of their nature is not as important as it always was. Rather, some parameters do not exist in the sense that they cannot, in theory or practice, be calculated. Hence, when speaking of a parameter, for example in a dependent means *t*-test, it is especially important to be clear about the meaning of a significant test statistic and the associated decision to reject the null hypothesis. Textbooks as well as many studies that use a dependent means protocol for data analysis typically give this matter only cursory consideration (e.g. [Mason, Lind, & Marchal, 1999](#); [Gravetter & Wallnau, 1995](#); [Wright, 1997](#); [Baillargeon, 2012](#); [Salkind, 2011](#)). Such superficial or dismissive treatment leads to inadequate control of time-dependent and potentially confounding variables, and as a consequence, to imprecise or erroneous conclusions.

In the quest to produce statistically significant results, data analysts frequently advocate dependent means designs to reduce a sampling distribution's variation and to create a greater *t*-value, arguing that fewer degrees of freedom is a price worth paying for a smaller test statistic denominator (e.g. [Gravetter & Wallnau, 1995](#)). However, those who are less concerned with statistical significance and more

interested in contextual and ethnographic elements of a problem often favor between- over within-subjects designs (e.g. Lewin & Somekh, 2011; Adams, Khan, Raeside, & White, 2007). For these latter theorists, carry-over effects and other, more general concerns about experimental control are especially important. Such researchers are mostly satisfied that, in lieu of a control group, a matched-pair design where subjects act as their own controls is practical despite being potentially a theoretically compromised solution (Lewin & Somekh, 2011; Alasuutari, Bickman, & Brannen, 2009).

Repeated measures procedures are sometimes viewed as being plagued by the problem of time-related confounding influences (e.g. Cousineau, 2009). This concern is more fruitfully analyzed as the manifestation of an inexistent parameter. Such a perspective makes clear that two options for interpreting a significant result are possible. To understand what is meant by an inexistent parameter, three representations of typical two-mean comparison situations are presented in the first row of Figure 1. Beside each representation is a depiction of the population distribution, the sampling distribution of the mean, and formulae for calculating a test result (to be compared with an appropriate critical value drawn from a Student *t*-distribution when testing hypotheses). To improve clarity and concision, confidence interval formulae are not presented in Figure 1, only hypothesis testing and rejection of the null hypothesis are discussed. The conclusions and insights offered are equally relevant to confidence interval applications. Furthermore, such findings can be extended to more advanced applications of the general linear model such as ANOVA-based procedures.

In row 1.1 of Figure 1, the one-sample case, a statistic (mean) is calculated and indirectly compared to a parameter which actually exists but which is ‘hidden’, difficult or impractical to discover. (The statistic is compared indirectly because it is compared with the mean for the sampling distribution of the mean that is equal to but not the same as the population mean.) The fact that the parameter is hiding in such situations can be appreciated with a thought experiment. Imagine that, at the same time that a statistic is being calculated using the formula in the last cell of row 1.1, another person is calculating the actual population parameter (μ). In such a case, the aim would be to see how close an obtained statistic falls from the calculated specified population value. Now also imagine that an analyst substituted the parameter (mean) for the population (μ) for the parameter (mean) of its sampling distribution of the mean (μ_M). This manipulation would allow an obtained statistic (M) to be compared with a parameter value of interest (μ) rather than with its proxy value, the mean for the sampling distribution of the mean (μ_M). If this

THE LANGUAGE OF DEPENDENT MEANS *t*-TEST INTERPRETATIONS

Graphic Depiction of key elements of analysis	Depiction of the Population Frequency Distribution ¹	Depiction of the Sampling Distribution of the Mean ²	Formula for calculating a test statistic
1.1 One-Sample <i>t</i>-test	Distribution exists as part of the natural order		$t = (M - \mu_M) / \delta_M$ (with $n - 1$ degrees of freedom)
1.2 Independent Means <i>t</i>-test	Distribution exists as part of the natural order		$t = \{(\bar{M}_1 - \bar{M}_2) - (\mu_{M1} - \mu_{M2})\} / \delta_{M1-M2}$ (with $n_1 + n_2 - 2$ degrees of freedom)
1.3 Dependent Means <i>t</i>-test	Distribution does not exist as part of the natural order		$t = (\bar{M}_d - \mu_{M(d)}) / \delta_{M(d)}$ (with $n - 1$ degrees of freedom)

Figure 1. Key elements of three kinds of two-mean comparison *t*-test applications

Note: ¹Depicted as normal although, due to the Central Limit Theorem, a normally distributed population frequency is not essential for applying a *t*-test procedure. ²Depicted as normal because it is assumed that samples used to create the sampling distribution of the mean are of a size $n > 12$ and $n < 30$ (Central Limit Theorem)

were to occur, $t = (M - \mu_M) / \delta_M$ would become $t = (M - \mu) / \delta_M$. Like the orthodox technique, this manipulation would yield the correct result. However, it would be bizarre because it would be standardizing a score (μ) from a non-related distribution,

the population distribution, with two elements (M and δ_M) from the sampling distribution of the mean. Aside from being conceptually unsound, this change in formula adds an arduous and unrealistic additional step to the procedure.

However, parameter substitution could be done and could be synchronized to coincide with calculation of the test statistic (t_{obtained}). In such a case, the analyst would make the rejection/non rejection decision at the same moment that they became aware of the real population mean. They would have used a procedure which bypasses the step which relies on the central limit theorem to prove that $\mu = \mu_M$ (Gravetter & Wallnau, 1995; Wright, 1997; Baillargeon, 2012). One-sample t -test situations of the kind just described (and depicted in row 1.1 of Figure 1) are somewhat rare but nonetheless occasionally important. For example, to measure the temperature of bath water, it is possible to put a thermometer in the bath itself or to take a cupful of water as a representative sample of the liquid and plunge a thermometer therein to record a reading. In this latter case, the researcher would subsequently make an inference about the bath temperature from the cup temperature.

Row 1.2 of Figure 1 is a depiction of an independent means application of a t -test. This technique has some conceptual similarity with the one-sample t -test case discussed above; for example, the parameter of interest here ($\mu_1 - \mu_2$) exists to be discovered before, during, and after the calculation of a test statistic. Once again, in a sense, the parameter may be thought of as hiding, but nonetheless part of the natural order. Another thought experiment makes this clear. Imagine that there was a binary independent variable, say gender, taking two possible values, male or female in this instance. Imagine that an interval-scaled dependent variable was height; the hypothesis for testing (H_1) was that males are taller than females in the population; and the null hypothesis (H_0) was that males are not taller than females in the population. The characteristics of this problem require analysis using an independent means t -test. Specifically, a test t -value (t -statistic) could be calculated to be compared with a critical value drawn from a Student t -distribution (with $n_1 + n_2 - 2$ degrees of freedom). In such a case, the statistic ($M_M - M_F$) is tangible. Further, at the instant of its calculation, there also exists a real, equally tangible population mean for height for males (μ_M) and a real, equally tangible population mean for height for females (μ_F). Although these two population values and their difference ($\mu_M - \mu_F$) may be difficult or impractical to calculate, they are a real part of the natural order that is perhaps hidden but nonetheless there at the moment that a t_{obtained} value is being compared to a t_{critical} value.

When it comes to the dependent means t -test – a protocol which mostly aims to assess the consequence of an intervention – the parameter does not exist at the

moment a statistic is calculated. It is not part of reality because each element of the population is not yet present in the post-intervention state. Hence, in such a case, a t_{obtained} cannot be calculated through substituting μ_d for $\mu_{d(\text{mean})}$. The broken-line depictions in the first cell of row 1.3 of Figure 1, where the distributions of (μ_2) and (μ_d) are stylized portrayals of the population in the post-intervention state, indicate this state of affairs. As such, together they represent one member of a family of future scenarios. Textbook authors are typically unclear about this point, with disparate recommendations offered for drawing a conclusion (e.g. Adams et al., 2011; Cousineau, 2009; Mason et al., 1999; Gravetter & Wallnau, 1995; Wright, 1997; Baillargeon, 2012; Salkind, 2011). In fact, scholars are inclined to offer one of either two antithetical ways of addressing this missing parameter and its associated missing sampling distribution. The first of these is technically and conceptually correct but impossible to operationalize (e.g. Gravetter & Wallnau, 1995; Wright, 1997; Levin & Fox, 2000); it will henceforth be referred to as the ‘parallel universe’ view.

For example, Gravetter and Wallnau (1995), in commenting on a significant result for a matched-pair intervention procedure aimed at controlling asthma symptoms using relaxation, concluded, “Relaxation training resulted in a decrease in the number of doses of medication needed to control asthma symptoms. This reduction was statistically significant, $t(4) = -3.72$, $p = 0.05$, two-tailed” (p. 256). Similarly, Wright (1997) interpreted the result of a significant dependent-means *t*-test by saying “an effect was detected” (p. 53). Furthermore, Levin and Fox (2000), in drawing a conclusion about a significant dependent-means *t*-statistic concerning the efficacy of a remedial math intervention, stated, “The remedial math program has produced a statistically significant improvement in math ability” (p. 227). In each of these cases, past tense conditional verb conjugation was used. Hence, in rejecting the null-hypothesis, a backward-looking conditional inference was invoked with the linguistic structure: If each element of the population had been subjected to the same intervention that was applied to the sample, there would have been a difference in the pre- and post-intervention population means.

The second way of dealing with the missing parameter conundrum is to assert or assume tacitly that an intervention on the population of the kind that was applied to the sample will affect the population as it affected the sample (e.g. Mason et al., 1999; Levin & Rubin, 1998; Elliott & Woodward, 2007). Such conclusions are often parsed in the present or future tense. For example, Mason et al. (1999), wrote “is a difference” (p. 369) to describe the state of a population following a significant dependent-means *t*-test result. This conclusion has a prospective focus. Similarly, Levin and Rubin (1998), in commenting on significantly improved typing speeds

for secretaries using new word-processing software in a pre-test/post-test design, concluded “The difference in typing speed can be attributed to the different word processors” (p. 473). A similar conclusion was offered by Elliott and Woodward (2007) who, in interpreting the effectiveness of a weight-loss regime following a significant dependent means *t*-test result stated, “We reject H_0 and conclude that the diet is effective; $t(14) = 2.567$, one-tailed, $p = 0.001$ ” (p. 73).

These interpretations of a dependent-means protocol make a claim about the future. Owing to control-related confounding influences arising from time-displacement, such ‘proven future’ inferences are technically and conceptually more spurious than those concerning a parallel universe. However, proven future interpretations are attractive because they are inclined to be practical. They are, in a sense, the *raison d’être* for a repeated measures protocol (Fortin, Côte, & Filion, 2006).

These examples were extracted from data analysis textbooks. Textbook authors typically do not give a clear rationale for the way their conclusions are formulated, which makes them appear arbitrary. It is therefore not surprising to find that the applied literature perpetuates this lack of clarity. However, this literature also reveals some patterns. For example, medical research tends to favor a proven future interpretation. Thus, Kutcher, Wei, and Morgan (2015) concluded their study by noting that “these results [i.e., those obtained from their intervention] suggest a simple but effective approach to improving MHL [mental health literacy] in young people” (p. 580). Comparable inferences are found in studies investigating educational-type interventions. According to Baykara, Demir, and Yaman (2015), “ethics education given to students enables them to distinguish ethical violations” (p. 661); Azarbarzin, Malekian, and Taleghani (2015) inferred “supportive-educative programs can enhance some aspects of quality of life” (p. 577). Additional examples are Lau, Li, Mak, and Chung (2004), Scott and Graham (2015), Garst and Ozier (2015), or again, Pritchard, Hansen, Scarboro, and Melnic (2015). This dominance of the proven future interpretation is understandable because medical or educational interventions are intended to be therapeutic and/or remedying. In such circumstances, incentives often exist strongly to push the case for a putative treatment.

Compared with proven future interpretations, parallel universe inferences are less commonly found in the applied literature, although they do exist. For example, Fee, Gray, and Lu (2013), in commenting on improvement in cross-cultural awareness following a stint in a foreign country, reported “expatriates’ level of cognitive complexity increased significantly during the 12-month study period” (p. 299).

Two Interpretations of a Significant Dependent-Means *t*-Test Result: Parallel Universe versus Proven Future

The parallel universe interpretation of a significant dependent-means *t*-test result uses past-tense conditional verb conjugation to describe what would have happened if each member of a population had been subjected to the intervention which was applied to the sample. This emphasis is on an imagined alternative reality in which each element of a population is subjected to a treatment protocol at the same moment that elements of the sample were so subjected. Remembering that such an interpretation is offered following the finding of a significant *t*-test result, it can only be hypothetical because it is not possible to go back in time. In this sense, analysts who rely on it offer a conclusion that has limited practical utility.

Alternatively, a significant dependent means *t*-test result may be used to reject the null hypothesis with the conclusion that, if an intervention is applied to all members of a population, the post-hoc mean would be different to the a-priori mean. This proven future statement has implications for practice. However, it is less likely than the parallel universe view to reflect reality, because it is vulnerable to a source of control-related error, namely time-displacement effects. Time-displacement is an umbrella term covering at least three circumscribed classes of phenomena: development/maturation, non-treatment-related learning, and historical events (Fortin et al., 2006). Each of these has implications for interpretation of a significant dependent means *t*-test finding that deserve further discussion.

Development/maturation is a relatively permanent change in behavior, values, or cognition that cannot be accounted for by experience or a health-related event such as illness or injury (Demetriou, 1998; Upton, 2011). For example, normally developing babies do not learn to walk (Upton, 2011); rather, regardless of whether they have been trained or otherwise instructed, babies typically take their first steps at between 11 to 15 months of age with a normal distribution of habitual infant bipedalism centered on a mean of 13 months (Upton, 2011). Suppose there were theoretical grounds for challenging the maturational perspective of the emergence of upright walking behavior in healthy human beings and that it was possible to train babies to walk earlier than they otherwise would. Further, suppose that there was a misconception about when infants generally walk upright and that it was believed that they mostly crawl until they are at least 18 months old. In such a case, it would be worthwhile to take a sample of 11 to 12-month old babies, measure their propensity towards bipedalism on a suitable scale, subject them to ‘mobility-training’, and then re-measure them on the same instrument in a matched-pair one-tailed hypothesis-test design.

If this analysis did not yield a significant dependent means t -test result, it would have if the three phases of the study (pre-test measurement, intervention, and post-test measurement) had been instituted over a longer period. What this suggests is that a significant test result can be caused by a third, maturation-related variable and thus have nothing to do with the intervention. However, it is also possible for a particular cohort of babies ('late walkers'), a treatment intervention decreases the mean age of bipedalism from 14 months to 13 months. To demonstrate such an effect through a repeated measures protocol, the delay between each phase of the study must be kept minimal; the longer the delay, the less informative would be the assessed post-intervention state. Hence, when there are two population frequency distributions depicted on a single axis with values representing points in time as levels of an independent variable (prior to walking training and following walking training in this example), a proven-future interpretation of a significant test statistic is likely to be misleading. Hence, claims about what will occur remain at best ambiguous and at worst spurious when maturation is an alternative explanation for change on the dependent variable. A key problem here is that the existence of such a competing explanation is not necessarily known.

Learning is a relatively permanent change in behavior, value, or cognition that occurs as a consequence of experience and not as a result of either maturation or a traumatic health-related event (Schacter, Gilbert, & Wegner, 2011). For example, school district officials may implement a stranger-danger initiative to discourage children from accepting lifts with adults who they do not know. The creators of such an initiative want to determine if their intervention changes behavior. They institute a pre-test/post-test repeated measures assessment protocol (before and after the stranger-danger intervention using a sample) and establish a suitable index of childhood propensity to accept rides from strangers as a dependent variable.

In a case like this, the null hypothesis is typically that the stranger-danger intervention does not influence children to be less inclined to accept a ride from people they do not know and the one-tailed test hypothesis is that the stranger-danger intervention makes children less inclined to accept a ride from people they do not know. Suppose that a significant t -test value for this analysis is not obtained at the orthodox Type-1 error rate of $\alpha = 0.01$, and therefore the null hypothesis cannot be rejected. However, the researchers notice that if they make the test slightly less conservative, say by adjusting the Type-1 error rate to $\alpha = 0.05$, they obtain a significant result and can therefore reject the null hypothesis. Further, assume that, on the night of the intervention, there is a lead news story about the abduction and murder of a child who accepted a ride from a stranger.

THE LANGUAGE OF DEPENDENT MEANS *t*-TEST INTERPRETATIONS

This story is followed up over the ensuing days. In such a case, the timing of the three elements of the study, once again, becomes important. From a practical standpoint, it is likely that the almost significant *t*-test result would underestimate the effect of the intervention if it were – in the future – carried out on all members of the population. Hence, in this example, a *t*-test that was non-significant on one day may be significant on the following day (i.e., after the nightly news). This phenomenon, once again, highlights the importance of keeping the amount of time between pre-measurement, intervention, and post-measurement to a minimum, although it is unclear how small that minimum should be. What is generally true is that, the more protracted this delay, the more the subjects (or sample elements) are exposed to stimuli which can elicit learning of an unplanned and uncontrolled, but nonetheless systematic, nature.

An historical occurrence is a time-displacement effect that can be viewed as a special case of learning; special in the sense that it is not cyclical or typical. (Certain events are rare or one-off in nature; they are not amenable to measurement even if they can be said to create a ‘new normal.’ The assassination of President Kennedy is a case in point: it was an unprecedented event in twentieth-century American history and created new and enduring anxieties about the welfare of political leaders. Conversely, child abductions and murders are unfortunately recurring events in large societies; new cases regularly appear and their occurrence can be quantified in probabilistic terms where a numerator is non-zero and is all instances where an event could have occurred.)

For example, suppose that on September 10, 2001, there is a study taking place in New York City that aims to determine if a particular intervention intended to help those who are anxious about flying in an airplane overcome their fear. Following a pre-test/post-test pairing protocol, a dependent means *t*-test is used to analyze data in a one-tailed improvement versus no-improvement hypothesis. A significant test result is obtained, the null hypothesis is rejected and the researcher concludes that the intervention is efficacious. As noted, however, two interpretations are possible. The first (the parallel universe view) says, all things being equal and if all members of the population of interest had received the intervention at the time the study was being carried out, there would have been a mean improvement on the dependent variable (fear of flying). Such an interpretation, although not especially useful because it is purely hypothetical, controls best for potential historical influences; indeed, its linguistic formulation makes it invulnerable to competing explanations arising from time-displacement.

The alternative interpretation (the proven future inference) is misrepresentative in the case described. On September 11, 2001, there were high-

profile terrorist attacks directed towards targets in New York and elsewhere in the United States involving hijacked commercial airplanes. A significant result for the efficacy of a fear of flying initiative obtained on September 10 would presumably not have been produced if the pre-test measurement was made on September 10 and the intervention (and measurement protocol) instituted on the afternoon of September 11.

Conclusion

Using the case of the t -test as an exemplar of other repeated measures designs such as ANOVA, the emphasis here is on interpreting matched-pair hypothesis test results. Questions concerning the strengths and weaknesses of between- versus within-subjects designs have not been addressed. Similarly, the focus has not been on what to do about time-displacement confounding effects, when they are likely to occur, or on techniques for controlling for them. Rather, the focus here has been merely on interpretation. In studies that investigate, by way of a dependent means protocol, the value of a parameter which does not exist as part of the natural order, two types of interpretations are possible when the t -test result is significant: parallel universe or proven future. Although each of these is equivalent in its description of what is being observed, each differs in its practical utility.

The parallel universe view is best suited to circumstances where it is reasonable to believe that time displacement effects are credible alternatives to an observed change in the dependent variable following manipulation of the independent variable. Such cases typically exist in (but are not limited to) the social sciences, even when researchers believe that the reality they study is unaffected by their investigations. Although within-subjects designs have the potential to reduce the standard error of the mean and hence increase the likelihood of statistical significance, they control for a more limited range of confounding influences than between-subjects designs. This disadvantage is not insurmountable unless and until time displacement effects can offset or compound the changes introduced when an independent variable is manipulated. In such instances, theorists typically propose the idea of a control group as the solution (Lewin & Somekh, 2011; Adams et al., 2011; Tajfel & Fraser, 1986). However, the creation of a control group is not always practicable.

The alternative is the proven future interpretation. In the physical sciences, the possibility of time-displacement effects are perhaps lesser than in the social sciences. In these kinds of situations, ‘proven future’ views are often easier to justify; intuitively this seems reasonable because the past is always the best

predictor of the future in the same sense that gravity, for example, can always be relied upon to occur. It is therefore in the social sciences where the choice between the two interpretations of a within-subjects finding becomes especially relevant.

Irrespective of their field of study, researchers, because of the nature of what they do (in particular when engaging in funded research where significant results are typically those that are rewarded), are inclined to favor the proven future interpretation of significant dependent-mean *t*-test results. Two possible factors influence this phenomenon. First (the focus here), not enough consideration is given to the protocol of interpretation. The second is more psychological and arises from researchers' desires to be consequential in their endeavors. Where parallel universe interpretations exist in the literature, they tend to be made implicitly as if researchers lacked confidence to make an inference about their population of interest.

Whatever the case, it is noteworthy that the choice between a parallel universe and a proven future interpretation represents a trade-off: the former is less prone to confounding influences but is less practical (no one can go back in time), the latter is more practical but more prone to confounding influences from time displacement effects. In reporting results, this compromise should be acknowledged and its consequences with respect to the research problem emphasized. When explaining procedures and depicting the elements of an analysis in diagrammatic form, the trade-off perspective can be highlighted through using broken-lines to indicate those frequency distributions which can only occur in the future. Such a convention could be used to signal that certain distributions exist only in a particular future, one which is contingent on the ubiquitous presence of the second level of the independent variable (i.e. the post-intervention state). Two such contingent within-subjects future distributions (which could be depicted with broken lines) would be the population distribution after all elements of the population have received an intervention (μ_2 ; σ_2) and the difference between each element of a population before and after an intervention (μ_d ; σ_d). Use of such a nomenclature would flag the distinctiveness of the repeated measures procedure and serve as a reminder that certain parameters (e.g. μ_2 ; σ_2 ; μ_d ; σ_d) are missing for the moment and contingent on a forthcoming population-wide intervention.

References

Adams, J., Khan, H., Raeside, R., & White, D. (2011). *Research methods for graduate business and social science students* (7th ed.). Los Angeles, CA: Sage. doi: [10.4135/9788132108498](https://doi.org/10.4135/9788132108498)

Ainsworth, S., & Hardy, C. (2012). Subjects of inquiry: Statistics, stories, and the production of knowledge. *Organization Studies*, 33(12), 1693-1714. doi: 10.1177/0170840612457616

Alasuutari, P., Bickman, L., & Brannen, J. (Eds.). (2009). *The SAGE handbook of social research methods*. London, UK: Sage. doi: 10.4135/9781446212165

Azarbarzin, M., Malekian, A., & Taleghani, F. (2015). Effects of supportive -educative program on quality of life of adolescents living with a parent with cancer. *Iranian Journal of Nursing and Midwifery Research*. 20(5), 577-581. Retrieved from <http://ijnmr.mui.ac.ir/index.php/ijnmr/article/view/1236>

Baillargeon, G. (2012). *Méthodologie et techniques statistiques avec applications en management et relations industrielles* [Statistical methodology and techniques with applications in management and industrial relations] (3rd ed.). Trois-Rivières, QC: Les Éditions SMG.

Baykara, Z. G., Demir, S. G., & Yaman, S. (2015). The effect of ethics training on students recognizing ethical violations and developing moral sensitivity. *Nursing Ethics*, 22(6), 661-675. doi: 10.1177/0969733014542673

Cousineau, D. (2009). *Panorama des statistiques pour psychologies: Introductions aux méthodes quantitatives* [Survey of statistics for psychology: Introduction to quantitative methods]. Brussels, Belgium: De Boeck.

Demetriou, A. (1998). Cognitive development. In A. Demetriou, W. Doise, & C. F. M. van Lieshout (Eds.), *Life-span developmental psychology*. Chichester, NY: J. Wiley & Sons.

Elliott, A. C., & Woodward, W. (2007). *Statistical analysis quick reference guidebook*. Thousand Oaks, CA: SAGE. doi: 10.4135/9781412985949

Fee, A., Gray, S. J., & Lu, S. (2013). Developing cognitive complexity from the expatriate experience: Evidence from a longitudinal field study. *International Journal of Cross Cultural Management*, 13(3), 299-318. doi: 10.1177/1470595813484310

Fortin, M.-F., Côte, J., & Filion, F. (2006). *Fondements et étapes du processus de recherche* [Foundations and stages of the research process]. Montréal, QC: Chenelière Éducation.

Garst, A. B., & Ozier, L. W. (2015). Enhancing youth outcomes and organizational practices through a camp-based reading program. *Journal of Experiential Education*, 38(4), 324-338. doi: 10.1177/1053825915578914

- Gravetter, F. J., & Wallnau, L. B. (1995). *Essentials of statistics for the behavioral sciences* (2nd ed.). New York: West Publishing Company.
- Kutcher, S., Wei, Y., & Morgan, C. (2015). Successful application of a Canadian mental health curriculum resource by usual classroom teachers in significantly and sustainably improving student mental health literacy. *The Canadian Journal of Psychiatry*, 60(12), 580-586. doi: [10.1177/070674371506001209](https://doi.org/10.1177/070674371506001209)
- Lau, E. Y. Y., Li, E. K. W., Mak, C. W. Y., & Chung, I. C. P. (2004). Effectiveness of conflict management training for traffic police officers in Hong Kong. *International Journal of Police Science & Management*, 6(2), 97-109. doi: [10.1350/ijps.6.2.97.34468](https://doi.org/10.1350/ijps.6.2.97.34468)
- Levin, J., & Fox, J. A. (2000). *Elementary statistics in social research* (8th ed.). Boston, MA: Allyn and Bacon.
- Levin, R. I., & Rubin, D. S. (1998). *Statistics for management* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Lewin, C., & Somekh, B. (2011). *Theory and methods in social research* (2nd ed.). Los Angeles, CA: SAGE.
- Mason, R. D., Lind, D. A., & Marchal, W. G. (1999). *Statistical techniques in business and economics* (10th ed.). New York: Irwin/McGraw Hill.
- Pritchard, T., Hansen, A., Scarboro, S., & Melnic, I. (2015). Effectiveness of the sport education fitness model on fitness levels, knowledge, and physical activity. *The Physical Educator*, 72(4), 577-600. doi: [10.18666/tpe-2015-v72-i4-6568](https://doi.org/10.18666/tpe-2015-v72-i4-6568)
- Salkind, N. (2011). *Statistics for people who (think they) hate statistics* (4th ed.). Thousand Oaks, CA: SAGE.
- Schacter, D. L., Gilbert, D. T., & Wegner, D. M. (2011). *Psychology* (2nd ed.). New York, NY: Worth Publishers.
- Scott, K. E., & Graham, J. A. (2015). Service-learning: Implications for empathy and community engagement in elementary school children. *Journal of Experiential Education*, 38(4), 354-372. doi: [10.1177/1053825915592889](https://doi.org/10.1177/1053825915592889)
- Tajfel, H., & Fraser, C. (1986) *Introducing social psychology*. London, UK: Penguin Books.
- Upton, P. (2011). *Developmental psychology*. Exeter, UK: Learning Matters.
- Wright, D. B. (1997). *Understanding statistics: An introduction for the social sciences*. Thousand Oaks, CA: Sage Publications.

Power and Sample Size Estimation for Nonparametric Composite Endpoints: Practical Implementation using Data Simulations

Paul M. Brown
University of Alberta
Edmonton, AB

Justin A. Ezekowitz
University of Alberta
Edmonton, AB

Composite endpoints are a popular outcome in controlled studies. However, the required sample size is not easily obtained due to the assortment of outcomes, correlations between them and the way in which the composite is constructed. Data simulations are required. A macro is developed that enables sample size and power estimation.

Keywords: Sample size, composite endpoints, simulations, nonparametric

Introduction

Nonparametric composite endpoints which combine individual study outcomes into a single univariate measure are becoming an increasingly popular primary endpoint in controlled studies; a recent survey showed approximately 50% of studies adopted a composite (Ferreira-Gonzalez et al., 2007). They may be favored due to the increase in power offered over the analysis of individual outcomes, or to calibrate potentially optimistic surrogate endpoints with clinical outcomes that show lower event rates, and to obtain an overall effect of the treatment or intervention.

Composites of the type described here were considered in various fields of research such as psychology (Pedersen, van Domburg, & Larsen, 2004), HIV (Finkelstein & Schoenfeld, 1999), oncology (Buyse, 2010), brain injury (Temkin et al., 2007), limb ischemia (Subherwal et al., 2012), and heart failure (Allen & Spertus, 2013). However, a review of endpoints in acute heart failure noted that the varied use of such endpoints “remains a major potential barrier to progress in the field” (Allen, Hernandez, O'Connor, & Felker, 2009, p. 1), thus some guidance and consistency in use is needed.

*Paul Brown works in the Canadian Vigour Centre. Email him at:
paul.brown@ualberta.ca.*

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

Several composites have been proposed and preference will depend on the purpose of the study. Sun, Davison, Cotter, Pencina, and Koch (2012) compared an eclectic mix of composites based on power estimates. But few papers have emphasized the limitations of composite endpoints (Chi, 2005; Neaton, Gray, Zuckerman, & Konstam, 2005) or described power calculations (Matsouaka & Betensky, 2015; Subherwal et al., 2012) and thorough power assessment that takes correlations among outcomes into account by using simulations may be lacking.

Programs for sample size estimation are not readily available to the researcher when designing a study that employs a composite of novel endpoints. Because construction of the composite is to an extent ad hoc (e.g. how to weight or prioritize outcomes, the number of outcomes etc.) the standard equations for sample size estimation do not apply. This is especially the case for those composite endpoints which are unrestricted in the number and type of outcomes they are composed of. Such composites are the focus of this paper.

The objective here is to describe SAS/IML macros developed which enable the derivation of two popular but quite different composite endpoints and employ data simulations to obtain power and sample size estimates and hence inform study design. With the use of the macros it becomes an easy matter to evaluate the sensitivity of power to changes in the assumptions made, e.g. about the size of the treatment effect on outcomes and the correlations among outcomes. This code, available for download, is used to plan a study in acute heart failure which is used to illustrate the use of the macros and provide example output. SAS/IML macros described in the following section are available to download here: digitalcommons.wayne.edu/jmasm/vol16/iss2/11/. They were developed using SAS 9.4 and SAS procs. The macros which derive the composite endpoints may also be used independently of the simulations macro i.e. to derive and analyze the composite endpoints at study completion.

Methodology

The composite endpoints of interest are the global rank (Felker & Maisel, 2010) and the average Z-score (Sun et al., 2012). They were used, for example, in the Functional Impact of GLP-1 for Heart Failure Treatment (FIGHT) study which compared Liraglutide and placebo groups using a global rank composite comprising mortality, hospital readmission and time-averaged proportional change in N-terminal pro-B-type natriuretic peptide (NTproBNP) level (Margulies et al., 2014). They were also used in the biased ligand of the angiotensin receptor study in acute heart failure (BLAST-AHF) study which used an average Z-score to

compare three dose groups and a placebo in acute patients with heart failure ([Felker et al., 2016](#)).

The global rank assigns each patient a rank according to their responses across a number of outcomes. A rank of 1 is allocated to the patient with the most severe response (an early death for example) and a rank of n (where n is the sample size) is allocated to the patient with the most favorable response. This is achieved by arranging the relevant outcomes in a meaningful way, for example with the most definitive (e.g. mortality) at the top and perhaps a surrogate endpoint at the bottom. If the patient dies they are ranked based on their survival time. If the patient does not die then they may be ranked according to their response on the next outcome in the hierarchy; if they do not fail on that outcome either, then we move to the next outcome, and so forth down the hierarchy of outcomes until the patient receives their rank.

The average Z -score, on the other hand, converts the response on each outcome to a Z -score before combining these scores by taking the average (Z -scores are obtained by subtracting the overall mean and dividing by the corresponding standard deviation). Before taking the average, the Z -scores for the different outcomes must be aligned so that e.g. a positive Z -score represents a beneficial outcome. Thus, the global rank prioritizes outcomes according to a hierarchy and thus weights them, while the average Z -score does not. Analysis for both composites is by the Wilcoxon rank sum test. The average Z -score, at least with regards power, seems superior ([Sun et al., 2012](#)).

The null hypothesis for the rank based composites is that the distribution of ranks are equal for the treatment groups and rejection of this hypothesis implies that the ranks are higher/lower for one of the treatments. Each composite produces a score or rank per patient that summarizes their response to treatment (in the case of the global rank all outcome data are not necessarily taken into account to determine the patient's score). These composites were chosen because their differences imply they will be apt or favored according to the circumstances or researcher, and comparable alternatives are scarce for the situation where various types of outcomes are to be combined.

Composites amenable to this situation must be unrestricted with regard to the number of outcomes they are derived from and therefore provide a broad summary of efficacy. These composites may combine outcomes of varying types, e.g. dichotomous, survival, log normal etc. Their nature implies difficulties not relevant for other composites, e.g. data simulations are required for the estimation of power and this is not straightforward when the outcomes must show certain correlations, i.e. iterations are needed. Our aim was to develop SAS macros flexible enough to

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

allow power estimation for the global rank and average Z-score which incorporate any number of outcomes of any type and in any order (as required by the hierarchical global rank), i.e. this is where SAS macros would prove most useful because other composites are easily coded or less open to ad hoc construction.

Data Simulations (%simul_data)

Assumed treatment differences for each outcome are input into the SAS/IML macro (%simul_data) which are converted to normal variates e.g. log(odds) for dichotomous outcomes, log(hazard) for survival endpoints etc. (using e.g. the delta method for the variance). Random samples of the normal variates are then generated from a multivariate normal distribution using proc iml and the randnormal function before being converted to the specified outcomes, e.g. exponential survival times are generated by

$$-\frac{\log(u)}{\text{hazard}} \quad (1)$$

where u is from the standard uniform distribution (Austin, 2012) and lognormal outcomes are converted to percentage change from baseline, i.e. $100 \times (\exp(x) - 1)$.

Correlations between outcomes are obtained via iteration (%iterat_simul) because the covariance specified for the normal variates using the randnormal function will not ultimately hold among the outcome variables of mixed type. To ensure the correlations between outcomes are those specified by the user, correlations among the normal variates are adjusted on subsequent iterations in order that they converge to the desired values within a certain precision specified by the user; iterations stop when the desired accuracy is achieved (the maximum absolute difference between desired and actual correlations) or the maximum number of iterations is reached. Correlations are determined using Pearson's correlation coefficient from proc corr (including binary outcomes because Pearson produces the same correlation as the apt biserial point correlation). During iteration, correlation matrices that are not positive definite are identified and the nearest correlation matrix is determined using Higham's method as per the NearestCorr function described by Wicklin (2012). Multiple sources may inform what values to assume for the correlations (see the illustrative example below).

The resulting dataset includes two sets of variables for the nominal 'active' and 'control' groups based on the treatment differences specified for each outcome, with the number of random samples and the size of the samples also dictated by the

user; it can easily be verified that the resulting outcomes have the properties specified, e.g. mean response, etc. The run time for convergence and the accuracy are outputted to a separate dataset containing the correlation matrices produced at each iteration.

Global Rank (%derive_GR)

As described above, the global rank is a hierarchical composite meaning that the outcomes are arranged according to importance, i.e. hard endpoints with low event rates such as mortality are at the top with surrogate endpoints with higher responses typically at the bottom. Patients proceed down the hierarchy until they fail on an outcome according to some criterion. A decision rule employing criteria for failure is not necessary for a global rank composite, but we follow the approach of Felker and Maisel (2010) here; ‘global rank’ is a generic term and various specifications could fall under this label (Califf, Harrelson-Woodlief, & Topol, 1990; Finkelstein & Schoenfeld, 1999; Lachin, 1999; Margulies et al., 2014; Pocock, Ariti, Collier, & Wang, 2012; Temkin et al., 2007). The intention is to assign every patient a rank which reflects the severity of response.

Computationally, it is straightforward: patients are ranked according to their response on an outcome if they are among the subset who fail on that outcome; the patient retains the rank that corresponds to the outcome highest in the hierarchy. There is a question of how to rank patients who do not fail on any outcomes, and Felker and Maisel (2010) suggest ranking them on the outcome positioned last in the hierarchy. There is a strong likelihood for tied ranks, e.g. a dichotomous outcome will generate ties; note that handling of ties will depend on the software used (Bergmann, Ludbrook, & Spooren, 2000).

A simple equation yielding arbitrary values that rank patients could be given as follows:

$$s_i = \min_j \left(\delta_{ij} \left[j + \frac{r_{ij}}{n} \right] \right) + \left(1 - \max_j \delta_{ij} \right) (G + r_{iG}/n) \quad (2)$$

where n is the total sample size, G is the total number of outcomes, $\delta_{ij} = 1$ if patient i failed on outcome j and 0 otherwise, and r_{ij} is the rank for patient i on outcome j (rank 1 being the worst response and n being the best). Patients who fail on the last outcome are included in the first term and those who do not are included in the second term, although it is not necessary to define a criterion for failure on the last outcome.

The global rank composite is becoming increasingly popular in phase II research; see the FIGHT study where the global rank was comprised of three outcomes (Margulies et al., 2014). Its appeal is the simplicity of construction and openness to input from researchers regarding prioritizing outcomes.

Average Z-Score (%derive_ZS)

The average Z-score, on the other hand, is computationally intensive and statistically rigorous more so than intuitive. It is an extension of O'Brien's well-known rank sum composite (O'Brien, 1984) for outcomes of different types which must be placed on par by first calculating Z-scores and then taking the average across outcomes (we should also ensure that Z-scores are aligned so that, e.g., bigger scores represent better outcomes).

For survival endpoints this means first transforming to log-rank scores which prolongs the run time of the program (we wrote a macro for this purpose called %lrscores). The LR scores are calculated as

$$1 - \hat{\Lambda}(t_j) \quad (3)$$

for uncensored survival times, and

$$-\hat{\Lambda}(t_j) \quad (4)$$

for censored survival times, where

$$\hat{\Lambda}(t) = -\log \hat{S}(t) \quad (5)$$

is the cumulative hazard and $\hat{S}(t)$ may be obtained from proc lifetest (see e.g. Collett, 2003; Zink & Koch, 2012). The code accounts for censoring by truncating the survival times generated (in order not to overestimate power, especially considering the low event rates often expected for clinical outcomes such as mortality, thus implying many tied Z-scores and reduced power). The log-rank scores thus calculated can be validated by checking they sum to the log-rank test statistic (also provided by proc lifetest).

Using the log rank scores, and for continuous and dichotomous variables too, Z-scores are obtained by subtracting the mean across treatment groups and dividing by the corresponding standard deviation; proc stdize is used for this purpose. For

dichotomous outcomes we want to avoid division by zero for small samples with low event rates (i.e. when all patients have the same response). This macro, as for %derive_GR, uses Wilcoxon and proc npar1way (an output dataset includes a *p*-value per random sample).

Results

Illustrative Power Calculation with Sample Output

When designing a clinical trial in acute heart failure we considered both the global rank and the average Z-score as candidates for the primary endpoint. Given the recruitment and funding feasibility of a pilot or phase II study and expected low event rates for clinical outcomes, an increase in power obtained by combining outcomes was obviously appealing. We deemed 80% power to be satisfactory and planned to measure the following five outcomes: mortality at 30 days, heart failure related hospital readmission at 30 days, worsening heart failure at day 7, dyspnea by 5-day area-under-the curve visual analogue scale, and percent change in NT-proBNP (N-terminal of the prohormone brain natriuretic peptide). It would not be necessary to combine all five outcomes in the chosen composite. Instead the intent is to evaluate how many outcomes would be needed to achieve sufficient power.

Thus, the data include two survival endpoints and single dichotomous, continuous and log-normal endpoints. The ordering of outcomes as listed above indicates the hierarchy employed for the global rank, i.e. mortality and hospital readmission at the top and the surrogate biomarker NT-proBNP, which will potentially show the greatest effect of treatment, at the bottom. The cut-offs employed for the global rank are also implied: for example, 30 days for mortality and hospital readmission and 7 days for worsening heart failure (as far as the code is concerned, the cut-off for dichotomous outcomes is merely 1 indicating presence of disease). These cut-offs and the order of outcomes for the global rank hierarchy are specified in the %derive_GR macro and the outcome type (i.e. dichotomous, survival etc.), and treatment differences are specified in the %simul_data macro.

Treatment responses on the control were based on available data, and modest treatment effect sizes were assumed for the outcomes (2% for mortality, readmission and worsening heart failure, 20% difference in change from baseline NT-proBNP, and 500 for dyspnea visual analogue scale area under the curve). Correlations between outcomes deemed plausible are shown in [Table 1](#). These were based on in-house and published data e.g. Sun et al. (2012, p. 742) noted that “there is a lack of correlation between treatment effects for surrogate endpoints and those

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

for symptom relief or outcome.” The correlation between dyspnea and worsening heart failure (WHF) is high because the latter is derived based on the former (among other data). Within the %iterat_simul macro, criterion=0.05, indicating that the maximum allowable difference between the resulting correlations and the desired correlations is 0.05. Initial working correlations are specified in %simul_data.

Table 1. Correlations assumed between component outcomes

	Mortality	Readmission	WHF	Dyspnea	NTproBNP
Mortality	1.00	0.10	-0.06	0.05	0.00
Readmission	0.10	1.00	-0.03	0.00	0.00
WHF	-0.06	-0.03	1.00	-0.60	0.00
Dyspnea	0.05	0.00	-0.60	1.00	0.00
NTproBNP	0.00	0.00	0.00	0.00	1.00

Order of outcomes: mortality, hospital readmission, WHF, dyspnea VAS AUC, NT-proBNP

Figure 1. Power versus number of outcomes by composite endpoint

With a composite endpoint, when contemplating power the question isn't merely: How many patients are needed?, but may also be: How many outcomes?, with additional outcomes possibly providing additional power (it is not infrequently the case that an outcome's priority is inversely proportional to its sensitivity, i.e. clinical outcomes such as mortality with low event rates are favored before sensitive biomarkers, thus power increases as outcomes are added). There is incentive to limit the outcomes contributing to the composite: missing data become more pervasive the more outcomes used, the interpretability of the composite may become murky, and in terms of data cleaning and validation the outcomes relevant for the primary endpoint ought to receive the most scrutiny which demands extra effort. Thus, in the following SAS code we vary the sample size and the number of outcomes to be incorporated in the composites, deriving for each patient their score for the two composites and then conducting the Wilcoxon test (proc npar1way) to compare the nominal treatment groups:

```
%do varyn = 100 %to 500 %by 100;
  %do varyvar = 3 %to 5 %by 1;
    %iterat_simul(n_=&varyn, numvar_=&varyvar, criterion=0.05, out=randsamp);
    %derive_GR(indata=randsamp, outdata=globrnk);
    %derive_ZS(indata=randsamp, outdata=zscores);
  %end;
%end;
```

Using 1000 simulated samples the power is then estimated as the percentage of samples yielding a p -value < 0.05 . The results are summarized in [Figure 1](#). We can see that to achieve 80% power we need to make use of all five outcomes and recruit 300 patients, if the average Z-score is adopted, or an additional 200 patients for the global rank. We should inflate these numbers to account for potential missing data, bearing in mind that the effect on power would be greater for the average Z-score (if a patient is missing on a single outcome then the average is incalculable and the patient falls out of the analysis, without imputation, which is not the case for the global rank). The addition of a fifth outcome results in a steeper increase in power for the average Z-score. It is obvious that the average Z-score is preferable with regard to power, however some researchers may have a strong preference for a global rank based statistic ([Felker, Anstrom, & Rogers, 2008](#)). The higher power for the Z-score is expected because it does not prioritize clinical outcomes with low event rates, as the global rank does, and by doing so using the

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

global rank we dampen the chances of an optimistic result; Neaton et al. (2005) discuss weighted versus unweighted composites.

With any sample size calculation it is important to examine how sensitive the power estimates are to changes in the assumptions made e.g. regarding the size of the treatment effect. The size of the treatment difference is varied on each outcome (including more pessimistic values), then re-evaluated power for the various scenarios. The results are summarized in Figure 2. In this way uncertainty in the assumptions is reflected in the spread of the box plots and we may now question whether 300 patients are sufficient, depending on our confidence in the anticipated treatment effect. We likewise varied the correlations assumed between mortality and the other outcomes, considering only plausible values, i.e. those with magnitude 0 and 0.1; the results are summarized in Figure 3. In this case uncertainty regarding the strength of correlations between outcomes has a less pronounced effect on power estimates, which might imply that a high degree of convergence (i.e. accuracy ~ 0.01) is not essential. Although we can only say that correlations

*1-2% for mortality, readmission and WHF, 10-20% for NT-proBNP change from baseline, and 400-500 for dyspnea VAS AUC

Figure 2. Power versus sample size when treatment effect size is varied on outcomes

*Correlations between mortality and other outcomes varied between 0.0 and 0.1

Figure 3. Power versus sample size when correlations between outcomes are varied

do not seem important in this case and cannot extrapolate to other potential scenarios (the correlation between e.g. mortality and readmission is necessarily limited given that patients who die have less opportunity to record hospital readmissions; although simulated data should reflect this, i.e. a patient is censored for hospital readmission after death). We could also easily change the order of outcomes in the hierarchy and assess what effect this has on power for the global rank, however the ordering is a clinical decision rather than a statistical one.

These plots can be time consuming to run because the number of scenarios increases with the number of outcomes and sample sizes considered (for Figure 2 there are 5×2^5 passes through the do-loop, with larger sample sizes consuming more time, owing especially to the derivation of log-rank scores). The above code for Figure 1 completes reasonably quickly however with a single pass through taking between 2.85 and 8.35 minutes (depending on the number of outcomes) for a sample size of 100. The maximum number of iterations was set to 50, although correlations often converge in less than 10 iterations, and in the absence of

convergence, i.e. at 50 iterations, reasonable accuracy (~ 0.05) was always achieved.

Conclusion

The code is limited to five outcomes. It could easily be extended to include an increased number of outcomes although this may not be advisable. Increasing the number of outcomes increases the possibility of opposing effects and this would adversely affect power. Also, the cogency and clarity of the composite may be weakened when disparate outcomes are combined. Five outcomes strikes the right balance as a maximum. Also, macros for other composites could be developed: in our study we considered a modification of Finkelstein and Schoenfeld (1999), although this was not included here because the approach and resulting power is similar to the global rank (it is a global rank method with a different decision rule) and the handling of survival and non-survival endpoints is sufficiently different to make generalizing code difficult, i.e. the flexibility of a general program has less value. A clinical composite may also be considered (Massie et al., 2010), although like the Finkelstein and Schoenfeld endpoint it is too ad hoc to make a general program useful, and it is easily coded. A macro was included for the unmatched win-ratio composite (derive_WR) at the web link above; see the supplementary material to Pocock et al. (2012) although note the small error in the variance equation which should sum U^2 from 1 to N .

The code was validated in a number of ways including reproducing power estimates for current trials such as FIGHT which uses a global rank of three outcomes (Margulies et al., 2014) and BLAST using an average Z-score for five outcomes (Felker et al., 2015), both of which used data simulations for sample size estimation. The macros have also been used to evaluate these composites (Brown, Anstrom, Felker, & Ezekowitz, 2016). It is not meant to be implied the construction of a composite should be based entirely on statistical reasoning, e.g. the power attained; first and foremost it will be guided by clinical reasoning (Senn, 1989). When power estimates are based on a composite of multiple endpoints it implies multiple assumptions about, e.g. event rates. It would be prudent to plan an interim, blinded reassessment of power.

The SAS macros described allow the user to readily obtain power estimates when designing a phase II trial based on an overall summary of efficacy, namely the global rank and average Z-score. It is thus easy to compare the composites and evaluate how sensitive power is to a change in their construction or assumptions about the anticipated treatment effects and correlations between the outcomes (such

uncertainty ought to be reflected in the power estimates). The order of the outcomes may be changed in the hierarchical global rank, although the order of outcomes is a clinical decision and should determine the power, rather than vice versa. Appropriate design of clinical trials is aided by a strong statistical framework accounting for assumptions, prior data, estimated treatment effect and our macro assists in that key design step.

References

- Allen, L. A., Hernandez, A. F., O'Connor, C. M., & Felker, G. M. (2009). End points for clinical trials in acute heart failure syndromes. *Journal of the American College of Cardiology*, 53(24), 2248-2258. doi: [10.1016/j.jacc.2008.12.079](https://doi.org/10.1016/j.jacc.2008.12.079)
- Allen, L. A., & Spertus, J. A. (2013). End points for comparative effectiveness research in heart failure. *Heart Failure Clinics*, 9(1), 15-28. doi: [10.1016/j.hfc.2012.09.002](https://doi.org/10.1016/j.hfc.2012.09.002)
- Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29), 3946-3958. doi: [10.1002/sim.5452](https://doi.org/10.1002/sim.5452)
- Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages. *The American Statistician*, 54(1), 72-77. doi: [10.1080/00031305.2000.10474513](https://doi.org/10.1080/00031305.2000.10474513)
- Brown, P. M., Anstrom, K. J., Felker, G. M., & Ezekowitz, J. A. (2016). Composite end points in acute heart failure research: Data simulations illustrate the limitations. *Canadian Journal of Cardiology*, 32(11), 1356.e1321-1356.e1328. doi: [10.1016/j.cjca.2016.02.067](https://doi.org/10.1016/j.cjca.2016.02.067)
- Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, 29(30), 3245-3257. doi: [10.1002/sim.3923](https://doi.org/10.1002/sim.3923)
- Califf, R. M., Harrelson-Woodlief, L., & Topol, E. J. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation*, 82(5), 1847-1853. doi: [10.1161/01.cir.82.5.1847](https://doi.org/10.1161/01.cir.82.5.1847)
- Chi, G. Y. H. (2005). Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology*, 19(6), 609-619. doi: [10.1111/j.1472-8206.2005.00370.x](https://doi.org/10.1111/j.1472-8206.2005.00370.x)

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

- Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Felker, G. M., Anstrom, K. J., & Rogers, J. G. (2008). A global ranking approach to end points in trials of mechanical circulatory support devices. *Journal of Cardiac Failure*, 14(5), 368-372. doi: [10.1016/j.cardfail.2008.01.009](https://doi.org/10.1016/j.cardfail.2008.01.009)
- Felker, G. M., Butler, J., Collins, S. P., Cotter, G., Davison, B. A., Ezekowitz, J. A., . . . Pang, P. S. (2015). Heart failure therapeutics on the basis of a biased ligand of the angiotensin-2 type 1 receptor. Rationale and design of the BLAST-AHF study (Biased Ligand of the Angiotensin Receptor Study in Acute Heart Failure). *JACC: Heart Failure*, 3(3), 193-201. doi: [10.1016/j.jchf.2014.09.008](https://doi.org/10.1016/j.jchf.2014.09.008)
- Felker, G. M., Butler, J., Collins, S. P., Cotter, G., Davison, B. A., Ezekowitz, J. A., . . . Pang, P. S. (2016). *Biased ligand of the angiotensin receptor in acute heart failure (BLAST-AHF)*. Paper presented at the Heart Failure Association of the ESC, Florence, Italy.
- Felker, G. M., & Maisel, A. S. (2010). A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure*, 3(5), 643-646. doi: [10.1161/circheartfailure.109.926030](https://doi.org/10.1161/circheartfailure.109.926030)
- Ferreira-Gonzalez, I., Busse, J. W., Heels-Ansdell, D., Montori, V. M., Akl, E. A., Bryant, D. M., . . . Guyatt, G. H. (2007). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*, 334(7597), 786. doi: [10.1136/bmj.39136.682083.ae](https://doi.org/10.1136/bmj.39136.682083.ae)
- Finkelstein, D. M., & Schoenfeld, D. A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, 18(11), 1341-1354. doi: [10.1002/\(SICI\)1097-0258\(19990615\)18:11<1341::AID-SIM129>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1097-0258(19990615)18:11<1341::AID-SIM129>3.0.CO;2-7)
- Lachin, J. M. (1999). Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials*, 20(5), 408-422. doi: [10.1016/s0197-2456\(99\)00022-7](https://doi.org/10.1016/s0197-2456(99)00022-7)
- Margulies, K. B., Anstrom, K. J., Hernandez, A. F., Redfield, M. M., Shah, M. R., Braunwald, E., & Cappola, T. P. (2014). GLP-1 agonist therapy for advanced heart failure with reduced ejection fraction: Design and rationale for the functional impact of GLP-1 for heart failure treatment study. *Circulation: Heart Failure*, 7(4), 673-679. doi: [10.1161/circheartfailure.114.000346](https://doi.org/10.1161/circheartfailure.114.000346)
- Massie, B. M., O'Connor, C. M., Metra, M., Ponikowski, P., Teerlink, J. R., Cotter, G., . . . Dittrich, H. C. (2010). Rolofylline, an adenosine A1-receptor

antagonist, in acute heart failure. *The New England Journal of Medicine*, 363(15), 1419-1428. doi: [10.1056/nejmoa0912613](https://doi.org/10.1056/nejmoa0912613)

Matsouaka, R. A., & Betensky, R. A. (2015). Power and sample size calculations for the Wilcoxon-Mann-Whitney test in the presence of death-censored observations. *Statistics in Medicine*, 34(3), 406-431. doi: [10.1002/sim.6355](https://doi.org/10.1002/sim.6355)

Neaton, J. D., Gray, G., Zuckerman, B. D., & Konstam, M. A. (2005). Key issues in end point selection for heart failure trials: Composite end points. *Journal of Cardiac Failure*, 11(8), 567-575. doi: [10.1016/j.cardfail.2005.08.350](https://doi.org/10.1016/j.cardfail.2005.08.350)

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4), 1079-1087. doi: [10.2307/2531158](https://doi.org/10.2307/2531158)

Pedersen, S. S., van Domburg, R. T., & Larsen, M. L. (2004). The effect of low social support on short-term prognosis in patients following a first myocardial infarction. *Scandinavian Journal of Psychology*, 45(4), 313-318. doi: [10.1111/j.1467-9450.2004.00410.x](https://doi.org/10.1111/j.1467-9450.2004.00410.x)

Pocock, S. J., Ariti, C. A., Collier, T. J., & Wang, D. (2012). The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, 33(2), 176-182. doi: [10.1093/eurheartj/ehr352](https://doi.org/10.1093/eurheartj/ehr352)

Senn, S. (1989). Combining outcome measures: Statistical power is irrelevant. *Biometrics*, 45(3), 1027-1028. doi: [10.2307/2531703](https://doi.org/10.2307/2531703)

Subherwal, S., Anstrom, K. J., Jones, W. S., Felker, M. G., Misra, S., Conte, M. S., . . . Patel, M. R. (2012). Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia. *American Heart Journal*, 164(3), 277-284. doi: [10.1016/j.ahj.2012.07.002](https://doi.org/10.1016/j.ahj.2012.07.002)

Sun, H., Davison, B. A., Cotter, G., Pencina, M. J., & Koch, G. G. (2012). Evaluating treatment efficacy by multiple end points in phase II acute heart failure clinical trials: Analyzing data using a global method. *Circulation: Heart Failure*, 5(6), 742-749. doi: [10.1161/circheartfailure.112.969154](https://doi.org/10.1161/circheartfailure.112.969154)

Temkin, N. R., Anderson, G. D., Winn, H. R., Ellenbogen, R. G., Britz, G. W., Schuster, J., . . . Dikmen, S. S. (2007). Magnesium sulfate for neuroprotection after traumatic brain injury: A randomised controlled trial. *The Lancet Neurology*, 6(1), 29-38. doi: [10.1016/s1474-4422\(06\)70630-5](https://doi.org/10.1016/s1474-4422(06)70630-5)

Wicklin, R. (2012). *Computing the nearest correlation matrix*. Retrieved from <http://blogs.sas.com/content/iml/2012/11/28/computing-the-nearest-correlation-matrix.html>

POWER ESTIMATION FOR COMPOSITE ENDPOINTS

Zink, R. C., & Koch, G. G. (2012). NParCov3: A SAS/IML macro for nonparametric randomization-based analysis of covariance. *Journal of Statistical Software*, 50(3), 1-17. doi: [10.18637/jss.v050.i03](https://doi.org/10.18637/jss.v050.i03)

A Monte Carlo Study of the Effects of Variability and Outliers on the Linear Correlation Coefficient

Hussein Yousif Eledum

University of Tabuk
Tabuk, Saudi Arabia

Monte Carlo simulations are used to investigate the effect of two factors, the amount of variability and an outlier, on the size of the Pearson correlation coefficient. Some simulation algorithms are developed, and two theorems for increasing or decreasing the amount of variability are suggested.

Keywords: association, correlation, Pearson correlation coefficient, outlier, sample size, variability, Monte Carlo

Introduction

A correlation describes the relationship between two variables. Although there are a number of different correlation statistics, the one that is used most often is the Pearson's correlation (*PC*) defined in terms of the population correlation rho, as

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

where $Cov(X,Y)$ is the correlation between X and Y , σ_X, σ_Y are the population standard deviations of X and Y respectively.

The corresponding sample correlation $r_{x,y}$ (or $\hat{\rho}_{x,y}$) defined by

$$r_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y} \quad (2)$$

Dr. Eledum is an Associate Professor in the Department of Statistics. Email him at heledum@yahoo.com.

EFFECTS OF VARIABILITY ON THE LCC

where, s_X and s_Y are the sample standard deviations of X and Y respectively. The term $\sum_{i=1}^n (x_i - \bar{x})(y - \bar{y})$ is the sample covariance. In terms of z -scores,

$$\rho_{x,y} = \frac{\sum_{i=1}^N z_X z_Y}{N} \quad (3)$$

where, z_X is the z -score of the X variable, calculate using the population μ_X , and standard deviation σ_X , z_Y is likewise the z -score of the Y variable, and N is the number of pairs of scores.

Many studies have been conducted to study factors affecting the size of the correlation coefficient. Goodwin & Leech (2006) discussed factors that affect the size of PC , and Bates et al. (1996) investigated the effects of variability as a function of sample size on the PC under assumption of perfect relationship between two variables. Osborne & Overbay (2004) used the NELS data set (a national longitudinal study of 8th Grade students attending 1,052 high schools across the United States) to see the effect of outliers on two different types of correlations. In the current study, a Monte Carlo simulation will be used to investigate the effects of variability and outliers on the size of PC . In order to generate such data some algorithms have been developed, and two theorems are suggested to increase (or decrease) the amount of variability.

Variability

Variability refers to how spread out a set of data is. The four main measures to describe variability in a data set are: range, interquartile range, variance, and standard deviation. Conceptually, the Pearson Correlation PC of [equation 2](#), is the ratio of the variation shared by X and Y to the variation of X and Y separately. That is:

$$r_{x,y} = \frac{\text{shared variability of } X \text{ and } Y}{\text{separate variability of } X \text{ and } Y} \quad (4)$$

When there is a perfect linear relationship, every change in the X variable is accompanied by a corresponding change in the Y variable. In this case, all variation in X is shared with Y , so $r_{x,y} = 1$. At the other extreme, when there is no linear relationship between X and Y , then the numerator is zero, so $r_{x,y} = 0$. So, [equation 4](#) indicates that definitely variability influence the size of PC . Looking at

equation 4 we observe that increase or decrease in variability of single variable X or Y increases or decreases the shared variability (numerator) and variability of X or Y (part of denominator). Also, increase or decrease in variability of both variables X and Y increases or decreases the shared variability (numerator) and separate variability of X and Y (denominator). Therefore, the size of PC increases if the nominator is greater (or decreases if less) than the denominator, and this depends only on the data set, sample size, and the amount of variability in X , Y , or both.

Glass & Hopkins (1996) noted the value of the correlation coefficient PC will be greater if there is more variability among the observations than if there is less variability. Peers (2006) mentioned a good sample design will minimize the amount of variability in observations. The reduction in variability of a variable has the effect of reducing the correlation a variable has with other variables. The simple correlation is impacted when the variances of two measures are different, such as might occur with a restricted range.

In terms of restriction of range, there are procedures available for the estimation of the correlation for the entire group from the correlation obtained with the selected group (Glass & Hopkins, 1996; Gulliksen, 1950; Nunnally & Bernstein, 1994; Thorndike, 1982). However, the equation used to estimate the unrestricted correlation requires knowledge of the standard deviations of X and Y for the entire group and also requires several assumptions that are rarely tenable in practical situations (Crocker & Algina, 1986). Furthermore, the obtained estimates are often imprecise unless the sample size n is very large (Gullickson & Hopkins, 1976; Linn, 1983). A way to increase or decrease variability is to concomitantly increase or decrease the range. The following two theorems were developed to reduce the variability in term of variance using the idea of reduction range of data set.

Theorem 1

Suppose $x_1, x_2, \dots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_x^2 , such that $x_1, x_2, \dots, x_{n-1} < x_n$, if x_n substituted by $\bar{x}$, let $\bar{x}^*$, s_x^{*2} be mean and variance for new data set respectively, then

$$(a) \bar{x}^* < \bar{x} \quad (b) s_x^{*2} < s_x^2$$

Proof.

Proof of part (a).

First, prove that $\bar{x} < x_n$. The mean of original data is defined by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (*)$$

According to above formula, $\bar{x} = x_n$ if and only if $x_1 = x_2 = \dots = x_n$, but $x_1, x_2, \dots, x_{n-1} < x_n$, therefore, $\bar{x} < x_n$.

Substituting x_n by $\bar{x}$ in formula *

$$\bar{x}^* = \frac{x_1 + x_2 + \dots + \bar{x}}{n}$$

since $\bar{x} < x_n$, then $\bar{x}^* < \bar{x}$.

Proof of part (b).

The sample variance of the original data is defined by

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

Replacing x_n by $\bar{x}$, obtain

$$s_x^{*2} = \frac{(x_1 - \bar{x}^*)^2 + (x_2 - \bar{x}^*)^2 + \dots + (\bar{x} - \bar{x}^*)^2}{n-1}$$

Suppose $x_1, x_2, \dots, x_{n-1} < \bar{x}$, because $\bar{x}^* < \bar{x}$ then,

$$(x_i - \bar{x})^2 > (x_i - \bar{x}^*)^2 \quad \forall i = 1, 2, \dots, n-1$$

then

$$\sum_{i=1}^{n-1} (x_i - \bar{x})^2 > \sum_{i=1}^{n-1} (x_i - \bar{x}^*)^2 \quad (a1)$$

Let $w = \frac{1}{n} \sum_{i=1}^{n-1} x_i$ then,

$$\bar{x} = w + \frac{x_n}{n} \quad (\text{b1})$$

$$\bar{x}^* = w + \frac{\bar{x}}{n} \quad (\text{b2})$$

Adding -1 times equation b2 to equation b1, $\bar{x} - \bar{x}^* = \frac{x_n - \bar{x}}{n}$, since $n > 1$, therefore, $x_n - \bar{x} > \bar{x} - \bar{x}^*$ implies

$$(x_n - \bar{x})^2 > (\bar{x} - \bar{x}^*)^2 \quad (\text{a2})$$

Combining two inequalities of a1 and a2

$$\sum_{i=1}^{n-1} (x_i - \bar{x})^2 + (x_n - \bar{x})^2 > \sum_{i=1}^{n-1} (x_i - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2$$

Dividing each side of above inequality by $n - 1$ to obtain

$$\frac{\sum_{i=1}^{n-1} (x_i - \bar{x})^2 + (x_n - \bar{x})^2}{n-1} > \frac{\sum_{i=1}^{n-1} (x_i - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2}{n-1} \text{ or } s_X^2 > s_X^{*2}$$

Therefore, $s_X^{*2} < s_X^2$.

Corollary

Suppose $x_1, x_2, \dots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_X^2 , such that $x_1, x_2, \dots, x_{n-1} > x_n$, if x_n substituted by $\bar{x}$, let $\bar{x}^*$, s_X^{*2} be mean and variance for new data set respectively, then

$$(a) \bar{x}^* > \bar{x} \quad (b) s_X^{*2} > s_X^2$$

Proof. Follow the same steps used for the proof of Theorem 1 above.

Theorem 2

Suppose $x_1, x_2, \dots, x_n$ are n real positive numbers with mean $\bar{x}$ and variance s_X^2 , such that $x_1, x_2, \dots, x_{n-2} < x_{n-1}, x_n$, and $x_{n-1} = x_n$. If x_n was substituted by $\bar{x}$, to get new data set 1 $x_1, x_2, \dots, x_{n-1}, \bar{x}$ with mean $\bar{x}^*$ and variance s_X^{*2} respectively. Suppose x_{n-1} in a new data set 1 substituted by $\bar{x}^*$, let $\bar{x}^{**}, s_X^{**2}$ be mean and variance for new data set 2 respectively, then

$$(a) \bar{x}^{**} < \bar{x}^* \quad (b) s_X^{**2} < s_X^{*2}$$

Proof.

Proof of part (a).

First, prove that $\bar{x}^* \leq x_{n-1}$. The mean of the new data set 1 is defined by

$$\bar{x}^* = \frac{x_1 + x_2 + \dots + \bar{x} + x_{n-1}}{n} \quad (**)$$

Because $x_{n-1} = x_n$ and $\bar{x} < x_n$ (Theorem 1) then $\bar{x} < x_{n-1}$.

According to formula **, $\bar{x}^* = x_{n-1}$ if and only if $x_1 = x_2 = \dots = \bar{x} = x_{n-1}$, but $x_1, x_2, \dots, x_{n-2}, \bar{x} < x_{n-1}$, therefore, $\bar{x}^* < x_{n-1}$. Because $\bar{x}^* < x_{n-1}$, then $\bar{x}^{**} < \bar{x}^*$.

Proof of part (b).

The variances of new data set 1 are defined by

$$s_X^{*2} = \frac{(x_1 - \bar{x}^*)^2 + (x_2 - \bar{x}^*)^2 + \dots + (x_{n-2} - \bar{x}^*)^2 + (x_{n-1} - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2}{n-1}$$

Replacing x_{n-1} by $\bar{x}^*$,

$$s_X^{**2} = \frac{(x_1 - \bar{x}^{**})^2 + (x_2 - \bar{x}^{**})^2 + \dots + (x_{n-2} - \bar{x}^{**})^2 + (x_{n-1} - \bar{x}^{**})^2 + (\bar{x} - \bar{x}^{**})^2}{n-1}$$

Suppose $x_1, x_2, \dots, x_{n-2}, \bar{x} < \bar{x}^*$, since $\bar{x}^{**} < \bar{x}^*$ then,

$$(x_i - \bar{x}^*)^2 > (x_i - \bar{x}^{**})^2 \quad \forall \quad i = 1, 2, \dots, n-2 \quad \text{and} \quad (\bar{x} - \bar{x}^*)^2 > (\bar{x} - \bar{x}^{**})^2$$

Therefore

$$\sum_{i=1}^{n-2} (x_i - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2 > \sum_{i=1}^{n-2} (x_i - \bar{x}^{**})^2 + (\bar{x} - \bar{x}^{**})^2 \quad (\text{c1})$$

Now let $w = \frac{1}{n} \sum_{i=1}^{n-2} x_i + \frac{\bar{x}}{n}$ then,

$$\bar{x}^* = w + \frac{x_{n-1}}{n} \quad (\text{d1})$$

$$\bar{x}^{**} = w + \frac{\bar{x}}{n} \quad (\text{d2})$$

Adding -1 times equation d2 to equation d1, $\bar{x}^* - \bar{x}^{**} = \frac{x_{n-1} - \bar{x}}{n}$, since $n > 1$, therefore, $x_{n-1} - \bar{x}^* > \bar{x}^* - \bar{x}^{**}$ implies that

$$(x_{n-1} - \bar{x}^*)^2 > (\bar{x}^* - \bar{x}^{**})^2 \quad (\text{c2})$$

Combining two inequalities of c1 and c2 obtain

$$\sum_{i=1}^{n-2} (x_i - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2 + (x_{n-1} - \bar{x}^*)^2 > \sum_{i=1}^{n-2} (x_i - \bar{x}^{**})^2 + (\bar{x} - \bar{x}^{**})^2 + (\bar{x}^* - \bar{x}^{**})^2$$

Dividing both sides of above inequality by $n - 1$ to obtain

$$\frac{\left\{ \sum_{i=1}^{n-2} (x_i - \bar{x}^*)^2 + (\bar{x} - \bar{x}^*)^2 \right\} + (x_{n-1} - \bar{x}^*)^2}{n-1} > \frac{\left\{ \sum_{i=1}^{n-2} (x_i - \bar{x}^{**})^2 + (\bar{x} - \bar{x}^{**})^2 \right\} + (\bar{x}^* - \bar{x}^{**})^2}{n-1} \text{ or } s_X^{*2} > s_X^{**2}$$

Therefore, $s_X^{**2} < s_X^{*2}$.

Outliers

Outliers can be defined as a data point far outside the norm for a variable or population (see, e.g. Jarrell, 1994; Rasmussen, 1988; Stevens, 1984). Hawkins (1980) described outlier as an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. Outliers have also been defined as values that are dubious in the eyes of the researcher (Dixon, 1950) and contaminants (Wainer, 1976). Generally, outliers can be defined as a score, case, or subject that falls outside the range of the rest of the scores, cases, or subjects.

Outlier can also be defined in terms of distributions rather than numerical distance between observations. Therefore, distribution of order statistics from independent non-identical random variables are closely related with the outlier models. Barnett and Lewis (1994) considered the single-outlier model. Balakrishnan (2007) focused on the multiple-outlier model. He presented many results on order statistics from multiple-outlier models and illustrated their use in robustness studies.

Balakrishnan (1988) derived recurrence relations among moments of Order Statistics from two related Outlier models. Balakrishnan (1994a, 1994b) obtained recurrence relations for the single and product moments from non-identical exponential distribution and its right truncated. Balakrishnan and Balasubramanian (1995) gave recurrence relations for moments from non-identical power function distribution. Childs and Balakrishnan (1998) obtained recurrence relations for moments from non-identical Pareto and truncated Pareto distribution. Childs (2001) gave recurrence relations for the single and product moments from non-identical right truncated Lomax distribution. Moshref (2000) established recurrence relations for moments from non-identical generalized power function. Mahmoud et al. (2005) derived order statistics from non-identical generalized Pareto random variables. Recurrence relations for moments for Logistic from non-identical random variables have obtained by Childs and Balakrishnan (2006).

Outliers are often caused by human error, such as errors in data collection, recording, or entry. Sampling errors is another reason for outliers to be occurred, it is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample (Osborne & Overbay, 2004). Outliers can also be caused by research methodology, or by incorrect assumptions about the distribution of the data (Iglewicz & Hoaglin, 1993). Barnett and Lewis (1994) explained not all outliers are illegitimate contaminants, and not all illegitimate scores show up as outliers. Generally, outliers can be classified into

two major categories, those due to errors in the data, and those from the inherent variability of the data.

The presence of outlier can result in an increase or decrease in the size of PC , depending on the location of the outlier (Glass & Hopkins, 1996). Stockburger (2013) demonstrated outlier that falls near where the regression line would normally fall would necessarily increase the size of the correlation coefficient. An outlier that falls some distance away from the original regression line would decrease the size of the correlation coefficient. They also illustrated that smaller the sample size, the greater the effect of the outlier, and at some point the outlier will have little or no effect on the size of the correlation coefficient.

There are various methods of outlier detection; one simple way is to examine the scatter diagram, another method is to use the rules of thumb (data points three or more standard deviations from the mean, or 1.5 IQR criterion). Some researchers prefer visual inspection of the data. Lornez (1987) argued outlier detection is a special case of the examination of data for influential data points.

If there exists an outlier on the dataset, first check for human error (errors in data collection, recording, or entry). If there are no justifications for categorizing the datum an outlier, it should not be removed from the analysis.

Monte Carlo Simulation

A computer program using R Version 3.3.3 was developed as follows.

Algorithm 1

- Step 1.1.** *Population 1 of size 1,000,000:* Generating random variable X follows normal distribution with a mean μ_X and a standard deviation of σ_X .
- Step 1.2.** *Population 2 of size 1,000,000:* Generating another random variable Y follows normal distribution with a mean μ_Y and standard deviation σ_Y , correlated with X with a particular population ρ .
- Step 1.3.** *Sample of size n :* Selecting sample of size n at randomly from each population. Then Algorithm 2 (or 3) is executed.

EFFECTS OF VARIABILITY ON THE LCC

Step 1.4. Replication: Procedures of [Step 1.3](#) were repeated 100,000 times, and the overall average of these repetitions is computed.

To examine the effect of this two factors on the size of PC , some different values of ρ were set, that is, 0.002 (weak correlation), 0.5 (moderate correlation) and 0.99 (strong correlation). Furthermore, sample of sizes 20, 60, 120, and 360 have been determined. [Equation 1](#) is used to create the variance covariance matrix Σ for the univariate normal distribution of the two variables X and Y ,

$$\Sigma = \begin{bmatrix} 1 & \text{cov}(X,Y) \\ \text{cov}(X,Y) & 1 \end{bmatrix} \text{ where } \text{cov}(X,Y) = \rho_{x,y} = \sigma_X \sigma_Y$$

Variability

To illustrate the relationship between variability and the size of PC , follow the two steps of [Algorithm 1](#) by setting $\mu_X = 10$, $\mu_Y = 20$, $\sigma_X = \sigma_Y = 1$, for the values of ρ we selected only the high correlation i.e. $\rho_{x,y} = 0.99$, then developed [Algorithm 2](#).

Algorithm 2

Step 2.1. After generating N pairs of data points (X,Y) , with population correlation ρ , the data were arranged in ascending order on X and Y . PC s for new variables were calculated and stored.

Step 2.2. To conduct the effect of variability on the size of PC , reduce the amount of variability using [Theorem 2](#) after some modifications. Reduction of variability included

1. Both variables X and Y gradually by deleting the highest 5%, 10%, and 20% values each time.
2. Single variable Y by deleting the highest 5%, 10%, and 20% values each time.

To avoid decrease of the sample size, substitute the deleted values by the averages of X and Y for (1) and the average of Y for (2).

Compiled in [Table 1](#) are the variance of X , variance of Y , and size of PC for original samples and samples (new samples) after (1) of [Step 2.2](#) applied, for each

sample size. Compiled in [Table 2](#) are the variance of Y , size of PC for original samples, and samples (new samples) after (2) of [Step 2.2](#) applied.

Note that

- The value between two brackets represents a percentage of reduction in variance of X , variance of Y , and size of PC for new samples with respect to original samples.
- Newsample1, Newsample2, and Newsample3 represent sample (new samples) after the highest 5%, 10%, and 20% values of original sample have been deleted and substituted by the average of specific variable respectively.

EFFECTS OF VARIABILITY ON THE LCC

Table 1. $\text{Var}(X)$, $\text{Var}(Y)$, and size of PC for original and new samples for each sample size.

n	Data	$\text{Var}(X)$	$\text{Var}(Y)$	PC
20	Original sample	0.999802	0.9998876	0.9940756
	Newsample1	0.7945104 (20.53323)	0.7946805 (20.52302)	0.9931182 (0.09631254)
	Newsample2	0.6647697 (33.50986)	0.6650363 (33.4889)	0.9923499 (0.1736026)
	Newsample3	0.4984001 (50.1501200)	0.4987495 (50.11945)	0.9910805 (0.301297)
60	Original sample	0.9983744	0.9971149	0.9968837
	Newsample1	0.7770588 (22.1676)	0.7764702 (22.12832)	0.996571 (0.04099317)
	Newsample2	0.6489883 (34.9955)	0.648791 (34.93318)	0.9961512 (0.07347576)
	Newsample3	0.4879877 (51.12178)	0.4878847 (51.07036)	0.9955679 (0.13199)
120	Original sample	1.004146	1.003993	0.9981221
	Newsample1	0.7768069 (22.64006)	0.7767548 (22.63348)	0.9979473 (0.01751774)
	Newsample2	0.6497357 (35.29471)	0.6496027 (35.29811)	0.9977732 (0.03496057)
	Newsample3	0.48919 (51.283)	0.4891632 (51.27825)	0.9974267 (0.06967345)
360	Original sample	0.9963796	0.9957264	0.9992268
	Newsample1	0.7699424 (22.726)	0.7690821 (22.7617)	0.9992059 (0.002093478)
	Newsample2	0.6445929 (35.3065)	0.644101 (35.31345)	0.9991336 (0.009329307)
	Newsample3	0.4861082 (51.21254)	0.4853523 (51.25645)	0.9989874 (0.0239634)

Table 2. Var(Y) and size of PC for original and new samples for each sample size

n	Original Sample		Newsample1	
	Var(Y)	PC	Var(Y)	PC
20	0.9998876	0.9940756	0.7946805 (20.52302)	0.8954083 (9.925539)
60	0.9971149	0.9968837	0.7764702 (22.12832)	0.8903554 (10.68614)
120	1.003993	0.9981221	0.7767548 (22.63348)	0.8894803 (10.88462)
360	0.9957264	0.9992268	0.7690821 (22.7617)	0.8902479 (10.90632)

n	Newsample2		Newsample3	
	Var(Y)	PC	Var(Y)	PC
20	0.6650363 (33.4889)	0.8442923 (15.0676)	0.4987495 (50.11945)	0.8095944 (18.5581)
60	0.648791 (34.93318)	0.8403698 (15.70031)	0.4878847 (51.07036)	0.8077327 (18.97423)
120	0.6496027 (35.29811)	0.8404212 (15.79976)	0.4891632 (51.27825)	0.8082734 (19.02059)
360	0.644101 (35.31345)	0.8415755 (15.7733)	0.4853523 (51.25645)	0.8097937 (18.95797)

For $n = 20$, a reduction of 20.5%, 33.5%, and 50.1% in the variances of both variables X and Y , results in reduction of 0.0963%, 0.1736%, and 0.3013% in the size of PC respectively. The same reductions in the variance of Y led to reduction of 9.925%, 15.068%, and 18.558% in the size of PC . When $n = 60$, a reduction of 22.2%, 34.9%, and 51.1% in the variances of both X and Y yields 0.0409%, 0.0735%, and 0.1319% reductions in the size of PC respectively. The same reductions in the variance of Y results in 10.686%, 15.700%, and 18.974% reductions in the size of PC .

The same reductions as $n = 20$ and 60 in the variances of both variables, yield a reduction of 0.0175%, 0.0349%, and 0.0697% for $n = 120$, and 0.0021%, 0.0093%, and 0.02396% for $n = 360$ in the size of PC respectively. Whereas, reductions in Y alone for these two sample sizes follow the same pattern of $n = 20$ and 60.

EFFECTS OF VARIABILITY ON THE LCC

Accordingly, the following conclusions are advanced:

1. As the percentage of deleting highest values from original sample increases, the percentage of reduction in PC increases for all sample sizes. This can be seen also from [Figure 1](#), and it means that as the amount of variability increases the size of PC decreases.
2. As the sample size increases the percentage of reduction in the size of PC decreases, also the effect of reduction in variances of two variables X and Y on the size of the PC decreases as the sample size increases (see [Figure 1](#)).
3. The effect of reduction in variance of Y alone on the size of PC is not affected by the sample size.
4. A reduction in the variance of Y alone has strong effect on the size of the PC than a reduction in the variances of two variables X and Y .

Figure 1. Percentage of reduction in PC of new samples for each sample size

The effect of an Outlier

To study the impact of an outlier on the size of PC, the two steps of [Algorithm 1](#) have been followed after setting $\mu_X = 10$, $\mu_Y = 10$, $\sigma_X = 1$, and $\sigma_Y = 1$, and some values of ρ have been set, that is, 0.002 (weak correlation), 0.05 (moderate correlation) and 0.99 (strong correlation), then [Algorithm 3](#) below has been designed.

Algorithm 3

Step 3.1. *Creating an outlier:* Add a single observation out of the samples ranges that were selected in [Step 1.3](#); this observation represents an outlier. This is done for all samples of each variable X and Y separately, and for both variables at the same time. Take into account the position of this observation from regression line and other observations. Then compute the size of PC between X and Y for each case. In [Table 3](#), outliers and their outlier distant from other observations of two variables X and Y are given.

Table 3. The created outliers for each variable X and Y .

X	Y	$\mu_X + 0\sigma_X$	$\mu_X + 4\sigma_X$	$\mu_X + 6\sigma_X$	$\mu_X + 8\sigma_X$	$\mu_X + 10\sigma_X$
$\mu_Y + 4\sigma_Y$		✓	✓			
$\mu_Y + 6\sigma_Y$		✓	✓	✓		
$\mu_Y + 8\sigma_Y$		✓	✓	✓	✓	
$\mu_Y + 10\sigma_Y$		✓	✓	✓	✓	✓

The check symbol (✓) in [Table 3](#) above indicates this data point (outlier) is done; for example, the shaded cell with the check symbol implies that the created outlier is $\mu_X + 6\sigma_X$ for variable X and $\mu_Y + 4\sigma_Y$ for variable Y , where μ_X, μ_Y are the averages of two populations, and σ_X, σ_Y are standard deviations. Set $\mu_X = \mu_Y = 10$ and $\sigma_X = \sigma_Y = 1$, and therefore, the data point (x,y) corresponding the shaded cell is (16,14).

Complied in [Tables 4 - 7](#) are an outlier and the size of PC for sample sizes 20, 60, 120 and 360 for each value of $\hat{\rho}$. The value between parentheses represents the percentages of increase in the size of PC after the existence of outlier.

EFFECTS OF VARIABILITY ON THE LCC

Table 4. An outlier and size of PC for each value of $\hat{\rho}$ when $n = 20$.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.000134	14	0.0004155 (207.961)	0.4540743 (336402.4)			
	16	0.0004376 (224.305)	0.5427325 (402104.6)	0.648574 (480540)		
	18	0.0004526 (235.482)	0.589405 (436692.4)	0.704355 (521879)	0.7649272 (566767)	
	20	0.0004652 (244.757)	0.6157849 (456241.8)	0.735883 (545243)	0.7991673 (592141)	0.834962 (618667)
0.517289	14	0.3832509 (-25.9117)	0.7396204 (42.97996)			
	16	0.3077681 (-40.5037)	0.7721365 (49.26581)	0.833112 (61.053)		
	18	0.251873 (-51.3089)	0.7771106 (50.22738)	0.855385 (65.359)	0.8886333 (71.7864)	
	20	0.2111501 (-59.1815)	0.7730817 (49.44853)	0.862465 (66.727)	0.9028568 (74.5361)	0.921919 (78.2211)
0.9989	14	0.7331424 (-26.6086)	0.999444 (0.049636)			
	16	0.5864489 (-41.2934)	0.9799945 (-1.897359)	0.999646 (0.06982)		
	18	0.4786875 (-52.0808)	0.946635 (-5.236816)	0.991416 (-0.75405)	0.9997641 (0.08168)	
	20	0.4005459 (-59.9032)	0.9147883 (-8.424848)	0.976297 (-2.2675)	0.9959278 (-0.30235)	0.999835 (0.08877)

Table 5. An outlier and size of PC for each value of $\hat{\rho}$ when $n = 60$.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.0032	14	0.002748 (-15.028)	0.216529 (6595.58)			
	16	0.002391 (-26.071)	0.286851 (8770.12)	0.380909 (11678.6)		
	18	0.002057 (-36.378)	0.335072 (10261.2)	0.445228 (13667.5)	0.5209696 (16009.6)	
	20	0.002057 (-45.231)	0.367746 (11271.5)	0.488834 (15015.9)	0.5721052 (17590.9)	0.6286688 (19339.94)
0.5278	14	0.467915 (-11.345)	0.629935 (19.3529)			
	16	0.415922 (-21.196)	0.654849 (24.0732)	0.708196 (34.1809)		
	18	0.365841 (-30.685)	0.659175 (24.8929)	0.733548 (38.9843)	0.7745937 (46.7612)	
	20	0.322084 (-38.975)	0.653423 (23.8031)	0.743026 (40.7801)	0.7957244 (50.7648)	0.8254869 (56.4038)
0.9989	14	0.884845 (-11.425)	0.999202 (0.02223)			
	16	0.785982 (-21.322)	0.982626 (-1.6371)	0.999371 (0.03919)		
	18	0.690923 (-30.837)	0.947056 (-5.1977)	0.989356 (-0.9633)	0.9995146 (0.05355)	
	20	0.607981 (-39.139)	0.906502 (-9.2572)	0.967956 (-3.1055)	0.9935535 (-0.54316)	0.9996243 (0.064532)

EFFECTS OF VARIABILITY ON THE LCC

Table 6. An outlier and size of PC for each value of $\hat{\rho}$ when $n = 120$.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.0067	14	0.006246 (-6.7973)	0.124657 (1759.99)			
	16	0.005812 (-13.266)	0.171539 (2459.53)	0.23751 (3443.88)		
	18	0.005336 (-20.380)	0.208615 (3012.74)	0.289602 (4221.14)	0.3538695 (5180.07)	
	20	0.004866 (-27.397)	0.237067 (3437.27)	0.329624 (4818.31)	0.4030966 (5914.59)	0.4596879 (6758.984)
0.498478	14	0.468111 (-6.0919)	0.558143 (11.9694)			
	16	0.437 (-12.333)	0.576238 (15.5995)	0.615193 (23.4142)		
	18	0.402374 (-19.279)	0.581287 (16.6123)	0.637398 (27.8688)	0.6739801 (35.2075)	
	20	0.368035 (-26.168)	0.577984 (15.9497)	0.647793 (29.9542)	0.6959424 (39.6134)	0.7274151 (45.92715)
0.999	14	0.937906 (-6.1159)	0.999124 (0.01188)			
	16	0.875407 (-12.372)	0.987769 (-1.1246)	0.999237 (0.02323)		
	18	0.805878 (-19.332)	0.960065 (-3.8979)	0.990903 (-0.8110)	0.9993538 (0.03492)	
	20	0.736952 (-26.231)	0.924295 (-7.4784)	0.971021 (-2.8012)	0.9934752 (-0.55351)	0.9994598 (0.045533)

Table 7. An outlier and size of PC for each value of $\hat{\rho}$ when $n = 360$.

$\hat{\rho}$	Outlier	10	14	16	18	20
0.000459	14	0.000409 (-10.815)	0.042961 (9255.71)			
	16	0.000379 (-17.301)	0.062564 (13524.7)	0.091253 (19772.4)		
	18	0.000348 (-24.137)	0.080469 (17424.1)	0.117436 (25474.3)	0.1511931 (32825.8)	
	20	0.000316 (-31.062)	0.096466 (20907.6)	0.140828 (30568.6)	0.1813409 (39391.1)	0.2175222 (47270.46)
0.499015	14	0.488268 (-2.1537)	0.520538 (4.31308)			
	16	0.475749 (-4.6623)	0.528039 (5.81613)	0.544929 (9.20103)		
	18	0.459742 (-7.8701)	0.530407 (6.29067)	0.556016 (11.4226)	0.5751654 (15.2601)	
	20	0.441361 (-11.554)	0.52852 (5.91259)	0.562015 (12.6249)	0.588533 (17.9389)	0.608581 (21.95636)
0.999	14	0.977556 (-2.1471)	0.999048 (0.00424)			
	16	0.952604 (-4.6448)	0.994232 (-0.4778)	0.999096 (0.00906)		
	18	0.92069 (-7.8393)	0.980909 (-1.8115)	0.994835 (-0.4174)	0.9991558 (0.01504)	
	20	0.884029 (-11.509)	0.96103 (-3.8013)	0.983257 (-1.5764)	0.9954962 (-0.35134)	0.9992218 (0.021644)

At data points (14,14), (16,16), (18,18), and (20,20), the size of PC increases for all sample sizes and all values of $\hat{\rho}$. Also, at this data points, the percentage of increase in the size of PC goes up as data points distant away from X and Y coordinates, that is, the percentage of increase at (20,20) greater than percentage of increase at the other data points. This can be seen in Figures 2a, 2b, and 2c. The reason is these data points follow the same pattern of the remainder of the data, in other words, some data points fall near the regression line and other fall in the same direction of the regression line if it is extended (see Figures 3a, 3b, and 3c).

At data points (x,10) where $x = 14, 16, 18, 20$, the size of PC decreases when $n = 60, 120$, and 360 for all values of $\hat{\rho}$, and when $n = 20$ for $\hat{\rho} = 0.000134$, whereas PC increases when $n = 20$ for $\hat{\rho} = 0.517, 0.998$, because these data points lie at the bottom of the regression line top of the x coordinate

EFFECTS OF VARIABILITY ON THE LCC

Figure 2a. Percentage of increase in PC for each sample size at $\hat{\rho}$ close to 0.

Figure 2b. Percentage of increase in PC for each sample size at $\hat{\rho}$ close to 0.5.

Figure 2c. Percentage of increase in PC for each sample size at $\hat{\rho}$ close to 1.

Figure 3a. The outliers on scatter diagram when ρ close to 0.

Figure 3b. The outliers on scatter diagram when ρ close to 0.5.

Figure 3c. The outliers on scatter diagram when ρ close to 1.

EFFECTS OF VARIABILITY ON THE LCC

when $n = 60, 120, 360$, and top of the regression line when $n = 20$ for $\hat{\rho} = 0.517$, 0.998 , and in the same direction of regression line at $n = 20$ for $\hat{\rho} = 0.000134$ (Figures 3a, 3b, and 3c). Moreover, the percentage of decrease in the size of PC at this data points goes down as the sample size increases ($n \neq 20$).

At data points (x,y) where $x \neq y$, the size of PC decreases for all sample sizes and $\hat{\rho}$ close to 1, whereas, for other two values of $\hat{\rho}$ the size of PC increases. This happens because the data points distant away from the rest of observations at the bottom of the regression line when $\hat{\rho}$ close to 1 (Figure 3c), and locates close to the rest of observations or in the same direction of the regression line when $\hat{\rho}$ close to zero and 0.5 (see Figures 3a and 3b).

The conclusions that may now be drawn are:

1. The existence of an outlier on data set might increase or decrease the size of PC , according to the position of this outlier from the rest of observations and the regression line.
2. The effect of an outlier on the size of PC decreases as the sample size increases.
3. Location of an outlier rather than its magnitude, determine the amount of its effect on the size of PC .
4. The effect of an outlier in the size of PC becomes more sensitive as the value of PC close to 1.

Conclusion

Several conclusions can be drawn from present study; the most important of them are: (a) As the amount of variability increases the size of PC decreases, (b) The effect of increase or decrease in the amount of variability in both variables on the size of PC decreases as the sample size increases, whereas the effect of increase or decrease in the amount of variability in single variable on the size of PC is not affected by the sample size, (c) a reduction in the variance of Y alone has a stronger effect on the size of the PC than a reduction in the variances of both variables X and Y , (d) everything else being equal, as the sample size increases the effect of an outlier on the size of PC decreases, and (e) the amount of effect of an outlier on the size of PC depends on factors including the amount of an outlier,

location of an outlier from the regression line or from the rest of observations, the size of PC itself, and the sample size.

References

- Balakrishnan, N. (1988). Recurrence relations among moments of order statistics from two related outlier models. *Biometrical Journal*, 30(6), 741–746. doi: [10.1002/bimj.4710300619](https://doi.org/10.1002/bimj.4710300619)
- Balakrishnan, N. (1994a). On order statistics from non-identical exponential random variables and some applications. *Computational Statistics and Data Analysis*, 18(2), 203 - 253. doi: [10.1016/0167-9473\(94\)90172-4](https://doi.org/10.1016/0167-9473(94)90172-4)
- Balakrishnan, N. (1994b). On order statistics from non-identical right-truncated exponential random variables and some applications. *Communication in Statistics - Theory & Methods*, 23(12), 3373 - 3393. doi: [10.1080/03610929408831453](https://doi.org/10.1080/03610929408831453)
- Balakrishnan, N. (2007). Permanents, order statistics, outliers, and robustness. *Revista Matemática Complutense*, 20(1), 7-107. doi: [10.5209/rev_rema.2007.v20.n1.16528](https://doi.org/10.5209/rev_rema.2007.v20.n1.16528)
- Balakrishnan, N. & Balasubramanian, K. (1995). Order statistics from non-identically power function random variables. *Communication in Statistics - Theory & Methods*, 24(6), 1443 - 1454. doi: [10.1080/03610929508831564](https://doi.org/10.1080/03610929508831564)
- Barnett, V. & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.
- Bates, B. T., Zhang, S., Dufek, J. S., Chen, F. C. (1996). The effects of sample and variability on the correlation coefficient. *Medicine and Science in Sports and Exercise*, 28(3), 386-391. doi: [10.1097/00005768-199603000-00015](https://doi.org/10.1097/00005768-199603000-00015)
- Childs, A. & Balakrishnan, N. (1998). Generalized recurrence relations for moments of order statistics from non-identical Pareto and truncated Pareto random variables with applications to robustness. In N. Balakrishnan and C. R. Rao, Eds. *Handbook of Statistics: Order Statistics: Theory and Methods* (Volume 16). Amsterdam, Netherlands: Elsevier Science B. V. doi: [10.1016/s0169-7161\(98\)16017-0](https://doi.org/10.1016/s0169-7161(98)16017-0)
- Childs, A., Balakrishnan, N., & Moshref, M. (2001). Order statistics from non-identical right-truncated Lomax random variables with applications. *Statistical Papers*, 42(2), 187–206. doi: [10.1007/s003620100050](https://doi.org/10.1007/s003620100050)

EFFECTS OF VARIABILITY ON THE LCC

Childs, A. & Balakrishnan, N. (2006). Relations for order statistics from non-identical logistic random variables and assessment of the effect of multiple outliers on bias of linear estimators. *Journal of Statistics Planning and Inference*, 136(7), 2227-2253. doi: [10.1016/j.jspi.2005.08.026](https://doi.org/10.1016/j.jspi.2005.08.026)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

Dixon, W. J. (1950). Analysis of extreme values. *Annals of Mathematical Statistics*, 21(4), 488-506. doi: [10.1214/aoms/1177729747](https://doi.org/10.1214/aoms/1177729747)

Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Goodwin, D. L. & Leech, L. N. (2006). Understanding correlation: factors that affect the size of r. *The Journal of Experimental Education*, 74(3), 251-266. doi: [10.3200/jexe.74.3.249-266](https://doi.org/10.3200/jexe.74.3.249-266)

Gullickson, A. R. & Hopkins, K. D. (1976). Interval estimation of correlation coefficients corrected for restriction of range. *Educational and Psychological Measurement*, 36(1), 9-25. doi: [10.1177/001316447603600102](https://doi.org/10.1177/001316447603600102)

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. doi: [10.1037/13240-000](https://doi.org/10.1037/13240-000)

Hawkins, D. M. (1980). *Identification of outliers*. London: Chapman and Hall. doi: [10.1007/978-94-015-3994-4](https://doi.org/10.1007/978-94-015-3994-4)

Iglewicz, B. & Hoaglin, D. C. (1993). *How to detect and handle outliers*. Milwaukee, WI.: ASQC Quality Press.

Jarrell, M. G. (1994). A comparison of two procedures, the Mahalanobis Distance and the Andrews-Pregibon Statistic, for identifying multivariate outliers. *Research in the Schools*, 1(1), 49-58.

Linn, R. L. (1983). Pearson selection formulas: implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, 20(1), 1-16. doi: [10.1111/j.1745-3984.1983.tb00185.x](https://doi.org/10.1111/j.1745-3984.1983.tb00185.x)

Lornez, F. O. (1987). Teaching about influence in simple regression. *Teaching Sociology*, 15(2), 173-177. doi: [10.2307/1318032](https://doi.org/10.2307/1318032)

Mahmoud, M., Moshref, M., & Sultan, K. (2005). Order statistics from non-identical generalized Pareto random variables with applications. *Journal of Applied Statistical Science*, 14(1), 85-98.

Moshref, M. (2000). Order statistics from non-identical generalized power function random variables and applications. *The Egyptian Statistical Journal*, 44(1), 99-111.

- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Osborne, J. W. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1-8.
- Peers, I. S. (2006). *Statistical analysis for education and psychology researchers*. Washington, DC: The Falmer Press. doi: [10.4324/9780203985984](https://doi.org/10.4324/9780203985984)
- Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research*, 23(2), 189-202. doi: [10.1207/s15327906mbr2302_4](https://doi.org/10.1207/s15327906mbr2302_4)
- Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, 95(2), 334-344. doi: [10.1037/0033-2909.95.2.334](https://doi.org/10.1037/0033-2909.95.2.334)
- Stockburger, D. W. (2013). *Introductory Statistics: Concepts, Models, and Applications* (3rd Web Ed.). Springfield, MO: Author. Available at <http://psychstat3.missouristate.edu/Documents/IntroBook3/sbk.htm>
- Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.
- Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*, 1(4), 285-312. doi: [10.2307/1164985](https://doi.org/10.2307/1164985)

Parameter Estimation In Weighted Rayleigh Distribution

M. Ajami

Vali-e-Asr University of Rafsanjan
Rafsanjan, Iran

S. M. A. Jahanshahi

University of Sistan and Baluchestan
Zahedan, Iran

A weighted model based on the Rayleigh distribution is proposed and the statistical and reliability properties of this model are presented. Some non-Bayesian and Bayesian methods are used to estimate the β parameter of proposed model. The Bayes estimators are obtained under the symmetric (squared error) and the asymmetric (linear exponential) loss functions using non-informative and reciprocal gamma priors. The performance of the estimators is assessed on the basis of their biases and relative risks under the two above-mentioned loss functions. A simulation study is constructed to evaluate the ability of considered estimation methods. The suitability of the proposed model for a real data is shown by using the Kolmogorov-Smirnov goodness-of-fit test.

Keywords: Bayesian estimators, estimation methods, goodness-of-fit, loss function, reliability, weighted model

Introduction

The Rayleigh distribution has been used in many areas of research, such as reliability, life-testing and survival analysis. Modeling the lifetime of random phenomena has been another area of study for which the Rayleigh distribution has been significantly used. Being first introduced by Rayleigh (1880), this statistical model was originally derived in connection with a problem in acoustics. More details on the Rayleigh distribution can be found in Johnson et al. (1994) and references therein.

The Rayleigh distribution has the following probability density function (pdf) and the cumulative distribution function (cdf), respectively,

M. Ajami is an Assistant Professor in the Department of Statistics. Email him at m.ajami@vru.ac.ir. S. M. A. Jahanshahi is a Professor in the Department of Statistics. Email him at mehdi.jahanshahi@gmail.com.

$$f(x) = \frac{x}{\beta} \exp\left(-\frac{x^2}{2\beta}\right), \quad x > 0, \beta > 0,$$

$$F(x) = 1 - \exp\left(-\frac{x^2}{2\beta}\right), \quad x > 0, \beta > 0.$$

Weighted distributions are employed mainly in research associated with reliability, bio-medicine, meta-analysis, econometrics, survival analysis, renewal processes, physics, ecology and branching processes which are found in Patil and Rao (1978), Gupta and Kirmani (1990), Gupta and Keating (1985), Oluyede (1999), Patil and Ord (1976) and Zelen and Feinleib (1969). A weighted form of Rayleigh distribution has been published by Reshi et al. (2014). They introduced a new class of Size-biased Generalized Rayleigh distribution and also investigated the various structural and characterizing properties of that model. In addition, they studied the Bayes estimator of the parameter of the Rayleigh distribution under the Jeffrey's and the extended Jeffrey's priors assuming two different loss functions. They compared four estimation methods by using mean square error through simulation study with varying sample sizes. In fact, weighted distributions arise in practice when observations from a sample are recorded with unequal probabilities

Suppose X is a non-negative random variable with its unbiased pdf $f(x, \beta)$, β is a parameter, then g distribution is weighted version of f and is defined as

$$g(x, \alpha, \beta) = \frac{w(x, \alpha) f(x, \beta)}{E(w(X, \alpha))},$$

where the weight function $w(x, \alpha)$ is a non-negative function and $0 < E(w(X, \alpha))$ is a normalizing constant which is $E(w(X, \alpha)) = \int w(x, \alpha) f(x, \beta) dx$. Furthermore, α is a parameter which may or may not depend on β and $E(w(X, \alpha)) = 1/E_g(1/w(X, \alpha))$ is the harmonic mean of $w(x, \alpha)$ with the pdf $g(\cdot)$.

When $w(x, \alpha) = x^\alpha$, $\alpha = 0$, the distribution is referred to as weighted distributions of order α .

$$g(x, \alpha, \beta) = \frac{x^\alpha f(x, \beta)}{E(X^\alpha)}. \quad (1)$$

For $\alpha = 1$ or 2 , the pdf (1) are referred to as length-biased (size-biased) and area-biased distributions, respectively.

A weighted Rayleigh (WR) distribution is proposed based on (1) and all calculations are done based upon this model, but in the sections of numerical simulations and application to real data a length-biased Rayleigh (LBR) distribution is used without loss of generality. Because determining the value of α depends on the sampling method so it is not necessary to estimate α in practice, therefore the focus on estimating the β parameter.

Weighted Rayleigh distribution

In the following, the $WR(\alpha, \beta)$ distribution is introduced and then, some properties including the r^{th} moment, the corresponding CDF and hazard rate function are calculated.

Definition 1. A nonnegative random variable X is said to have the $WR(\alpha, \beta)$ distribution provided that the variable's density function is given by

$$g(x, \alpha, \beta) = \frac{1}{2^{\alpha/2} \beta^{\alpha/2+1} \Gamma(\alpha/2+1)} x^{\alpha+1} e^{-x^2/2\beta}, \quad x > 0, \alpha, \beta > 0. \quad (2)$$

Remark 1. Suppose that X follows $WR(\alpha, \beta)$ and let $U = X^2/2\beta$, then U follows $\Gamma(\alpha/2+1, 1)$ distribution.

Remark 2. The $WR(\alpha, \beta)$ distribution belongs to the exponential family. Therefore, $T = \sum_{i=1}^n X_i^2$ is a sufficient complete statistic.

The r^{th} moments are useful for inference and model fitting. A result that allows us to compute the moments of the $WR(\alpha, \beta)$ distribution is given in the following lemma.

Lemma 1. If X be a random variable with density function (2), then the r^{th} moment is given by

$$E(X^r) = 2^{r/2} \beta^{r/2} \frac{\Gamma\left(\frac{\alpha+r}{2}+1\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)},$$

where r is a positive integer.

Proof.

According to (2)

$$E(X^r) = \int_0^\infty \frac{x^{\alpha+r+1}}{2^{\alpha/2} \beta^{\alpha/2+1} \Gamma(\alpha/2+1)} e^{-x^2/2\beta} dx,$$

let $x^2/2\beta = u^2$, then we have

$$E(X^r) = \frac{2\beta(\sqrt{2\beta})^{\alpha+r}}{2^{\alpha/2} \beta^{\alpha/2+1} \Gamma(\alpha/2+1)} \int_0^\infty u^{\alpha+r+1} e^{-u^2} du = 2^{r/2} \beta^{r/2} \frac{\Gamma\left(\frac{\alpha+r}{2}+1\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)}.$$

■

Lemma 1 concludes

$$E(X) = 2^{1/2} \beta^{1/2} \frac{\Gamma\left(\frac{\alpha+1}{2}+1\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)},$$

$$V(X) = 2\beta \left[\frac{\Gamma\left(\frac{\alpha+2}{2}+1\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)} - \left(\frac{\Gamma\left(\frac{\alpha+1}{2}+1\right)}{\Gamma\left(\frac{\alpha}{2}+1\right)} \right)^2 \right],$$

$$CV(X) = \frac{\sqrt{\Gamma\left(\frac{\alpha}{2}+1\right) \Gamma\left(\frac{\alpha+2}{2}+1\right) - \Gamma^2\left(\frac{\alpha+1}{2}+1\right)}}{\Gamma\left(\frac{\alpha+1}{2}+1\right)}.$$

The corresponding CDF of the $WR(\alpha, \beta)$ distribution is as follows:

$$G(x) = \frac{1}{\Gamma(\alpha/2 + 1)} \int_0^{x^2/2\beta} t^{\alpha/2} e^{-t} dt = \frac{1}{\Gamma(\alpha/2 + 1)} \gamma(\alpha/2 + 1, x^2/2\beta),$$

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

In addition, the survival and the hazard rate functions of the $WR(\alpha, \beta)$ distribution are

$$\bar{G}(x) = \frac{1}{\Gamma(\alpha/2 + 1)} \int_{x^2/2\beta}^{\infty} t^{\alpha/2} e^{-t} dt = \frac{1}{\Gamma(\alpha/2 + 1)} \gamma_1(\alpha/2 + 1, x^2/2\beta),$$

and

$$h(x) = \frac{x^{\alpha+1} e^{-x^2/2\beta}}{2^{\alpha/2} \beta^{\alpha/2+1} \gamma_1(\alpha/2 + 1, x^2/2\beta)},$$

respectively, where $\gamma_1(a, z) = \int_z^{\infty} t^{a-1} e^{-t} dt$ denotes upper incomplete gamma function.

In special cases, if $\alpha = 1$, corresponding length-biased distribution is

$$g(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{\beta^{3/2}} e^{-x^2/2\beta}, \quad x > 0, \beta > 0,$$

and if $\alpha = 2$ corresponding area-biased distribution is

$$g(x) = \frac{x^3}{2\beta^2} e^{-x^2/2\beta}, \quad x > 0, \beta > 0.$$

Plots of length-biased and area-biased (ABR) distributions for some parameter values are displayed in [Figure 1](#). Some possible shapes of the LBR and ABR hazard rate functions are displayed in [Figure 2](#)

Figure 1. The LBR(β) (left panel) and ABR(β) (right panel) density functions for some parameter values.

Figure 2. The LBR(β) (left panel) and ABR(β) (right panel) hazard rate functions for some parameter values.

Parameter estimation

In this section, the method of moments, the maximum likelihood method, uniformly minimum variance unbiased method, maximum goodness-of-fit method and some Bayesian methods are used to estimate the β parameter of the model.

Method of moments estimator

Hereafter, let $X_1, \dots, X_n$ be a random sample from the $WR(\alpha, \beta)$ distribution. The method of moments estimator (MME) is

$$\hat{\beta}_{MME} = \frac{\bar{X}^2}{2} \left(\frac{\Gamma\left(\frac{\alpha}{2} + 1\right)}{\Gamma\left(\frac{\alpha+1}{2} + 1\right)} \right)^2.$$

Maximum likelihood estimator

The likelihood function can be written as

$$L(\beta; x_1, \dots, x_n) = \left(\frac{1}{2^{\alpha/2} \beta^{\alpha/2+1} \Gamma(\alpha/2 + 1)} \right)^n \prod_{i=1}^n X_i^{\alpha+1} e^{-\frac{\sum_{i=1}^n X_i^2}{2\beta}}, \quad x > 0, \alpha, \beta > 0.$$

One can easily calculate maximum likelihood estimator (MLE) of β by taking natural logarithm and derivative relative to β as

$$\hat{\beta}_{MLE} = \frac{T}{n(\alpha + 2)},$$

where $T = \sum_{i=1}^n X_i^2$.

To study asymptotic normality of $\hat{\beta}_{MLE}$, calculate the Fisher information $I(\beta)$ as

$$I(\beta) = \frac{n}{\beta^2} \left(\frac{2\Gamma\left(\frac{\alpha}{2} + 2\right)}{\Gamma\left(\frac{\alpha}{2} + 1\right)} - \left(\frac{\alpha}{2} + 1\right) \right), \quad (3)$$

So according to theorem 18 of Ferguson (1996)

$$\sqrt{I(\beta)}(\hat{\beta}_{MLE} - \beta) \xrightarrow{D} N(0, 1) \text{ as } n \rightarrow \infty.$$

Therefore, an $100(1 - \alpha)\%$ approximate confidence interval of β can be obtained as

$$\hat{\beta}_{MLE} \pm Z_{\alpha/2} \sqrt{I^{-1}(\beta)},$$

where $Z_{\alpha/2}$ is the $\alpha/2^{\text{th}}$ percentile point of the standard normal distribution.

Uniformly minimum variance unbiased estimator

Based upon [Lemma 1](#),

$$E(X^2) = 2\beta \frac{\Gamma\left(\frac{\alpha+2}{2} + 1\right)}{\Gamma\left(\frac{\alpha}{2} + 1\right)}.$$

Therefore,

$$E(T) = 2n\beta \frac{\Gamma\left(\frac{\alpha+2}{2} + 1\right)}{\Gamma\left(\frac{\alpha}{2} + 1\right)},$$

and

$$E\left(\frac{T}{2n} \frac{\Gamma\left(\frac{\alpha}{2} + 1\right)}{\Gamma\left(\frac{\alpha+2}{2} + 1\right)}\right) = \beta,$$

which is a function of the sufficient and complete statistic T that is unbiased for β . Thus based on Lehman-Scheffe theorem we have

$$\hat{\beta}_{UMVUE} = \frac{T}{2n} \frac{\Gamma\left(\frac{\alpha}{2} + 1\right)}{\Gamma\left(\frac{\alpha + 2}{2} + 1\right)}.$$

Maximum goodness-of-fit estimators

Maximum goodness-of-fit estimators (otherwise known as minimum distance estimators) of the parameters of the CDF can be calculated by minimizing any distance of the empirical distribution function (EDF) statistics regarding to the unknown parameters. As other research has shown there is no unique EDF statistic which can be considered the most efficient for all situations (Alizadeh and Arghami, 2011). Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling statistics seem to be momentous in situations are

$$\begin{aligned} D_n &= \sqrt{n} \sup_{1 \leq i \leq n} \left| G_n(x_{(i)}) - G(x_{(i)}) \right|, \\ W_n^2 &= n \sum_{i=1}^n \left[G_n(x_{(i)}) - G(x_{(i)}) \right]^2 p(x_{(i)}), \\ A_n^2 &= n \sum_{i=1}^n \frac{\left[G_n(x_{(i)}) - G(x_{(i)}) \right]^2}{G(x_{(i)})(1 - G(x_{(i)}))} p(x_{(i)}), \end{aligned}$$

where $p(x_{(i)}) = G(x_{(i)}) - G(x_{(i)} - 1)$ is the probability under H_0 and considering that $G_n(\cdot)$ is EDF for $G(\cdot)$.

Bayes estimators of β

Considering β as a random variable, two different priors, namely Jeffreys and reciprocal gamma are considered for β . Taking into account the priors, two different loss functions are used for the $WR(\alpha, \beta)$ model, the first one is the squared error loss (SEL) function and the second one is linear exponential (LINEX) loss function.

Bayes estimator based on Jeffreys' prior

Based on (3) the Jeffreys' prior is

$$\pi(\beta) \propto \sqrt{I(\beta)} = 1/\beta, \beta > 0,$$

and then, the posterior density will be

$$\pi(\beta | \underline{x}) \propto \frac{1}{\beta^{n(\alpha/2+1)+1}} e^{-\sum_{i=1}^n X_i^2 / 2\beta}, \quad (4)$$

which follows reciprocal gamma distribution as

$$\beta | \underline{x} : r\gamma\left(n\left(\frac{\alpha}{2} + 1\right), \frac{\sum_{i=1}^n X_i^2}{2}\right).$$

The Bayesian estimator of β under the SEL function is

$$\hat{\beta}_{SEL} = \frac{T}{n(\alpha + 2) - 2},$$

where the SEL function is

$$L(\beta, \hat{\beta}) = (\hat{\beta} - \beta)^2,$$

and $T = \sum_{i=1}^n X_i^2$.

In the following, Bayesian estimator is calculated under the LINEX loss function. This loss function was proposed by Varian (1975) and Zellner (1986). The LINEX loss function for scale parameter β is given by

$$L(\Delta) = e^{a\Delta} - a\Delta - 1, a \neq 0, \quad (5)$$

where $\Delta = \frac{\hat{\beta}}{\beta} - 1$ and $\hat{\beta}$ is an estimator of β . The sign and magnitude of “a” represent the direction and degree of asymmetry respectively (see Soliman, 2000, and Sanku, 2012). Under LINEX loss function (5) and using the posterior (4), the posterior mean of loss function, $L(\Delta)$, is

$$E[L(\Delta)] = e^{-a} E \left[e^{a \left(\frac{\hat{\beta}}{\beta} \right)} \right] - a E \left[\left(\frac{\hat{\beta}}{\beta} \right) - 1 \right] - 1,$$

one can easily obtain $\hat{\beta}$ which minimizes the posterior expectation of the loss function (5), denoted by $\hat{\beta}_{LJ}$ as

$$\hat{\beta}_{LJ} = \frac{T}{2a} \left(1 - \exp \left(- \frac{a}{n(\alpha/2 + 1) + 1} \right) \right).$$

Bayes estimator based on reciprocal gamma prior

Suppose β follows reciprocal gamma distribution as prior distribution which is

$$\pi(\beta) = \frac{b^\sigma}{\Gamma(\sigma)} \frac{e^{-b/\beta}}{\beta^{\sigma+1}}, \quad \beta, \sigma > 0.$$

Then, the posterior density satisfies

$$\pi(\beta | \underline{x}) \propto \frac{1}{\beta^{n(\alpha/2+1)+\sigma+1}} e^{-(b + \sum_{i=1}^n X_i^2/2)/\beta},$$

so the Bayesian estimator of β under the SEL function is

$$\hat{\beta}_{SERG} = \frac{T + 2b}{c},$$

where $c = n(\alpha + 2) + 2\sigma - 2$.

In special case, if we suppose $\sigma = 1$, $b = 0$ then Bayesian estimator of β is

$$\hat{\beta}_{SERG} = \frac{T}{n(\alpha + 2)},$$

which is equal to MLE.

In addition, Bayesian estimator of β under the LINEX loss function is

$$\hat{\beta}_{LRG} = d(T + 2b),$$

$$\text{where } d = \frac{1 - \exp\left(-\frac{a}{n(\alpha/2 + 1) + \sigma + 1}\right)}{2a}.$$

The risk efficiency of $\hat{\beta}_{SEJ}$ regarding to $\hat{\beta}_{LJ}$ under LINEX and squared errors loss function based on Jeffreys' prior

If random variable X follows the distribution function (2), so X^2 obeys $\Gamma((\alpha/2+1), 2\beta)$ then $T: \Gamma(n(\alpha/2+1), 2\beta)$ as

$$h_T(t) = \frac{1}{(2\beta)^{n(\alpha/2+1)} \Gamma(n(\alpha/2+1))} t^{n(\alpha/2+1)-1} e^{-\frac{t}{2\beta}} \quad \alpha, \beta > 0.$$

Because the risk functions of estimators $\hat{\beta}_{SEJ}$ and $\hat{\beta}_{LJ}$ are important, calculate these risk functions which are denoted by $R_L(\hat{\beta}_{LJ})$, $R_L(\hat{\beta}_{SEJ})$, $R_S(\hat{\beta}_{LJ})$, and $R_S(\hat{\beta}_{SEJ})$ where the subject L denotes risk relative LINEX loss function and the subject S denotes risk relative to SEL.

Lemma 2. Let $X: WR(\alpha, \beta)$, then risk function of $\hat{\beta}_{SEJ}$ under LINEX loss function with respect to the Jeffreys' prior is

$$R_L(\hat{\beta}_{SEJ}) = e^{-a} \left(1 - \frac{a}{n(\alpha/2 + 1) - 1}\right)^{-n(\alpha/2+1)} - \frac{an(\alpha/2 + 1)}{n(\alpha/2 + 1) - 1} + a - 1.$$

Proof. By definition,

$$\begin{aligned}
 R_L(\hat{\beta}_{SEJ}) &= E[L(\Delta)] = \int_0^\infty \left\{ e^{a \left[\left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} \right]} - a \left[\left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} \right] - 1 \right\} h(t) dt \\
 &= \int_0^\infty e^{a \left[\left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} \right]} h(t) dt - a \int_0^\infty \left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} h(t) dt + a - 1.
 \end{aligned} \tag{6}$$

It is easy to verify

$$\text{(I). } \int_0^\infty e^{a \left[\left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} \right]} h(t) dt = e^{-a} \left(1 - \frac{a}{n(\alpha/2+1)-1} \right)^{-n(\alpha/2+1)},$$

$$\text{(II). } \int_0^\infty \left(\frac{\hat{\beta}_{SEJ}}{\beta} \right)^{-1} h(t) dt = \frac{n(\alpha/2+1)}{n(\alpha/2+1)-1}.$$

Substituting (I)-(II) into (6), the result desired follows. ■

Corollary 1.

Based on Lemma 2, one can conclude that

$$R_L(\hat{\beta}_{LJ}) = e^{-\frac{a}{n(\alpha/2+1)+1}} - n(\alpha/2+1) \left(1 - e^{-\frac{a}{n(\alpha/2+1)+1}} \right) + a - 1.$$

Lemma 3.

Let $X: WR(\alpha, \beta)$, then the risk function of $\hat{\beta}_{LJ}$ under SEL function with respect to the Jeffreys' prior is

$$R_S(\hat{\beta}_{LJ}) = \frac{\beta^2 \left(1 - e^{-\frac{a}{n\left(\frac{\alpha}{2}+1\right)}} \right)}{a} \left(\frac{\left(n\left(\frac{\alpha}{2}+1\right) \right) \left(n\left(\frac{\alpha}{2}+1\right) + 1 \right) \left(1 - e^{-\frac{a}{n\left(\frac{\alpha}{2}+1\right)}} \right)}{a - n(\alpha+2) + a \left(1 - e^{-\frac{a}{n\left(\frac{\alpha}{2}+1\right)}} \right)^{-1}} \right).$$

Proof.

By definition,

$$R_S(\hat{\beta}_{LJ}) = \int_0^\infty (\hat{\beta}_{LJ} - \beta)^2 h(t) dt = E(\hat{\beta}_{LJ}^2) - 2\beta E(\hat{\beta}_{LJ}) + \beta^2. \quad (7)$$

However,

$$(I). E(\hat{\beta}_{LJ}^2) = \frac{\beta^2 (n(\alpha/2 + 1))(n(\alpha/2 + 1) + 1) \left(1 - e^{-\frac{a}{n(\alpha/2 + 1)}}\right)^2}{a^2}$$

$$(II). 2\beta E(\hat{\beta}_{LJ}) = \frac{\beta^2 n(\alpha + 2) \left(1 - e^{-\frac{a}{n(\alpha/2 + 1)}}\right)}{a}.$$

Substituting (I)-(II) into (7), the proof is completed. ■

Corollary 2.

In the same procedure of Lemma 3 the $R_S(\hat{\beta}_{SEJ})$ under the SEL is

$$R_S(\hat{\beta}_{SEJ}) = \beta^2 \left[\frac{(n(\alpha/2 + 1))(n(\alpha/2 + 1) + 1)}{(n(\alpha/2 + 1) - 1)^2} - \frac{2n(\alpha/2 + 1)}{(n(\alpha/2 + 1) - 1)} + 1 \right].$$

Definition 2.

The risk efficiency of $\hat{\beta}_2$ regarding to $\hat{\beta}_1$ under L loss function is defined as

$$RE_L(\hat{\beta}_1, \hat{\beta}_2) = \frac{R_L(\hat{\beta}_2)}{R_L(\hat{\beta}_1)}.$$

The risk efficiency of $\hat{\beta}_{SERG}$ regarding to $\hat{\beta}_{LRG}$ under LINEX and SEL functions based on reciprocal gamma's prior

In the following, the risk functions of estimators $\hat{\beta}_{SERG}$ and $\hat{\beta}_{LRG}$ are calculated. Therefore, they are denoted by $R_L(\hat{\beta}_{LRG})$, $R_L(\hat{\beta}_{SERG})$, $R_S(\hat{\beta}_{LRG})$, and $R_S(\hat{\beta}_{SERG})$.

Corollary 3. Let $X: WR(\alpha, \beta)$, then the risk function of $\hat{\beta}_{SERG}$ under the LINEX and the SEL functions and reciprocal gamma prior are

$$R_L(\hat{\beta}_{SERG}) = \frac{e^{-a\left(1-\frac{2b}{\beta c}\right)}}{(1-2c)^{n(\alpha/2+1)}} - \frac{a}{\beta c}(2b + n\beta(\alpha + 2)) + a - 1,$$

$$R_L(\hat{\beta}_{SERG}) = \frac{2n\beta^2(\alpha + 2) + n^2\beta^2(\alpha + 2)^2 + 4b^2 + 4nb\beta(\alpha + 2)}{c^2} - \frac{2\beta(2b + n\beta(\alpha + 2))}{c} + \beta^2.$$

Corollary 4. Similar to Corollary 3 under the LINEX and the SEL functions and reciprocal gamma prior we have

$$R_L(\hat{\beta}_{LRG}) = \frac{e^{\frac{b}{\beta}(1-k)-a}}{(2k-1)^{n(\alpha/2+1)}},$$

and

$$R_S(\hat{\beta}_{LRG}) = d^2 \left[2n\beta^2(\alpha + 2) + n^2\beta^2(\alpha + 2)^2 + 4b^2 + 4nb\beta(\alpha + 2) \right] - 2\beta^2 dn(\alpha + 2) - 4bd\beta + \beta^2,$$

where $k = \exp\left(\frac{-a}{n(\alpha/2+1) + \sigma + 1}\right).$

Numerical simulations

In the following, some experimental results are presented to investigate the effectiveness of the different estimation methods which have been so far performed. Bias and MSE for non-Bayesian estimators are mostly compared for different estimation methods. In this study, different sample sizes of $n = 10, 20$ (small), 30, 40 (moderate), 50 (large) and 100 (very large) are considered. In Table 1, the average estimates of β based on 10,000 replications are presented for different estimation methods in which the MSEs are noted in the parentheses.

As can be seen in Table 1, among simple estimators the MLE and UMVUE have the smallest values of bias and MSE for various values of sample size so MLE and UMVUE are the best estimation methods in terms of bias and MSE. In addition, the other two good methods of estimation in priority of order are MME and CVM.

Table 1. Bias and MSE values of simple estimators for β parameter

n	MLE	MME	UMVUE	KS	CVM	AD
10	-0.002550 (0.000007)	0.014100 (0.000995)	-0.002550 (0.000007)	0.024430 (0.000601)	0.023930 (0.000572)	0.030840 (0.000952)
20	-0.002770 (0.000006)	0.006340 (0.000040)	-0.002770 (0.000006)	0.012050 (0.000145)	0.011170 (0.000125)	0.014600 (0.000213)
30	-0.000260 (0.000003)	0.005690 (0.000032)	-0.000260 (0.000003)	0.009430 (0.000088)	0.009170 (0.000084)	0.011480 (0.000131)
40	-0.001070 (0.000002)	0.003430 (0.000012)	-0.001070 (0.000002)	0.006660 (0.000044)	0.005620 (0.000031)	0.007670 (0.000058)
50	-0.003730 (0.000001)	-0.000380 (0.000000)	-0.003730 (0.000001)	0.001920 (0.000000)	0.001050 (0.000000)	0.002810 (0.000000)
100	0.000550 (0.000000)	0.002840 (0.000000)	0.000550 (0.000000)	0.003610 (0.000001)	0.003110 (0.000001)	0.004190 (0.000001)

Bias values and risk functions are computed to compare considered Bayesian estimators. A comparison of this type is needed to check whether an estimator is inadmissible under some loss function. Therefore, if it is so, the estimator would not be used for the losses specified by that loss function. For this purpose, the risks of the estimators and the efficiency of them are computed. In each case, $a = 1$, $a = -1$, $b = 2$ and $\sigma = 2$ are taken without loss of generality.

Because comparing different loss functions is not reasonable, compare the results in similar loss function, but in different priors. According to results compiled in Tables 2, 3, 5 and 6, all the four considered Bayesian estimators

PARAMETER ESTIMATION IN WEIGHTED RAYLEIGH DISTRIBUTION

based on reciprocal gamma prior have small values of bias. Further, the $\hat{\beta}_{SERG}$ estimator has smaller bias than $\hat{\beta}_{LRG}$ estimator for $a = 1$ while $\hat{\beta}_{LRG}$ estimator has smaller bias than $\hat{\beta}_{SERG}$ estimator for $a = -1$.

According to Tables 2, 3, 5 and 6, among the four considered Bayesian risks based SEL the $R_s(\hat{\beta}_{LRG})$ has the smallest values of risk for various values of sample size.

Also among the four considered Bayesian risks based LINEX, the $R_L(\hat{\beta}_{LJ})$ has the smallest values of risk for various values of sample size.

Table 2. Bias and risk values of Bayesian estimators for β parameter and $a = 1$

n	$bias(\hat{\beta}_{SEJ})$	$bias(\hat{\beta}_{LJ})$	$R_s(\hat{\beta}_{SEJ})$	$R_s(\hat{\beta}_{LJ})$	$R_L(\hat{\beta}_{SEJ})$	$R_L(\hat{\beta}_{LJ})$
10	0.072	-0.090	0.100	0.056	0.047	0.031
20	0.035	-0.047	0.041	0.031	0.020	0.016
30	0.024	-0.031	0.025	0.021	0.012	0.011
40	0.016	-0.025	0.018	0.016	0.009	0.008
50	0.011	-0.022	0.014	0.013	0.007	0.007
100	0.005	-0.011	0.007	0.007	0.003	0.003

Table 3. Bias and risk values of Bayesian estimators for β parameter and $a = -1$

n	$bias(\hat{\beta}_{SEJ})$	$bias(\hat{\beta}_{LJ})$	$R_s(\hat{\beta}_{SEJ})$	$R_s(\hat{\beta}_{LJ})$	$R_L(\hat{\beta}_{SEJ})$	$R_L(\hat{\beta}_{LJ})$
10	0.065	-0.039	0.099	0.071	0.037	0.032
20	0.034	-0.017	0.041	0.035	0.018	0.016
30	0.022	-0.011	0.025	0.023	0.012	0.011
40	0.018	-0.007	0.018	0.017	0.009	0.008
50	0.013	-0.007	0.014	0.014	0.007	0.007
100	0.007	-0.003	0.007	0.007	0.003	0.003

Table 4. Relative risk values of Bayesian estimators for β parameter

n	10	20	30	40	50	100
$RE_{s(a=1)}(\hat{\beta}_{LJ}, \hat{\beta}_{SEJ})$	1.524	1.233	1.149	1.110	1.087	1.043
$RE_{L(a=1)}(\hat{\beta}_{LJ}, \hat{\beta}_{SEJ})$	1.788	1.338	1.214	1.157	1.124	1.060
$RE_{s(a=-1)}(\hat{\beta}_{LJ}, \hat{\beta}_{SEJ})$	1.164	1.078	1.051	1.038	1.031	1.015
$RE_{L(a=-1)}(\hat{\beta}_{LJ}, \hat{\beta}_{SEJ})$	1.382	1.174	1.112	1.083	1.066	1.032

Table 5. Bias and risk values of Bayesian estimators for β parameter and $a =$

n	$bias(\hat{\beta}_{SERG})$	$bias(\hat{\beta}_{LRG})$	$R_s(\hat{\beta}_{SERG})$	$R_s(\hat{\beta}_{LRG})$	$R_L(\hat{\beta}_{SERG})$	$R_L(\hat{\beta}_{LRG})$
10	0.061	-0.083	0.073	0.045	0.504	2.317
20	0.031	-0.046	0.035	0.027	0.501	2.487
30	0.022	-0.031	0.023	0.019	0.500	2.556
40	0.015	-0.025	0.017	0.015	0.500	2.593
50	0.011	-0.021	0.014	0.012	0.500	2.617
100	0.007	-0.009	0.007	0.006	0.500	2.666

Table 6. Bias and risk values of Bayesian estimators for β parameter and $a = -1$

n	$bias(\hat{\beta}_{SERG})$	$bias(\hat{\beta}_{LRG})$	$R_s(\hat{\beta}_{SERG})$	$R_s(\hat{\beta}_{LRG})$	$R_L(\hat{\beta}_{SERG})$	$R_L(\hat{\beta}_{LRG})$
10	0.060	-0.031	0.073	0.050	4.916	0.474
20	0.034	-0.014	0.035	0.029	5.365	0.425
30	0.022	-0.011	0.023	0.020	5.531	0.407
40	0.015	-0.009	0.017	0.015	5.616	0.397
50	0.014	-0.006	0.014	0.013	5.670	0.392
100	0.007	-0.003	0.007	0.006	5.778	0.380

Table 7. Relative risk values of Bayesian estimators for β parameter

n	10	20	30	40	50	100
$RE_{S(a=1)}(\hat{\beta}_{LRG}, \hat{\beta}_{SERG})$	1.743	1.331	1.211	1.156	1.123	1.060
$RE_{L(a=1)}(\hat{\beta}_{LRG}, \hat{\beta}_{SERG})$	0.218	0.201	0.196	0.193	0.192	0.188
$RE_{S(a=-1)}(\hat{\beta}_{LRG}, \hat{\beta}_{SERG})$	1.511	1.241	1.157	1.117	1.093	1.046
$RE_{L(a=-1)}(\hat{\beta}_{LRG}, \hat{\beta}_{SERG})$	10.372	12.625	13.593	14.131	14.474	15.205

Application to real data

Here, in order to display the usage of proposed model in real data, it is needed to analyze two sets of the seven from the afore presented data in paper by Bennett and Filliben (2000). Reportedly, they have notified minority electron mobility for p-type $\text{Ga}_{1-x}\text{Al}_x\text{As}$ with seven different values of mole fraction. To do so, two data sets are employed relating to the mole fractions of 0.25 and 0.30. The data values are as followed:

PARAMETER ESTIMATION IN WEIGHTED RAYLEIGH DISTRIBUTION

Data Set 1 (belongs to mole fraction 0.25): 3.051, 2.779, 2.604, 2.371, 2.214, 2.045, 1.715, 1.525, 1.296, 1.154, 1.016, 0.7948, 0.7007, 0.6292, 0.6175, 0.6449, 0.8881, 1.115, 1.397, 1.506, 1.528.

Data Set 2 (belongs to mole fraction 0.30): 2.658, 2.434, 2.288, 2.092, 1.959, 1.814, 1.530, 1.366, 1.165, 1.041, 0.9198, 0.7241, 0.6403, 0.576, 0.5647, 0.5873, 0.8013, 1.002, 1.250, 1.347, 1.368.

To evaluate the fitting quality of the Rayleigh and LBR distributions, the Kolmogorov-Smirnov (K-S) tests and AIC and BIC's criterions are used. The information about comparing both models are given in Table 8. Since probability values of the LBR model are greater than corresponding values of the Rayleigh model and the AIC and BIC criterions of the LBR model are less than corresponding values of the Rayleigh model. Although the values of considered statistics are not significantly different but we it can be infered that the LBR distribution fits better than the Rayleigh distribution in both considered data.

The MLEs of β are 0.9322 and 0.7309 and the 95 percent confidence intervals of β based on MLEs as suggested above under heading [Parameter Estimation](#), can be obtained as (0.6067,1.2577) and (0.4757,0.9861) respectively.

Table 8. Comparing related statistics for Rayleigh and LBR

Data	Model	D	p.value	AIC	BIC
1	Rayleigh	0.1411	0.7458	46.0090	47.0540
1	LBR	0.1275	0.8427	45.9160	46.9610
2	Rayleigh	0.1354	0.7883	40.3870	41.4320
2	LBR	0.1311	0.8180	39.7820	40.8260

Conclusion

Different estimation procedures were studied for estimating the unknown scale parameter of the $WR(\alpha, \beta)$ distribution being the maximum likelihood estimator, the method of moment estimator, uniformly minimum variance unbiased estimator, maximum goodness-of-fit estimators and the Bayes estimators. Since it is not possible to compare different methods theoretically, some simulations were used for comparison of different estimators with respect to biases, mean squared errors and risks.

All the four considered Bayesian estimators based on reciprocal gamma prior have small values of bias. In addition, the $\hat{\beta}_{SERG}$ estimator has smaller bias

than $\hat{\beta}_{LRG}$ estimator for $a = 1$ but $\hat{\beta}_{LRG}$ estimator has smaller bias than $\hat{\beta}_{SERG}$ estimator for $a = -1$.

Among the four considered Bayesian risks based SEL the $R_S(\hat{\beta}_{LRG})$ has the smallest values of risk and based LINEX, the $R_L(\hat{\beta}_{LJ})$ has the smallest values of risk for various values of sample size. Thus from a Bayesian perspective we suggest using $\hat{\beta}_{LRG}$ estimator based on SEL and using $\hat{\beta}_{LJ}$ based on LINEX loss function.

The performance of the MLE and UMVUE is also quite satisfactory and in overall non-Bayesian estimators are better than Bayesian estimators, thereby employing of the MLE and UMVUE estimators can be recommend for all practical purposes.

Acknowledgements

Portions of this paper are developed from the authors' earlier conference presentation (Ajami & Jahanshahi, 2016).

References

- Ajami, M. & Jahanshahi, S. M. A. (2016, August 24-26). Comparison of various estimation methods for size-biased Rayleigh distribution. Paper presented at the 13th Iranian Statistical Conference, Shahid Bahonar University of Kerman, Iran.
- Alizadeh, N. H. & Arghami, N. R. (2011). Monte Carlo comparison of five exponentiality tests using different entropy estimates. *Journal of Statistical Computation and Simulation*, 81(11), 1579–1592. doi: 10.1080/00949655.2010.496368
- Bennett, H. S. & Filliben, J. J. A. (2000). A systematic approach for multidimensional, closed form analytic modeling: minority electron mobilities in $\text{Ga}_{1-x}\text{Al}_x\text{As}$ heterostructures, *Journal of Research of the National Institute of Standards and Technology*, 105(3), 441–452. doi: 10.6028/jres.105.037
- Ferguson, T. S. (1996). *A Course In Large Sample Theory*. New York: Chapman and Hall.
- Gupta, R. C. & Keating, J. P. (1985). Relations for reliability measures under length biased sampling, *Scandinavian Journal of Statistics*, 13, 49–56.

- Gupta, R. C. & Kirmani, S. N. U. A. (1990). The role of weighted distributions in stochastic modeling, *Communications in Statistics - Theory and Methods*, 19(9), 3147–3162. doi: [10.1080/03610929008830371](https://doi.org/10.1080/03610929008830371)
- Johnson N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, Vol 1* (2nd Ed.). New York: Wiley.
- Oluyede, B. O. (1999). On inequalities and selection of experiments for length-Biased Distributions. *Probability in the Engineering and Informational Sciences*, 13(2), 169–185. doi: [10.1017/s0269964899132030](https://doi.org/10.1017/s0269964899132030)
- Patil, G. P. & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34(2), 179–184. doi: [10.2307/2530008](https://doi.org/10.2307/2530008)
- Patil, G. P. & Ord, J. K. (1976). On size-biased sampling and related form-invariant weighted distribution. *Sankhya: The Indian Journal Of Statistics B*, 38, 48–61.
- Rayleigh, J. W. S. (1880). On the resultant of a large number of vibrations of the some pitch and of arbitrary phase, *Philosophical Magazine, 5th Series*, 10(60), 73–78. doi: [10.1080/14786448008626893](https://doi.org/10.1080/14786448008626893)
- Reshi, J. A., Ahmed, A. & Mir, K. A. (2014). Characterizations and estimation in the length-biased generalized Rayleigh distribution. *Mathematical Theory and Modeling*, 4(6), 87-98.
- Sanku, D. (2012). Bayesian estimation of the parameter and reliability function of an inverse Rayleigh distribution, *Malaysian Journal of Mathematical Sciences*, 6(1), 113–124.
- Soliman, A. A. (2000). Comparison of LINEX and quadratic Bayes estimators for the Rayleigh distribution. *Communications in Statistics - Theory and Methods*, 29(1), 95–107. doi: [10.1080/03610920008832471](https://doi.org/10.1080/03610920008832471)
- Varian, M. M. (1975). A Bayesian approach to real estate assessment. In S. E. Fienberg and A. Zellner, Ed. *Studies in Bayesian Econometrics and Statistics (In Honor of Leonard J. Savage)*, pp. 195–208. Amsterdam: North Holland/Elsevier.
- Zelen, M. & Feinleib, M. (1969). On the theory of chronic diseases, *Biometrika*, 56(3), 601–614. doi: [10.2307/2334668](https://doi.org/10.2307/2334668)
- Zellner, A. (1986). Bayesian estimation and prediction using loss functions, *Journal of the American Statistical Association*. 81(394), 446–451. doi: [10.2307/2289234](https://doi.org/10.2307/2289234)

Modeling Agreement between Binary Classifications of Multiple Raters in R and SAS

Aya A. Mitani
Boston University
Boston, MA

Kerrie P. Nelson
Boston University
Boston, MA

Cancer screening and diagnostic tests often are classified using a binary outcome such as diseased or not diseased. Recently large-scale studies have been conducted to assess agreement between many raters. Measures of agreement using the class of generalized linear mixed models were implemented efficiently in four recently introduced R and SAS packages in large-scale agreement studies incorporating binary classifications. Simulation studies were conducted to compare the performance across the packages and apply the agreement methods to two cancer studies.

Keywords: Agreement, binary classifications, Cohen's kappa, Fleiss' kappa, generalized linear mixed model, multiple raters

Introduction

Assessing the strength of agreement between physicians' ratings of screening test results is of primary interest because an effective diagnostic procedure is dependent upon high levels of consistency between raters. However, in practice, substantial discrepancies are often observed between physicians' ratings and is considered a major issue in many common screening tests including mammography and diagnosis of invasive bladder cancer (Beam, Conant, & Sickles, 2002; Comprat et al., 2013; Elmore, Wells, Lee, Howard, & Feinstein, 1994; Onega et al., 2012). This has motivated large-scale studies to examine accuracy and agreement between physicians' ratings and to investigate factors that may play an influential role on the consistency of ratings, precipitating a pressing need for statistical methods of agreement that can flexibly accommodate classifications of a large number of raters.

Aya Mitani is a graduate student in the Department of Biostatistics. Email her at: amitani@bu.edu. Kerrie Nelson is a Research Associate Professor of Biostatistics. Email her at: kerrie@bu.edu.

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

The outcome of a patient's screening test may be classified using a binary categorical scale (for example, diseased or not diseased) based upon the physician's (subjective) interpretation of the screening test result. For example, mammographic results are often categorized as requiring recall or no recall of a patient for further testing and bladder cancer images may be classified as indicating invasive or non-invasive cancer (Comprat et al., 2013). In this paper we focus on large-scale agreement studies where more than two raters' classifications are made using a binary categorical scale.

When multiple raters participate in a large-scale agreement study, only a limited number of methods are available to assess agreement between their binary ratings in a unified and comprehensive approach. Summary measures include Fleiss' measure of agreement and Shrout and Fleiss' intraclass correlation coefficient (ICC) (Fleiss & Cuzick, 1979; Fleiss, 1971; Shrout & Fleiss, 1979). Modeling approaches include a Bayesian generalized linear mixed model (GLMM) with nested random effects and an approach based upon GLMMs with crossed random effects (Hsiao, Chen, & Kao, 2011; Nelson & Edwards, 2008, 2010). Log linear models, another modeling approach, are best-suited for modeling agreement between two or three raters (Agresti, 1989; Tanner & Young, 1985).

Due to a lack of statistical methods that can easily be implemented in practice for studies with multiple raters, clinical research papers tend to instead focus on comparing agreement using pairwise approaches (i.e. comparing between each pair of raters at a time) which can be inefficient, lending itself to several summary measures and often complex or disjointed interpretation of agreement (Ciatto et al., 2005; Comprat et al., 2013; Epstein, Allsbrook, Amin, Egevad, & ISUP Grading Committee, 2005; Ooms et al., 2007).

Until recently, various modeling approaches such as Nelson and Edwards' (2008) GLMM-based method have been challenging to implement due to a lack of availability in standard statistical software packages for modeling GLMMs and a necessity for sophisticated programming skills. However, recent advances in statistical software packages including R (R Core Team, 2014) and SAS (Cary, NC: SAS Institute) have led to much improved and efficient procedures for fitting complex models including GLMMs with crossed random effects. In this paper we demonstrate how Fleiss' kappa for multiple raters and Nelson and Edwards' GLMM modeling approach can easily be implemented in four R packages and in SAS software to assess agreement in large-scale studies with binary classifications.

The aim of this study is to compare the performance of the different software packages using extensive simulation studies to assess the impact of normally and non-normally distributed (symmetric and skewed) random effects and sample size

on parameter estimation and the calculation of the agreement statistics. It is motivated by two large-scale agreement studies. The first is a study of 119 community radiologists assessing 109 mammograms as recall or no recall conducted by the Breast Cancer Surveillance Consortium (BCSC) (Onega et al., 2012). The second study conducted by Comprat et al. (2013) involved 8 pathologists reviewing 25 bladder cancer specimens for the presence or absence of invasive cancer. For each of these two studies we implement the different agreement approaches described above in each of the four statistical software packages and assess levels of agreement between the multiple raters. We also demonstrate how the classifications of individual raters can be assessed from their random effect terms.

Models and Measures of Agreement for Multiple Ratere

GLMM Approach An approach based upon GLMMs with a crossed random effects structure can be implemented to assess levels of agreement between multiple raters' binary classifications (Nelson & Edwards, 2008, 2010). This approach, unlike many others, is intended to accommodate the ratings of multiple raters, does not grow increasingly complex as the number of raters increases, and can accommodate missing data where some raters do not classify every test result (Ibrahim & Molenberghs, 2009). Derived from this model is a chance-corrected measure of agreement which incorporates data from the entire sample of subjects. Its value, unlike Cohen's kappa statistics, is robust to the underlying prevalence of the disease. A brief description of the method is following; full details can be found in Nelson and Edwards (2008, 2010). Our setup assumes a sample of J raters ($j = 1, \dots, J$) each independently classifying a sample of I subjects ($i = 1, \dots, I$) generating the set of binary outcomes Y_{ij} , each taking the value 0 or 1.

The binary GLMM with a probit link function and crossed random effects models the probability that a subject's test result is classified as a success, $\Pr(Y_{ij} = 1)$ as follows:

$$\Phi^{-1}\left(\Pr\left(Y_{ij} = 1 \mid u_i, v_j\right)\right) = \eta + u_i + v_j \quad (1)$$

where η is the intercept and u_i and v_j are the random effects for the i^{th} subject and the j^{th} rater, respectively. The subject random effects u_i ($i = 1, \dots, I$) and the rater random effects v_j ($j = 1, \dots, J$) are assumed normally distributed with mean 0 and variances σ_u^2 and σ_v^2 , respectively. A positive random effect value for u_i indicates a test result that is more likely than other test results to be classified as a success

over many raters. A positive value for v_j suggests a rater who is liberal in classifying a subject as a success over their classification of many such test results. The chance-corrected model-based kappa has been derived previously and takes the form

$$\kappa_m = 1 - 4 \int_{-\infty}^{+\infty} \Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right) \left\{1 - \Phi\left(\frac{z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right\} \phi(z) dz \quad 0 \leq \kappa_m \leq 1 \quad (2)$$

with its variance derived using the multivariate delta method as

$$\begin{aligned} \text{var}(\hat{\kappa}_m) = 16 \left\{ \left[\int_{-\infty}^{+\infty} \left(\frac{1}{2\hat{\rho}(1-\hat{\rho})} \right) \left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}} \right) \phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right) \left[1 - 2\Phi\left(\frac{z\sqrt{\hat{\rho}}}{\sqrt{1-\hat{\rho}}}\right) \right] \phi(z) dz \right]^2 \right. \\ \left. \times \left[\left(\frac{\hat{\sigma}_v^2 + 1}{(\hat{\sigma}_T^2)^2} \right)^2 \left(\frac{2\hat{\sigma}_u^4}{I} \right) + \left(\frac{\hat{\sigma}_u^2}{(\hat{\sigma}_T^2)^2} \right)^2 \left(\frac{2\hat{\sigma}_v^4}{I} \right) \right] \right\} \end{aligned}$$

where $\sigma_T^2 = \sigma_u^2 + \sigma_v^2 + 1$ and $\rho = \sigma_u^2 / \sigma_T^2$. Full details on the derivation of κ_m and its variance can be found in Nelson and Edwards (2008, 2010). The summary measure of agreement κ_m takes values between 0 and 1 and is interpreted in a similar manner to Cohen's original kappa where a value close to 0 indicates little or no chance-corrected agreement and values closer to 1 reflect strong chance-corrected agreement between raters (Cohen, 1968; Landis & Koch, 1977).

The marginal likelihood function for the GLMM model takes the form:

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{Y}) = \\ \int \int_{\mathbf{u} \mathbf{v}} \left\{ \prod_{i=1}^I \prod_{j=1}^J \Phi(\eta + u_i + v_j) [1 - \Phi(\eta + u_i + v_j)] \right\} \times \left(\prod_{i=1}^I \frac{e^{-\frac{u_i^2}{2\sigma_u^2}}}{\sqrt{2\pi\sigma_u^2}} \right) \left(\prod_{j=1}^J \frac{e^{-\frac{v_j^2}{2\sigma_v^2}}}{\sqrt{2\pi\sigma_v^2}} \right) d\mathbf{u} d\mathbf{v} \end{aligned}$$

where $\mathbf{Y}$ is the vector of all the binary classifications of all raters.

The inclusion of the crossed random effects leads to a high-dimensional likelihood function, thus no closed form solution for maximizing the marginal likelihood function is available. Hence, approximate maximum likelihood methods are explored for estimating the parameters. Adaptive Gaussian quadrature is not a

viable technique for obtaining approximate maximum likelihood estimates due to the large number of random effects. Instead, estimates of the parameters $\boldsymbol{\theta} = (\eta, \sigma_u^2, \sigma_v^2)$ can be obtained by fitting the GLMM using an approximate maximum likelihood approach such as the Monte-Carlo expectation-maximum (MCEM) algorithm provided in McCulloch (1997) and Kuk and Cheng (1997). These methods based on Monte-Carlo Markov-Chain (MCMC) (Karim & Zeger, 1992; Kuk & Cheng, 1997; McCulloch, 1997) are feasible in obtaining approximate maximum likelihood estimates for these GLMM models, however they often take a large amount of computational programming and running time and are sometimes unstable, not reaching convergence. Recently a multivariate Laplacian approximation technique, which is computationally very efficient and stable, has been implemented in R and SAS for fitting GLMMs with crossed random effects. In the multivariate Laplacian approximation method, large-sample approximate standard errors are estimated by taking the square-roots of the diagonals of matrix $\mathbf{H}$ at convergence, i.e.

$$\text{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\text{diag}\left\{-[\mathbf{H}(\hat{\boldsymbol{\theta}})]^{-1}\right\}}$$

where

$$\mathbf{H} = \frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{u}, \mathbf{v}, \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

is the second-order derivative of the log-likelihood function $l(\boldsymbol{\theta}; \mathbf{u}, \mathbf{v}, \mathbf{y})$ evaluated at the approximate maximum likelihood estimates of $\boldsymbol{\theta}$ and is generated during the model-fitting process.

Fleiss Kappa for Multiple Raters Fleiss (1971) described a generalized Kappa statistic which extends Scott's pi (Scott, 1955) in order to accommodate multiple raters and multiple categories. Later, Fleiss and Cuzick (1979) introduced a version of their kappa statistic for binary classifications with unequal number of ratings per test result. Briefly, it is structured as follows: For I subjects ($i = 1, \dots, I$) under study, let n_i denote the number of raters rating the i^{th} subject and let x_i denote the number of positive ratings on the i^{th} subject. Defining $p_i = x_i / n_i$ as the proportion of positive ratings for each subject,

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

$$\bar{n} = \frac{\sum_i n_i}{I}$$

as the mean number of raters for each subject, and

$$\bar{p} = \frac{\sum_i x_i}{(I\bar{n})}$$

as the overall proportion of positive ratings, the Fleiss' kappa for agreement takes the form

$$\hat{\kappa}_F = 1 - \frac{\sum_i n_i p_i (1 - p_i)}{I(\bar{n} - 1) \bar{p} (1 - \bar{p})} \quad (3)$$

with variance

$$\text{Var}(\hat{\kappa}_F) = \frac{2(\bar{n}_H - 1)}{I\bar{n}_H (\bar{n} - 1)^2} + \frac{(\bar{n} - \bar{n}_H) [1 - 4\bar{p}(1 - \bar{p})]}{I\bar{n}\bar{n}_H (\bar{n} - 1)^2 \bar{p}(1 - \bar{p})} \quad (4)$$

where $\bar{n}_H$ is defined as the harmonic mean number of raters for each subject,

$$\bar{n}_H = \frac{I}{\sum_i \frac{1}{n_i}}$$

When the number of raters per subject is constant, $\hat{\kappa}_F$ is equivalent to the Fleiss kappa statistic introduced by Fleiss in 1971 (1971; Fleiss, Nee, & Landis, 1979; Fleiss & Cuzick, 1979). Fleiss' kappa take values between 0 and 1 and are interpreted in a similar manner to Cohen's original kappa (Cohen, 1968), where 0 indicates no chance-corrected agreement and values closer to 1 suggest strong chance-corrected agreement between the raters. For further details on this summary agreement measures, see Fleiss (1971) and Fleiss and Cuzick (1979). A potential drawback of Fleiss' kappa includes vulnerability to marginal prevalence issues in a similar manner to Cohen's kappa.

Statistical Software Packages in SAS and R

Until recently GLMMs with crossed random effects have been challenging to implement in standard software packages, instead requiring sophisticated programming skills and often computationally intensive algorithms (Kuk & Cheng, 1997; McCulloch, 1997). However, recent advances in SAS and R allow for these models to be fit efficiently by using packages or procedures that do not require programming skills. Four of the available procedures that are capable of fitting GLMMs with crossed random effects allowing for a probit link function in R and SAS are (we will briefly discuss each in turn):

- a) R – `clmm` function in ORDINAL package
- b) R – `glmer` function in LME4 package
- c) R – MCMCglmm package
- d) SAS – GLIMMIX procedure

ORDINAL Package in R The ORDINAL package (Christensen, 2013) was recently added to R and is primarily intended for fitting cumulative mixed models such as ordered regression models, proportional odds and proportional hazards models for grouped survival times, and ordered logit/probit models. The `clmm` function in the ORDINAL package allows GLMMs with crossed random effects to be fitted with a probit link function. Estimation procedures include the Laplace approximation and Gaussian quadrature but we are restricted to the Laplace method to fit our model of interest with crossed random effects. While this package is primarily intended to fit ordinal models, it also provides an efficient approach for estimating parameters in a binary GLMM. For fitting our GLMM of interest, the probit link function and the random effects structure can be specified in the model formula. Solutions to the random effects for subjects and raters are computed based on the conditional modes, the points at which the conditional density of the estimated random effects are maximized. We are not aware of any studies comparing the performance of the ORDINAL package to that of other packages such as LME4.

LME4 Package in R The `glmer` function in LME4 package is perhaps the most widely-used function to fit GLMMs in R. Its default approximation method is the Laplace approximation and the function accommodates crossed random effects. To fit the model of interest, `family = binomial(link = "probit")` and the random effects structure are specified in the model formula. Similarly to the ORDINAL package, the solution to the random effects are computed based on the conditional modes.

MCMCglmm Package in R The above packages use a frequentist approach to fit GLMMs. The MCMCglmm package uses a Bayesian approach and can fit GLMMs with crossed random effects and a probit link function (Hadfield, 2010). Priors for the fixed effects and variance structures for the random effects and residuals need to be specified. In MCMCglmm, the prior distribution for the fixed effects are assumed multivariate normal with the user specifying the parameters, and the prior distribution for both the R-structure for the error distribution and the G-structure for the random effects variance covariance matrices are assumed inverse-Wishart, again with the user specifying the parameters (Hadfield, 2015). The function *posterior.mode* or *posterior.mean* is used to obtain solutions to the random effects for each subject and rater.

GLIMMIX Procedure in SAS In a similar manner to the ORDINAL and LME4 packages in R, the GLIMMIX procedure in SAS relies on the Laplace approximation for estimation of GLMMs with crossed random effects. The solution to the random effects are again computed based on the conditional modes.

Another procedure in SAS that fits GLMMs is the NLMIXED procedure. The NLMIXED procedure estimates the parameters by integral approximation methods through adaptive Gaussian quadrature. However, at present, the procedure cannot accommodate a crossed random effects structure so it will not be examined here.

Methodology

Although the LME4, MCMCglmm, and PROC GLIMMIX packages were described for estimation in various binary GLMM models (Kim, Choi, & Emery, 2013; Li, Lingsma, Steyerberg, & Lesaffre, 2011; Zhang et al., 2011), the performance of the ORDINAL package has not yet been reported for binary outcomes nor for the calculation of agreement measures. Our focus in this paper is to explore the use of these four aforementioned packages in R and SAS to calculate the measures of agreement for multiple raters classifying test results using a binary scale. To achieve this, we conducted extensive simulation studies to compare the performance of the four packages with regards to estimation of GLMM model parameters and the summary agreement measures. One important motivation for conducting these simulation studies is to ensure that reasonably unbiased estimates of the model-based measure of agreement κ_m are obtained from the existing packages.

Simulation studies were conducted under scenarios that varied in sample size (number of subjects and raters), random effects components, distributions of the random effects, and the choice of priors for the MCMCglmm Bayesian method. The various simulation scenarios we explored are displayed in [Table 1](#). Part I of the simulations had normally-distributed random effects, while parts II and III had non-normally-distributed random effects. In part II, the random effects were symmetric (mixture of two normal distributions and uniform distribution) and, in part III, at least one of the random effects were skewed (exponential, Gamma or chi-squared distribution). For each part of the simulations, we evaluated four scenarios. The first scenario (Scenario 1) resembled the BCSC breast cancer data set to verify that our methods perform well in this setting and others ($\eta = -0.1$, $\sigma_u^2 = 1.5$, $\sigma_v^2 = 0.2$). In Scenario 2, the variance of the rater random effects was set to be larger than the variance of the subject random effects ($\eta = 1$, $\sigma_u^2 = 1$, $\sigma_v^2 = 5$). In Scenario 3, the variance of the subject random effects was set to be larger than the variance of the rater random effects ($\eta = 1$, $\sigma_u^2 = 5$, $\sigma_v^2 = 1$). In Scenario 4, the variances of both random effects were set as large ($\eta = 1$, $\sigma_u^2 = 10$, $\sigma_v^2 = 10$). Regardless of the random effects distribution, the variances of the subject and rater random effects were kept constant for each scenario (i.e. for Scenario 1, the variance of the subject random effects was set as 1.5 for normal, non-normal symmetric, and skewed distributed random effects). Within each scenario, one was larger in sample size with 150 subjects and 100 raters (Scenario #a) while the other was smaller with 100 subjects and 50 raters (Scenario #b). [[Table 1](#)]

For each simulation scenario in part I (normally distributed random effects), one thousand datasets were generated using R in the following manner: First, I subject random effects and J rater random effects were randomly generated from $N(0, \sigma_u^2)$ and $N(0, \sigma_v^2)$ distributions, respectively. For each $(ij)^{\text{th}}$ observation, the probability of the j^{th} rater correctly classifying the i^{th} subject was generated according to the ordinal probit GLMM

$$p_{ij} = \Pr(Y_{ij} = 1 | u_i, v_j) = \Phi(\eta + u_i + v_j), \quad i = 1, \dots, I; j = 1, \dots, J$$

using the *qnorm* function in R.

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Table 1. List of parameters used to generate simulated data sets for each scenario
(Number of simulations per scenario = 1,000)

Scenario	<i>I</i>	<i>J</i>	True η	Distribution of u_i^*	Distribution of v_j^*
I. Normally distributed random effects					
1a	150	100	-0.1	N(0, 1.5)	N(0, 0.2)
1b	100	50			
2a	150	100	1	N(0, 1)	N(0, 5)
2b	100	50			
3a	150	100	1	N(0, 5)	N(0, 1)
3b	100	50			
4a	150	100	1	N(0, 10)	N(0, 10)
4b	100	50			
II. Non-normally distributed random effects (Symmetric)					
1a	150	100	-0.1	0.5N(-1, 0.5) + 0.5N(1, 0.5)	Unif(-0.775, 0.775)
1b	100	50			
2a	150	100	1	0.5N(-0.8, 0.36) + 0.5N(0.8, 0.36)	Unif(-3.87, 3.87)
2b	100	50			
3a	150	100	1	0.5N(-2, 1) + 0.5N(2, 1)	Unif(-1.73, 1.73)
3b	100	50			
4a	150	100	1	0.5N(-3, 1) + 0.5N(3, 1)	Unif(-5.48, 5.48)
4b	100	50			
III. Non-normally distributed random effects (Skewed)					
1a	150	100	-0.1	$\text{Exp}\left(\sqrt{1/1.5}\right)$	$\text{Gamma}\left(4, \sqrt{20}\right)$
1b	100	50			
2a	150	100	1	N(0, 1)	Gamma(5, 1)
2b	100	50			
3a	150	100	1	Gamma(5, 1)	Unif(-1.73, 1.73)
3b	100	50			
4a	150	100	1	$\chi^2_{df=5}$	N(0, 10)
4b	100	50			

Note: * Mean and variance shown for normal distributions, N(μ , σ^2)

A binary classification Y_{ij} was then randomly generated for each observation from the corresponding Bernoulli distribution with probability p_{ij} . To assess the impact of a misspecified random effects distribution in GLMM, we also generated

data with non-normally-distributed subject and rater random effects (Litière, Alonso, & Molenberghs, 2008). In the symmetric non-normal random effects scenarios, the random effects of the subjects were randomly sampled from a symmetric mixture of two normal distributions with mean of 0 and the same variance as the corresponding normal distribution. Each i^{th} subject was assigned a number generated from uniform distribution with (0, 1) support. If the assigned number was less than 0.5, the random effect of the subjects was sampled from the first of the two normal distributions. Otherwise, the random effect of the subjects was sampled from the second of the two normal distributions. The rater random effects were randomly sampled from a uniform distribution with mean of 0 and the same variance as the corresponding normally distributed random effects. In the skewed random effects scenarios, the random effects of the subjects and raters were randomly sampled from a combination of various skewed distributions (exponential, Gamma, and chi-squared) and normal and uniform distributions. For the true random effects distribution to have mean 0, an assumption of GLMM, each of the skewed random effects distributions was centered by subtracting its true mean value. See Table 1 for the parameters and distributions of random effects used in each set of scenarios.

The binary GLMM in equation (1) was then fitted to each of the one thousand simulated datasets using each of the four statistical packages (PROC GLIMMIX, LME4, ORDINAL, and MCMCglmm). With the MCMCglmm package, two different sets of priors were used for each scenario. We specified the variances of the subject and rater random effect terms to follow an inverse-Wishart (IW) distribution, which is comprised of two parameters: the scale parameter V , and the degree of freedom parameter ν , also referred to as the degree of belief parameter. For the first set of priors, denoted by “MCMCglmm1”, we let the variance of the random effects follow an IW distribution with $V = 1$ and $\nu = 1$, and for the second set of priors, denoted by “MCMCglmm10”, we let the variance of the random effects follow an IW distribution with $V = 10$ and $\nu = 1$. Under Scenario 1a, we also used the uninformative prior specification with $V = 1$ and $\nu = 0.002$ which is used frequently for variance structures (Hadfield, 2015).

The GLMM parameters of interest estimated for each dataset were η , σ_u^2 , and σ_v^2 . These parameter estimates were then used to compute the model-based measure of agreement, $\hat{\kappa}_m$, and its variance, $\text{Var}(\hat{\kappa}_m)$. Fleiss’ agreement measure $\hat{\kappa}_F$ was also calculated for each dataset.

Results

Simulation results from normally-distributed and symmetric non-normally-distributed random effects datasets are presented in Tables 2 and 3, respectively, for large sample size ($I = 150, J = 100$). Results from skewed random effects are presented in Supplementary Table 3. For each simulation scenario, the mean of the 1,000 estimates (Mean Estimate) and the mean of the 1,000 model-based standard errors (Mean SE) estimated for each of the model parameters, η , σ_u^2 , and σ_v^2 from each of the four software packages are reported. The mean of the 1,000 estimates and standard errors for measure of agreement κ_m are also reported for each set of simulations. The coverage probability (the percent of times the 95% confidence interval for $\hat{\kappa}_m$ included the true κ_m value) of κ_m over the 1,000 simulated datasets is also reported for each of the four statistical packages, as well as the convergence rate of the GLMM based on the number of times the model was able to produce the standard errors for σ_u^2 , and σ_v^2 estimates. Also, the mean estimated Fleiss' kappa ($\hat{\kappa}_F$) and the mean standard error for each simulation scenario are reported. The focus is on results from scenarios with large sample size ($I = 150, J = 100$). Simulation results from scenarios with small sample size ($I = 100, J = 50$) followed a similar pattern to those from scenarios with large sample size. Full details of the simulation results of small sample size can be viewed in Supplementary Tables 1, 2, and 4.

GLMM Parameter Estimates

Minimal biases were observed in the estimation of η across the four packages when the random effects were normally distributed. Slightly larger biases were observed under the scenarios with non-normal random effects and when one of the variance components, σ_u^2 or σ_v^2 , was 5 and the other was 1 (Scenarios 2 and 3). These biases tended to be larger under the MCMCglmm package for both sets of priors. Biases in the estimation of η were largest under the scenarios with skewed random effects but varied little among the different packages. [Supplementary Table 3] Due to the model format used in its package, the η estimates produced from the ORDINAL package have an opposite sign from those produced from other packages. To make the comparison between packages easier, we present η estimates with consistent signs in the tables.

Observe more variability in biases of the random effects variance component estimates between the different packages. Generally, with normally-distributed random effects, ORDINAL, LME4, and PROC GLIMMIX tended to slightly

underestimate σ_u^2 and σ_v^2 while MCMCglmm1 and MCMCglmm10 tended to overestimate them. For example, under Scenario 1a, σ_u^2 were 1.492, 1.500, and 1.493 for ORDINAL, LME4, and PROC GLIMMIX, respectively, while they were 1.530 and 1.613 for MCMCglmm1 and MCMCglmm10, respectively. [Table 2] For the symmetric non-normal random effects, most packages overestimated σ_u^2 , and σ_v^2 under Scenarios 1 and 2. Under Scenario 3 ($\sigma_u^2 = 5$ and $\sigma_v^2 = 1$), ORDINAL, LME4, and PROC GLIMMIX estimated σ_v^2 with minimal bias (0.993, 0.999, and 0.993, respectively) but overestimated σ_u^2 (5.816, 5.758, and 5.816, respectively). MCMCglmm1 and MCMCglmm10 also overestimated σ_u^2 (6.263 and 6.386, respectively). [Table 3] For the skewed random effects, all packages tended to overestimate the larger of the two variances under Scenarios 1, 2, and 3. Under Scenario 4, all packages underestimated σ_u^2 while σ_v^2 was estimated with smaller biases. [Supplementary Table 3]

The ORDINAL package and the GLIMMIX procedure produced identical GLMM parameter estimates to the third decimal place confirming that these two packages employ virtually identical multivariate Lapacian procedures. With the exception of LME4, the other three packages exhibited very stable estimation procedures with usually a 100% convergence success rate over each set of 1,000 simulated data sets, for both normally- and non-normally-distributed random effects. The LME4 package proved to be consistently less stable compared to all the other packages, with convergence rates ranging from 79.8% to 99.9%. In particular, convergence rate for LME4 tended to be worse for simulation scenarios with large random effects variances and for non-normally-distributed random effects distribution (symmetric and skewed). The average time to fit one GLMM for the larger data set was 9, 8, 109, 104, and 27 seconds for ORDINAL, LME4, MCMCglmm1, MCMCglmm10, and PROC GLIMMIX, respectively, indicating that all four packages were able to fit these models in a computationally efficient manner.

Agreement Measures

The parameter κ_m was estimated with minimal bias in all simulation scenarios and across all four packages and various values of σ_u^2 and σ_v^2 when the random effects were normally distributed. In general, observe slightly larger bias under simulations with non-normally-distributed random effects compared to those with normally-distributed random effects (symmetric and skewed).

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Shown in Figure 1 is the relationship between mean absolute bias and coverage probability based on the 1,000 κ_m estimates for each statistical package and for each simulation scenario. The results from PROC GLIMMIX are omitted because they were identical to those from ORDINAL. The dotted line across the horizontal axis represent bias at 0 and the dotted line across the vertical axis represent coverage probability at 95%. An ideal situation is when the estimate falls on the intersection between the two dotted lines. In general, the mean absolute bias was lowest under scenarios with normal random effects, slightly larger under scenarios with symmetric non-normal random effects, and largest under scenarios with skewed random effects. For scenarios with normally-distributed random effects, the coverage probabilities were consistently close to the anticipated 95% (90-95% for all packages). [Table 2] For scenarios with symmetric non-normal random effects, coverage probabilities were slightly higher than the anticipated 95% under Scenario 1 (97.1%-98.1%) and Scenario 4 (98.7%-98.7%), while they were slightly lower than anticipated under Scenario 2 (82.6%-93.0%) and Scenario 3 (64.8%-87.5%). [Table 3] For scenarios with skewed random effects, the coverage probabilities were lower, especially under the extreme case scenarios, Scenario 1 and Scenario 3, where both the random effects distributions were highly skewed. More specifically, under Scenario 1 where the subject and rater random effects followed an exponential distribution and a Gamma distribution respectively, coverage probability ranged from 37.5% to 40.9% amongst all packages. Under Scenario 3 where the subject and rater random effects followed a Gamma distribution and a uniform distribution respectively, coverage probability ranged from 52.9% to 65.2% amongst all packages. [Figure 1; Supplementary Table 3]

Note the largest differences in mean absolute bias and coverage probability between the four packages under Scenario 3, when $\sigma_u^2 = 5$ and $\sigma_v^2 = 1$. For symmetric non-normal random effects, ORDINAL (same as PROC GLIMMIX) and LME4 yielded lower mean absolute biases (0.007 and 0.010, respectively) and higher coverage probabilities (84.1% and 87.5%, respectively) compared with MCMCglmm1 (mean absolute bias = 0.019, coverage probability = 64.8%) and MCMCglmm10 (mean absolute bias = 0.121, coverage probability = 74.0%). However, for skewed random effects, MCMCglmm1 and MCMCglmm10 yielded lower mean absolute biases (0.041 and 0.045, respectively) and higher coverage probabilities (65.2% and 59.9%, respectively) compared to ORDINAL/PROC GLIMMIX (mean absolute bias = 0.048, coverage probability = 54.5%) and LME4 (mean absolute bias = 0.049, coverage probability = 52.9%).

Table 2. Mean estimates and mean standard errors (SEs) from 1,000 simulations for the probit GLMM and agreement statistics computed from each statistical package with normally distributed random effects, $I = 150$ and $J = 100$

Scenario	Parameter	Truth	Statistical Package				
			ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
			Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)
1a	GLMM parameters:						
	η	-0.1	-0.103 (0.110)	-0.103 (0.110)	-0.103 (0.112)	-0.104 (0.119)	-0.103 (0.110)
	σ_u^2	1.5	1.492 (0.186)	1.500 (0.184)	1.530 (0.193)	1.613 (0.203)	1.493 (0.186)
	σ_v^2	0.2	0.198 (0.031)	0.199 (0.031)	0.213 (0.033)	0.316 (0.047)	0.198 (0.031)
	Agreement measures:						
	Model-based Kappa, κ_m	0.375	0.373 (0.022)	0.374 (0.033)	0.376 (0.022)	0.370 (0.023)	0.373 (0.022)
	Fleiss Kappa, κ_F				0.373 (0.001)		
	Coverage probability of κ_m (%)		93.2	93.3	92.4	94.2	93.2
2a	GLMM parameters:						
	η	1	1.012 (0.238)	1.010 (0.236)	1.015 (0.243)	1.020 (0.247)	1.012 (0.238)
	σ_u^2	1	0.999 (0.125)	1.006 (0.127)	1.014 (0.128)	1.085 (0.135)	0.999 (0.125)
	σ_v^2	5	4.791 (0.774)	4.771 (0.698)	5.061 (0.827)	5.214 (0.852)	4.791 (0.774)
	Agreement measures:						
	Model-based Kappa, κ_m	0.091	0.095 (0.013)	0.096 (0.013)	0.093 (0.013)	0.096 (0.014)	0.095 (0.013)
	Fleiss Kappa, κ_F				0.083 (0.001)		
	Coverage probability of κ_m (%)		94.1	94.6	93.1	94.7	94.1
	GLMM convergence rate (%)		100.0	99.0	100.0	1000.0	100

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Table 2 (continued)

Scenario	Parameter	T ruth	Statistical Package				
			ORDINAL	LME4	MCMCglimm1	MCMCglimm10	PROC GLIMMIX
			Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)
3a	GLMM parameters:						
	η	1	0.995 (0.211)	0.992 (0.208)	0.999 (0.215)	1.002 (0.219)	0.995 (0.211)
	σ_u^2	5	4.849 (0.657)	4.815 (0.637)	5.122 (0.703)	5.230 (0.718)	4.849 (0.657)
	σ_v^2	1	0.998 (0.149)	1.005 (0.151)	1.023 (0.155)	1.124 (0.169)	0.998 (0.149)
	Agreement measures:						
	Model-based Kappa, κ_m	0.506	0.500 (0.025)	0.498 (0.025)	0.507 (0.025)	0.502 (0.026)	0.500 (0.025)
	Fleiss Kappa, κ_F				0.497 (0.001)		
	Coverage probability of κ_m (%)		91.9	92.1	90	90.8	91.9
	GLMM convergence rate (%)		100	96.3	100	100	100
	4a	GLMM parameters:					
η		1	0.999 (0.409)	1.003 (0.409)	0.999 (0.412)	0.999 (0.415)	0.999 (0.409)
σ_u^2		10	10.013 (1.273)	10.101 (1.361)	10.191 (1.302)	10.275 (1.305)	10.013 (1.273)
σ_v^2		10	9.913 (1.501)	10.009 (1.563)	10.151 (1.558)	10.258 (1.566)	9.912 (1.501)
Agreement measures:							
Model-based Kappa, κ_m		0.316	0.319 (0.031)	0.319 (0.031)	0.318 (0.031)	0.318 (0.031)	0.319 (0.031)
Fleiss Kappa, κ_F					0.312 (0.001)		
Coverage probability of κ_m (%)			94.6	94.7	93.6	93.9	94.6
GLMM convergence rate (%)			100	93.5	100	100	100

Table 3. Mean estimates and mean standard errors from 1,000 simulations for the probit GLMM and agreement statistics computed from each statistical package with symmetric non-normally distributed random effects, $I = 150$ and $J = 100$

Scenario	Parameter	Truth	Statistical Package				
			ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
			Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)
1a	GLMM parameters:						
	η	-0.1	-0.103 (0.112)	-0.104 (0.112)	-1.104 (0.114)	-0.104 (0.120)	-0.103 (0.112)
	σ_u^2	1.5	1.554 (0.189)	1.564 (0.188)	1.588 (0.196)	1.669 (0.205)	1.554 (0.189)
	σ_v^2	0.2	0.200 (0.031)	0.201 (0.031)	0.214 (0.033)	0.317 (0.048)	0.200 (0.031)
	Agreement measures:						
	Model-based Kappa, κ_m	0.375	0.381 (0.022)	0.382 (0.022)	0.383 (0.022)	0.377 (0.023)	0.381 (0.022)
	Fleiss Kappa, κ_F				0.421 (0.001)		
	Coverage probability of κ_m (%)		97.5	97.3	97.1	98.1	97.5
GLMM convergence rate (%)		100.0	100.0	100.0	100.0	100.0	
2a	GLMM parameters:						
	η	1	1.104 (0.260)	1.095 (0.254)	1.112 (0.265)	1.119 (0.269)	1.104 (0.260)
	σ_u^2	1	0.999 (0.125)	1.006 (0.132)	1.015 (0.128)	1.086 (0.136)	0.999 (0.125)
	σ_v^2	5	5.655 (0.921)	5.609 (0.804)	6.073 (0.995)	6.222 (1.018)	5.655 (0.921)
	Agreement measures:						
	Model-based Kappa, κ_m	0.091	0.084 (0.012)	0.085 (0.012)	0.081 (0.012)	0.084 (0.012)	0.084 (0.012)
	Fleiss Kappa, κ_F				0.063 (0.001)		
	Coverage probability of κ_m (%)		91.3	93	82.6	91.1	91.3
GLMM convergence rate (%)		100.0	92.3	100.0	100.0	100.0	

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Table 3 (continued)

Scenario	Parameter	Truth	Statistical Package				
			ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
			Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)	Mean Estimate (Mean SE)
3a	GLMM parameters:						
	η	1	0.995 (0.211)	0.992 (0.208)	0.999 (0.215)	1.002 (0.219)	0.995 (0.211)
	σ_u^2	5	4.849 (0.657)	4.815 (0.637)	5.122 (0.703)	5.230 (0.718)	4.849 (0.657)
	σ_v^2	1	0.998 (0.149)	1.005 (0.151)	1.023 (0.155)	1.124 (0.169)	0.998 (0.149)
	Agreement measures:						
	Model-based Kappa, κ_m	0.506	0.500 (0.025)	0.498 (0.025)	0.507 (0.025)	0.502 (0.026)	0.500 (0.025)
	Fleiss Kappa, κ_F				0.497 (0.001)		
	Coverage probability of κ_m (%)		91.9	92.1	90	90.8	91.9
GLMM convergence rate (%)		100.0	96.3	100.0	100.0	100.0	
4a	GLMM parameters:						
	η	1	1.025 (0.394)	1.030 (0.395)	1.028 (0.397)	1.027 (0.400)	1.025 (0.394)
	σ_u^2	10	8.970 (1.141)	9.058 (1.242)	9.091 (1.158)	9.173 (1.164)	8.970 (1.141)
	σ_v^2	10	9.413 (1.434)	9.493 (1.488)	9.666 (1.487)	9.789 (1.505)	9.413 (1.434)
	Agreement measures:						
	Model-based Kappa, κ_m	0.316	0.307 (0.031)	0.307 (0.031)	0.305 (0.031)	0.305 (0.031)	0.307 (0.031)
	Fleiss Kappa, κ_F				0.324 (0.001)		
	Coverage probability of κ_m (%)		98.7	98.7	97.8	98.1	98.7
GLMM convergence rate (%)		100.0	92.1	100.0	100.0	100.0

Figure 1. Absolute mean bias and coverage probability of estimated model-based kappa for each statistical package by scenario

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Figure 2. Density of model-based kappa measure of agreement estimates from each statistical package by varying sample size and random effects distribution for scenario 1

Interestingly, small to moderate biases in the GLMM parameter estimates had little noticeable impact on the estimates of the agreement measure κ_m . For example, under one of the scenarios with normally-distributed random effects (Scenario 1a), the estimates for σ_u^2 and σ_v^2 under the ORDINAL package were 1.492 and 0.198, respectively, while under MCMCglmm10, they were 1.613 and 0.316. Even with

such seemingly different estimates, both packages produced similar κ_m estimates (0.373 under ORDINAL and 0.370 under MCMCglmm10).

Shown in Figure 2 are the density of κ_m estimates from the simulation scenario with $\sigma_u^2 = 1.5$ and $\sigma_v^2 = 0.2$ (Scenario 1; normal, symmetric non-normal, and skewed random effects distributions). Again, the results from PROC GLIMMIX are omitted because they were identical to those from ORDINAL. The densities of κ_m estimates obtained from all set of simulations were examined using plots, and found to be symmetric and reasonably bell-shaped, centered around the true value of κ_m for normal and symmetric non-normal random effects distributions. For skewed random effects distribution, the density of κ_m estimates appeared to be symmetric and bell-shaped but off-centered with a wider spread. Within each type of random effects distributions, the densities of κ_m estimates were extremely similar across the four packages. Similar densities of κ_m estimates were obtained from other simulation scenarios.

The empirical standard errors, computed as the standard deviation of the 1000 estimated κ_m , were comparable to the means of the model-based standard errors (Mean SE) presented in Tables 2 and 3. In general, when the random effects distribution was normal or skewed, the empirical standard errors were equal to or slightly larger than the model-based standard errors. On the other hand, when the random effects distribution was symmetric non-normal, the empirical standard errors were equal to or smaller than the model-based standard errors.

Fleiss' kappa estimates ($\hat{\kappa}_F$) were comparable to model-based kappa estimates ($\hat{\kappa}_m$) in the majority of scenarios under normally distributed random effects. When the random effects distribution was symmetric non-normal, we observed slightly larger differences between $\hat{\kappa}_F$ and $\hat{\kappa}_m$. For example, under symmetric non-normal Scenario 1a ($\sigma_u^2 = 1.5$ and $\sigma_v^2 = 0.2$), the mean of $\hat{\kappa}_F$ was 0.421, while the means of $\hat{\kappa}_m$ ranged from 0.377 to 0.383 depending on the package. [Table 3] Under the scenarios with skewed random effects, the mean $\hat{\kappa}_F$ and $\hat{\kappa}_m$ were also comparable except under Scenario 3 ($\sigma_u^2 = 5$ and $\sigma_v^2 = 1$) where the mean of $\hat{\kappa}_F$ was 0.438 while the means of $\hat{\kappa}_m$ ranged from 0.459 to 0.466 depending on the package. The mean standard errors of $\hat{\kappa}_F$ computed using equation (4) were extremely small, ranging from 0.001 to 0.003 depending on the sample size. However, the empirical standard errors for Fleiss' kappa ranged from

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

0.026 to 0.055, suggesting that the theoretical standard error potentially underestimates the variability of Fleiss' kappa statistic. This is a topic that needs to be further examined.

Applications to Large-Scale Cancer Studies

Mammogram Screening Study One of the two data sets used for illustration is from a previously-published study conducted by the BCSC, the Assessing and Improving Mammography (AIM) study, where radiologists evaluated whether a subject should be recalled or not based upon their screening mammogram results (Onega et al., 2012). In brief, the AIM study recruited 119 radiologists and obtained a set of 130 mammograms from 6 breast screening registries. The investigators developed 4 mammogram test sets, each containing 109 mammograms sampled from a set of 130 mammograms. Each test set varied by cancer prevalence and case difficulty, and included more cancer cases than a standard screening set; thus recall rates cannot be compared to a standard screening study. Participating radiologists were randomly assigned to one of the test sets and classified the mammograms in their test set. The primary outcome measured on each patient was a binary measure of whether the patient should be recalled for further testing versus no recall. See Onega et al. for further details on the AIM study design.

The aims are to assess the levels of agreement between the study radiologists using the two measures of agreement and to compare these results between the four available statistical packages. The data set was fit in all four packages.

Table 4. Estimates and standard errors for the probit GLMM and agreement statistics computed from each statistical package on the AIM data set

Parameters	Statistical Package				
	ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
η	-0.124 (0.114)	-0.125 (0.114)	-0.121 (0.113)	-0.116 (0.125)	-0.124 (0.114)
σ_u^2	1.431 (0.192)	1.444 (0.189)	1.494 (0.205)	1.559 (0.218)	1.431 (0.192)
σ_v^2	0.195 (0.029)	0.195 (0.029)	0.207 (0.033)	0.295 (0.040)	0.195 (0.029)
κ_m (95% CI)	0.367 (0.321-0.413)	0.368 (0.322-0.414)	0.373 (0.326-0.420)	0.368 (0.321-0.415)	0.367 (0.321-0.413)
κ_F (95% CI)			0.358 (0.356-0.361)		

Table 5. Estimates and standard errors for the probit GLMM and agreement statistics computed from each statistical package on bladder cancer data set

Parameters	Statistical Package				
	ORDINAL	LME4	MCMCglmm1	MCMCglmm10	PROC GLIMMIX
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
η	0.490 (0.460)	0.499 (0.461)	0.622 (0.502)	0.613 (0.763)	0.490 (0.460)
σ_u^2	3.137 (1.492)	3.156 (1.452)	6.114 (1.898)	5.853 (3.345)	3.137 (1.492)
σ_v^2	0.369 (0.274)	0.366 (0.275)	0.723 (0.575)	2.508 (1.587)	0.369 (0.274)
κ_m (95% CI)	0.490 (0.375-0.605)	0.492 (0.377-0.607)	0.570 (0.449-0.691)	0.430 (0.259-0.601)	0.490 (0.375-0.605)
κ_F (95% CI)			0.465 (0.391-0.539)		

Table 4 presents the estimated parameters with the standard errors from the GLMM model, the model-based kappa values with 95% CI, and the Fleiss kappa value ($\hat{\kappa}_F$) with 95% CI for this study. The version of Fleiss' kappa for unequal number of raters per subject was used because subjects' mammograms were classified by different number of raters. The model-based kappa $\hat{\kappa}_m$ produced slightly higher estimates compared to Fleiss' kappa in all four packages. For the model-based approaches, ORDINAL, LME4, MCMCglmm10, and PROC GLIMMIX produced extremely comparable results ($\hat{\kappa}_m = 0.367, 0.368, 0.368$, and 0.367 , respectively) indicating fair agreement between the radiologists. The kappa value obtained from MCMCglmm1 was slightly higher ($\hat{\kappa}_m = 0.373$), but not enough to alter the inference and conclusion of the agreement. Fleiss' kappa ($\hat{\kappa}_F = 0.358$) was estimated slightly lower than the model-based kappa estimates $\hat{\kappa}_m$.

One of the simulation scenarios (Scenario 1) was designed to resemble the BCSC breast cancer data set. Under normally distributed random effects, the biases and coverage probabilities of $\hat{\kappa}_m$ were comparable between the packages. [Figure 1] Slightly more variability in bias was observed under non-normally-distributed random effects. Bias of $\hat{\kappa}_m$ obtained from MCMCglmm1 was the highest (0.008) while the bias obtained from MCMCglmm10 was the lowest (0.002). [Figure 1]

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Figure 3. Rater- and subject-specific random effects from breast cancer data set

Bladder Cancer Study

The second data set used for illustration is a study carried out by Comprat et al. (2013) which assessed agreement among eight genitourinary pathologists reviewing twenty-five bladder cancer specimens. Each pathologist provided a binary classification for each specimen according to whether or not they considered the sample to be non-invasive or invasive bladder cancer.

Figure 4. Rater- and subject-specific random effects from bladder cancer data set

This data set was fit using the four packages and calculated the two agreement measures ($\hat{\kappa}_m, \hat{\kappa}_F$). Model-based kappa estimates $\hat{\kappa}_m$ obtained from ORDINAL, LME4, and MCMCglmm packages with the smaller prior were higher compared to the Fleiss' kappa estimate ($\hat{\kappa}_F = 0.465$), which corroborate the original study value of moderate agreement between study pathologists. [Table 5] Results from the MCMCglmm package yielded an especially higher kappa estimate with the smaller

prior ($\hat{\kappa}_m = 0.570$) and a lower kappa estimate with the larger prior ($\hat{\kappa}_m = 0.430$) relative to the estimates from the other packages. Compared to the previous AIM data set example, this data set provided a wider range of $\hat{\kappa}_m$ computed by the different packages, with the lowest and highest kappa values as 0.430 (MCMCglmm1) and 0.570 (MCMCglmm10), respectively. In a similar manner to our simulations, ORDINAL, LME4, and PROC GLIMMIX provided equivalent kappa estimates. However, all packages indicated that the pathologists had moderate agreement.

Unique Characteristics of Raters and Test Results Each statistical package can generate subject- and rater-specific random effects based on the conditional modes of the conditional distributions for the random effects. These solutions to the random effects are useful in understanding the behavior of individual raters if, for example, a rater is liberal or conservative in their classification of the test results. We present the solutions to the random effects from the ORDINAL package, and similar solutions were obtained from PROC GLIMMIX.

Presented in [Figure 3](#) are the rater-specific random effects with 95% CI and the subject-specific random effects with 95% CI for the AIM study. Radiologists with large positive random effects values tended to recall mammograms more aggressively compared to other raters. However, radiologists with large negative random effects values were less likely to recall mammograms relative to other raters. For example, the radiologist with ID 22 who had the largest rater random effect ($\hat{v}_{22} = 1.07$) recalled 71% of the mammograms that he/she classified while the average recall rate among all radiologists was 43%. [[Figure 3a](#)] The subject-specific random effects ranged from -2.08 to 2.82. Large positive random effects values indicate mammograms with a high probability of recall while large negative values indicate mammograms with low probability of recall. Values that are close to 0 indicate mammograms with ambiguous results and suggest that the disease status on these mammograms was less well-defined than others. For example, subjects with IDs 136 and 147 had the largest random effects ($\hat{u}_{136} = \hat{u}_{147} = 2.82$) and they both had a recall rate of 100% while subject with ID 103 with the smallest random effect ($\hat{u}_{103} = -2.08$) had a recall rate of 2%. [[Figure 3b](#)]

Displayed in [Figure 4](#) are the random effects conditional modes for the bladder cancer study. The rater-specific random effects were all moderate in value, ranging from -0.527 to 0.657. Relative to other pathologists, pathologists 1 and 3 were more likely to categorize the specimens as invasive (more liberal) while

pathologists 8 and 4 were less likely to do so (more conservative). [Figure 4a] The subject-specific random effects ranged from -2.345 to 1.614. Subjects with large positive values of random effects (IDs 6-25) suggest having a more clear indication of invasive cancer compared to other subjects. On the other hand, subjects with large negative values of random effects (IDs 4-14) suggest that their samples indicate a non-invasive cancer. [Figure 4b] Note that many rater- and subject-specific random effects are equal to others due to the small number of raters and test results in this study.

Conclusion

The performance of four different packages in R and SAS was compared in the estimation of parameters for the binary GLMM and for two available measures of agreement between multiple raters. The GLMM parameter estimates were similar between the four packages when the random effects were normally distributed, especially between the packages that use a frequentist approach (ORDINAL, LME4, and PROC GLIMMIX). For one of the scenarios (Scenario 1a), the Bayesian package (MCMCglmm) was explored further by altering the belief parameter (ν) to 0.002 which is used regularly in the prior specification of the random effects variance structure (Hadfield, 2015). Changing the specification of the priors had a minimal impact on the estimation of the random effects parameters and on the agreement statistic in the Bayesian package (MCMCglmm). When the random effects were non-normally distributed (both symmetric and skewed), we observed more variability in the GLMM parameter estimates between the four packages. However, we observed considerably smaller variability in the model-based agreement estimates even when the difference in the GLMM parameter estimates between the packages were relatively large.

It was shown in many studies misspecification of the random effects distributions do not seriously affect the estimation of the fixed effects. In computing the model-based kappa statistic from GLMM, however, the interest is in estimating the variances of the subject and rater random effects. Fewer studies have evaluated the impact of model misspecification on the random effects estimates and variance components. Through simulation, Agresti, Caffo, and Ohman-Strickland (2004) showed that extreme departure from Gaussian of the random effects may lead to loss of efficiency in the estimated variance of the random effects when fitting binary GLMM. If the true variance of the random effects is small, however, the problem of misspecification is negligible even if the true distribution is not Gaussian. In their simulation study, Litierie et al. (2008) assessed the impact of misspecified random

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

effects distribution under binary GLMM on the maximum likelihood estimate of the random effects variance component. They observed that substantial bias can occur under misspecification even if the true variance of the random effects is small. On the other hand, McCulloch and Neuhaus (2011) showed that the estimation of random effects variance components is robust to misspecification of the random effects distribution. In our simulation study, we did observe slightly higher bias in the estimated variance of the random effects when the true random effects distribution were skewed compared to when the true random effects distribution was normal. This was more pronounced under the extreme scenarios where both the subject and rater random effects were non-normally distributed. Litierie et al. (2008) also noted that a more serious bias can be observed with more than one random effects in the model. However, the absolute bias in the model-based kappa estimates, which takes values between 0 and 1, was generally low (0.06 or less) even for these extreme scenarios across the four packages.

Typically used as an approach to measure reliability among multiple judges, the intra-class correlation coefficient (ICC) is another popular summary statistic for assessing agreement. Fleiss and Cuzick (1979) show that if the sample size is moderately large, ICC is “virtually identical” to kappa.” (p. 539) Indeed, in our simulation study, we observed that Fleiss’ kappa and ICC were identical to the second decimal place and hence only report the Fleiss’ kappa as a comparison measure to the model-based agreement statistic.

In general, under normally distributed random effects, Fleiss’ kappa estimates were smaller compared to the model-based kappa estimates, except in one scenario where Fleiss’ kappa estimate was considerably larger than the model-based kappa estimates. Fleiss’ kappa has several restrictions: First, it requires a constant number of ratings per subject. If the number of ratings per subject differs, then an alternate form of Fleiss’ kappa is required to compute agreement. Second, Fleiss’ kappa is prone to prevalence of success. If the success rate is low, Fleiss’ kappa will underestimate the agreement between raters (Nelson & Edwards, 2008). Furthermore, although not discussed here, Fleiss’ kappa cannot be extended to incorporate information about rater characteristics that may impact agreement. Lastly, in the simulation study, the standard errors of estimated Fleiss’ kappa statistics computed using equation (4) were much smaller compared to the empirical standard errors. However, this issue needs to be further examined.

This study has some limitations. The assessment was restricted to four packages in R and SAS because of their popularity and accessibility. Other packages available in estimating GLMM with a crossed random effects structure such as MLwiN, WinBUGS, and Stata were not included.

This study has several strengths. First, the data generated for these simulation studies included realistic scenarios including the implementation of non-normally distributed random effects. In fact, the data set generated for one of the simulation scenarios was based on a real-life data set from the AIM study. Second, to our knowledge, this is the first study where the relatively new ORDINAL package was compared with existing packages on the performance of fitting GLMM with a crossed random effects structure for binary responses. The ORDINAL package is extremely stable, unlike the LME4 package, computationally efficient, and its parameter estimates were identical to those of PROC GLIMMIX in SAS. Lastly, the straightforward and reliable implementation of model-based measure of agreement ($\hat{\kappa}_m$) using existing packages was demonstrated. Model-based measure of agreement is robust to missing and unbalanced data, where not every subject's test result is rated by each rater.

Among frequentist R users, the ORDINAL package is recommended over the LME4 package for its stability and computational efficiency regardless of sample size and distribution of random effects. The GLIMMIX procedure in SAS produced nearly identical results to the ORDINAL package. For those who prefer Bayesian analysis, the MCMCglmm package performs well in fitting binary GLMM with a crossed random effects structure and for computing model-based agreement statistics. Although there was very little variability in the model-based agreement measures using different sets of priors, performing sensitivity analyses is recommended by altering the prior specification of the random effects distribution. A useful advantage of the Bayesian package implemented here (MCMCglmm) is its flexibility in incorporating a known characteristic of the data set to the model through the use of priors and its robustness to model misspecification when random effects distribution is skewed. Programs for fitting the binary GLMM with a crossed random effects structure for each of the four packages and an example data set are provided in supplementary materials. Full code for computing $\hat{\kappa}_m$ and its variance from GLMM parameter estimates for each package described in this paper is also included in the programs.

Overall, existing statistical software offer satisfactory packages or procedures for fitting binary GLMMs with a crossed random effects structure, and for estimation of agreement measures in large-scale agreement studies based upon multiple raters' binary classifications.

Acknowledgements

The authors are grateful for the support provided by grant R01-CA-17246301 from the United States National Institutes of Health. The AIM study was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope@Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274, SIRSG-07-275, SIRSG-06-281, SIRSG-09-270, SIRSG-09-271), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (HHSN261201100031C). The cancer data used in the AIM study was supported in part by state public health departments and cancer registries in the U.S., see <http://www.bcsc-research.org/work/acknowledgement.html>. We also thank participating women, mammography facilities, radiologists, and BCSC investigators for their data. A list of the BCSC investigators is provided at: <http://www.bcsc-research.org/>.

Funding

This study was funded by the United States National Institutes of Health (grant number 1R01CA17246301-A1).

References

- Agresti, A. (1989). A model for agreement between ratings on an ordinal scale. *Mathematical and Computer Modelling*, 12(9), 1188. doi: 10.1016/0895-7177(89)90272-0
- Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47(3), 639-653. doi: 10.1016/j.csda.2003.12.009
- Beam, C. A., Conant, E. F., & Sickles, E. A. (2002). Factors affecting radiologist inconsistency in screening mammography. *Academic Radiology*, 9(5), 531-540. doi: 10.1016/S1076-6332(03)80330-6
- Christensen, R. H. B. (2013). ordinal: Regression models for ordinal data (Version 2013.9-30) [R software package]. Retrieved from <http://www.cran.r-project.org/package=ordinal>
- Ciatto, S., Houssami, N., Apruzzese, A., Bassetti, E., Brancato, B., Carozzi, ..., Scorsolini, A. (2005). Categorizing breast mammographic density:

Intra- and interobserver reproducibility of BI-RADS density categories. *The Breast*, 14(4), 269-275. doi: [10.1016/j.breast.2004.12.004](https://doi.org/10.1016/j.breast.2004.12.004)

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213-220. doi: [10.1037/h0026256](https://doi.org/10.1037/h0026256)

Comprat, E., Egevad, L., Lopez-Beltran, A., Camparo, P., Algaba, F., Amin, M.,... Van der Kwast, T. H. (2013). An interobserver reproducibility study on invasiveness of bladder cancer using virtual microscopy and heatmaps. *Histopathology*, 63(6), 756-766. doi: [10.1111/his.12214](https://doi.org/10.1111/his.12214)

Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. *The New England Journal of Medicine*, 331(22), 1493-1499. doi: [10.1056/NEJM199412013312206](https://doi.org/10.1056/NEJM199412013312206)

Epstein, J. I., Allsbrook, W. C. J., Amin, M. B., Egevad, L. L., & ISUP Grading Committee. (2005). The 2005 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. *The American Journal of Surgical Pathology*, 29(9), 1228-1242. doi: [10.1097/01.pas.0000173646.99337.b1](https://doi.org/10.1097/01.pas.0000173646.99337.b1)

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382. doi: [10.1037/h0031619](https://doi.org/10.1037/h0031619)

Fleiss, J. L., & Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, 3(4), 537-542. doi: [10.1177/014662167900300410](https://doi.org/10.1177/014662167900300410)

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5), 974-977. doi: [10.1037/0033-2909.86.5.974](https://doi.org/10.1037/0033-2909.86.5.974)

Hadfield, J. D. (2010). MCMC methods for multi-response generalised linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2), 1-22. doi: [10.18637/jss.v033.i02](https://doi.org/10.18637/jss.v033.i02)

Hadfield, J. D. (2015). *MCMCglmm course notes*. Retrieved from: <https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf>

Hsiao, C. K., Chen, P.-C., & Kao, W.-H. (2011). Bayesian random effects for interrater and test-retest reliability with nested clinical observations. *Journal of Clinical Epidemiology*, 64(7), 808-814. doi: [10.1016/j.jclinepi.2010.10.015](https://doi.org/10.1016/j.jclinepi.2010.10.015)

AGREEMENT BETWEEN RATERS' BINARY CLASSIFICATIONS

Ibrahim, J., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *TEST*, 18(1), 1-43. doi: [10.1007/s11749-009-0138-x](https://doi.org/10.1007/s11749-009-0138-x)

Karim, M. R., & Zeger, S. L. (1992). Generalized linear models with random effects; Salamander mating revisited. *Biometrics*, 48(2), 631-644. doi: [10.2307/2532317](https://doi.org/10.2307/2532317)

Kim, Y., Choi, Y.-K., & Emery, S. (2013). Logistic regression with multiple random effects: A simulation study of estimation methods and statistical packages. *The American Statistician*, 67(3), 37-41. doi: [10.1080/00031305.2013.817357](https://doi.org/10.1080/00031305.2013.817357)

Kuk, A. Y. C., & Cheng, Y. W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation*, 59(3), 233-250. doi: [10.1080/00949657708811858](https://doi.org/10.1080/00949657708811858)

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. doi: [10.2307/2529310](https://doi.org/10.2307/2529310)

Li, B., Lingsma, H. F., Steyerberg, E. W., & Lesaffre, E. (2011). Logistic random effects regression models: A comparison of statistical packages for binary and ordinal outcomes. *BMC Medical Research Methodology*, 11(77). doi: [10.1186/1471-2288-11-77](https://doi.org/10.1186/1471-2288-11-77)

Litière, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine*, 27(16), 3125-3144. doi: [10.1002/sim.3157](https://doi.org/10.1002/sim.3157)

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437), 162-170. doi: [10.2307/2291460](https://doi.org/10.2307/2291460)

McCulloch, C. E., & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1), 270-279. doi: [10.1111/j.1541-0420.2010.01435.x](https://doi.org/10.1111/j.1541-0420.2010.01435.x)

Nelson, K. P., & Edwards, D. (2008). On population-based measures of agreement for binary classifications. *Canadian Journal of Statistics*, 36(3), 411-426. doi: [10.1002/cjs.5550360306](https://doi.org/10.1002/cjs.5550360306)

Nelson, K. P., & Edwards, D. (2010). Improving the reliability of diagnostic tests in population-based agreement studies. *Statistics in Medicine*, 29(6), 617-626. doi: [10.1002/sim.3819](https://doi.org/10.1002/sim.3819)

Onega, T., Smith, M., Miglioretti, D. L., Carney, P. A., Geller, B. A., Kerlikowske, K.,... Yankaskas, B. (2012). Radiologist agreement for mammographic recall by case difficulty and finding type. *Journal of the American College of Radiology*, 9(11), 788-794. doi: [10.1016/j.jacr.2012.05.020](https://doi.org/10.1016/j.jacr.2012.05.020)

Ooms, E. A., Zonderland, H. M., Eijkemans, M. J. C., Kriege, M., Mahdavian Delavary, B., Burger, C. W., & Ansink, A. C. (2007). Mammography: Interobserver variability in breast density assessment. *The Breast*, 16(6), 568-576. doi: [10.1016/j.breast.2007.04.007](https://doi.org/10.1016/j.breast.2007.04.007)

R Core Team (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical. Retrieved from: <http://www.R-project.org/>

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, 19(3), 321-325. doi: [10.1086/266577](https://doi.org/10.1086/266577)

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428. doi: [10.1037/0033-2909.86.2.420](https://doi.org/10.1037/0033-2909.86.2.420)

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80(389), 175-180. doi: [10.2307/2288068](https://doi.org/10.2307/2288068)

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., & Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, 30(20), 2562-2572. doi: [10.1002/sim.4265](https://doi.org/10.1002/sim.4265)

Characterizations of Distributions by Expected Values of Lower Record Statistics with Spacing

M. Faizan

Aligarh Muslim University
Aligarh, India

Ziaul Haque

National Sample Survey Office
New Delhi, India

M. A. Ansari

Aligarh Muslim University
Aligarh, India

The characterizations of a certain class of probability distributions are established through conditional expectation of lower record values when the conditioned record value may not be the adjacent one. Some of its important deductions are also discussed.

Keywords: Characterization, continuous distributions, conditional expectation, lower record values

Introduction

Record values have been extensively studied in literature. For some excellent reviews, see Ahsanullah (1995), Arnold, Balakrishnan, and Nagaraja (1998), and Nevzorov (2001). Characterization of a probability distribution plays an important role in the determination of distributions by using certain characteristics in the given data. Different methods were used to identify several types of distributions. Conditional expectations of record values were extensively used in characterizing the continuous probability distributions. For examples, consider Malinowska and Szynal (2008), Shawki and Bakoban (2009), and, recently, Yanev (2012), Ahsanullah, Shakil, and Golam Kibria (2013), Azedine (2013), and Nadarajah, Teimouri, and Shih (2014), among others.

Let $X_1, X_2, \dots$ be a sequence of independent, identically-distributed continuous random variables with distribution function (df) $F(x)$ and probability density function (pdf) $f(x)$. Let $X_{L(r)}$ be the r^{th} lower record value; then the conditional pdf of $X_{L(s)}$ given $X_{L(r)} = x$, $1 \leq r < s$, is (Ahsanullah, 1995)

M. Faizan is in the Department of Statistics and Operations Research. Email them at: mdfaizan02@gmail.com. Ziaul Haque is in the Field Operations Division. Email them at: ziaulstats@gmail.com. M. A. Ansari is in the Department of Statistics and Operations Research.

$$f\left(X_{L(s)} | X_{L(r)} = x\right) = \frac{1}{\Gamma(s-r)} \left[-\ln F(y) + \ln F(x)\right]^{s-r-1} \frac{f(y)}{F(x)} \quad (1)$$

and the conditional pdf of $X_{L(r)}$ given $X_{L(s)} = y$, $1 \leq r < s$, is (Ahsanullah, 1995)

$$\begin{aligned} f\left(X_{L(r)} | X_{L(s)} = y\right) \\ = \frac{\Gamma(s)}{\Gamma(r)\Gamma(s-r)} \frac{\left[-\ln F(x)\right]^{r-1}}{\left[-\ln F(y)\right]^{s-1}} \left[-\ln F(y) + \ln F(x)\right]^{s-r-1} \frac{f(x)}{F(x)} \end{aligned} \quad (2)$$

Let X be a continuous random variable with df $F(x)$ defined by:

$$(i) \quad F(x) = \left[ah(x) + b\right]^c, \quad x \in (\alpha, \beta) \quad (3)$$

$$(ii) \quad F(x) = \exp\left[-e^{-ah(x)+b}\right], \quad x \in (\alpha, \beta) \quad (4)$$

where a , b , and c are constants and $h(x)$ is a monotonic and differentiable function of x defined on (α, β) such that $F(x)$ is a df.

Here, the aim is to characterize a family of distributions defined in (3) and (4) by considering conditional expectation of functions of lower record values when the conditioning is on any record value, not necessarily the adjacent one. Various well-known distributions (e.g., Power Function, Pareto, Inverse Weibull, Cauchy) arise from the above family of distributions by suitable choices of $h(x)$ and the constants a , b and c .

Characterization Theorems

Theorem 1: Let X be an absolutely continuous random variable with df $F(x)$ and pdf $f(x)$ on support (α, β) , where α and β may be finite or infinite. Then, for $1 \leq j < s$,

$$E\left[h\left(X_{L(s)}\right) | X_{L(j)} = x\right] = a_{s|j}h(x) + b_{s|j}, \quad j = r, r+1 \quad (5)$$

if and only if

CHARACTERIZATIONS OF DISTRIBUTIONS

$$F(x) = [ah(x) + b]^c, x \in (\alpha, \beta) \quad (6)$$

where

$$a_{s|j} = \left(\frac{c}{c+1} \right)^{s-j}, \quad b_{s|j} = -\frac{b}{a}(1 - a_{s|j})$$

and $h(x)$ is a monotonic and differentiable function of x such that $h(x) \rightarrow 0$ as $x \rightarrow \beta$ and $h(x)F(x) \rightarrow 0$ as $x \rightarrow \alpha$.

Proof: First, prove (6) implies (5):

$$F(x) = [ah(x) + b]^c, f(x) = ac h'(x) [ah(x) + b]^{c-1}$$

From (1) and (5),

$$\begin{aligned} E \left[h(X_{L(s)}) | X_{L(r)} = x \right] &= \frac{1}{\Gamma(s-r)} \int_{\alpha}^x h(y) [-\ln F(y) + \ln F(x)]^{s-r-1} \frac{f(y)}{F(x)} dy \\ &= \frac{1}{\Gamma(s-r) [ah(x) + b]^c} \int_{\alpha}^x h(y) \left[c \ln \left(\frac{ah(x) + b}{ah(y) + b} \right) \right]^{s-r-1} \\ &\quad \times ac h'(y) [ah(y) + b]^{c-1} dy \end{aligned} \quad (7)$$

Let

$$t = \ln \left(\frac{ah(x) + b}{ah(y) + b} \right)^c$$

then (7) is

$$E \left[h(X_{L(s)}) | X_{L(r)} = x \right] = \frac{1}{a\Gamma(s-r)} \int_0^{\infty} [(ah(x) + b)e^{-t/c} - b] t^{s-r-1} e^{-t} dt$$

which reduces to

$$E\left[h\left(X_{L(s)}\right) \mid X_{L(r)}=x\right]=\left(\frac{c}{c+1}\right)^{s-r} h(x)+\frac{b}{a}\left[\left(\frac{c}{c+1}\right)^{s-r}-1\right]$$

and hence the ‘if’ part.

To prove (5) implies (6),

$$\frac{1}{\Gamma(s-r)} \int_{\alpha}^x h(y)\left[-\ln F(y)+\ln F(x)\right]^{s-r-1} \frac{f(y)}{F(x)} d y=a_{s|r} h(x)+b_{s|r}=g_{s|r}(x)$$

or

$$\frac{1}{\Gamma(s-r)} \int_{\alpha}^x h(y)\left[-\ln F(y)+\ln F(x)\right]^{s-r-1} f(y) d y=g_{s|r}(x) F(x) \quad (8)$$

Differentiating both the sides of (8) with respect to x and re-arranging the terms,

$$\frac{f(x)}{F(x)}=\frac{g'_{s|r}(x)}{g_{s|r+1}(x)-g_{s|r}(x)}=A(x) \quad (9)$$

Now

$$g_{s|r+1}(x)-g_{s|r}(x)=\frac{\left(\frac{c}{c+1}\right)^{s-r-1}[a h(x)+b]}{a(c+1)}$$

Therefore

$$A(x)=\frac{a c h'(x)}{[a h(x)+b]}$$

and hence

$$\frac{f(x)}{F(x)} = \frac{ac h'(x)}{[ah(x)+b]} \quad (10)$$

Thus

$$F(x) = [ah(x)+b]^c$$

and hence the sufficiency part.

Theorem 2: Under the conditions given in the Theorem 1 and for $1 \leq j \leq r < s$,

$$E \left[h(X_{L(s)}) | X_{L(j)} = x \right] = a_{s|j} E \left[h(X_{L(r)}) | X_{L(j)} = x \right] + b_{s|j}, \quad j = m, m+1 \quad (11)$$

if and only if (6) holds, where $a_{s|j}$ and $b_{s|j}$ are defined as in Theorem 1.

Proof: In the view of Theorem 1,

$$E \left[h(X_{L(s)}) | X_{L(m)} = x \right] = a_{s|m} h(x) + b_{s|m} \quad (12)$$

and

$$E \left[h(X_{L(r)}) | X_{L(m)} = x \right] = a_{r|m} h(x) + b_{r|m} \quad (13)$$

Now,

$$\begin{aligned} a_{s|m} &= \left(\frac{c}{c+1} \right)^{s-m} \\ &= \left(\frac{c}{c+1} \right)^{s-r} \left(\frac{c}{c+1} \right)^{r-m} = a_{s|r} a_{r|m} \end{aligned}$$

Therefore

$$\begin{aligned}
 E\left[h\left(X_{L(s)}\right) \mid X_{L(m)}=x\right] &= a_{s|m} h(x) + b_{s|m} \\
 &= a_{s|r} a_{r|m} h(x) + b_{s|m} \\
 &= a_{s|r} \left[a_{r|m} h(x) + b_{r|m} \right] + b_{s|r} \\
 &= a_{s|r} E\left[h\left(X_{L(r)}\right) \mid X_{L(m)}=x\right] + b_{s|r}
 \end{aligned}$$

and hence the necessary part.

For the sufficiency part,

$$\begin{aligned}
 &\frac{1}{\Gamma(s-m)} \int_a^x h(y) [-\ln F(y) + \ln F(x)]^{s-m-1} f(y) dy \\
 &= a_{s|r} \frac{1}{\Gamma(r-m)} \int_\alpha^x h(y) [-\ln F(y) + \ln F(x)]^{r-m-1} f(y) dy + b_{s|r} F(x)
 \end{aligned} \tag{14}$$

Differentiating both the sides of (14) with respect to x ,

$$\begin{aligned}
 \frac{1}{\Gamma(s-r)} \int_\alpha^x h(y) [-\ln F(y) + \ln F(x)]^{s-r-1} \frac{f(y)}{F(x)} dy &= a_{s|r} h(x) + b_{s|r} \\
 &= g_{s|r}(x)
 \end{aligned} \tag{15}$$

Proceeding as in Theorem 1 gives the result.

Table 1. Examples based on the df $F(x) = [ah(x) + b]^c$

Distribution	F(x)	a	b	c	h(x)
Power function	$a^{-p}x^p, 0 < x \leq a$	a^{-q}	0	p/q	x^q
		a^{-1}	0	p	x
Pareto	$1 - a^p x^{-p}, a \leq x < \infty$	$-a^p$	1	1	$x^{-p}, p > 0$
Inverse Weibull	$e^{-\theta x^\rho}, 0 \leq x < \infty$	1	0	1	$e^{-\theta x^\rho}$
		1	0	θ	e^{-x^ρ}
Burr type III	$[1 + \theta x^{-p}]^{-\lambda}, 0 \leq x < \infty$	θ	1	$-\lambda$	$x^{-p}, p, \lambda \neq 1$
		1	1	$-\lambda$	$\theta x^{-p}, p, \lambda \neq 1$
Cauchy	$\frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left(\frac{x-\theta}{\lambda}\right),$ $-\infty < x < \infty$	$\frac{1}{\pi}$	$\frac{1}{2}$	1	$\tan^{-1}\left(\frac{x-\theta}{\lambda}\right),$ $-\infty < \theta < \infty, \lambda > 0$

CHARACTERIZATIONS OF DISTRIBUTIONS

Remark 1: At $r = j$, Theorem 2 reduces to Theorem 1.

Remark 2: At $a = -a / c$, $b = 1$, and $c \rightarrow \infty$, $F(x) = [ah(x) + b]^c \rightarrow e^{-ah(x)}$ as obtained by Faizan and Khan (2011).

Theorem 3: Let X be an absolutely continuous random variable with df $F(x)$ and pdf $f(x)$ on support (α, β) , where α and β may be finite or infinite. Then, for $r \leq s < t$,

$$E[h(X_{L(s)}) - h(X_{L(r)}) | X_{L(t)} = y] = -\frac{1}{a} \sum_{j=r}^{s-1} \frac{1}{j} \quad (16)$$

if and only if

$$F(x) = \exp[-e^{-ah(x)+b}], x \in (\alpha, \beta) \quad (17)$$

Proof: In view of (17), it follows

$$\begin{aligned} & E[h(X_{L(s)}) | X_{L(t)} = y] \\ &= \frac{\Gamma(t)}{\Gamma(s)\Gamma(t-s)} \int_y^\beta h(x) \frac{\left[-\ln e^{-e^{-ah(x)+b}} \right]^{s-1}}{\left[-\ln e^{-e^{-ah(y)+b}} \right]^s} \left[1 - \frac{\ln e^{-e^{-ah(x)+b}}}{\ln e^{-e^{-ah(y)+b}}} \right]^{t-s-1} a h'(x) e^{-ah(y)+b} dx \\ &= \frac{\Gamma(t)}{\Gamma(s)\Gamma(t-s)} \int_y^\beta h(x) \left[\frac{e^{-ah(x)+b}}{e^{-ah(y)+b}} \right]^{s-1} \left[1 - \frac{e^{-ah(x)+b}}{e^{-ah(y)+b}} \right]^{t-s-1} \frac{a h'(x) e^{-ah(y)+b}}{e^{-ah(y)+b}} dx \end{aligned}$$

Setting

$$u = \frac{e^{-ah(x)+b}}{e^{-ah(y)+b}}$$

obtains

$$E[h(X_{L(s)}) | X_{L(t)} = y] = \frac{\Gamma(t)}{\Gamma(s)\Gamma(t-s)} \int_0^1 \left(h(y) - \frac{1}{a} \ln u \right) u^{s-1} (1-u)^{t-s-1} du$$

or

$$\mathbb{E}\left[\mathbf{h}\left(X_{L(s)}\right) \mid X_{L(t)}=y\right]=\mathbf{h}(y)-\frac{1}{a} \frac{\Gamma(t)}{\Gamma(s) \Gamma(t-s)} \int_0^1 \frac{1}{a} \ln u u^{s-1}(1-u)^{t-s-1} d u$$

From Gradshteyn and Ryzhik (2007, p. 540)

$$\int_0^1 x^{u-1}\left(1-x^r\right)^{v-1} \ln x d x=\frac{1}{r^2} \mathbf{B}\left(\frac{u}{r}, v\right)\left[\psi\left(\frac{u}{r}\right)-\psi\left(\frac{u}{r}+v\right)\right]$$

where

$$\mathbf{B}(a, b)=\frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}, \quad \psi(a)=\frac{d}{d a} \ln \Gamma(a)$$

$$\frac{\Gamma(t)}{\Gamma(s) \Gamma(t-s)} \int_0^1 \ln u u^{s-1}(1-u)^{t-s-1} d u=\psi(s)-\psi(t)$$

Therefore

$$\mathbb{E}\left[\mathbf{h}\left(X_{L(s)}\right) \mid X_{L(t)}=y\right]=\mathbf{h}(y)-\frac{1}{a}\left[\psi(s)-\psi(t)\right] \quad (18)$$

Thus

$$\mathbb{E}\left(\mathbf{h}\left(X_{L(s)}\right)-\mathbf{h}\left(X_{L(r)}\right) \mid X_{L(t)}=y\right)=-\frac{1}{a}\left[\psi(s)-\psi(r)\right]$$

Using the result (Medina & Moll, 2009), for $n \in \mathbb{N}$,

$$\psi(n)=-\gamma+\sum_{k=1}^{n-1} \frac{1}{k}$$

Therefore,

CHARACTERIZATIONS OF DISTRIBUTIONS

$$E\left(h\left(X_{L(s)}\right)-h\left(X_{L(r)}\right) \mid X_{L(t)}=y\right)=-\frac{1}{a} \sum_{j=r}^{s-1} \frac{1}{j}$$

which proves the necessary part.

For the sufficiency part, let

$$k=\frac{1}{a} \sum_{j=r}^{s-1} \frac{1}{j}$$

$$\begin{aligned} & \frac{\Gamma(t)}{\Gamma(s) \Gamma(t-s)} \int_y^{\beta} h(x) \frac{\left[-\ln F(x)\right]^{s-1}}{\left[-\ln F(y)\right]^{t-1}}\left[-\ln F(y)+\ln F(x)\right]^{t-s-1} \frac{f(x)}{F(x)} d x \\ & -\frac{\Gamma(t)}{\Gamma(r) \Gamma(t-r)} \int_y^{\beta} h(x) \frac{\left[-\ln F(x)\right]^{r-1}}{\left[-\ln F(y)\right]^{t-1}}\left[-\ln F(y)+\ln F(x)\right]^{t-r-1} \frac{f(x)}{F(x)} d x \quad (19) \\ & =-k \end{aligned}$$

Differentiating both sides of (19) with respect to y ,

$$\begin{aligned} & \frac{\Gamma(s)}{\Gamma(r) \Gamma(s-r)} \int_y^{\beta} h(x) \frac{\left[-\ln F(x)\right]^{r-1}}{\left[-\ln F(y)\right]^{s-1}}\left[-\ln F(y)+\ln F(x)\right]^{s-r-1} \frac{f(x)}{F(x)} d x \quad (20) \\ & =h(y)+k=g_{r|s}(y) \end{aligned}$$

Again, differentiating (20) with respect to y and simplifying,

$$-\frac{f(y)}{F(y) \ln F(y)}=\frac{g'_{r|s}(y)}{\left[(s-1) g_{r|s-1}(y)-g_{r|s}(y)\right]}=A(y)$$

Using the result (Khan, Anwar, & Chisti, 2010)

$$F(x)=\exp \left[-e^{\int_p^x A(t) d t}\right]$$

where $-\ln F(p) = 1$. Thus $F(x) = \exp[-e^{-ah(x)+b}]$ and hence the theorem.

Examples

Proper choice of a , b , and $h(x)$ characterize the distributions as given below:

(i) Power Function Distribution

$$a = 1, b = \ln p, h(x) = -\ln(-\ln(x/a))$$

$$F(x) = (x/a)^p, 0 < x < a$$

(ii) Inverse Weibull Distribution

$$a = 1, b = \ln \theta, h(x) = -\ln x^{-p}$$

$$F(x) = \exp(-\theta x^p), 0 < x < \infty$$

(iii) Gumbel Distribution

$$a = 1, b = 0, h(x) = x$$

$$F(x) = \exp(-e^{-x}), -\infty < x < \infty$$

(iv) Extreme Value-II Distribution

$$a = 1, b = \ln \theta^p, h(x) = \ln x^p$$

$$F(x) = \exp(-\theta/x)^p, -\infty < x < \infty$$

(v) Logistic Distribution

$$a = 1, b = 0, h(x) = -\ln(\ln(1 + e^{-c}))$$

$$F(x) = (1 + x^{-c})^{-k}, -\infty < x < \infty$$

(vi) Burr Type-II Distribution

$$a = 1, b = \theta, h(x) = -\ln(\ln(1 + e^{-x}))$$

$$F(x) = (1 + e^{-x})^{-\theta}, -\infty < x < \infty$$

(vii) Burr Type-III Distribution

$$a = 1, b = \ln k, h(x) = -\ln(\ln(1 + x^{-c}))$$

$$F(x) = (1 + x^{-c})^{-k}, 0 < x < \infty$$

(viii) Burr Type-IV Distribution

$$a = 1, b = \ln k, h(x) = -\ln \left(\ln \left(1 + \left(\frac{c-x}{x} \right)^{\frac{1}{c}} \right) \right)$$

$$F(x) = \left(1 + \left(\frac{c-x}{x} \right)^{\frac{1}{c}} \right)^{-k}, 0 < x < c$$

(ix) Burr Type-V Distribution

$$a = 1, b = \ln k, h(x) = -\ln(\ln(1 + ce^{\tan x}))$$

$$F(x) = (1 + ce^{\tan x})^{-k}, -\pi/2 < x < \pi/2$$

(x) Burr Type-VI Distribution

$$a = 1, b = \ln k, h(x) = -\ln(\ln(1 + ce^{-k \sinh x}))$$

$$F(x) = (1 + ce^{-k \sinh x})^{-k}, -\infty < x < \infty$$

Acknowledgements

The authors are thankful to the anonymous Reviewer and Editor for their comments and suggestions, which resulted in an improvement in the presentation of this manuscript.

References

- Ahsanullah, M. (1995). *Record statistics*. New York, NY: Nova Science Publishers.
- Ahsanullah, M., Shakil, M., & Golam Kibria, B. M. (2013). A characterization of power function distribution based on lower record values. *ProbStat Forum*, 6, 68-72. Retrieved from <http://probbstat.org.in/PSF-2013-08.pdf>
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1998). *Records*. New York, NY: Wiley. doi: [10.1002/9781118150412](https://doi.org/10.1002/9781118150412)
- Azedine, G. (2013). Characterization of the power distribution based on the lower records. *Applied Mathematical Sciences*, 7(106), 5259-5267. doi: [10.12988/ams.2013.36326](https://doi.org/10.12988/ams.2013.36326)

- Faizan, M., & Khan, M. I. (2011). A characterization of continuous distributions through lower record statistics. *ProbStat Forum*, 4, 39-43.
- Gradshteyn, I. S., & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (7th ed.). Amsterdam: Academic Press.
- Khan, A. H., Anwar, Z., & Chisti, S. (2010). Characterization of continuous distributions through conditional expectation of functions of dual generalized order statistics. *Pakistan Journal of Statistics*, 26(4), 615-628.
- Malinowska, I., & Szynal, D. (2008). On characterization of certain distribution of k th lower (upper) record values. *Applied Mathematics Computation*, 202(1), 338-347. doi: [10.1016/j.amc.2008.02.022](https://doi.org/10.1016/j.amc.2008.02.022)
- Medina, L. A., & Moll, V. H. (2009). The integrals in Gradshteyn and Ryzhik. Part 10: The digamma function. *SCIENTIA. Series A: Mathematical Sciences*, 17, 45-66. Retrieved from <http://www.mat.utfsm.cl/scientia/archivos/vol17/ar6.pdf>
- Nadarajah, S., Teimouri, M., & Shih, S. H. (2014). Characterizations of the Weibull and uniform distributions using record values. *Brazilian Journal of Probability and Statistics*, 28(2), 209-222. doi: [10.1214/12-bjps202](https://doi.org/10.1214/12-bjps202)
- Nevzorov, V. B. (2001). *Records: Mathematical theory*. Providence, RI: American Mathematical Society.
- Shawki, A. I., & Bakoban, R. A. (2009). Conditional expectation of certain distributions of record values. *International Journal of Mathematical Analysis*, 3(17), 829-838. Retrieved from <http://www.m-hikari.com/ijma/ijma-password-2009/ijma-password17-20-2009/bakobanIJMA17-20-2009.pdf>
- Yanev, G. P. (2012). Characterization of exponential distribution via regression of one record value on two non-adjacent record values. *Metrika*, 75(6), 743-760. doi: [10.1007/s00184-011-0350-z](https://doi.org/10.1007/s00184-011-0350-z)

Semi-Parametric Method to Estimate the Time-to-Failure Distribution and its Percentiles for Simple Linear Degradation Model

Laila Naji Ba Dakhn
Yarmouk University
Irbid, Jordan

Mohammed Al-Haj Ebrahim
Yarmouk University
Irbid, Jordan

Omar Eidous
Yarmouk University
Irbid, Jordan

Most reliability studies obtained reliability information by using degradation measurements over time, which contains useful data about the product reliability. Parametric methods like the maximum likelihood (ML) estimator and the ordinary least square (OLS) estimator are used widely to estimate the time-to-failure distribution and its percentiles. In this article, we estimate the time-to-failure distribution and its percentiles by using a semi-parametric estimator that assumes the parametric function to have a half-normal distribution or an exponential distribution. The performance of the semi-parametric estimator is compared via simulation study with the ML and OLS estimators by using the mean square error and length of the 95% bootstrap confidence interval as the basis criteria of the comparison. An application to real data is given. In general, if there are assumptions on the random effect parameter, the ML estimator is the best; otherwise the kernel semi-parametric estimator with half-normal distribution is the best.

Keywords: Degradation model, semi-parametric estimator, maximum likelihood estimator, ordinary least square estimator

Introduction

Meeker and Escobar (1998) defined the reliability of a unit as the probability that a unit will perform its intended function until a specified point of time under encountered use conditions. There are many proposed applications to measure the reliability of any product. One of these applications is the estimation of the time-to-failure distribution and its percentiles. In estimation, traditional life tests are

Ms. Laila Naji Ba Dakhn is a graduate student in the Department of Statistics. Email her at: la2010_math@yahoo.com. Prof. Mohammed Al-Haj Ebrahim is a Professor of Statistics. Email them at: m_hassanb@hotmail.com. Prof. Omar Eidous is a Professor of Statistics. Email them at: omarm@yu.edu.jo.

often not the most efficient way to obtain reliability information because few failure time data are observed by the end of the test; it is then difficult to use the traditional reliability analysis that records only failure time data to analyze life time data. Thus, it is possible to get failure data by degradation measurements over time which may contain useful data about product reliability.

Degradation is usually measured as a function of time T . Let $D(t)$ denote the actual sequence or path of the degradation of a particular unit over time t for each sample unit that will be observed, and let D_f denote the critical level for the degradation path where failure has occurred. The focus is on the linear degradation model for estimating the $100r^{\text{th}}$ percentile of the time-to-failure distribution.

Gertsbakh and Kordonskiĭ (1966/1969) discussed the degradation problem from an engineering point of view. They presented the Bernstein distribution, which describes the time-to-failure distribution for a simple linear model with random intercept and random slope. Amster and Hooper (1983) proposed a simple degradation model for single, multiple, and step-stress life tests. They explain how to use this model to estimate the central tendency of the time-to-failure distribution. Lu, Meeker, and Escobar (1996) compared the degradation analysis and traditional failure time analysis in terms of asymptotic efficiency. They demonstrated that the degradation analysis gives more precision than the traditional failure time analysis in general.

Al-Haj Ebrahim, Eidous, and Kmail (2009) proposed the nonparametric classical kernel method to estimate the time-to-failure distribution and its percentiles for the simple linear degradation model. They compared the performance of this method with the existing parametric methods like ML and OLS. They gave the time-to-failure distribution based on the classical kernel method (by assuming Gaussian kernel), which is

$$F_T(t) = 1 - \frac{1}{n} \sum_{i=1}^n \Phi \left(\frac{\frac{D_f}{t} - \beta_i}{h} \right)$$

where Φ is the distribution function of the standard normal distribution. They compute the bandwidth using the formula (Silverman, 1986)

$$h_{\text{int}} = 0.9An^{-1/5} \quad (1)$$

where $A = \min\{SD, IQR/1.34\}$, SD is the sample standard deviation, and IQR is the sample inter-quartile range.

The kernel function K is taken to be the Gaussian function

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad -\infty < u < \infty \quad (2)$$

and the smoothing parameter of the nonparametric estimator is computed using the formula in (1).

Model and Time-to-Failure Distribution

Consider the following simple linear degradation model to estimate the time-to-failure distribution:

$$y_{ij} = \beta_i t_{ij} + \varepsilon_{ij} \quad (3)$$

where y_{ij} is the observed degradation measurement of the i^{th} unit at time t_{ij} , β_i is the random effect parameter (the slope of the linear degradation model for unit i), t_{ij} is the failure time for the degradation model, and ε_{ij} is the random error term, where the ε_{ij} are iid with $N(0, \sigma_\varepsilon^2)$.

In general, the time-to-failure distribution can be written as a function of the degradation model parameters. The failure time T is defined as the time when the actual path $D(t)$ crosses the critical degradation level D_f , i.e. T is the solution of

$$D_f = D(t)$$

By considering the simple linear degradation model (3),

$$D_f = \beta T$$

Then the distribution function of the time-to-failure is

$$\begin{aligned}
 F_T(t) &= P(T \leq t) \\
 &= P\left(\frac{D_f}{\beta} \leq t\right) \\
 &= P\left(\beta \geq \frac{D_f}{t}\right) \\
 &= 1 - G_\beta\left(\frac{D_f}{t}\right), \quad t > 0
 \end{aligned} \tag{4}$$

where β is a random effect parameter and $G(\cdot)$ is the distribution function of β .

Let the 100^rth percentile of the time-to-failure distribution be denoted by t_r . To find t_r , we need to solve

$$r = F_T(t_r) = 1 - G_\beta\left(\frac{D_f}{t_r}\right)$$

with respect to t_r . This gives

$$t_r = \frac{D_f}{G_\beta^{-1}(1-r)} \tag{5}$$

It is clear that, for a fixed value of D_f , the distribution of T and the 100^rth percentile depend on the distribution of β , the random effect parameter. In some simple cases, a closed-form expression for $F(t)$ could be obtained, but for most practical path models, it is necessary to evaluate $F(t)$ using numerical methods. For example, consider the linear degradation model (3) with random effect distributed as $N(\mu, \sigma^2)$.

From equation (4),

$$F_T(t) = \Phi\left(\frac{t_r\mu - D_f}{t_r\sigma}\right)$$

and, from equation (5),

$$t_r = \frac{D_f}{\mu - \sigma\Phi^{-1}(r)}$$

where, $\Phi(z)$ is the standard normal cumulative distribution function. As another example, if $\beta \sim \exp(\theta)$, then

$$F_T(t) = e^{-D_f/t^\theta}$$

and

$$t_r = \frac{-D_f}{\theta \ln(r)}$$

For the above two examples, the parameters μ , σ , and θ can be estimated using the ML or OLS methods, or even by any good statistical method.

Estimating Percentiles of Time-to-Failure Distribution Using Semi Parametric Density Method

If $\hat{\theta}$ is an estimator of θ , we can construct the following semi-parametric estimator of $f(x)$:

$$\tilde{f}_{SP}(x) = \frac{f(x; \hat{\theta})}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \left(f(X_i; \hat{\theta})\right)^{-1}, \quad x \in \mathbb{R} \quad (6)$$

This estimator is semi-parametric because it combines a nonparametric estimator, the classical kernel estimator, and a parametric estimator, $f(x, \theta)$. In this work, two versions of $\tilde{f}_{SP}(x)$ are considered and studied. The first one assumes $f(x, \theta)$ follows a half-normal distribution and the other assumes $f(x, \theta)$ follows an exponential distribution.

If the degradation model is a simple linear as in (3), and if $\beta_1, \beta_2, \dots, \beta_n$ is a random sample from unknown pdf ($g_\beta(b)$), then we proposed – in this section – the semi-parametric estimator for the time-to-failure distribution and its percentiles.

The Half-Normal Distribution

The semi-parametric estimator of $g_\beta(b)$ that depends on the half-normal distribution is

$$\hat{g}_{\text{SPH}}(b) = \frac{v_1(b)}{nh} \sum_{j=1}^n K\left(\frac{b-\beta_j}{h}\right) \left(v_1(\beta_j)\right)^{-1}$$

where

$$v_1(b) = \sqrt{\frac{2}{\pi\sigma^2}} e^{-\frac{b^2}{2\hat{\sigma}^2}}$$

with

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \beta_i^2}{n}$$

Taking the kernel function $K(u)$ to be a Gaussian function,

$$\begin{aligned} \hat{g}_{\text{SPH}}(b) &= \frac{\sqrt{\frac{2}{\pi\sigma^2}} e^{-\frac{b^2}{2\hat{\sigma}^2}}}{nh} \sum_{j=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b-\beta_j}{h}\right)^2} \left(\sqrt{\frac{2}{\pi\sigma^2}} e^{-\frac{\beta_j^2}{2\hat{\sigma}^2}}\right)^{-1} \\ &= \frac{e^{-\frac{b^2}{2\hat{\sigma}^2}}}{nh} \sum_{j=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b-\beta_j}{h}\right)^2} \left(e^{-\frac{\beta_j^2}{2\hat{\sigma}^2}}\right)^{-1} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j^2}{2\hat{\sigma}^2}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b-\beta_j}{h}\right)^2 - \frac{b^2}{2\hat{\sigma}^2}} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j^2}{2\hat{\sigma}^2}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2\hat{\sigma}^2 h^2}(\hat{\sigma}^2(b-\beta_j)^2 + h^2 b^2)} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j^2}{2\hat{\sigma}^2}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2\hat{\sigma}^2 h^2} \left(h^2 + \hat{\sigma}^2 \right) \left(b - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2} \right)^2 + \frac{\hat{\sigma}^2 h^2 \beta_j^2}{h^2 + \hat{\sigma}^2}} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{j=1}^n \frac{e^{-\frac{1}{2\hat{\sigma}^2 h^2} \left(\frac{\hat{\sigma}^2 h^2 \beta_j^2}{h^2 + \hat{\sigma}^2} \right)}}{e^{-\frac{\beta_j^2}{2\sigma^2}}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{h^2 + \hat{\sigma}^2}{2\hat{\sigma}^2 h^2} \left(b - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2} \right)^2} \\
 &= \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2 \beta_j^2}{2\hat{\sigma}^2 (h^2 + \hat{\sigma}^2)}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{h^2 + \hat{\sigma}^2}{2\hat{\sigma}^2 h^2} \left(b - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2} \right)^2}
 \end{aligned}$$

The estimation of the distribution function of the time-to-failure is

$$\begin{aligned}
 \hat{F}_T(t) &= 1 - \hat{G}_{SPH} \left(\frac{D_f}{t} \right) \\
 &= 1 - \int_{-\infty}^{\frac{D_f}{t}} \hat{g}_{SPH}(b) db \\
 &= 1 - \int_{-\infty}^{\frac{D_f}{t}} \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2 \beta_j^2}{2\hat{\sigma}^2 (h^2 + \hat{\sigma}^2)}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{h^2 + \hat{\sigma}^2}{2\hat{\sigma}^2 h^2} \left(b - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2} \right)^2} db \\
 &= 1 - \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2 \beta_j^2}{2\hat{\sigma}^2 (h^2 + \hat{\sigma}^2)}} \int_{-\infty}^{\frac{D_f}{t}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{h^2 + \hat{\sigma}^2}{2\hat{\sigma}^2 h^2} \left(b - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2} \right)^2} db \\
 &= 1 - \frac{1}{n} \sqrt{\frac{\hat{\sigma}^2}{h^2 + \hat{\sigma}^2}} \sum_{j=1}^n e^{\frac{h^2 \beta_j^2}{2\hat{\sigma}^2 (h^2 + \hat{\sigma}^2)}} \text{Pro} \left(Q \leq \frac{D_f}{t} \right)
 \end{aligned}$$

where Q is a random variable distributed as

$$N \left(\frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2}, \frac{\hat{\sigma}^2 h^2}{(h^2 + \hat{\sigma}^2)} \right)$$

Then

$$\hat{F}_T(t) = 1 - \frac{1}{n} \sqrt{\frac{\hat{\sigma}^2}{h^2 + \hat{\sigma}^2}} \sum_{j=1}^n e^{\frac{h^2 \beta_j^2}{2\hat{\sigma}^2(h^2 + \hat{\sigma}^2)}} \Phi \left(\frac{\frac{D_f}{t} - \frac{\hat{\sigma}^2 \beta_j}{h^2 + \hat{\sigma}^2}}{\sqrt{\frac{\hat{\sigma}^2 h^2}{h^2 + \hat{\sigma}^2}}} \right) \quad (7)$$

where $\Phi(\cdot)$ is a standard normal distribution. To estimate the $100r^{\text{th}}$ percentiles (denoted by $\hat{t}_{r\text{-SPH}}$), we should solve $\hat{F}_{T\text{-SPH}}(\hat{t}_{r\text{-SPH}}) = r$ numerically for $\hat{t}_{r\text{-SPH}}$.

The Exponential Distribution

The semi-parametric estimator of $g_{\beta}(b)$ based on the exponential distribution is

$$\hat{g}_{\text{SPE}}(b) = \frac{v_2(b)}{nh} \sum_{j=1}^n K\left(\frac{b - \beta_j}{h}\right) (v_2(\beta_j))^{-1}$$

where

$$v_2(b) = \frac{e^{-\frac{b}{\hat{\alpha}}}}{\hat{\alpha}}$$

with

$$\hat{\alpha} = \frac{\sum_{i=1}^n \beta_i}{n}$$

By taking $K(u)$ to be a Gaussian function, we obtain

$$\begin{aligned} \hat{g}_{\text{SPE}}(b) &= \frac{e^{-\frac{b}{\hat{\alpha}}}}{\hat{\alpha}nh} \sum_{j=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b - \beta_j}{h}\right)^2} \left(\frac{e^{-\frac{\beta_j}{\hat{\alpha}}}}{\hat{\alpha}} \right)^{-1} \\ &= \frac{1}{nh} \sum_{j=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b - \beta_j}{h}\right)^2} e^{-\frac{b}{\hat{\alpha}}} \left(e^{-\frac{\beta_j}{\hat{\alpha}}} \right)^{-1} \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j}{\hat{\alpha}}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{b-\beta_j}{h}\right)^2 - \frac{b}{\hat{\alpha}}} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j}{\hat{\alpha}}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left((b-\beta_j)^2 + 2h^2 \frac{b}{\hat{\alpha}}\right)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j}{\hat{\alpha}}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left(\left(b-\left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)\right)^2 + \frac{2h^2}{\hat{\alpha}}\left(\beta_j - \frac{h^2}{2\hat{\alpha}}\right)\right)} \\
 &= \frac{1}{n} \sum_{j=1}^n \frac{\left(e^{-\frac{\beta_j}{\hat{\alpha}}}\right)^{-1}}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left(b-\left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)\right)^2 - \frac{1}{\hat{\alpha}}\left(\beta_j - \frac{h^2}{2\hat{\alpha}}\right)} \\
 &= \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2}{2\hat{\alpha}^2}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left(b-\left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)\right)^2}
 \end{aligned}$$

The estimation of the distribution function of the time-to-failure by using $\hat{\mathbf{g}}_{\text{SPE}}(b)$ is

$$\begin{aligned}
 \hat{\mathbf{F}}_T(t) &= 1 - \hat{\mathbf{G}}_{\text{SPE}}\left(\frac{D_f}{t}\right) \\
 &= 1 - \int_{-\infty}^{\frac{D_f}{t}} \hat{\mathbf{g}}_{\text{SPE}}(b) db \\
 &= 1 - \int_{-\infty}^{\frac{D_f}{t}} \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2}{2\hat{\alpha}^2}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left(b-\left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)\right)^2} db \\
 &= 1 - \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2}{2\hat{\alpha}^2}} \int_{-\infty}^{\frac{D_f}{t}} \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2h^2}\left(b-\left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)\right)^2} db \\
 &= 1 - \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2}{2\hat{\alpha}^2}} \text{Pro}\left(S \leq \frac{D_f}{t}\right)
 \end{aligned}$$

where S is a random variable distributed as

$$N\left(\beta_j - \frac{h^2}{\hat{\alpha}}, h^2\right)$$

Therefore

$$\hat{F}_T(t) = 1 - \frac{1}{n} \sum_{j=1}^n e^{\frac{h^2}{2\hat{\alpha}^2}} \Phi\left(\frac{\frac{D_f}{t} - \left(\beta_j - \frac{h^2}{\hat{\alpha}}\right)}{h}\right) \quad (8)$$

where $\Phi(\cdot)$ is a standard normal distribution. To estimate the $100r^{\text{th}}$ percentiles (denoted by $\hat{t}_{r\text{-SPH}}$) we should solve $\hat{F}_{T\text{-SPH}}(\hat{t}_{r\text{-SPH}}) = r$ numerically with respect to $\hat{t}_{r\text{-SPH}}$.

Estimating Percentiles of Time-to-Failure Distribution Using Maximum Likelihood (ML) Estimator Method

Consider the simple linear degradation model

$$y_{ij} = \beta_i t_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, n; j = 1, \dots, m$$

where y_{ij} is the observed degradation measurement of the i^{th} unit at time t_{ij} , β_i is the random effect parameter (the slope of the linear degradation model for unit i), t_{ij} is the soft failure for the degradation model, and ε_{ij} is the random error term, where the ε_{ij} are iid with $N(0, \sigma_\varepsilon^2)$. By using the formula of the time-to-failure distribution in (4), we will construct the ML estimator of t_r for the following distributions:

The Half-Normal Distribution

If $\beta_i \sim \text{half normal}(\sigma^2)$, the time-to-failure distribution is

$$F_T(t) = 1 - \int_0^{\frac{D_f}{t}} \sqrt{\frac{2}{\pi\sigma^2}} e^{-\frac{b^2}{2\sigma^2}} db \quad (9)$$

By the Leibniz integral rule, and by differentiating both sides of (9) with respect to t , we obtain the pdf of the time-to-failure distribution, which is

$$f_T(t; \sigma^2) = \frac{D_f}{t^2} \sqrt{\frac{2}{\pi \sigma^2}} e^{-\frac{D_f^2}{2\sigma^2 t^2}} \quad (10)$$

Now, to find the ML estimator of σ^2 , let $t_1, t_2, \dots, t_n$ be a random sample from (10); then the natural logarithm of the likelihood function of σ^2 is

$$\begin{aligned} \ln(L(\sigma^2; t_1, t_2, \dots, t_n)) &= \ln\left(\prod_{i=1}^n f_T(t_i; \sigma^2)\right) \\ &= \ln\left(\left(\sqrt{\frac{2}{\pi \sigma^2}} D_f\right)^n \left(\prod_{i=1}^n \frac{1}{t_i^2}\right) e^{-\frac{D_f^2}{2\sigma^2} \sum_{i=1}^n \frac{1}{t_i^2}}\right) \\ &= n \ln\left(\sqrt{\frac{2}{\pi}} D_f\right) - \frac{n}{2} \ln \sigma^2 - \frac{D_f^2}{2\sigma^2} \sum_{i=1}^n \frac{1}{t_i^2} - \ln\left(\prod_{i=1}^n t_i^2\right) \end{aligned}$$

Now, by differentiating with respect to σ^2 , we obtain

$$\frac{d}{d\sigma^2} \left(\ln(L(\sigma^2; t_1, t_2, \dots, t_n)) \right) = -\frac{n}{2\sigma^2} + \frac{D_f^2}{2\sigma^4} \sum_{i=1}^n \frac{1}{t_i^2}$$

By solving

$$\frac{d}{d\sigma^2} \left(\ln(L(\sigma^2; t_1, t_2, \dots, t_n)) \right) = 0$$

we obtain the ML estimator of σ^2 , which is

$$\sigma_{MLH}^2 = \frac{D_f^2}{n} \sum_{i=1}^n \frac{1}{t_i^2} \quad (11)$$

The ML estimator of t_r (denoted by $\hat{t}_{MLH}$) is obtained by solving the following equation with respect to $\hat{t}_{MLH}$:

$$r = 1 - \int_0^{\frac{D_f}{\hat{t}_{MLH}}} \sqrt{\frac{2}{\pi \sigma^2_{MLH}}} e^{-\frac{b^2}{2\sigma^2_{MLH}}} db \quad (12)$$

The Exponential Distribution

If $\beta_i \sim \exp(\alpha)$, then the time-to-failure distribution is

$$F_T(t) = e^{-\frac{D_f}{\alpha t}} \quad (13)$$

By differentiating both sides of (13) with respect to t , we obtain the pdf of the time-to-failure distribution, which is

$$f_T(t; \alpha) = \frac{D_f}{\alpha t^2} e^{-\frac{D_f}{\alpha t}} \quad (14)$$

To find the ML estimator of α let $t_1, t_2, \dots, t_n$ be a random sample from (14); then the natural logarithm of the likelihood function of α is

$$\begin{aligned} \ln(L(\alpha; t_1, t_2, \dots, t_n)) &= \ln\left(\prod_{i=1}^n f_T(t_i; \alpha)\right) \\ &= \ln\left(\left(\frac{D_f}{\alpha}\right)^n \left(\prod_{i=1}^n \frac{1}{t_i^2}\right) e^{-\frac{D_f}{\alpha} \sum_{i=1}^n \frac{1}{t_i}}\right) \\ &= n \ln(D_f) - n \ln(\alpha) - \frac{D_f}{\alpha} \sum_{i=1}^n \frac{1}{t_i} - \ln\left(\prod_{i=1}^n t_i^2\right) \end{aligned}$$

Therefore

$$\frac{d}{d\alpha} \left(\ln(L(\alpha; t_1, t_2, \dots, t_n)) \right) = -\frac{n}{\alpha} + \frac{D_f}{\alpha^2} \sum_{i=1}^n \frac{1}{t_i}$$

By solving

$$\frac{d}{d\alpha} \left(\ln \left(L(\alpha; t_1, t_2, \dots, t_n) \right) \right) = 0$$

we obtain the ML estimator of α , which is given by

$$\hat{\alpha}_{\text{MLE}} = \frac{D_f}{n} \sum_{i=1}^n \frac{1}{t_i} \quad (15)$$

Then the ML estimator of t_r (denoted by $\hat{t}_{\text{MLE}}$) is

$$\begin{aligned} \hat{t}_{\text{MLE}} &= -\frac{D_f}{\hat{\alpha}_{\text{MLE}} \ln(r)} \\ &= -\frac{n}{\ln(r) \sum_{i=1}^n \frac{1}{t_i}} \end{aligned} \quad (16)$$

Estimating Percentiles of Time-to-Failure Distribution Using OLS Estimator Method

By considering the same degradation model that was studied in the previous section, and by letting $\beta_1, \beta_2, \dots, \beta_n$ be a random sample of size n from the probability density function $g_\beta(b; \mu)$ and the distribution function $G_\beta(b; \mu)$, the OLS estimator of μ (denoted by $\hat{\mu}_{\text{OLS}}$) will be obtained as follows:

$$\begin{aligned} Q(\mu) &= \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - E(y_{ij}) \right)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - t_{ij} E(\beta_i) \right)^2 \end{aligned} \quad (17)$$

where $E(\beta_i)$ is a function of μ .

By minimizing (17) with respect to μ we get the OLS estimator of μ . Then the OLS estimator for the time-to-failure distribution is

$$\hat{F}_{\text{OLS}}(t) = 1 - G_\beta \left(\frac{D_f}{\hat{t}_{r-\text{OLS}}} \right) \quad (18)$$

where $\hat{t}_{r-OLS}$ is the OLS estimator of t_r that is given by solving

$$r = 1 - G_{\beta} \left(\frac{D_f}{\hat{t}_{r-OLS}} \right)$$

By using the formula of the time-to-failure distribution (18) we obtain the OLS estimator for the following distributions:

The Half-Normal Distribution

If $\beta_i \sim \text{half normal}(\sigma^2)$, then the OLS estimator of σ^2 (denoted by σ^2_{OLSH}) is obtained by minimizing

$$Q(\sigma^2) = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - E(y_{ij}))^2$$

where

$$E(y_{ij}) = \sqrt{\frac{2}{\pi}} \sigma^2 t_{ij}$$

Therefore

$$Q(\sigma^2) = \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - \sqrt{\frac{2}{\pi}} \sigma^2 t_{ij} \right)^2$$

and

$$\frac{dQ(\sigma^2)}{d\sigma^2} = \frac{-2}{\sqrt{2\pi}\sigma^2} \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - \sqrt{\frac{2}{\pi}} \sigma^2 t_{ij} \right) t_{ij}$$

Equating the above derivative to zero, we get

$$\sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - \sqrt{\frac{2}{\pi}} \sigma^2 t_{ij} \right) t_{ij} = 0$$

Now, solving the last equation with respect to σ^2 , we obtain

$$\sigma^2_{\text{OLSH}} = \frac{\pi}{2} \left(\frac{\sum_{i=1}^n \sum_{j=1}^m y_{ij} t_{ij}}{\sum_{i=1}^n \sum_{j=1}^m t_{ij}^2} \right)^2 \quad (19)$$

and the $\hat{t}_{r-\text{OLSH}}$ is given by solving the following equation with respect to $\hat{t}_{r-\text{OLSH}}$:

$$r = 1 - G_{\beta} \left(\frac{D_f}{\hat{t}_{r-\text{OLSH}}} \right) \quad (20)$$

The Exponential Distribution

If $\beta_i \sim \exp(\alpha)$, then the OLS estimator of α (denoted by $\hat{\alpha}_{\text{OLSE}}$) is obtained by minimizing

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - E(y_{ij}) \right)^2$$

where $E(y_{ij}) = \alpha t_{ij}$. Thus

$$Q(\alpha) = \sum_{i=1}^n \sum_{j=1}^m \left(y_{ij} - \alpha t_{ij} \right)^2$$

and

$$\frac{dQ(\alpha)}{d\alpha} = -2 \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \alpha t_{ij}) t_{ij}$$

By equating the above derivative to zero, we get

$$\sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \alpha t_{ij}) t_{ij} = 0$$

The OLS estimator of α is obtained by solving the last equation with respect to α . Thus,

$$\hat{\alpha}_{\text{OLSE}} = \frac{\sum_{i=1}^n \sum_{j=1}^m y_{ij} t_{ij}}{\sum_{i=1}^n \sum_{j=1}^m t_{ij}^2} \quad (21)$$

The estimator $\hat{t}_{r-\text{OLSE}}$ of t_r is given by

$$\begin{aligned} \hat{t}_{r-\text{OLSE}} &= -\frac{D_f}{\hat{\alpha}_{\text{OLSE}} \ln(r)} \\ &= -\frac{D_f}{\ln(r)} \frac{\sum_{i=1}^n \sum_{j=1}^m t_{ij}^2}{\sum_{i=1}^n \sum_{j=1}^m y_{ij} t_{ij}} \end{aligned} \quad (22)$$

Simulation Study and Results

Consider the performance of the four estimators of t_r . The bandwidth for each estimator is computed by using formula (1). The bias (B), the mean square error (MSE), and the length of 95% bootstrap confidence interval using the bootstrap percentile method (Efron & Tibshirani, 1993; Racine & MacKinnon, 2007) of each estimator are computed from the data of size n that is simulated from the selected distributions, half normal(σ^2) or exp(α). Let $\beta_1, \beta_2, \dots, \beta_n$ be a random sample of size n generated from one of the above distributions. To compute B and MSE for each of the four estimators of t_r , find the exact value of t_r .

Case 1

If $\beta_i \sim \text{half normal}(\sigma^2)$, then by using equation (9), the distribution function of the time-to-failure is

$$F_T(t) = 1 - \int_0^{\frac{D_f}{t}} \sqrt{\frac{2}{\pi\sigma^2}} e^{-\frac{b^2}{2\sigma^2}} db$$

SEMI-PARAMETRIC ESTIMATION FOR LINEAR DEGRADATION MODEL

and, based on equation (5), the exact value of t_r is

$$t_r = \frac{D_f}{H^{-1}(1-r)}$$

where $H^{-1}(\cdot)$ is the inverse distribution function of the half-normal distribution.

Case 2

If $\beta_i \sim \exp(\alpha)$, then based on equation (13),

$$F_T(t) = 1 - \left(1 - e^{-\frac{D_f}{\alpha t}} \right) = e^{-\frac{D_f}{\alpha t}}$$

and, based on equation (5), the exact value of t_r is

$$t_r = -\frac{D_f}{\alpha \ln(r)}$$

In simulation, the initial values are,

- Sample size $n = 20, 40$, or 60 , $r = 0.1, 0.2$, or 0.3 , and σ^2 or $\alpha = 10$.
- The critical level of degradation $D_f = 20$ for each sample of size n .
- The number of iterations to compute B and MSE is $N = 2000$.
- The number of bootstrap iterations are $M = 1000$.

Simulation results are presented in Tables 1-6.

Table 1. B , MSE, and length of 95% bootstrap confidence interval of estimate t_r from half normal(10) with $n = 20$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	1.2159	0.0514	0.0566	0.2687	0.1518	0.0436	0.0463	0.0359	0.0502	1.3591	1.6369	0.8280	1.2687
0.2	1.5606	0.0579	0.0865	0.3128	0.2269	0.0665	0.0789	0.0474	0.0800	1.7436	2.1057	0.9522	1.6558
0.3	1.9297	0.0670	0.1389	0.4047	0.4195	0.0773	0.1176	0.0627	0.1257	2.3842	3.0097	1.3241	1.9469

Table 2. B , MSE and length of 95% bootstrap confidence interval of estimate t_r from half normal(10) with $n = 40$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	1.2159	0.0250	0.0253	0.1975	0.0721	0.0232	0.0208	0.0152	0.0242	0.9616	1.1542	0.5720	0.8045
0.2	1.5606	0.0376	0.0418	0.2326	0.1097	0.0256	0.0325	0.0214	0.0390	1.0938	1.3996	0.7565	1.0598
0.3	1.9297	0.0271	0.0613	0.2710	0.1695	0.0397	0.0535	0.0196	0.0496	1.4412	1.7483	0.8858	1.3863

Table 3. B , MSE and length of 95% bootstrap confidence interval of estimate t_r from half normal(10) with $n = 60$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	1.2159	0.0156	0.0159	0.1689	0.0485	0.0184	0.0136	0.0072	0.0139	0.7482	0.8255	0.4627	0.6795
0.2	1.5606	0.0146	0.0267	0.1820	0.0689	0.0177	0.0228	0.0185	0.0234	0.9139	1.0511	0.5979	0.8825
0.3	1.9297	0.0186	0.0449	0.2301	0.1182	0.0243	0.0347	0.0171	0.0363	1.3115	1.5285	0.7201	1.0552

SEMI-PARAMETRIC ESTIMATION FOR LINEAR DEGRADATION MODEL

Table 4. B, MSE and length of 95% bootstrap confidence interval of estimate t_r from $\exp(10)$ with $n = 20$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	0.8690	0.0248	0.0641	0.3105	0.1937	0.0525	0.0527	0.0553	0.0535	1.4330	1.7700	0.8370	1.2080
0.2	1.2430	0.0359	0.1100	0.3664	0.2933	0.0732	0.0970	0.0663	0.0999	2.0460	2.4300	1.2230	1.7210
0.3	1.6610	0.0214	0.1794	0.4704	0.5168	0.0998	0.1892	0.0821	0.1815	2.8010	3.7720	1.4590	2.4160

Table 5. B, MSE and length of 95% bootstrap confidence interval of estimate t_r from $\exp(10)$ with $n = 40$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	0.8690	0.0061	0.0269	0.2424	0.0972	0.0260	0.0219	0.0239	0.0227	0.9510	1.1250	0.5590	0.7840
0.2	1.2430	0.0064	0.0545	0.2783	0.1522	0.0312	0.0449	0.0210	0.0455	1.3440	1.5050	0.8460	1.1600
0.3	1.6610	-0.0086	0.0889	0.3434	0.2480	0.0193	0.0734	0.0435	0.0762	1.7670	2.1250	1.0930	1.4220

Table 6. B, MSE and length of 95% bootstrap confidence interval of estimate t_r from $\exp(10)$ with $n = 60$

r	t_r	SPH estimator		SPE estimator		ML estimator		OLS estimator		Length of 95% bootstrap CI			
		B	MSE	B	MSE	B	MSE	B	MSE	SPH	SPE	ML	OLS
0.1	0.8690	-0.0017	0.0183	0.2038	0.0683	0.0150	0.0131	0.0160	0.0136	0.7760	0.8680	0.4490	0.6230
0.2	1.2430	-0.0001	0.0333	0.2404	0.0994	0.0230	0.0284	0.0217	0.0285	1.0520	1.2010	0.6710	0.8720
0.3	1.6610	-0.0224	0.0592	0.2801	0.1590	0.0300	0.0512	0.0242	0.0479	1.3600	1.5970	0.8650	1.1750

From Tables 1-6, the following conclusions may be made:

- The MSE for each estimator decrease as n increases.
- The MSE for each estimator increase as r increases.
- The length of the 95% bootstrap confidence interval is decrease as n increases.
- By comparing the MSE of the four estimators, the ML and OLS estimators have the smallest values of the MSE and the semi parametric half normal estimator close to them. ML estimator has the smallest value of the MSE and the shortest length of a 95% bootstrap confidence interval for each distribution and different sample size, so ML estimator has the best performance.

Estimating $t_{0.5}$ Using Data with Misspecified Density

In this section, we will study and compare the performance of the semi-parametric method, OLS, and ML estimators for t_r when the distribution of the random effect is not chosen correctly. To perform this comparison, generate the random effect from Weibull(2, 15) and assume this generated sample is from half normal(15) or exp(15). The true value of $t_{0.5}$ is

$$t_{0.5} = \frac{20}{W^{-1}(0.5)} = 1.6015$$

where, $W^{-1}(\cdot)$ is the inverse distribution function of the Weibull(2, 15) distribution. Under this misspecification, estimate $t_{0.5}$ using four estimators.

Table 7. Estimating $t_{0.5}$ for a sample from a Weibull(2, 15) distribution that is misspecified as a half normal(15) distribution

Estimator	Bias (B)	Mean Square Error (MSE)
$\hat{t}_{0.5-SPH}$	0.0353	0.0188
$\hat{t}_{0.5-SPE}$	0.0977	0.0319
$\hat{t}_{0.5-ML}$	0.3889	0.1681
$\hat{t}_{0.5-OLS}$	0.1874	0.0500

SEMI-PARAMETRIC ESTIMATION FOR LINEAR DEGRADATION MODEL

Table 8. Estimating $t_{0.5}$ for a sample from a Weibull(2, 15) distribution that is misspecified as an exp(15) distribution

Estimator	Bias (B)	Mean Square Error (MSE)
$\hat{t}_{0.5\text{-SPH}}$	0.03530	0.01880
$\hat{t}_{0.5\text{-SPE}}$	0.09770	0.03190
$\hat{t}_{0.5\text{-ML}}$	0.57644	0.35365
$\hat{t}_{0.5\text{-OLS}}$	0.58674	0.36570

In this simulation, the initial values are

- Sample size $n = 60$, σ^2 or $\alpha = 15$, and $r = 0.5$.
- The critical level of degradation $D_f = 20$.
- The number of iterations to compute B and MSE is $N = 2000$.

Simulation results are presented in Tables 7 and 8.

From Tables 7 and 8 we conclude the following:

- The ML and OLS estimators perform poorly when the random effect distribution is misspecified.
- The semi-parametric half-normal estimator has the best performance.

Real Data Application

The laser degradation data gives the percent increase in laser operating current for GaAs lasers tested at 80°C which is presented in Table c.17 of Meeker and Escobar (1998, p. 642). In this article, failure is assumed to occur at the critical degradation level $D_f = 5$. Figure 1 shows percent increase in operating current for GaAs lasers tested at 80°C.

Data Analysis

Consider the data to estimate the percentiles of the time-to-failure distribution for estimators which have been discussed previously (semi-parametric estimators (SPH & SPE), maximum likelihood estimator (ML), and ordinary least square estimator (OLS)). These estimators will be compared by computing the mean square error (MSE) and the length of the 95% bootstrap confidence interval of the percentiles ($r = 0.5$) of the time-to-failure distribution.

Figure 1. Percent increase in laser operating current for GaAs lasers tested at 80°C

Table 9. Failure time and slope for each unit i

Unit i	t_i	$\hat{\beta}_i$
1	7.1479	0.694
2	8.0192	0.599
3	11.6707	0.445
4	12.0667	0.418
5	10.6066	0.460
6	7.5732	0.692
7	12.5926	0.407
8	12.6000	0.390
9	10.2093	0.493
10	6.6667	0.756
11	10.5000	0.475
12	9.4884	0.505
13	8.9730	0.523
14	11.2931	0.430
15	12.4933	0.408

To compute the MSE, the true value of t_r and the values of $\beta_1, \beta_2, \dots, \beta_n$, the slopes of the linear model (3), must be known. From the laser degradation data, scale the times of the degradation measurements by dividing each time by 250. To get the failure time, use linear interpolation and, by fitting simple linear regression

between t_i and y_{ij} , obtain the estimation of $\beta_1, \beta_2, \dots, \beta_n$. Table 9 contains the time-to-failure t_i and the values of the slope estimate $\hat{\beta}_i$.

Estimating the 50th Percentile of the Time-to-Failure Distribution

Under the assumption that the random effect parameter is distributed as half normal(σ^2) or exp(α), we estimate the 50th percentile of the time-to-failure distribution, $t_{0.5}$, for the semi-parametric estimators using formulas (7) and (8), OLS estimator, and ML estimator as follows:

To estimate $t_{0.5}$ using the semi-parametric estimators:

1. Take a random sample, with replacement, of size 15 from the slopes in Table 9.
2. Depending on this sample, obtain $\hat{t}_{0.5\text{-SPH}}$ and $\hat{t}_{0.5\text{-SPE}}$ by solving $\hat{F}_{T\text{-SPH}}(\hat{t}_{0.5\text{-SPH}}) = 0.5$ and $\hat{F}_{T\text{-SPE}}(\hat{t}_{0.5\text{-SPE}}) = 0.5$, respectively.

To estimate $t_{0.5}$ using the ML estimator:

1. Take a random sample, with replacement, of size 15 from the failure times t_i in Table 9.
2. Depending on this sample, obtain $\hat{t}_{0.5\text{-ML}}$ using (12) or (16) according to the assumed distribution.

Table 10. The results of the real data under the assumption $\beta_i \sim \text{half normal}(\sigma^2)$

Estimator	Bias	MSE	Length of 95% bootstrap CI
Semi parametric half normal	-0.1345	0.4562	2.6621
Semi parametric exponential	-0.1003	0.4302	2.6232
ML	3.6277	13.8605	3.1407
OLS	1.5333	7.7073	7.3704

Table 11. The results of the real data under the assumption $\beta_i \sim \exp(\sigma^2)$

Estimator	Bias	MSE	Length of 95% bootstrap CI
Semi parametric half normal	-0.1345	0.4562	2.6621
Semi parametric exponential	-0.1003	0.4302	2.6232
ML	3.5381	13.1561	3.0727
OLS	4.2186	25.4938	8.9887

To estimate $t_{0.5}$ using the OLS estimator:

1. Take a random sample, with replacement, of size 15 of vectors from the data, where each vector consists of y_{ij} and time t_j for $i = 1, 2, \dots, 15, j = 1, 2, \dots, 16$.
2. Depending on this sample, obtain $\hat{t}_{0.5-OLS}$ using (20) or (22) according to the assumed distribution.

From Tables 10 and 11, the following may be concluded:

- The estimation of the median of the time-to-failure distribution using the semi-parametric exponential estimator has the smallest MSE value and smallest 95% confidence interval length.
- The performance of the semi-parametric half-normal estimator and semi-parametric exponential estimator are comparable.
- The ML and OLS estimators perform poorly compared to the semi-parametric estimators.

Conclusions

When the distribution of the random effect is assumed to be known, the ML and OLS estimators of t_r perform better than the semi-parametric estimators in terms of the MSE and the length of the 95% bootstrap confidence interval. Otherwise, the semi-parametric estimators perform best.

References

Al-Haj Ebrahim, M., Eidous, O., & Kmail, G. (2009). Estimating percentiles of time-to-failure distribution obtained from a linear degradation

model using the kernel density method. *Communication in Statistics – Simulation and Computation*, 38(9), 1811-1822. doi: [10.1080/03610910903145130](https://doi.org/10.1080/03610910903145130)

Amster, S. J., & Hooper, J. H. (1983, August). *Accelerated life tests with measured degradation data and growth models*. Paper presented at the American Statistical Association Annual Meeting, Toronto, Canada.

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

Gertsbakh, I. B., & Kordonskiĭ, Kh. B. (1969). *Models of failure* (Scripta Technica Inc., Trans.). New York, NY: Springer-Verlag. (Original work published 1966)

Lu, C. J., Meeker, W. Q., & Escobar, L. A. (1996). A comparison of distribution and failure time analysis methods for estimating a time-to-failure distribution. *Statistica Sinica*, 6(3), 531-546. Available from <http://www.jstor.org/stable/24305606>

Meeker, Q. W., & Escobar, L. A. (1998). *Statistical method for reliability data*. New York, NY: John Wiley and Sons, Inc.

Racine, J. S. & MacKinnon, J. G. (2007). Inference via kernel smoothing of bootstrap *P* values. *Computational Statistics & Data Analysis*, 51(12), 5949-5957. doi: [10.1016/j.csda.2006.11.013](https://doi.org/10.1016/j.csda.2006.11.013)

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London, UK: Chapman and Hall.

Prediction of Percent Change in Linear Regression by Correlated Variables

Stan Lipovetsky
GfK North America
Minneapolis, MN

Multiple linear regression can be applied for predicting an individual value of dependent variable y by the given values of independent variables x . But it is not immediately clear how to estimate percent change in y due to changes in predictors, especially when those are correlated. This work considers several approaches to this problem, including its formulation via predictors adjusted by their correlation structure. Ordinary least squares regression is used, together with Shapley value regression and another model based on solving some system of differential equations. Numerical estimations performed for a real marketing research data demonstrate meaningful results. The considered techniques can be very useful in practical estimations of the percent change of dependent variable by the change in predictors.

Keywords: Multiple regression, percent change in outcome, predictors percent change, correlated structure, OLS, Shapley value regression, differential equations

Introduction

Ordinary least squares (OLS) multiple linear regression is one of the main tools of statistical modeling widely used for estimation the dependent variable (DV) value by the values of independent variables (IVs), or predictors. In applied studies, often the need is to estimate not a DV value itself but a percent change in the outcome due to percent changes in IVs. Several problems should be solved in such estimations, beginning from finding a good criterion for evaluation of the percent change in the outcome. The mean values of DV and IVs satisfy a linear regression model with intercept, and this relation is employed for measuring percentage change in the output due to changes in the input variables.

Also, the predictors are only called independent in contrast to the dependent variable, but IVs in a sampled data are always correlated and not statistically

Dr. Stan Lipovetsky is the Senior Research Director in the Marketing Sciences department. Email him at: stan.lipovetsky@gfk.com.

PREDICTION OF PERCENT LIFT IN LINEAR REGRESSION

independent. So the correlation structure should be accounted in finding the IVs' values for adequate prediction by regression. And there is a problem of multicollinearity among IVs which produces regressions with inflated values of coefficients, yields their signs opposite to the signs of the pair correlations, makes theoretically important variables to get small coefficients, causes a reduction in statistical power, and leads to wider confidence intervals for the coefficients so they could be incorrectly identified as being insignificant (Grapentine, 1997; Mason & Perreault, 1991). For instance, in marketing research, it is often known in advance that the influence of each IV on the DV of customer satisfaction should be positive, and it is supported by the pair correlations. But in the OLS regression many coefficients can occur to be negative, so for predicting a change in the output should we increase or decrease a presumably beneficial variable which, however, has a negative coefficient in the model?

To overcome deficiencies of multicollinearity and produce meaningful regression parameters, various modifications of OLS have been developed. Among those are: ridge regressions (Hoerl & Kennard, 1970; Lipovetsky, 2010), Shapley value regression (SVR) based on cooperative game theory used for finding predictors' importance and adjusting the regression coefficients (Shapley, 1953; Roth, 1988; Lipovetsky & Conklin, 2001), nonlinear parameterization of linear regression coefficients by multinomial shares, using elasticity criterion for building regression coefficients by data gradients, Gibson-Johnson and Johnson indices of predictor importance, and other techniques (for more detail and references within: Gibson, 1962; R. Johnson, 1966; J. Johnson, 2000; Lipovetsky, 2013; Lipovetsky & Conklin, 2014). Those techniques produce very similar models so we can employ one of them, SVR, as the model with interpretable coefficients.

Several approaches to the problem of estimating a percent change in the DV due to the percent changes in the IVs are considered here. OLS and SVR are used as the models for predicting change in the outcome for a given set of the predictors' values. These values are adjusted due to the structure of correlations among the predictors. For this aim, the mutual regressions of each predictor by all the other predictors are used. In another approach, a system of linear differential equations is considered as well.

Change Estimation by Multiple Regression

Consider several main relations of the OLS regression that will be needed further. A multiple linear regression can be presented as a model:

$$y_i = a_0 + a_1x_{i1} + \dots + a_nx_{in} + \varepsilon_i \equiv \hat{y}_i + \varepsilon_i \quad (1)$$

where x_{ij} and y_i are i^{th} observations ($i = 1, \dots, N$ – number of observations) by each j^{th} IV ($j = 1, \dots, n$ – number of variables) and by the DV, a_j are the coefficients of regression, a_0 is its intercept, $\hat{y}_i$ denotes theoretical linear aggregate of the predictors, and ε_i are the deviations from the theoretical relationship. Least squares (LS) objective for finding coefficients consists in minimization of the total of squared deviations:

$$S^2 = \sum_{i=1}^N \varepsilon_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - a_0 - a_1x_{i1} - \dots - a_nx_{in})^2 \quad (2)$$

Derivatives $\partial S^2 / \partial a_j = 0$ yield the normal system of equations for finding parameters of the model. Such a derivative by a_0 equaled zero produces the expression:

$$\bar{y} = a_0 + a_1\bar{x}_1 + \dots + a_n\bar{x}_n \quad (3)$$

where a bar above variables denotes their mean values. This relation is used to calculate the intercept when other coefficients of regression are found in the solution of the normal system which can be expressed in the matrix form as follows:

$$\mathbf{a} = \mathbf{C}_{xx}^{-1} \mathbf{c}_{xy} \quad (4)$$

where $\mathbf{a}$ denotes the vector of coefficients $a_1, \dots, a_n$ of regression, $\mathbf{C}_{xx}$ and $\mathbf{C}_{xx}^{-1}$ are the n -by- n covariance matrix between x s and its inverse matrix, respectively, and $\mathbf{c}_{xy}$ is the n^{th} order vector of covariance between x s and y . The results (3)-(4) present the OLS regression parameters.

Suppose the mean level of each driver $\bar{x}_j$ in the model can be increased by different portions f_j (100 f_j percent), so the absolute change (denoted by d) and new values (denoted by prime) can be written as follows:

$$d\bar{x}_j = \bar{x}_j f_j, \quad \bar{x}_j' = \bar{x}_j + d\bar{x}_j = \bar{x}_j (1 + f_j) \quad (5)$$

PREDICTION OF PERCENT LIFT IN LINEAR REGRESSION

Multiplying the parameters (4) by the vector of new values $\bar{x}'$ and adding the intercept, we find a predicted DV value by the regression model. Subtracting (3) from it yields the similar relation for the changes in mean levels:

$$d\bar{y} = a_1 d\bar{x}_1 + \dots + a_n d\bar{x}_n \quad (6)$$

Dividing both sides of (6) by the mean level of y , and dividing each change in x_j by the corresponding mean value, we transform (6) to the following relation

$$\frac{d\bar{y}}{\bar{y}} = \left(a_1 \frac{\bar{x}_1}{\bar{y}} \right) \frac{d\bar{x}_1}{\bar{x}_1} + \dots + \left(a_n \frac{\bar{x}_n}{\bar{y}} \right) \frac{d\bar{x}_n}{\bar{x}_n} = \left(a_1 \frac{\bar{x}_1}{\bar{y}} \right) f_1 + \dots + \left(a_n \frac{\bar{x}_n}{\bar{y}} \right) f_n \quad (7)$$

which presents the relative (%) change in the DV outcome via % changes f_j in the IVs (5) and regression coefficients modified to the form $a_j \bar{x}_j / \bar{y}$.

Instead of OLS regressions (4) it is possible to use Shapley value regression (Lipovetsky & Conklin, 2001). Coefficients of SVR have a more adequate meaning (similar to elasticity but in absolute changes) as a change in the output due to the unit change in each predictor holding other predictors constant. The modified coefficients in (7) also have a clear interpretation if considered via the SVR parameters.

Adjusting Independent Variables by Correlation Structure

Suppose the percentage changes f_j for x s are given so they can be used in (5) for finding the new values $\bar{x}'_j$. However, these values for prediction should be adjusted due to the correlations among the predictors given in the matrix $\mathbf{C}_{xx}$. A convenient way to such an adjustment is as follows: The diagonal elements of inverted correlation matrix $\mathbf{C}_{xx}^{-1}$ used in (4) (called variance inflation factors, VIF) equal the reciprocal values of the residual sums of squares in the regressions of each variable x_j by the rest of IVs,

$$\text{VIF}_j = (\mathbf{C}_{xx}^{-1})_{jj} = 1 / (1 - R_{j.}^2) \quad (8)$$

where $R_{j.}^2$ are the coefficients of multiple determination in the models of each x_j by the other x s. These models in the so-called Yule's notations are:

$$x_j = a_{j0} + a_{j1}x_1 + a_{j2}x_2 + \dots + a_{j,j-1}x_{j-1} + a_{j,j+1}x_{j+1} + \dots + a_{jn}x_n \quad (9)$$

where a_{jk} denotes a parameter of j^{th} regression by k^{th} variable among all $n - 1$ other x s (Kendall & Stuart, 1973; Lipovetsky & Conklin, 2004). The non-diagonal elements in any j^{th} row of $\mathbf{C}_{xx}^{-1}$ divided by the diagonal element in the same row and taken with opposite signs coincide with coefficients of regression x_j by all other x s (9) that can be presented as the following matrix $\mathbf{A}$:

$$\mathbf{A} \equiv \text{diag}(\mathbf{C}_{xx}^{-1})^{-1} \mathbf{C}_{xx}^{-1} = \begin{pmatrix} 1 & -a_{12.} & -a_{13.} & \dots & -a_{1n.} \\ -a_{21.} & 1 & -a_{23.} & \dots & -a_{2n.} \\ \dots & \dots & \dots & \dots & \dots \\ -a_{n1.} & -a_{n2.} & -a_{n3.} & \dots & 1 \end{pmatrix} \quad (10)$$

The product $\mathbf{A}\bar{\mathbf{x}}$ of the matrix (10) and the vector of mean values $\bar{\mathbf{x}}$ of all predictors coincides with the vector of intercepts $\mathbf{a}_0$ for the mutual regressions (9):

$$a_{j0} = \bar{x}_j - (a_{j1}\bar{x}_1 + a_{j2}\bar{x}_2 + \dots + a_{j,j-1}\bar{x}_{j-1} + a_{j,j+1}\bar{x}_{j+1} + \dots + a_{jn}\bar{x}_n) \quad (11)$$

Multiplying (10) by the vector of the new values $\bar{\mathbf{x}}'$ (with elements $\bar{x}'_j$ (5)) and subtracting the vector of intercepts (11) yields the difference of these values $\bar{\mathbf{x}}'$ and their predictions $\bar{\mathbf{x}}'_{\text{pred}}$ by the models (9) of mutual correlation structure

$$\mathbf{A}\bar{\mathbf{x}}' - \mathbf{A}\bar{\mathbf{x}} = \bar{\mathbf{x}}' - \bar{\mathbf{x}}'_{\text{pred}}$$

From this equation, the vector of predictions adjusted by the correlation structure is presented as follows:

$$\bar{\mathbf{x}}'_{\text{pred}} = (\mathbf{I} - \mathbf{A})\bar{\mathbf{x}}' + \mathbf{A}\bar{\mathbf{x}} \quad (12)$$

where $\mathbf{I}$ is the n^{th} order identity matrix. Then vector of updated relative changes equals

$$\mathbf{f}' = (\bar{\mathbf{x}}'_{\text{pred}} / \bar{\mathbf{x}}) - 1 \quad (13)$$

and can be used in (7) for estimating the relative change in the DV.

System of Differential Equations for Adjusting Independent Variables

Another way of performing predictor mutual adjustment consists in modeling with a system of equations defining change in each predictor as an aggregate of the values of all the predictors:

$$\left\{ \begin{array}{l} \frac{dx_1}{dz} = b_{10} + b_{11}x_1 + b_{12}x_2 + \dots + b_{1n}x_n \\ \frac{dx_2}{dz} = b_{20} + b_{21}x_1 + b_{22}x_2 + \dots + b_{2n}x_n \\ \vdots \\ \frac{dx_n}{dz} = b_{n0} + b_{n1}x_1 + b_{n2}x_2 + \dots + b_{nn}x_n \end{array} \right. \quad (14)$$

Each predictor derivative on the left-hand side of (14) is presented as a linear model by all the variables on the right-hand side. The derivatives are taken by some variable z identifying a general direction in which all predictors vary. For instance, if observations are gathered in time, it can serve as this profiling variable; price can be another example. Otherwise, if there is no evident variable which can be used for trending other variables by it, the principal component analysis (PCA) approach can be applied to the predictors:

$$\mathbf{C}_{xx}\mathbf{a} = \mu\mathbf{a} \quad (15)$$

where μ and $\mathbf{a}$ are the eigenvalues and eigenvectors of the covariance matrix $\mathbf{C}_{xx}$ (15). Scores by the first component $\mathbf{a}_1$ with maximum variance μ_1 can be taken as the variable z defining direction of the main variability of all the x s combined:

$$\mathbf{z} = \mathbf{X}\mathbf{a}_1 \quad (16)$$

Then in practical terms, construct a smoothed nonlinear trend of each $\mathbf{x}$ by the new variable $\mathbf{z}$ (for instance, using the “loess” function available in the R software), order all the observation points by ascending values of $\mathbf{z}$, and find the derivatives on the left-hand side of (14) as change in each two subsequent values of each

predictor $\mathbf{x}_j$ by increment in $\mathbf{z}$ values. In the next step, find coefficients in each equation in (14) separately as a linear regression. Having all the coefficients b_{jk} in (14), consider simultaneously solving of this system of linear differential equations.

Similar systems of homogeneous (without intercepts) equations are used in statistics for description of discrete state and continuous time Markov models presented as Chapman-Kolmogorov differential equations describing stochastic processes. As it is well known (e.g., Bellman, 1960; Pennisi, 1972), the solution of a homogeneous linear system of differential equations with constant coefficients can be presented as follows:

$$\mathbf{x}(\mathbf{z}) = \mathbf{P} \text{diag}(\exp(\lambda_j \mathbf{z})) \mathbf{c} \quad (17)$$

where $\mathbf{c}$ is a vector of constants, and λ_j are the eigenvalues and $\mathbf{P}$ is the corresponding matrix of columns $\mathbf{p}_j$ of eigenvectors obtained in solving the problem:

$$\mathbf{B}\mathbf{p} = \lambda\mathbf{p} \quad (18)$$

The eigenproblem (18) is considered for the matrix $\mathbf{B}$ of the coefficients b_{jk} (without intercepts b_{j0}) at the right-hand side of (14). For the value $\mathbf{z} = 0$ (corresponding to the mean of PCA scores) the solution (17) reduces to the vector of initial conditions $\mathbf{x}(0) = \mathbf{P}\mathbf{c}$, so we obtain the vector of the constants $\mathbf{c} = \mathbf{P}^{-1}\mathbf{x}(0)$. Using it in (17), the general solution of a homogeneous system of differential equations can be represented as

$$\mathbf{x}(\mathbf{z}) = \mathbf{P} \text{diag}(\exp(\lambda_j \mathbf{z})) \mathbf{P}^{-1}\mathbf{x}(0) \quad (19)$$

The expression $\mathbf{P} \text{diag}(\exp(\lambda_j \mathbf{z})) \mathbf{P}^{-1}\mathbf{x}(0)$ in (19) is known as matrix exponent. Each component of the vector $\mathbf{x}(\mathbf{z})$ is a linear combination of the exponents $\exp(\lambda_j \mathbf{z})$ which behave in accordance with the specific values of λ_j in the eigenproblem (18). For a real matrix $\mathbf{B}$ the eigenvalues (18) are real numbers or conjugated pairs of complex numbers, which correspond to exponents and to oscillating sine and cosine parts of functions. There also can be polynomial items corresponded to equal eigenvalues, although in practical numerical evaluations such cases are rare. The eigenvectors $\mathbf{p}$ corresponded to the complex eigenvalues are also complex, but the total expression (19) yields real values.

PREDICTION OF PERCENT LIFT IN LINEAR REGRESSION

When a homogeneous system of equations is solved in (19), the system (14) with intercepts can be solved as well. Besides exponents, a solution of a nonhomogeneous system can contain additional polynomial by $\mathbf{z}$ items (see the handbook by Kamke (1959), an example eq. 8.10). Taking exponents with already-known parameters λ_j with a polynomial part as a theoretical model of each predictor dependence on the aggregate $\mathbf{z}$, we construct the regressions $\mathbf{x}_j(\mathbf{z})$. For a set of given values $\bar{\mathbf{x}}'$ for predictors (5), we find the corresponding values of PCA scores $\mathbf{z}$ (16) and use them in the models $\mathbf{x}_j(\mathbf{z})$ to adjust predictors by their mutual structure:

$$\bar{\mathbf{x}}'_{j,\text{pred}} = \mathbf{x}_j(\mathbf{z}) = \mathbf{x}_j(\bar{\mathbf{x}}'\alpha_1) \quad (20)$$

The adjusted values (20) are used in estimation of the vector of updated relative changes (13) and then in (7) for finding the relative change in the DV.

Numerical Example

For a numerical example, a data set of 242 respondents from a real marketing research project on customer satisfaction with a service center is considered (this data was used in Lipovetsky and Conklin (2001) for SVR modeling). The variables are measured in a Likert 7-point scale and they are: y – overall satisfaction of clients with the company; x_1 – customer satisfaction with service representatives; x_2 – service representatives are courteous; x_3 – they provide all the needed information; x_4 – they give quick response; x_5 – they show care with customer problems; x_6 – they are accurate in the answers; x_7 – they take all the necessary actions.

Table 1. Correlations and regressions

Variable	Mean	cor(y, x)	Coefficients	OLS	SVR
y	5.8554	1.0000	a_0	1.5584	1.1359
x_1	6.3223	0.5432	a_1	0.2831	0.1305
x_2	6.5909	0.4503	a_2	-0.0292	0.1013
x_3	6.2727	0.5451	a_3	0.1792	0.1126
x_4	6.4339	0.4335	a_4	-0.0508	0.0955
x_5	6.1777	0.5110	a_5	0.0506	0.0892
x_6	6.3843	0.5462	a_6	0.2226	0.1292
x_7	6.2562	0.5045	a_7	0.0270	0.0849
			R^2	0.3560	0.3447

Table 2. Correlations among the predictors

	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.7485	0.6981	0.5891	0.8274	0.7058	0.8023
x2	0.7485	1.0000	0.6277	0.5275	0.7283	0.6853	0.6587
x3	0.6981	0.6277	1.0000	0.7691	0.7156	0.8952	0.7600
x4	0.5891	0.5275	0.7691	1.0000	0.6225	0.7592	0.5912
x5	0.8274	0.7283	0.7156	0.6225	1.0000	0.7161	0.7527
x6	0.7058	0.6853	0.8952	0.7592	0.7161	1.0000	0.6915
x7	0.8023	0.6587	0.7600	0.5912	0.7527	0.6915	1.0000

Table 3. Change in DV by the given and adjusted changes in predictors, in %

Variable	Given increase	Adjusted increase calculated by		
		OLS	SVR	Diff
x1	5.000	5.509	4.219	6.045
x2	4.000	3.195	2.788	4.142
x3	2.000	2.181	4.281	0.517
x4	3.000	1.614	2.450	5.716
x5	10.000	4.802	3.980	0.570
x6	1.000	2.799	3.971	5.068
x7	4.000	5.046	4.720	8.373
y _{pred} OLS	2.505	2.989		2.993
y _{pred} SVR	3.161		3.047	3.513

Presented in Table 1 are mean values of the variables, the pair correlations of y with x s, and coefficients of OLS and SVR models with their coefficients of multiple determination R^2 . The DV is correlated with all IVs rather evenly, so each variable could be useful in customer satisfaction impact. In the OLS model some predictor parameters are very close to zero and two of them are negative, but in SVR all parameters are more evenly distributed and all positive. In contrast to OLS, the SVR results are meaningful by all individual coefficients of regression. Judging by the mean values, it could be possible to change the overall satisfaction by improving the predictors.

Correlations between the predictors are shown in Table 2, and they are rather high, so the structure of correlated changes should be taken into account in prediction.

Suppose managers elaborate an improvement program that can result in increasing the mean values in each of the seven predictors by 5, 4, 2, 3, 10, 1, and 4 percent, respectively. The natural question is – what percent of change can be reached for the overall satisfaction by taking these measures? Table 3 in the first

PREDICTION OF PERCENT LIFT IN LINEAR REGRESSION

numerical column presents these given predictors' increase, and the corresponding results y_{pred} of the DV change estimated as 2.51% and 3.16% by the OLS and SVR models, respectively, are in the two bottom lines.

Shown in the last three columns in Table 3 are the predictor changes adjusted by the correlation structure using OLS, SVR, and differential equations (denoted Diff in the table). With the predictor increments found in OLS and SVR adjustment by correlation structure, the change in DV equals 2.99% and 3.05%, respectively. Using predictors from the last column of Diff adjustment, we can estimate by both OLS and SVR the yield in DV as 2.99% and 3.51%, respectively. In general, the adjustment increase the output prediction y_{pred} , especially in Diff estimations. To accept a more conservative expectation of the yield among all those adjusted by correlation structure, we can take the SVR prediction y_{pred} of 3.05%. The mean increase is then $5.8554(1 + 0.0305) = 6.034$, which is already above the next level of the seven-point Likert scale of overall satisfaction of clients with company.

Summary

A problem of estimation of a percent change in the dependent outcome variable due to changes in predictors, especially when those are correlated, was considered. Several questions were studied, including formulation of the problem via relation between mean values of all variables, and adjustment of predictors by their correlation structure in the ordinary least squares regression, Shapley value regression, and a model based on solving a system of differential equations. Numerical estimations performed for a real marketing research data set demonstrate meaningful results. Future research can include estimation of the percent change of the outcome due to changes in only a subset of predictors which can be reset almost precisely to some new values (control variables) whereas others cannot be manipulated; however, these would change anyway as a result of their natural correlations with control variables. Lift in the binary outcome and measures improvement in some utility due to selecting observations based on predicted performance (e.g. percent of top performers as predicted by logistic regression) versus a percent change by random selection can be studied as well. In the context of linear regression it is also possible to select, say, the 10% of units with the highest predicted outcome and compare with that 10% of units selected at random, then select the top 20% vs. another randomly selected 20%, etc, which would allow constructing a lift curve. The considered approach can be very useful in practical applications required estimation of the percent change of dependent variable by the

change in predictors in various problems of applied statistical modeling and prediction.

References

- Bellman, R. (1960). *Introduction to matrix analysis*. New York, NY: McGraw-Hill.
- Gibson, W. A. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychological Reports*, 11(1), 32-34. doi: [10.2466/pr0.1962.11.1.32](https://doi.org/10.2466/pr0.1962.11.1.32)
- Grapentine, A. (1997). Managing multicollinearity. *Marketing Research*, 9(3), 11-21.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. doi: [10.2307/1267351](https://doi.org/10.2307/1267351)
- Johnson, R. M. (1966). The minimal transformation to orthonormality. *Psychometrika*, 31(1), 61-66. doi: [10.1007/bf02289457](https://doi.org/10.1007/bf02289457)
- Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1), 1-19. doi: [10.1207/s15327906mbr3501_1](https://doi.org/10.1207/s15327906mbr3501_1)
- Kamke, E. (1959). *Differentialgleichungen losungsmethoden und losungen, I: Gewöhnliche differentialgleichungen*. Leipzig, Germany: B. G. Teubner.
- Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. 2). New York, NY: Hafner Publishing.
- Lipovetsky, S. (2010). Enhanced ridge regressions. *Mathematical and Computer Modelling*, 51(5-6), 338-348. doi: [10.1016/j.mcm.2009.12.028](https://doi.org/10.1016/j.mcm.2009.12.028)
- Lipovetsky, S. (2013). How good is best? Multivariate case of Ehrenberg-Weisberg analysis of residual errors in competing regressions. *Journal of Modern Applied Statistical Methods*, 12(2), 242-255. doi: [10.22237/jmasm/1383279180](https://doi.org/10.22237/jmasm/1383279180)
- Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4), 319-330. doi: [10.1002/asmb.446](https://doi.org/10.1002/asmb.446)
- Lipovetsky, S., & Conklin, M. (2004). Enhance-synergism and suppression effects in multiple regression. *International Journal of Mathematical Education in Science and Technology*, 35(3), 391-402. doi: [10.1080/00207390310001658492](https://doi.org/10.1080/00207390310001658492)

PREDICTION OF PERCENT LIFT IN LINEAR REGRESSION

Lipovetsky, S., & Conklin, M. (2014). Predictor relative importance and matching regression parameters. *Journal of Applied Statistics*, 42(5), 1017-1031. doi: [10.1080/02664763.2014.994480](https://doi.org/10.1080/02664763.2014.994480)

Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28(3), 268-280. doi: [10.2307/3172863](https://doi.org/10.2307/3172863)

Pennisi, L. L. (1972). *Elements of ordinary differential equations*. New York, NY: Holt.

Roth, A. E. (Ed.). (1988). *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge, UK: Cambridge University Press. doi: [10.1017/cbo9780511528446](https://doi.org/10.1017/cbo9780511528446)

Shapley, L. S. (1953). A value for n -person games. In H. W. Kuhn & A. W. Tucker (Eds.). *Contribution to the theory of games (AM-28) II* (pp. 307-318). Princeton, NJ: Princeton University Press. doi: [10.1515/9781400881970-018](https://doi.org/10.1515/9781400881970-018)

Approximating the Distribution of Indefinite Quadratic Forms in Normal Variables by Maximum Entropy Density Estimation

Ghasem Rekabdar

Abadan Branch, Islamic Azad University
Abadan, Iran

Rahim Chinipardaz

Shahid Chamran University of Ahvaz
Ahvaz, Iran

The quadratic form of non-central normal variables is presented based on a sum of weighted independent non-central chi-square variables. This presentation provides moments of quadratic form. The maximum entropy method is used to estimate the density function because distribution moments of quadratic forms are known. A Euclidean distance is proposed to select an appropriate maximum entropy density function. In order to compare with other methods some numerical examples were evaluated. Also, for discrimination between two groups by the Euclidean distances, we obtained a stochastic representation for the linear discriminant function using the quadratic form. The maximum entropy estimation was an acceptable method to approximate the distribution of quadratic forms in normal variables.

Keywords: Quadratic forms, maximum entropy density estimation, non-central chi-square distribution, linear discriminant analysis

Introduction

The distribution of quadratic forms in normal vectors or sums of weighted independent non-central chi-square variables are considered in some of applied statistical problems (Mathai & Provost, 1992). Researchers have introduced different methods to approximate the distribution of a weighted sum of chi-square variables. Distribution approximation based on moments is a simple method with suitable accuracy. This method is also used frequently for approximate distribution of the quadratic form. For the distribution of non-negative quadratic forms in non-central normal variables, Patnaik's two moments (Patnaik, 1949) and Pearson's

Ghasem Rekabdar is an Assistant Professor in the Department of Mathematics. Email them at: ghasem_rekabdar@yahoo.com. Rahim Chinipardaz is Faculty of Mathematical and Computer Science in the Department of Statistics. Email them at: chinipardaz_r@scu.ac.ir.

three moments (Pearson, 1959) central chi-square approximation are the simplest methods that are used in abundance. Recently, Liu, Tang and Zhang (2009) proposed a non-central chi-square approximation with the unknown degrees of freedom and non-centrality parameter determined by the first four cumulants of the quadratic forms. These methods are accurate for the approximate upper tail of definite (non-negative) quadratic forms and, in the case of indefinite quadratic forms, are not suitable. In the case of indefinite forms, the indefinite quadratic form can be written as the difference of two independent definite quadratic forms. In this case, the density function of a positive definite quadratic form can be approximated according to polynomial gamma or generalized gamma density functions (Mohsenipour & Provost, 2011). In light of this method the density function of the indefinite quadratic form can be approximated by the distribution of the difference of two polynomial gammas. In gamma-polynomial density approximation, we can use more than four moments to approximate the distribution of quadratic forms.

The maximum entropy density estimation is a flexible method to assign values of probability distributions based on limited information such as moments. In general, more limited information such as the percentiles of distribution are recommended. The maximum entropy approach, proposed by Jaynes (1957), is a flexible and powerful tool for density estimation. It was proposed for solving problems with little information about distribution. The maximum entropy method is a suitable tool to estimate properly all special distributions such as normal, exponential, Cauchy, etc. Expressed in Jaynes' language, all known special distributions represent an unbiased probability distribution when some information is not available (Zong, 2006). In this study, the maximum entropy method will be used to approximate indefinite quadratic forms distributions because the moments of weighted sums of non-central chi-square variable are known. As a result, this information can be used to maximize Shannon's entropy. Also, a new approach based on the distance between distributions is proposed to select the number of constraints and compared with the conventional method.

Stochastic Representation of Quadratic Forms

Recall some definitions and basic properties of indefinite quadratic forms in non-central normal variables: Let $\mathbf{X} = (X_1, \dots, X_d)'$ be a multivariate normal random vector $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. The quadratic form in the random variable $\mathbf{X}$ associated with a $d \times d$ real symmetric matrix $\mathbf{A}$ is defined by

$$Q = \mathbf{X}'\mathbf{A}\mathbf{X} \quad (1)$$

Because (1) is a nonlinear combination of correlated univariate normal random variables, it is hard to obtain any approximation directly from it. Instead, we can derive a stochastic representation for quadratic forms in terms of some simpler distribution, i.e.

$$Q = \sum_{i=1}^n \lambda_i \chi_i^2(\gamma_i^2) \quad (2)$$

where the $\chi_i^2(\gamma_i^2)$ are independent non-central chi-square random variables with one degree of freedom and non-centrality parameters γ_i^2 . The weights $\lambda_1 \geq \dots \geq \lambda_d$ are obtained by the spectral decomposition theorem, i.e.

$$\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{A}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}'$$

where $\mathbf{\Lambda}$ is a diagonal matrix for which the diagonal elements $\lambda_1, \dots, \lambda_d$ are the eigenvalues of matrix $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{A}\mathbf{\Sigma}^{\frac{1}{2}}$, $\mathbf{\Sigma}^{\frac{1}{2}}$ denotes the symmetric square root of matrix $\mathbf{\Sigma}$, and $\mathbf{H}$ is an orthogonal matrix. The non-centrality parameters γ_i^2 are obtained by taking the square elements of the vector

$$\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)' = \mathbf{H}'\mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}$$

According to this stochastic representation, obtain the cumulant generating function of Q by

$$C(t) = \frac{1}{2} \sum_{i=1}^d \log(1 - 2\lambda_i t) + \sum_{i=1}^d \frac{\lambda_i \gamma_i^2 t}{(1 - 2\lambda_i t)}$$

The formula for the r^{th} cumulant of the quadratic form is

$$\kappa_r = 2^{r-1} (r-1)! \sum_{i=1}^d \lambda_i^r (1 + r\gamma_i^2) \quad (3)$$

DISTRIBUTION OF QUADRATIC FORMS BY DENSITY ESTIMATION

The moments of the quadratic form can be obtained from its cumulants by means of the recursive relationship obtained by Smith (1995). According to this formula, the r^{th} moment of the quadratic form is given by

$$\mu'_r = \sum_{j=0}^{r-1} c_j^{r-1} \kappa_r \mu'_j \quad (4)$$

where $\mu'_j = E(X^j)$ and $c_j^{r-1} = (r-1)!/j!(r-j-1)!$ are the j^{th} non-central moment and the combination j of $r-1$, respectively.

An Application

The linear discriminant function is used when the d -dimensional observation $\mathbf{y} = (y_1, \dots, y_d)'$ in two independent populations Π_1 and Π_2 has multivariate normal densities $N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. We use the notation $P(i|j)$ to denote the probability of misclassification of an observation $\mathbf{y}$ into group i when, in fact, it belongs to the group j , where $i, j = 1, 2$. For simplicity, suppose that the prior probabilities are taken to be equal, i.e. $p_1 = p_2 = 1/2$. By the Bayes optimal classification rule, the linear discriminant function is defined as

$$W = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right) \quad (5)$$

Future observations $\mathbf{y}$ are assigned into the group Π_1 when $W \geq 0$. In the case of $W < 0$, this observation is assigned into the group Π_2 . The total probability of misclassification (TMP) is given by

$$\begin{aligned} \text{TMP} &= \frac{1}{2} (P(2|1) + P(1|2)) \\ &= \frac{1}{2} (P(W < 0) + P(W \geq 0)) \end{aligned}$$

Because the W distribution is univariate normal, the TPM is obtained as (Johnson & Wichern, 2007, p. 297)

$$\text{TPM}_W = \frac{1}{2} \left(\Phi \left(-\frac{\Delta}{2} \right) + \Phi \left(-\frac{\Delta}{2} \right) \right) = \Phi \left(-\frac{\Delta}{2} \right) \quad (6)$$

where $\Phi(\cdot)$ is the cumulative probability function of the standard normal distribution and Δ is the Mahalanobis distance between the two mean vectors μ_1 and μ_2 , i.e.

$$\Delta^2(\mu_1, \mu_2) = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \quad (7)$$

Clearly, the Bayes classification rule is equivalent to classification between the two groups by the minimum Mahalanobis distance. In this case, the discrimination variable W can be obtained if we equate the squared Mahalanobis distance between group means and observation y , i.e. $\Delta^2(y, \mu_1) = \Delta^2(y, \mu_2)$.

The squared Euclidean distances between means of the two groups is defined by

$$\delta^2(\mu_1, \mu_2) = (\mu_1 - \mu_2)' (\mu_1 - \mu_2) \quad (8)$$

The minimum Euclidean distance of observation y from the group means, i.e. $\delta^2(y, \mu_1) = \delta^2(y, \mu_2)$, can be used for discrimination. In this case the discriminant variable is given by

$$W' = (\mu_1 - \mu_2)' \left(y - \frac{1}{2}(\mu_1 + \mu_2) \right) \quad (9)$$

The TPM of W' is given by

$$\text{TPM}_{W'} = \frac{1}{2} \left(\Phi \left(\frac{-\delta^2}{2\Delta_1} \right) + \Phi \left(\frac{-\delta^2}{2\Delta_1} \right) \right) = \Phi \left(\frac{-\delta^2}{2\Delta_1} \right) \quad (10)$$

where the notation Δ_1^2 is defined in equation (7) by replacing Σ^{-1} with Σ (for more details see Appendix A).

In practice, the parameters μ_1 , μ_2 , and Σ are unknown. Estimate these parameters by means of independent random “training samples”. Suppose we have N_1 observations $y_1^{(1)}, \dots, y_{N_1}^{(1)}$ drawn from Π_1 and N_2 observations $y_1^{(2)}, \dots, y_{N_2}^{(2)}$ drawn

from Π_2 , where $N_1, N_2 > d$. We estimate μ_1, μ_2 by the unbiased sample means $\bar{y}_1$ and $\bar{y}_2$, respectively, and estimate Σ by the pooled sample covariance matrix S_p . Then the discriminant functions (5) and (9) can be modified as $\hat{W}$, and $\hat{W}'$ yields a “plug-in” discriminant functions. The sample distribution of $\hat{W}$ has been studied by several authors (Atakan, 2009).

The stochastic representations for the exact distribution of $\hat{W}$ in terms of elements of two independent 2×2 central and non-central Wishart matrices has been studied by Bowker (1961). Similarly, we can obtain stochastic representations for the plug-in discriminant function $\hat{W}'$ in terms of a sum of weighted independent non-central chi-square random variables where the weights are paired with different sign (see Appendix B).

Density Estimation Based on Moments

Maximum Entropy Density Estimation

Let X be a continuous variable with probability density function $f(x)$. Then Shanon's entropy is defined by

$$H(q) = -\int_{-\infty}^{\infty} f(q) \log f(q) dq \quad (11)$$

where

$$\int_{-\infty}^{\infty} f(q) dq = 1$$

A maximum entropy density function can be obtained by maximizing the Shanon's entropy subject to known moment constraints, i.e.

$$\mu'_j = \int_{-\infty}^{\infty} q^j f(q) dq, \quad j = 0, 1, \dots, k \quad (12)$$

where k is the number of known moments. So, to get the maximum entropy density, we have a non-linear system of equations. The solution of this system is a maximum entropy density function that can be obtained by a variational principle used for $f(q)$ according to the Lagrange method (Singh, 2013). The maximum entropy density is an exponential function is given by

$$\hat{f}(x|k) = \exp(-\theta_0 - \theta_1 q - \dots - \theta_k q^k) \quad (13)$$

where $\theta_1, \theta_2, \dots, \theta_k$ are determined so that (13) is a proper density function and satisfies all the $k + 1$ moment constraints. With the use of Lagrange's method to solve the non-linear equation system, $k + 1$ Lagrange parameters are obtained by solving $k + 1$ non-linear equations which maximum entropy density function (13) substituting into equation (12), i.e.

$$g_j(\theta) = \int_{-\infty}^{\infty} q^j \hat{f}(q|k) dq = \mu'_j, \quad j = 0, 1, \dots, k \quad (14)$$

The solution of equation (14) can be obtained through numerical methods such as the Newton-Raphson algorithm and dual approach (Golan, Judge, & Miller, 1996). In this paper, the Newton-Raphson's algorithm is used for obtaining the Lagrange vector $\theta = (\theta_0, \dots, \theta_k)'$. Let θ^0 be the initial vector, which may have elements that are zero. The first step is to compute the first order Taylor expansion of $g_j(\theta)$ around the initial trial θ^0 ; we have

$$\theta^1 = \theta^0 + G^{-1}v \quad (15)$$

where the elements of vector v are given by

$$v_j = \mu'_j - g_j(\theta^0), \quad j = 0, 1, \dots, k$$

and the elements g_{ij} of the Hessian matrix G are

$$g_{ij} = g_{ji} = \int_{-\infty}^{\infty} x^{i+j} \hat{f}(q|k) dq, \quad i, j = 0, 1, \dots, k$$

Because the Hessian matrix is positive definite in each iteration of the algorithm, equation (15) has a unique solution for θ^1 . In each iteration, the current estimation replaces θ^0 and the non-linear equations will be solved again. The iterative algorithm stops when the Euclidean norm of the difference between the current and previous parameter values is less than a small value, for example $\|\theta^1 - \theta^0\| \leq 10^{-5}$.

The determination of the number of constraints is important to the selection of the optimal density. If the number of constraints are taken to be over-large or over-small, then the accuracy of the estimated density function will be lost (Zong,

2006). Theoretically, among calculated densities, the function with the lowest entropy is selected. It should be mentioned, for a variable that is known information such as its moments, lower entropy indicates less uncertainty to the probability distribution of the variable. However, an increase in the number of constraints leads to a decrease of entropy (Singh, 2013), i.e.

$$H_{k_2}(q) \leq H_{k_1}(q), \quad k_2 > k_1$$

The Kullback-Leibler and information discrimination are other criteria that can be used to select a maximum entropy density function (Soofi, Ebrahimi, & Habibullah, 1995). The stochastic representation of the quadratic form in equation (2) can be performed by a Monte Carlo simulation to estimate percentiles of the distribution. Therefore, to determine the constraints k , the density is chosen in such a way that its percentiles have the minimum distance to the empirical percentiles obtained by Monte Carlo simulations. In other words, for $\alpha = 0.01, 0.02, \dots, 0.99$, the percentiles t_α of the empirical distribution function are calculated by Monte Carlo simulations. Percentiles of distribution for the known constraints k are calculated by

$$F_k(t_\alpha) = \int_{q_1}^{t_\alpha} \hat{f}(q|k) dq \quad (16)$$

where $\hat{f}(q|k)$ is estimated by the density function (13). In this study, the lowest integral bound is determined by the minimum observation in the Monte Carlo simulation. Therefore, the number of constraints k is chosen so that the following Euclidean distance has the minimum value:

$$\hat{k} = \min \|F_k(t) - G(t)\|, \quad k = 1, 2, \dots \quad (17)$$

where $G(t)$ is the empirical distribution.

In summary, to approximate the distribution of quadratic forms by maximum entropy estimation, the following algorithm can be utilized to approximate the density function:

- Step 1. Using the right-hand side of equation (2), we can perform a Monte Carlo simulation with 10^6 iterations.
- Step 2. The theoretical moments are obtained by equation (3).
- Step 3. The maximum entropy density function is obtained by (13).

- Step 4. The number of constraints $\hat{k}$ will be estimated by equation (17).
 Step 5. Cumulative probability (16) of the point t is calculated using numerical integration.

Modified Pearson's Approximation

When the quadratic form is indefinite some weights can be negative and, by equation (4), the third cumulant κ_3 will have a negative value. The skewness coefficient, i.e. $\beta_1 = \kappa_3 / \kappa_2^{3/2}$, is negative. Therefore, it is unsuitable to use gamma and chi-square distributions to approximate the distribution of indefinite quadratic forms directly. In this case, the proposed methods by Houshmand (1993) are remarkable. However, such positive distributions may be modified to approximate the distribution of indefinite quadratic forms. For example, in modified Pearson's tree moments, we write $Q; b\chi_\nu^2 + a$, where the symbol $(;)$ means "is approximately distributed as". Equate the first three cumulants on both sides and determine a , b , and ν , respectively. If we define

$$c_r = \sum_{j=1}^d \lambda_j^r (1 + r\delta_j^2)$$

then $a = -(c_2^2 / c_3) + c_1$, $b = c_3 / c_2$, and the degrees of freedom is estimated as $\nu = c_2^3 / c_3^2 = 8 / \beta_1^2$. Since the degrees of freedom ν can be fractional, for computational purposes we use a gamma distribution with shape and scale parameters $\nu / 2$ and $|b|$. As a result, the tail probability is given by (Houshmand, 1993)

$$P(Q \leq t) \approx \begin{cases} P(Y \leq t - a) & \text{if } b > 0 \\ 1 - P(Y < a - t) & \text{if } b < 0 \end{cases} \quad (18)$$

where $Y = |b| \chi_\nu^2(\delta_i^2)$.

Numerical Examples

Two artificial examples are provided to illustrate the usefulness of the maximum entropy density estimation method. For the well-known Egyptian skull data (Manly, 1994), the estimation methods of misclassification probabilities were compared. It

is further noted that Mathematica software was used to write program codes for numerical calculation. The package is available from the authors upon request.

Model Selection

In this example, a quadratic form is demonstrated as a weighted sum of non-central chi-square variables. The criteria (11) and (17) are compared to obtain the most appropriate probability density of the maximum entropy. Suppose in equation (1) the variable $\mathbf{X}$ has the multivariate normal distribution with mean $\boldsymbol{\mu} = (7, 12, -3)$, covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 6 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

and matrix of quadratic form given by

$$\mathbf{A} = \begin{bmatrix} 2 & -5 & -1 \\ -5 & 1 & 1 \\ -1 & 1 & 6 \end{bmatrix}$$

The weights of the non-central chi-square variables are $\lambda_1 = 28.3786$, $\lambda_2 = 9.4582$, and $\lambda_3 = -5.8833$. The non-centrality parameters of the non-central chi-square variables are $\gamma_1^2 = 0.161495$, $\gamma_2^2 = 28.5401$, and $\gamma_3^2 = 144.965$. In Table 1, different models are given to obtain of the maximum entropy density function. The Shanon's entropy decreased with an increasing number of constraints. Minimal change is between 4 to 9 constraints. From the table, $k = 12$ minimizes H, meaning that the best density is given by 12 constraints. The distance D is minimized by 9 constraints ($D = 0.00265$). Clearly, by obtained information about H and D , the constraints $k = 9$ provides the suitable maximum entropy density function.

An Exact Density of Quadratic Form

Imhof (1961) obtained the exact distribution of the indefinite quadratic form where the weights of the central chi-square variables are paired and, in each pair, the weights are equal. In this particular case, suppose that the eigenvalues are $\lambda_1 = \lambda_2 = 16$, $\lambda_3 = \lambda_4 = 7$, $\lambda_5 = \lambda_6 = 3$, $\lambda_7 = \lambda_8 = 1$, $\lambda_9 = \lambda_{10} = -2$, $\lambda_{11} = \lambda_{12} = -6$,

$\lambda_{13} = \lambda_{14} = -14$, $\lambda_{15} = \lambda_{16} = -32$, and $\lambda_{17} = \lambda_{18} = -70$. The stochastic representation of the indefinite quadratic form is given by

$$Q_1 = 16\chi_2^2 + 7\chi_2^2 + 3\chi_2^2 + \chi_2^2 - 2\chi_2^2 - 6\chi_2^2 - 14\chi_2^2 - 32\chi_2^2 - 70\chi_2^2$$

Table 1. Model selection values for number of constraints k ; Shanon's entropy H ; and proposed method D

k	H	D	k	H	D
1	7.49458	1.91482	7	6.61249	0.00283
2	6.61282	0.01816	8	6.61248	0.00294
3	6.61276	0.01754	9	6.61248	0.00265
4	6.61249	0.00424	10	6.61245	0.0036
5	6.61249	0.00281	11	6.61243	0.00458
6	6.61249	0.00277	12	6.61198	0.01108

Figure 1. Exact density (solid line) and approximated methods (dotted): maximum entropy (ME) and Pearson's method (P)

DISTRIBUTION OF QUADRATIC FORMS BY DENSITY ESTIMATION

Considering the absolute value of weights there is a definite quadratic form Q_2 . The best maximum entropy density function of Q_1 is obtained by the constraints $k = 10$, where the estimated criteria are $H = 6.35872$ and $D = 0.029954$. The discrepancy criterion D are estimated to be 0.109882 and 0.110786 for the modified Pearson's methods. The definite quadratic form Q_2 has minimum Shannon's entropy at $k = 11$, while the minimum value of the discrepancy criterion is 0.020173 for $k = 10$. The discrepancies are estimated for the Pearson's methods to be 0.006198 and 0.088602. Figure 1 displays the exact density functions of the quadratic forms Q_1 and Q_2 alongside the three methods of approximation. For the indefinite quadratic form Q_1 , the density approximated the maximum entropy method is quite accurate given the skewness of the exact density. The maximum entropy density is more accurate than the Pearson's method. In the definite form, the Pearson's method for upper tail is a suitable approximation.

Egyptian Skulls Data

The Egyptian skull data consists of four measurements: maximal breadth (X_1), basibregmatic height (X_2), basalveolar length (X_3), and nasal height (X_4), which were measured on skulls of ancient Egyptian males from five different time periods (4000 BC, 3300 BC, 1850 BC, 200 BC, 150 AD). Also, each time period consists of 30 observations. This data can be found in many books on applied multivariate statistical methods, e.g. Manly (1994). In this study, the observations are categorized into two groups: BC and AD. To determine whether or not the underlying assumption of normality of the groups is satisfied, Mardia's multivariate skewness statistic (Mardia, 1974) applied to test the hypothesis that each training sample is drawn from a multivariate normal population. By this statistic, the two groups BC and AD are normal at the 5% level. The test failed to reject the null hypothesis in either case. The smallest p -value for the Mardia's skewness statistic for either Π_1 or Π_2 was $p = 0.181$. The homogeneity of covariance matrices is evaluated with Box's M test, which was not significant at the 5% level ($M = 11.015$, $p = 0.403$). Therefore, the assumption of homogeneity of covariance matrices was satisfied. Hotelling's T^2 statistic is used to test for differences between means of two groups. The equality of means is rejected at the 1% level ($T^2 = 23.4985$, $p = 0.00025$). Therefore, the three assumptions of the linear discriminant analysis are satisfied.

Summarized in Table 2 are the misclassification probabilities of the Egyptian skull data. The apparent error rate (APER) is defined as the fraction of observations in the training sample that are classified by the plug-in discriminant function. APER

Table 2. Estimated misclassification probabilities for Egyptian skull data

Method	P(AD BC)	P(BC AD)	TPM
APER of $\hat{W}$	0.291667	0.400000	0.345833
APER of $\hat{W}'$	0.275000	0.366667	0.320833
Normal			0.312231
Monte-Carlo simulation of $\hat{W}'$	0.317874	0.317848	0.317861
ME	0.317860	0.318204	0.318032

does not depend on the distribution of the groups and that can be calculated for any classification procedure. Unfortunately, APER tends to underestimate the actual error rate and the problem does not disappear unless the sample sizes N_1 and N_2 are very large (Jhonson & Wichern, 2007, p. 598 ff.). From the Table 2, we see that the estimates of $P(AD | BC)$, $P(BC | AD)$, and TPM obtained by the APER of the discriminant function $\hat{W}$ are 0.275, 0.366667 and 0.320833, respectively. These estimates are less than the corresponding estimates obtained by $\hat{W}$.

The sample variance and squared Euclidean distance were estimated to be $\hat{\Delta}_1 = 0.989498$ and $\hat{\delta}^2 = 29.1546$, respectively. Estimate the misclassification probability of the plug-in discriminant function $\hat{W}'$ by normal approximation, $TPM = 0.312231$. The stochastic representation of $\hat{W}'$, if $y \in \Pi_1$, is given by

$$\begin{aligned} \hat{W}'_1 = & 3.224\chi_1^2(12.079) + 2.426\chi_1^2(1.4498) + 1.952\chi_1^2(0.4005) + 0.954\chi_1^2(0.327) \\ & - 0.954\chi_1^2(0.218) - 1.952\chi_1^2(0.266) - 2.426\chi_1^2(0.965) - 3.224\chi_1^2(8.036) \end{aligned}$$

and, if $y \in \Pi_2$, then

$$\begin{aligned} \hat{W}'_1 = & 3.224\chi_1^2(8.036) + 2.426\chi_1^2(1.4498) + 1.952\chi_1^2(0.266) + 0.954\chi_1^2(0.218) \\ & - 0.954\chi_1^2(0.327) - 1.952\chi_1^2(0.4005) - 2.426\chi_1^2(0.965) - 3.224\chi_1^2(12.079) \end{aligned}$$

since $\hat{W}'_1 = -\hat{W}'_2$, we have $P(AD | BC) = P(BC | AD)$. Estimate the TPM by performing 10^6 iterations of a Monte Carlo simulation of the stochastic representation $\hat{W}'_1$. From Table 2, note the estimate of TPM through Monte Carlo simulations is 0.317861. The maximum entropy method approximated the best density for $\hat{W}'_1$ and $\hat{W}'_2$ with $k = 9$. As can be seen from the table, the estimation of

misclassification probabilities obtained by maximum entropy method and Monte Carlo simulation are close to each other.

Conclusion

Determining the exact distribution of the indefinite quadratic form in normal variables is complicated, and its distribution remains an area of investigation. Fortunately, higher order moments of the quadratic form can be computed, and then the approximate distribution based on the moments approach. The maximum entropy density approximation method can be an alternative method to approximate the distribution. Despite its versatility, the maximum entropy density has not been widely used in empirical studies. One possible reason is that there is generally no analytical solution for the maximum entropy density problem and numerical estimation, such as the Newton-Raphson's algorithm, must be used. However, new computer systems are provided and computational speed has increased significantly. We proposed a criterion for selecting the number of constraints by the discrepancy between the approximated maximum entropy density and the empirical distribution of the stochastic representation of the quadratic form. This criterion can be used to select the maximum entropy density when it has minimum value. The results of the examples reveal that the approximated density of indefinite quadratic forms via maximum entropy are suitable.

References

- Atakan, C. (2009). Bootstrap percentile confidence intervals for actual error rate in linear discriminant analysis. *Hacettepe Journal of Mathematics and Statistics*, 38(3), 357-372. Retrieved from <http://www.hjms.hacettepe.edu.tr/uploads/6a0741d6-290b-4f18-b16d-a7b0351fba9b.pdf>
- Bowker, A. H. (1961). A representation of Hotelling's T^2 and Anderson's classification statistic W in terms of simple statistics. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 285-292). Stanford, CA: Stanford University Press.
- Golan, A., Judge, G. G., & Miller, D. (1996). *Maximum entropy econometrics: Robust estimation with limited data*. New York, NY: John Wiley.

- Houshmand, A. A. (1993). Misclassification probabilities for quadratic discriminant function. *Communication in Statistics – Simulation and Computation*, 22(1), 81-98. doi: [10.1080/03610919308813082](https://doi.org/10.1080/03610919308813082)
- Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4), 419-426. doi: [10.2307/2332763](https://doi.org/10.2307/2332763)
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review*, 106(4), 620-630. doi: [10.1103/physrev.106.620](https://doi.org/10.1103/physrev.106.620)
- Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis. Upper Saddle River, NJ: Pearson Prentice Hall.
- Liu, H., Tang, Y., & Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4), 853-856. doi: [10.1016/j.csda.2008.11.025](https://doi.org/10.1016/j.csda.2008.11.025)
- Manly, B. F. J. (1994). *Multivariate statistical methods: A primer*. New York, NY: Chapman & Hall.
- Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, 36(2), 115-128. Available from <http://www.jstor.org/stable/25051892>
- Mardia, K. V., Kent, T. J., & Bibby, J. (1979). *Multivariate analysis*. London, UK: Academic Press.
- Mathai, A. M., & Provost, S. B. (1992). *Quadratic forms in random variables, Theory and applications*. New York, NY: Marcel Dekker.
- Mohsenipour, A. A., & Provost, S. B. (2011). On approximating the distribution of indefinite quadratic expressions in singular normal vectors. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 15(1), 61-86. Retrieved from <http://math.ut.ee/acta/15-1/MohsenipourProvost1.pdf>
- Patnaik, P. B. (1949). The non-central χ^2 - and F -distributions and their application. *Biometrika*, 36(1/2), 202-232. doi: [10.2307/2332542](https://doi.org/10.2307/2332542)
- Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2 . *Biometrika*, 46(3/4), 364. doi: [10.2307/2333533](https://doi.org/10.2307/2333533)
- Singh, V. P. (2013). *Entropy theory and its application in environmental and water engineering*. New York, NY: John Wiley.
- Smith, P. J. (1995). A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician*, 49(2), 217-219. doi: [10.2307/2684642](https://doi.org/10.2307/2684642)

DISTRIBUTION OF QUADRATIC FORMS BY DENSITY ESTIMATION

Soofi, E., Ebrahimi, N., & Habibullah, M. (1995). Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association*, 90(43), 657-668. doi: [10.2307/2291079](https://doi.org/10.2307/2291079)

Zong, Z. (2006). *Information-theoretic methods for estimating complicated probability distributions*. Boston, MA: Elsevier.

Appendix A: TPM of W

Let $\mathbf{y}$ be from the population Π_1 . The variable $x = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{y}$ has univariate normal distribution with mean $\mu_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\mu}_1$ and variance $\sigma^2 = \Delta_1^2$, where Δ_1^2 is the squared Mahalanobis distance in (7) by substituting $\boldsymbol{\Sigma}$ for $\boldsymbol{\Sigma}^{-1}$. In this case,

$$\begin{aligned}
 P(2|1) &= P(W' < 0) = P\left(x < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) \\
 &= P\left(\frac{x - \mu_1}{\sigma} < \frac{\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\mu}_1}{\Delta_1}\right) \\
 &= P\left(z < -\frac{\frac{1}{2}\delta^2}{\Delta_1}\right) \\
 &= \Phi\left(-\frac{\delta^2}{2\Delta_1}\right)
 \end{aligned} \tag{A1}$$

where δ^2 is the squared Euclidean distance in (8). Similarly,

$$\begin{aligned}
 P(1|2) &= P(W' \geq 0) \\
 &= P\left(z \geq \frac{\frac{1}{2}\delta^2}{\Delta_1}\right) \\
 &= 1 - \Phi\left(\frac{\delta^2}{2\Delta_1}\right) \\
 &= \Phi\left(-\frac{\delta^2}{2\Delta_1}\right)
 \end{aligned} \tag{A2}$$

substituting equations (A1) and (A2) in (5),

$$\text{TPM}_{W'} = \Phi\left(-\frac{\delta^2}{2\Delta_1}\right)$$

In the linear discriminant analysis $\mu_1 \neq \mu_2$, then $\delta^2 > 0$. The TPM of W' is decreasing in Δ_1 if δ^2 is fixed. Therefore it follows that $\text{TPM}_{W'}$ is tending to zero if $\Delta_1 \rightarrow 0$.

Appendix B: Stochastic Representations of $\hat{W}'$

The discriminant function $\hat{W}'$ can be written as a quadratic form $\hat{W}' = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{x} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2, \mathbf{y} - \frac{1}{2}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2))$ and

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \frac{1}{2}\mathbf{I}_d \\ \frac{1}{2}\mathbf{I}_d & \mathbf{0} \end{bmatrix}$$

If the vector $\mathbf{y}$ is from the population Π_j , i.e. $\mathbf{y} \in \Pi_j, j = 1, 2$, then the vector $\mathbf{x}$ is distributed as $N_{2d}(\boldsymbol{\eta}_j, \boldsymbol{\Omega})$ with means $\boldsymbol{\eta}_1 = (\mu_1 - \mu_2, \frac{1}{2}(\mu_1 - \mu_2))$ and $\boldsymbol{\eta}_2 = (\mu_1 - \mu_2, -\frac{1}{2}(\mu_1 - \mu_2))$ and covariance matrix

$$\boldsymbol{\Omega} = \begin{bmatrix} \left(\frac{N_1 + N_2}{N_1 N_2}\right)\boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \left(\frac{N_1 + N_2 + 4N_1 N_2}{4N_1 N_2}\right)\boldsymbol{\Sigma} \end{bmatrix}$$

Therefore, apply the results of the quadratic form stochastic representation to estimate the TPM of W' . The eigenvalues of $\boldsymbol{\Omega}^{\frac{1}{2}}\mathbf{A}\boldsymbol{\Omega}^{\frac{1}{2}}$ are equal with the eigenvalues of matrix

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{0} & \left(\frac{N_1 + N_2}{2N_1N_2} \right) \mathbf{\Sigma} \\ \left(\frac{N_1 + N_2 + 4N_1N_2}{8N_1N_2} \right) \mathbf{\Sigma} & \mathbf{0} \end{bmatrix} \quad (\text{B1})$$

If $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$ are the eigenvalues of $\mathbf{\Sigma}$, then the characteristic equation of (B1), by applying equation (A.2.3.j) in Mardia, Kent, and Bibby (1979), has the eigenvalues

$$\pm \frac{\lambda_j \sqrt{(N_1 + N_2)(N_1 + N_2 + 4N_1N_2)}}{4N_1N_2}$$

where $j = 1, 2, \dots, d$.

Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques

Paola Zaninotto
University College London
London, UK

Amanda Sacker
University College London
London, UK

Using a simulation study, the performance of complete case analysis, full information maximum likelihood, multivariate normal imputation, multiple imputation by chained equations and two-fold fully conditional specification to handle missing data were compared in longitudinal surveys with continuous and binary outcomes, missing covariates, and an interaction term.

Keywords: Chained equations, longitudinal data, maximum likelihood, missing data, random intercepts, two-fold fully conditional specification

Introduction

Full information maximum likelihood (FIML) estimation and multiple imputation (MI) are considered “state of the art” missing data techniques (Schafer & Graham, 2002, p. 147) and are highly recommended (Baraldi & Enders, 2010). They are superior to ad-hoc procedures, because they require less strict assumptions and yield unbiased estimates with missing completely at random (MCAR) and missing at random (MAR) missing data mechanisms. Recently, the use of FIML, multivariate normal imputation (MVNI), and imputation by chained equations (MICE) (also known as sequential regressions, regression switching, and fully conditional specification, FCS) has become increasingly popular. A number of comparisons using linear models, generalized models, and structural equation models were published (Collins, Schafer, & Kam, 2001; Newman, 2003; Acock, 2005; Ibrahim, Chen, Lipsitz, & Herring, 2005; van Buuren, 2007; Buhi, Goodson, & Neilands, 2008; Marshall, Altman, Royston, & Holder, 2010; Peyre, Lepage, & Coste, 2011; Ferro, 2014).

Dr Paola Zaninotto is a Senior Lecturer in Medical Statistics in the Department of Epidemiology and Public Health. Email them at: p.zaninotto@ucl.ac.uk.

The FIML approach has been widely used in structural equation modelling analysis; MICE and MVNI can be easily extended to repeated measures, and the two-fold fully conditional specification (two-fold FCS) was proposed as a special case of the fully conditional specification for repeated measures (Nevalainen, Kenward, & Virtanen, 2009). In the presence of many waves and many variables with missing data, MICE, as opposed to the two-fold FCS, can potentially cause numerical problems because of over-fitting and collinearity. The algorithm might not converge because of high correlations among variables with missing values. It might be required to reduce the categories of a discrete variable; furthermore, temporally-ordered data are not exploited. Two-fold FCS can handle imputation of missing data of both continuous and discrete variables in many waves. Each wave is imputed separately, using the past and/or future information for each of the variables to be imputed.

The comparative performance of these techniques simultaneously in a longitudinal setting remains unclear. This comparison is needed because an increasing number of longitudinal datasets are now becoming available, which are subject to missing data due to item non-response and attrition. The collected longitudinal data are therefore often incomplete with a non-monotone pattern (Minini & Chavance, 2004). MAR mechanisms and non-monotone patterns are common in most realistic settings (Horton & Kleinman, 2007). Good imputation techniques for dealing with missing values in longitudinal data structured in a long format (1 record per observation-wave) are still missing. FIML and the MI techniques (MVNI, MICE, and two-fold FCS) can create valid multiple imputations under MAR (Enders, 2001a; van Buuren, 2007; Schafer & Graham, 2002; Nevalainen et al., 2009) and for monotone and non-monotone patterns, when data are structured in a wide format (i.e. unstacked data, one observation row per each subject with the measures for each occasion in separate columns). Data can then be converted into a long format file for analysis.

The aim of this study is to investigate the usefulness of the ML and MI methods in the context of longitudinal analysis with a large number of missing data points. The realistic situation of missing data in more than one outcome is considered, as well as accommodation of an interaction term and data missing on more than one independent variable. Generally, analyses should exclude individuals with imputed data on the outcome (von Hippel, 2007). However, in settings where one outcome is also the exposure for the other outcome, excluding individuals with imputed data may result in selection bias (Young & Johnson, 2010). Therefore, it was suggested that including imputed data on the outcome is feasible when auxiliary variables (or as in our case outcome variables) are highly correlated

MISSING DATA IN LONGITUDINAL SURVEYS

with the outcome (White, Royston, & Wood, 2011). Hence, the purpose of this study is to explore the performance of the missing data techniques in recovering the true parameters estimated from two random intercepts models, one with a continuous outcome and the other with a binary outcome. A simulation study is set up, where FIML is compared to three MI techniques: multivariate normal imputation (MVNI), multiple imputations by chained equations (MICE) and the recently proposed two-fold fully conditional specification (two-fold FCS). The missingness pattern is non-monotone and the missing data mechanism is missing at random (MAR). We show results of complete case analysis because it is an approach commonly used.

Methodology

Empirical Data

The aim of the simulation study was to evaluate the four techniques for handling missing data (FIML, MVNI, MICE, and two-fold FCS) in a longitudinal survey of ageing to explore gender differences in trajectories of quality of life and depression among people with coronary heart disease (CHD), adjusting for covariates. For this purpose the first three waves (2002-03 to 2006-07) of the English Longitudinal Study of Ageing, described elsewhere (Steptoe, Breeze, Banks, & Nazroo, 2013) was used. ELSA was established in 2002-2003 (wave 1), is a biannual, ongoing, nationally representative, prospective cohort study of people aged 50 years and over living in private households in England. In the first wave 11,391 respondents were interviewed.

The analytical sample of this study consisted of 4,496 participants in wave 1 with CHD and healthy participants (without CHD and known longstanding conditions at baseline); 3,465 in wave 2 (2004-05); and 3,031 in wave 3 (2006-07). A total of 1,998 participants had complete data on all variables across the three waves. The data consisted of two completely observed exposures (CHD and sex) and their interaction (CHD*sex), and two incomplete dependent variables: Quality of life (QoL) score ranging from 0 to 57 (approximately normally distributed), with higher scores indicating better quality of life; and a binary variable measuring depression ("0" no; "1" yes). The correlation between quality of life and depression is -0.45. The following covariates were also included: age (completely observed, normally distributed); marital status (categorical, three categories); wealth (categorical, three categories); physical activity (categorical, three categories); smoking status (categorical, three categories); and alcohol consumption

(categorical, three categories). With the exception of sex, all variables are time-varying. The ELSA dataset is publicly available via registration with the UK data service (<https://www.ukdataservice.ac.uk/>).

Obtaining the True Parameters from the Empirical Data

Random intercepts models (Goldstein, 2003) were used to explore gender differences in quality of life and depression among older people with coronary heart disease (CHD), adjusting for covariates. The 1,998 individuals with complete data were treated as if they were the underlying population, and the true parameters were obtained from the two random intercepts models estimated using the complete data. The first model is for the continuous outcome and was estimated as follows:

$$y_{ij} = \beta_0 + \sum_{p=1}^3 \beta_p x_{pj} + \sum_{p=1}^8 \beta_p x_{pij} + u_j + e_{ij} \quad (1)$$

where y_{ij} is the quality-of-life for individual j at time i , p denotes the number of time-varying and time invariant variables. x_{pj} are the time-invariant factors ($p = 3$) gender, CHD (at wave 1), and the interaction term between CHD and gender; x_{pij} are the time-varying factors ($p = 8$) age (a linear and quadratic term), cohabitation status, depression, wealth, smoking status, alcohol consumption and physical activity. u_j denotes the random effect accounting for the individual level variation and e_{ij} is a combination of random variation and measurement error specific to each occasion i for an individual j . The random effects have variances equal to σ_u^2 and σ_e^2 at the individual and occasion levels. The true parameters, obtained from (1) on the complete data, are: coefficient for CHD $\beta_1 = -1.64$ (s.e. 0.47); coefficient for sex $\beta_2 = 0.84$ (s.e. 0.28), coefficient for the interaction term between CHD and sex $\beta_3 = -0.19$ (s.e. 0.71), $\sigma_u^2 = 24.2$ (s.e. 1.10), $\sigma_e^2 = 24.8$ (s.e. 0.57), adjusted for the covariates described above (true parameters for the covariates are presented in supplemental resources).

A logit model was estimated for the binary outcome as follows:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \sum_{p=1}^3 \beta_p x_{pj} + \sum_{p=1}^6 \beta_p x_{pij} + u_j \quad (2)$$

where

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right)$$

is the log-odds that $y_{ij} = 1$ (i.e. the probability of having depression) at occasion i for individual j . All the other variables are the same as in model (1) with the exception of depressive symptoms and the quadratic term for age, which are not included as covariates; therefore $p = 6$ in the second part of the equation, u_j denotes the random individual level effect with variance σ_u^2 . The true parameters obtained from (2) on the complete data are: coefficient for CHD $\beta_1 = 0.65$ (s.e. 0.25); coefficient for sex $\beta_2 = 0.74$ (s.e. 0.16), coefficient for the interaction term between CHD and sex $\beta_3 = 0.03$ (s.e. 0.36), $\sigma_u^2 = 4.02$ (s.e. 0.44), adjusted for the covariates described above, (true parameters for the covariates are presented in online resources).

Data Simulation

We used a non-parametric simulation study based on the observed real data consisting of 1,998 ELSA participants with complete data at all three waves. Starting with the data of complete observations, missingness was imposed on the two dependent variables (quality of life and depression) and on five covariates (marital status, wealth, physical activity, smoking status and alcohol consumption). Generation of missing data was performed using the data in wide format, in which there is one observation row for each subject with each measurement represented as a different variable, therefore each time-varying variable was present three times for each subject. Random uniform numbers were used to reproduce the same probabilities of missingness as those occurring in the real data: if the rank of the random number was equal to or less than the probability observed, a missing value was generated for the variable of interest. Variables from the first wave only required generation of missing data due to item non-response, variables from waves 2 and 3 required the generation of missing data due to attrition in addition to item non-response, which were performed separately according to the specific probabilities observed in the real data. In order to set a MAR scenario, within each wave, missingness probabilities were allowed to depend on observed variables in the dataset, independently of the missing values themselves. Deletion was repeated for each of the 1,000 data sets. The resulting amount of missingness ranged between 54% and 57% (resulting in the sample size of each replicate ranging from 859 to 919).

Missing Data Methods for the Simulated Data

Missing data on the 1,000 data sets were treated using the following methods: FIML, MVNI, MICE, and two-fold FCS. For the latter three MI techniques, five data sets for each replicate were imputed. Most literature (Rubin, 1987; van Buuren et al., 1999) suggests that good inferences can be made with the number of imputed data sets (m) as few as $m = 5$. Rubin (1987) showed the efficiency of an estimate based on m imputations, relative to one based on an infinite number, is $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information. In this setting, with approximately 56% missing information, $m = 5$ imputations is 90% efficient. Complete case analysis (average sample size across the simulated data 887) was also conducted. All analyses were carried out in Stata version 12

Full Information Maximum Likelihood (FIML)

The FIML technique does not impute, or fill in missing values, but directly estimates model parameters and standard errors using all available raw data (Enders, 2001a). The FIML estimator maximizes a likelihood function that is the sum of n casewise likelihood functions (where n is the number of respondents). As noted by Enders (2001a), assuming multivariate normality, the following function is maximized:

$$\log L_i = K_i - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_i) \quad (3)$$

such that $\mathbf{x}_i$ is the vector of complete data for case i , $\boldsymbol{\mu}_i$ is the vector of mean estimates for those variables that are observed for case i , and K_i is a constant that depends on the number of complete points for case i . The determinant and inverse of the covariance matrix Σ_i are based only on those variables that are observed for case i . This likelihood function measures the discrepancy between the observed data and current parameter estimates using all available data for a given case.

Summing over the n casewise functions yields the discrepancy function for the entire sample:

$$\log L(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^n \log L_i \quad (4)$$

MISSING DATA IN LONGITUDINAL SURVEYS

To illustrate how FIML works, consider a model with four observed variables: X_1 , X_2 , X_3 , and X_4 . The parameters of interest are

$$\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4] \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

The likelihood value for a subject with missing X_1 would be a function of the values on the observations for the other three variables, X_2 , X_3 , and X_4 , as well as the parameter estimates that involved these three variables. The relevant parameters are shown in the following:

$$\boldsymbol{\mu} = [., \mu_2, \mu_3, \mu_4] \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} . & . & . & . \\ . & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ . & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ . & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

By contrast, the likelihood value for a subject with missing X_2 and X_4 would be a function of the two other observations (X_1 and X_3) as well as the parameter estimates that involved X_1 and X_3 . The relevant parameters are shown in the following:

$$\boldsymbol{\mu} = [\mu_1, ., \mu_3, .] \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & . & \sigma_{13} & . \\ . & . & . & . \\ \sigma_{31} & . & \sigma_{33} & . \\ . & . & . & . \end{bmatrix}$$

Then the value of the overall discrepancy function is obtained by summing the likelihood functions for each individual.

Enders (2001a) explained that, at a more conceptual level, it is assumed that missing values on a variable X are conditionally dependent on other variables in the data (missing at random, MAR), and incorporating vectors of partially complete data in the individual level likelihood functions (3) implies probable values for the missing data during the parameter estimation process. Conceptually this is analogous to generating predicted scores for the missing data by regressing X on

other variables used in the analysis. Although the method assumes multivariate normality, it has been shown that it can also be extended to situations in which multivariate normality does not hold (Enders, 2001b).

Multivariate Normal Imputation (MVNI)

In a multivariate normal imputation model, missing data are imputed using simultaneous linear regression models in which each variable potentially depends on all other variables (Schafer & Olsen, 1998). MVNI assumes a joint multivariate normal distribution for all variables. When dealing with an arbitrary missing data pattern it is possible to use data augmentation (DA) to generate imputed values assuming an underlying multivariate normal distribution. DA is an iterative Markov Chain Monte Carlo (MCMC) procedure which is suitable for arbitrary missing data patterns. Suppose $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ are multivariate data decomposed into an observed part $\mathbf{Y}_{\text{obs}}$ and a missing part $\mathbf{Y}_{\text{mis}}$ from a normal distribution $P(\mathbf{Y} | \boldsymbol{\theta}) = N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. In many incomplete-data problems, the observed-data posterior $P(\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}})$ is intractable and cannot be simulated easily. When $\mathbf{Y}_{\text{obs}}$ is augmented by an assumed latent value of the $\mathbf{Y}_{\text{mis}}$, the resulting conditional posterior distribution $P(\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ becomes much easier to handle. DA for missing data consists of two steps, the Imputation step (I-step) and the Posterior step (P-step). In the I-step, given a current $\boldsymbol{\theta}^{(t)}$, a value for the missing data-point is first drawn from the conditional predictive distribution of $\mathbf{Y}_{\text{mis}}^{(t+1)} \sim P(\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(t)})$ given the observed data $\boldsymbol{\theta}^{(t)}$. In the P-step, a new value of $\boldsymbol{\theta}^{(t+1)}$ is drawn from its conditional posterior $P(\boldsymbol{\theta} | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(t+1)})$ given $\mathbf{Y}_{\text{mis}}^{(t+1)}$. These new estimates are used in the next I-step. Without prior information about the parameters a non-informative prior distribution is used. Iterating these two steps creates a Markov chain of length $t \left\{ (\mathbf{Y}_{\text{mis}}^{(t)}, \boldsymbol{\theta}^{(t)}) : t = 1, 2, \dots \right\}$ which converges to $P(\boldsymbol{\theta}, \mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}})$. The two steps are iterated long enough for the results to be reliable for a multiply imputed data set.

Here, the posterior mode, the highest observed-data posterior density with a non-informative prior, is computed from the Expectation Maximization (EM) algorithm, and is used as the starting value for the chain. The MI procedure takes 200 burn-in iterations before the first imputation and 100 iterations between imputations. In a Markov chain, the information in the current iteration has influence on the state of the next iteration. The burn-in iterations are iterations at the beginning of each chain that are used to eliminate the dependence on the starting

value of the chain and to achieve a stationary distribution. Although the regression and MCMC methods assume multivariate normality, inferences based on multiple imputation can be robust to departures from the multivariate normality assumption if the amount of missing information is not large. It makes sense to use a normal model to create multiple imputations even when the observed data are somewhat non-normal, as supported by simulation studies described in Schafer (1997) and the original references therein. The imputation model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which were completely observed).

Multiple Imputation by Chained Equations (MICE)

MICE does not start with the construction of a well-defined joint distribution for the variables to be imputed. Instead, it starts with a collection of univariate conditional distributions for variables with missing data in terms of all other variables. The main idea is that a sequence of univariate conditional models is constructed for each potentially missing variable (dependent and/or explanatory) with fully conditional specifications of prediction equations. The other variables (potentially missing or complete) are used as explanatory variables in each univariate imputation model. The standard procedure for creating multiple imputations of each potential missing variable $Y_1, Y_2, \dots, Y_p$ using complete predictors $\mathbf{X}$ as independent variables for the t^{th} iteration of the method is described as follows:

$$\begin{aligned} Y_1^{(t+1)} &\sim f_1(Y_1 | Y_2^{(t)}, \dots, Y_p^{(t)}, \mathbf{X}, \theta_1) \\ Y_2^{(t+1)} &\sim f_2(Y_2 | Y_1^{(t+1)}, Y_3^{(t)}, \dots, Y_p^{(t)}, \mathbf{X}, \theta_2) \\ &\vdots \\ Y_p^{(t+1)} &\sim f_p(Y_p | Y_1^{(t+1)}, Y_2^{(t+1)}, \dots, Y_{p-1}^{(t+1)}, \mathbf{X}, \theta_p) \end{aligned} \tag{5}$$

Imputed values are drawn from (5) for iterations $t=0, 1, \dots, T$ until convergence is reached at $t=T$. The $\theta_1, \dots, \theta_p$ are the model parameters with a uniform prior, the univariate imputation models $f_1, \dots, f_p$ are chosen to be appropriate for imputing $Y_1, \dots, Y_p$. This means that logistic regression can be used for binary variables, linear regression for continuous, ordinal logistic regression for categorical variables and so forth (van Buuren, 2007). Univariate posterior draws are made one variable at a time by cycling through all p models given current values

of the other variables (Molenberghs & Kenward, 2007). After sufficient cycles (10-20), the imputations are taken from one final cycle through the univariate model. The ordering of univariate imputations is not relevant.

MICE differs from the MVNI in that it does not start with the construction of a well-defined joint distribution for the variables to be imputed.

The imputation model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which are completely observed). For the variable quality of life we used predictive mean matching (Royston & White, 2011), thus constraining imputed values to the set of observed values. We number of cycles used in the imputation was 100.

Two-fold Fully Conditional Specification (FCS)

Two-fold FCS (Nevalainen et al., 2009) is a special case of MICE described above. MICE for Y_l ($l = 1, \dots, p$) is extended to q repeated waves, as follows:

$$f\left(\mathbf{Y}_u^{\text{mis}} \mid Y_{u-1}^*, Y_{u+1}^*, X_{u1}, \dots, Z_{up}\right), \quad u = 1, \dots, q$$

At time i , Y_u is imputed conditional on the same variable observed at time $u - 1$ and $u + 1$, and the other p variables Z at time u . Similarly, at time u , Z_u with missing data is imputed conditional on the same variable observed at time $u - 1$ and $u + 1$, the other p variables Z , and the Y . One iteration runs over the variables $l = 1, \dots, p$, called within-time iteration. The past and future observations (Y_{u-1} and Y_{u+1}) are not imputed at this stage, they serve only in the role of predictors in the imputation model. There is also a second imputation iteration over waves ($u = 1, \dots, q$), called among-time iteration. The method is fully described in Nevalainen et al.

Two-fold FCS differs from MICE in that each time point is imputed separately. The method uses information recorded before and after the time point to impute missing data. The past and future values might have missing data which are imputed by default, but are only used as predictors at that given time.

The application of two-fold FCS to the simulation data was performed as follows and described graphically in Figure 1:

- 1) Variables with missing data at wave 1 were imputed using as predictors all other variables in the same wave, plus the future observation (at wave 2) of the same variable. For example, to impute

MISSING DATA IN LONGITUDINAL SURVEYS

missing values for the continuous outcome (QoL) at wave 1, the following linear regression model is used:

$$\begin{aligned} \text{QoL}_{\text{wave 1}} = & \text{Sex}_{\text{wave 1}} + \text{CHD}_{\text{wave 1}} + (\text{Sex} * \text{CHD})_{\text{wave 1}} \\ & + \text{Age}_{\text{wave 1}} + \text{Age}_{\text{wave 1}}^2 + \text{Marital}_{\text{wave 1}} \\ & + \text{Wealth}_{\text{wave 1}} + \text{Smoking}_{\text{wave 1}} + \text{Physicalact}_{\text{wave 1}} \\ & + \text{Alcohol}_{\text{wave 1}} + \text{Depression}_{\text{wave 1}} + \text{QoL}_{\text{wave 2}} \end{aligned}$$

In this example the variable QoL at wave 2 is used as a predictor to impute missing data for QoL at wave 1. Because QoL at wave 2 has missing data as well, MICE by default imputes missing data for that variable too. However, the imputed values of QoL at wave 2 are then dropped from the imputed data set as, in the imputation of wave 1, it serves only the role of a predictor and it will be imputed in the next stage.

- 2) Variables with missing data at wave 2 are imputed using as predictors all other variables in the same wave, plus the past (wave 1 including values imputed in the previous step) and future (wave 3) observations of the same variable. The imputed values of these variables from wave 3 are then dropped after the imputation as at this stage they serve only the role of predictors in the imputation model.
- 3) Variables with missing data at wave 3 are imputed using as predictors all other variables in the same wave, plus the past observations (wave 2 imputed in the previous step) of each variable to be imputed.

Figure 1 gives a graphical explanation of steps 1), 2), and 3).

Steps 1) to 3) form one among-times iteration. It must be decided how many among-times iterations are needed. Nevalainen et al. (2009) showed that increasing the number of iterations from one to five improved the performance of the estimators although the gain due to the increase was relatively small. We used three among-times iterations and compared the means of the imputed variables at each wave with the means of the complete case data. It was decided that three among-times iterations were enough because the estimates from the five imputed data sets were very close to those of the underlying population.

To summarize, steps 1) to 3) were repeated 3 times (three among-times iterations) to generate 1 imputed dataset, the procedure was then repeated four more times to obtain five imputed datasets each for the 1000 data sets. The imputation

model included the same variables as the substantive models including the interaction term between CHD and gender and the linear and quadratic effects of age (which are completely observed).

Figure 1. Two-fold fully conditional specification; * indicates the variable with missing data at the specific wave that is to be imputed

Data Analysis of Simulated Data

After the imputation stage had been completed on the 1,000 data sets for each MI technique (MVNI, MICE, and two-fold FCS), analyses based on the imputed data sets were carried out in Stata version 12. Linear random intercepts models were estimated for the continuous outcome (QoL) and random intercepts logit models were estimated for the binary outcome (depression), described earlier. The estimates from the analysis were stored. From the stored estimates, some summary measures (Burton, Altman, Royston, & Holder, 2006) were calculated to assess each missing data strategy as follows:

The (average) estimate of interest:

$$\bar{\hat{\beta}} = \frac{\sum_{k=1}^r \hat{\beta}_k}{n}$$

where r is the number of data sets (1,000), and $\hat{\beta}_k$ is the estimate of interest within each of the $k = 1, \dots, r$ data sets. When MI is performed, each $\hat{\beta}_k$ is the overall estimate obtained according to Rubin's formula (Rubin, 1987), which is just the average of the 5 combined estimates within each of the $k = 1, \dots, r$ data sets.

The (average) standard error of the estimate of interest:

$$SE(\hat{\beta}) = \frac{\sum_{k=1}^r SE(\hat{\beta}_k)}{n}$$

where $SE(\hat{\beta}_k)$ is the standard error of the estimate of interest within each of the $k = 1, \dots, r$ replications. When MI is performed, each $SE(\hat{\beta}_k)$ is the overall standard error of the estimate of interest obtained from the five combined estimates according to Rubin's formula (Rubin, 1987), within each of the $k = 1, \dots, r$ data sets.

Evaluation Criterion

In order to evaluate the performance of each procedure employed to deal with missing data, we used assessments of accuracy and precision. Accuracy indicates the degree of closeness of the estimated value to the true parameter; precision refers to the repeatability or reproducibility of the measurement.

For the assessment of accuracy the following were used:

Bias: $\left(\bar{\hat{\beta}} - \beta\right)$ which is the difference between the average estimate and the true parameter (Schafer & Graham, 2002; Sinharay, Stern, & Russell, 2001).

The Mean Square Error (MSE): $\left(\bar{\hat{\beta}} - \beta\right)^2 + \left(\text{SD}(\hat{\beta})\right)^2$ is the average squared difference between the estimate and its true parameter plus its variance; therefore it can be seen as a summary of both bias and variability. A value of the MSE close to zero indicates that the average estimator predicts the true parameter with good accuracy.

For the assessment of precision the following was used:

Standardized bias percent:

$$100 \times \frac{\left(\bar{\hat{\beta}} - \beta\right)}{\text{SE}(\hat{\beta})}$$

which is the bias as a percentage of the standard error. A standardized bias is considered to have a large impact on the precision if its absolute value exceeds 40 per cent (Collins et al., 2001).

Results

The results of the comparison of the missing data techniques for the analysis of the continuous outcome (quality of life) are shown in Table 1. The bias of the coefficients for CHD and sex are small and close to zero for the three MI techniques, which seem to give estimates of the parameter for sex closer to the true parameter compared to FIML. The largest bias in the coefficient for the CHD by sex interaction term was obtained under FIML and least bias was obtained under MVNI and MICE, although the values of the MSE were similar for all techniques. The MSE is the sum of the squared bias and the variance, therefore it can be evinced that the variability across replicates is lower for each of the three parameters estimated under FIML than under the three MI methods.

MISSING DATA IN LONGITUDINAL SURVEYS

The three MI techniques produced a more precise estimate of the coefficient for the interaction term than FIML, as shown by the smaller values of the Standardized bias percent. This is probably due to the fact that imputation using MICE, two-fold FCS, and MVNI models included the interaction term and in that sense they reflected the substantive model. Estimates of the between (i.e. individual) variance obtained from MVNI and MICE were closer to the true parameter compared to the estimates obtained from the other methods. The estimates of the within (i.e. wave specific) variance were close to the true parameter for all methods, except for the estimates produced by the two-fold FCS. Estimates obtained from complete case analysis (average sample size across simulated data 887) showed good precision but did not achieve good levels of accuracy. Results comparing the true parameters with each method for missing data for the time-varying variables are presented in supplemental Tables S1 to S5. All methods performed equally well in producing estimates close to the true parameters for most of the time-varying covariates. MVNI and Two-fold FCS produced estimates slightly less precise the true parameters of cohabiting status and wealth. MICE outperformed the other methods in producing estimates that were close to the true parameter of depression.

Reported in Table 2 are the results of the comparison of the missing data techniques for the analysis of the binary outcome (depression). The large values of the biases and of the standardized bias percent suggest that FIML did not produce estimates that were close to the true parameters for sex, CHD, the interaction term between sex, and CHD and the between variance. Whereas the estimates produces by MICE were very close to the true coefficients for sex and CHD, and overall showed good levels of precision and accuracy as demonstrated by the MSE and standardized bias percent. Estimates of the interaction term between sex and CHD obtained from MICE and MVNI were slightly less close to the true parameter compared to the estimated obtained from two-fold FCS. Estimates of the coefficients for sex and CHD obtained from MVNI were not as close to the true parameters as those obtained from MICE and two-fold FCS. The estimate of the between variance obtained from MICE showed good accuracy and precision. Estimates obtained under complete case analysis (average sample size across simulated data 887) showed good accuracy as evinced from the small bias and good precision according to the values of the standardized bias percent. Results comparing the true estimates for the time-varying variables and each method for missing data are presented in supplemental Tables S6 to S10. Estimates obtained from all methods were close to the true parameters and showed good accuracy and precision.

Table 1. Summary of validity criteria for true parameters compared to the methods for missing data applied to the model of the continuous outcome quality of life

		β				
True parameters	CHD	-1.64				
	Sex	0.84				
	CHD*Sex	-0.19				
	Between variance	24.21				
	Within variance	24.79				
		$\hat{\beta}$	$SE(\hat{\beta})$	Bias	Stand. Bias %	MSE
Complete Case	CHD	-1.60	0.70	0.70	6.7	0.321
	Sex	0.87	0.42	0.42	6.4	0.104
	CHD*Sex	-0.26	1.06	1.06	-6.9	0.600
	Between variance	24.80	1.03	1.03	0.8	1.053
	Within variance	24.01	1.07	1.07	-18.0	1.724
FIML	CHD	-1.54	0.48	0.10	20.8	0.041
	Sex	1.14	0.29	0.30	104.3	0.015
	CHD*Sex	-0.54	0.72	-0.35	-48.5	0.093
	Between variance	23.09	1.22	-1.11	91.2	1.861
	Within variance	25.05	0.72	0.25	-34.9	0.538
MVNI	CHD	-1.59	0.51	0.05	10.1	0.046
	Sex	0.79	0.31	-0.05	-14.8	0.019
	CHD*Sex	-0.24	0.77	-0.05	-6.7	0.104
	Between variance	23.59	1.06	-0.61	-57.9	1.492
	Within variance	24.85	1.03	0.05	5.1	1.169
MICE	CHD	-1.64	0.52	0.00	0.7	0.049
	Sex	0.86	0.31	0.02	6.1	0.019
	CHD*Sex	-0.22	0.79	-0.03	-4.4	0.115
	Between variance	24.33	1.06	0.13	12.1	0.818
	Within variance	25.01	1.03	0.21	20.7	0.556
Two-fold FCS	CHD	-1.63	0.51	0.01	1.7	0.066
	Sex	0.87	0.31	0.03	10.7	0.023
	CHD*Sex	-0.24	0.77	-0.06	-7.2	0.123
	Between variance	22.79	1.06	-1.41	-133.4	2.787
	Within variance	25.85	1.03	1.05	101.9	1.628

Note: Estimates adjusted for age, age squared, marital status, wealth, physical activity, smoking, alcohol consumption, depression, for which missing values have been imputed. CHD=Coronary heart disease. FIML=Full information maximum likelihood. MVNI= Multivariate normal imputation. MICE=Multiple imputation by chained equation. FCS=Fully conditional specification. SE=Standard error. MSE=Mean square error.

MISSING DATA IN LONGITUDINAL SURVEYS

Table 2. Summary of validity criteria for true parameters compared to the methods for missing data applied to the model of the binary outcome depression

		β				
True parameters	CHD	0.65				
	Sex	0.74				
	CHD*Sex	0.03				
	Between variance	4.02				
		$\hat{\beta}$	$SE(\hat{\beta})$	Bias	Stand. Bias %	MSE
Complete Case	CHD	0.64	0.38	-0.01	-3.0	0.079
	Sex	0.75	0.24	0.01	3.7	0.030
	CHD*Sex	0.03	0.55	0.00	0.6	0.160
	Between variance	3.98	1.18	-0.04	-3.8	0.227
FIML	CHD	0.44	0.24	-0.21	-90.1	0.061
	Sex	0.50	0.14	-0.24	-176.3	0.067
	CHD*Sex	0.20	0.34	0.18	51.7	0.048
	Between variance	2.94	0.16	-1.08	-675.0	1.526
MVNI	CHD	0.58	0.26	-0.07	-25.7	0.015
	Sex	0.65	0.16	-0.09	-53.6	0.012
	CHD*Sex	0.04	0.37	0.01	3.7	0.018
	Between variance	3.47	1.14	-0.55	-48.4	0.381
MICE	CHD	0.66	0.28	0.01	2.3	0.013
	Sex	0.74	0.17	0.00	-1.4	0.005
	CHD*Sex	0.01	0.39	-0.02	-6.2	0.018
	Between variance	3.86	1.14	-0.16	-14.1	0.099
Two-fold FCS	CHD	0.62	0.26	-0.03	-12.5	0.013
	Sex	0.70	0.16	-0.04	-22.4	0.007
	CHD*Sex	0.03	0.37	0.00	-0.4	0.024
	Between variance	3.40	1.14	-0.62	-54.7	0.465

Note: Estimates adjusted for age, marital status, wealth, physical activity, smoking, alcohol consumption, for which missing values have been imputed. CHD=Coronary heart disease. FIML=Full information maximum likelihood. MVNI= Multivariate normal imputation. MICE=Multiple imputation by chained equation. FCS=Fully conditional specification. SE=Standard error. MSE=Mean square error

Conclusion

This simulation study was based on a large, national longitudinal survey to assess the problem of handling a non-monotone pattern of missing data. The aim of this study was to obtain valid and efficient estimates of regression coefficients from random intercepts models fitted to longitudinal data (three waves). The data set for

this study had incomplete time-dependent outcomes (one continuous and one binary) and time-dependent and time-independent covariates (of different types), therefore it was necessary to accommodate missingness within each wave, as well as unit non-response at all follow-up waves. In order to investigate which technique could be suitable with this structure of the data, FIML was compared with three MI techniques: MVNI, MICE and the recently proposed two-fold FCS. The performance of each of the methods appeared to vary according to the type of outcome.

The continuous outcome variable was the variable with the largest proportion of missing data, and yet all four missing data techniques performed well, although the MI techniques showed better accuracy and precision than FIML. Complete case analysis did not achieve good levels of accuracy. Additionally, the three MI techniques produced estimates the interaction term that showed better precision and smaller bias values compared to the estimate obtained from FIML. This is an advantage of multiple imputation techniques: the interaction term can and should be accommodated in the imputation model thus reflecting the substantive model. MICE and MVNI produced estimates of the within (wave) variance close to the true estimates, although all methods except two-fold FCS produced good estimates of the between (individual) variance.

A different picture was given by the results involving the binary outcome (depression). Estimates obtained under complete case analysis achieved good levels of precision and accuracy, according to the values of the bias. Not surprisingly the methods that assume a joint normal distribution for the non-normally distributed outcome FIML and MVNI did not produce estimates that had the same level of bias and precision as those obtained with more flexible chained equations methods. Levels of accuracy and precision were less good for the estimates obtained under FIML. Also, estimates of the MVNI showed less precision compared to other MI methods. However, estimates obtained from MICE and two-fold FCS, compared to both FIML and MVNI, showed better accuracy and precision. The flexibility of MICE and two-fold FCS becomes obvious in the presence of a non-normal outcome for which an appropriate conditional distribution (logistic) was specified in the imputation stage. The estimates of the between variance of the binary outcome obtained from MICE were slightly closer to the true parameters and showed better precision than those obtained from two-fold FCS.

FIML is relatively easy to implement and it performed almost as well as MI techniques with the continuous outcome. However, the estimate of the interaction term was less precise and not as close to the true parameters as those obtained from the MI techniques.

MISSING DATA IN LONGITUDINAL SURVEYS

In general, good estimates in terms of accuracy and precision were obtained from MVNI for the continuous outcome. However, less precise and accurate estimates were obtained from MVNI in the case of the binary outcome. This is not surprising, given that the amount of missing data was relatively large (between 54% and 57%). It has been suggested that inferences based on multiple imputation can be robust to departures from the multivariate normality assumption if the amount of missing information is below 50% (Schafer, 1997).

The advantage of MICE and two-fold FCS is that each type of variable (continuous, binary, unordered, and ordered categorical) is modelled separately (Molenberghs & Kenward, 2007). For both outcomes, MICE showed slightly better precision and accuracy than two-fold FCS and the other two techniques.

The MICE procedure used all the longitudinal data to estimate missingness whereas two-fold FCS imputed each time point separately and used only information on the same variable recorded prior to and post a given time point. Therefore it was anticipated that estimating missingness using MICE would result in a slightly better performance.

In the simulation study consisting of three waves, estimates obtained from MICE were in general more accurate and precise; however, when the longitudinal data to be imputed have many time points (waves) per observation, using MICE may not be computationally feasible, especially for discrete variables. Researchers may opt for two-fold FCS which uses a doubly-iterative procedure: each wave is imputed separately using only the relevant past and future observations, therefore the intensiveness of the computation is reduced but the benefit of using past and future observation is retained.

Another approach to impute longitudinal data with many waves per observation is to structure the data in a long format and perform imputation in MICE ignoring the dependence of data across waves. Although this method is conceptually wrong, it is appealing to many researchers. We anticipate that data structured in long format and imputed with MICE would perform more poorly than two-fold FCS. The main reason is that ignoring the dependence between observations might result in inefficient estimates of the parameters of interest. Two-fold FCS on the other hand accounts for the correlation between individual's responses over time and therefore should (and in our simulation did) recover the parameters of interest well. Further studies are needed to test the performance of two-fold FCS, especially in the presence of many follow-up points, when the implementation of MICE is not feasible.

A major strength of this study is the use of a real data set to provide a suitable structure for simulating the 1,000 data sets, which simplifies the data generation

procedures and avoids arbitrary choices. Also, by replicating the patterns of missingness seen in the original incomplete data set, a realistic framework was provided for simulating the missing data (Marshall et al., 2010). Another major strength is that the recently proposed two-fold FCS was applied to longitudinal data from a national survey and compared with FIML and MVNI. Previously, only FIML with MVNI and FCS was compared with cross-sectional data. Also, the data time-invariant variables were completely observed. Nevertheless, the MI methods used are suitable for the imputation of missing data in time-invariant variables too. Time-invariant variables only appear once for each individual because the data are structured in a wide format. Therefore, strength of imputing longitudinal data in wide format is that imputed values of time invariant variables cannot end out time varying.

One of the possible limitations of this simulation study is that missing data were generated under MAR. The plausibility of the MAR assumption could have been affected by the fact that auxiliary variables were by design not included in the imputation model. In analysis not shown here, auxiliary variables found to be associated with the values of variables with missing data (such as housing tenure, number of people in the household, and Government Office Region) were used to impute missing data using the three MI techniques. The ability to recover the true parameters by MVNI, MICE, and two-fold FCS techniques did not depend upon the addition of these variables; rather they helped reinforce the MAR assumption. However, it was necessary to exclude auxiliary variables in order to make the four techniques comparable. It may be appropriate to strengthen the MAR assumptions and reduce the chance obtaining multiple imputations with similar bias and precision to a complete case analysis may decide to opt for one of the MI techniques presented here rather than FIML which does not allow the inclusion of auxiliary variables.

Auxiliary variables will also reduce bias compared to complete case analysis. Although it is generally argued that analyses should exclude individuals with imputed data on the outcome (von Hippel, 2007), this approach is unlikely to produce bias, because we used one imputation model for imputing our dependent variables, depression and quality of life; furthermore, the dependent variables were highly correlated and were auxiliary variables for each other.

Another possible limitation is the use of five imputed datasets. Although five imputed datasets have been suggested to be sufficient it is now believed that using a larger number of imputation increases the precision of the results (Sterne et al., 2009). Another possible limitation of our study is that Stata does not allow random effects to be introduced in the imputation stage, which could be the reason for the

poor performance in estimating the between variance. Longitudinal data can be thought of as clustered or two-level data (Goldstein, 2003). It was suggested if a data set to be imputed is multilevel, then the imputation model should be multilevel too (Carpenter & Goldstein, 2004). There is no definitive recommendation in the literature on the best way to impute clustered data, but one of the strategies that has been suggested in Stata and that we have implemented is to impute all clusters simultaneously. The multiple imputations methods used in this work were performed using data in wide format, i.e. unstacked data with one observation row per subject, thus implicitly allowing an unrestricted covariance structure. Alternatives include the use of MICE package in R software (van Buuren & Groothuis-Oudshoorn, 2011), which allows for random effects in the imputation model. More recently, REALCOM-IMPUTE software, which performs multilevel multiple imputation, has been improved in order to allow missing values in covariates, including models where there are interactions, ordered and unordered categorical variables, or other functions of covariates such as polynomials (Goldstein, Carpenter, & Browne, 2014). Future work should consider comparisons of performance with this method.

It is possible that the approach adopted for the simulation study might have excluded variability from repeated sampling of the data values. To evaluate whether this would compromise the relative performance of the approaches used for missing data we have run a sensitivity analysis using a simulated data, which randomly sampled (with replacement) 1998 rows from our complete cases data set to create a new complete-cases sample. We have then applied our missing-value mechanism as before and stored the results. This process was repeated 100 times and the estimates were stored and summarized (results available on request). As expected, the ranking of each missing data approach did not change.

Although MICE and two-fold FCS procedures are useful when the specification of a joint multivariate distribution of all the variables with missing values is difficult, from a theoretical standpoint these techniques can be problematic because the sequence of regression models might not be consistent with a true joint distribution (Shafer & Graham, 2002). This means that the iterative algorithm might never converge because the joint distribution to which they might converge does not exist. Despite the lack of a satisfactory theory, MICE seems to work quite well in many applications. A number of simulation studies provide evidence that MICE and two-fold FCS generally yield estimates that are unbiased and that possess appropriate coverage, at least in the variety of cases investigated (Brand, van Buuren, Groothuis-Oudshoorn, & Gelsema, 2003; Raghunathan, Lepkowski, van

Hoewyk, & Solenberger, 2001; van Buuren, 2007; Nevalainen et al., 2009; Welch, Bartlett, & Petersen, 2014).

An anonymous reviewer suggested an extension to this study which employs a simulation study with a longer follow-up, and in which the following methods for dealing with missing data are included in the comparisons: multilevel imputation, imputation using long format, longitudinal weighting combined with complete case analysis, and longitudinal weights combined with multiple imputation. A further suggestion was an extension to this study which compares several missing at random mechanisms plausible in longitudinal studies and in cross-sectional studies, since the causes leading to attrition are different from those leading to item non-response. Future studies on multiple imputation in the context of panel data might usefully consider incorporating these suggestions.

It was shown when dealing with non-monotone missing data in longitudinal studies where a continuous outcome is involved, FIML and MI techniques all perform well. MI techniques compared to FIML might be more suitable for accommodating interaction terms. It was also shown MICE and two-fold FCS produced estimates that were more accurate and precise than those obtained from FIML and MVNI techniques, especially when dealing with non-continuous variables and interaction terms. The results of this study showed that MICE in general showed slightly better precision and accuracy better than two-fold FCS. More studies are needed to test the performance of two-fold FCS especially in the presence of many waves.

References

- Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, 67(4), 1012-1028. doi: [10.1111/j.1741-3737.2005.00191.x](https://doi.org/10.1111/j.1741-3737.2005.00191.x)
- Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37. doi: [10.1016/j.jsp.2009.10.001](https://doi.org/10.1016/j.jsp.2009.10.001)
- Brand, J. P. L., van Buuren, S., Groothuis-Oudshoorn, K., & Gelsema, E. S. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1), 36-45. doi: [10.1111/1467-9574.00219](https://doi.org/10.1111/1467-9574.00219)
- Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior*, 32(1), 83-92. doi: [10.5993/ajhb.32.1.8](https://doi.org/10.5993/ajhb.32.1.8)

MISSING DATA IN LONGITUDINAL SURVEYS

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279-4292. doi: [10.1002/sim.2673](https://doi.org/10.1002/sim.2673)

Carpenter, J., & Goldstein, H. (2004). Multiple imputation using MLwiN. *Multilevel Modelling Newsletter*, 16(2), 9-18. Retrieved from <http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/new16-2.pdf>

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330-351. doi: [10.1037/1082-989x.6.4.330](https://doi.org/10.1037/1082-989x.6.4.330)

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, 61(5), 713-740. doi: [10.1177/0013164401615001](https://doi.org/10.1177/0013164401615001)

Enders, C. K. (2001b). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4), 352-370. doi: [10.1037/1082-989x.6.4.352](https://doi.org/10.1037/1082-989x.6.4.352)

Ferro, M. A. (2014). Missing data in longitudinal studies: Cross-sectional multiple imputation provides similar estimates to full-information maximum likelihood. *Annals of Epidemiology*, 24(1), 75-77. doi: [10.1016/j.annepidem.2013.10.007](https://doi.org/10.1016/j.annepidem.2013.10.007)

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, UK: Edward Arnold. doi: [10.1002/9780470973394](https://doi.org/10.1002/9780470973394)

Goldstein, H., Carpenter, J., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 177(2), 553-564. doi: [10.1111/rssa.12022](https://doi.org/10.1111/rssa.12022)

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79-90. doi: [10.1198/000313007x172556](https://doi.org/10.1198/000313007x172556)

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469), 332-346. doi: [10.1198/016214504000001844](https://doi.org/10.1198/016214504000001844)

- Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, 10(7). doi: [10.1186/1471-2288-10-7](https://doi.org/10.1186/1471-2288-10-7)
- Minini, P., & Chavance, M. (2004). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biostatistics*, 5(4), 531-544. doi: [10.1093/biostatistics/kxh006](https://doi.org/10.1093/biostatistics/kxh006)
- Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Chichester, UK: John Wiley & Sons. doi: [10.1002/9780470510445](https://doi.org/10.1002/9780470510445)
- Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine*, 28(29), 3657-3669. doi: [10.1002/sim.3731](https://doi.org/10.1002/sim.3731)
- Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6(3), 328-362. doi: [10.1177/1094428103254673](https://doi.org/10.1177/1094428103254673)
- Peyre, H., Leplege, A., & Coste, J. (2011). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, 20(2), 287-300. doi: [10.1007/s11136-010-9740-3](https://doi.org/10.1007/s11136-010-9740-3)
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27(1), 85-96. Retrieved from <http://www5.statcan.gc.ca/olc-cel/olc.action?ObjId=12-001-X20010015857&ObjType=47&lang=en>
- Royston, P., & White, I. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software*, 45(4). doi: [10.18637/jss.v045.i04](https://doi.org/10.18637/jss.v045.i04)
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons. doi: [10.1002/9780470316696](https://doi.org/10.1002/9780470316696)
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman and Hall/CRC.

MISSING DATA IN LONGITUDINAL SURVEYS

- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177. doi: [10.1037/1082-989x.7.2.147](https://doi.org/10.1037/1082-989x.7.2.147)
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571. doi: [10.1207/s15327906mbr3304_5](https://doi.org/10.1207/s15327906mbr3304_5)
- Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4), 317-329. doi: [10.1037/1082-989x.6.4.317](https://doi.org/10.1037/1082-989x.6.4.317)
- Stephens, A., Breeze, E., Banks, J., & Nazroo, J. (2013). Cohort profile: The English longitudinal study of ageing. *International Journal of Epidemiology*, 42(6), 1640-1648. doi: [10.1093/ije/dys168](https://doi.org/10.1093/ije/dys168)
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *The BMJ*, 338, b2393. doi: [10.1136/bmj.b2393](https://doi.org/10.1136/bmj.b2393)
- van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219-242. doi: [10.1177/0962280206074463](https://doi.org/10.1177/0962280206074463)
- van Buuren, S., & Groothuis-Oudshoorn, C. G. M., (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3). doi: [10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)
- von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1), 83-117. doi: [10.1111/j.1467-9531.2007.00180.x](https://doi.org/10.1111/j.1467-9531.2007.00180.x)
- Welch, C., Bartlett, J., & Petersen, I. (2014). Application of multiple imputation using the two-fold fully conditional specification algorithm in longitudinal clinical data. *The Stata Journal*, 14(2), 418-431. Retrieved from <http://www.stata-journal.com/article.html?article=st0345>
- White, I., Royston, P., Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399. doi: [10.1002/sim.4067](https://doi.org/10.1002/sim.4067)
- Young, R., & Johnson, D. R. (2010). Imputing the missing Y's: Implications for survey producers and survey users. In *Proceedings of the AAPOR conference abstracts* (pp. 6242-6248).

On Poisson Quasi-Lindley Distribution and its Applications

Razika Grine
Badji-Mokhtar University
Annaba, Algeria

Halim Zeghdoudi
Badji-Mokhtar University
Annaba, Algeria

This paper proposes a recent version of compound Poisson distributions named the Poisson quasi-Lindley (PQL) distribution by compounding Poisson and quasi-Lindley distributions. Some properties of the distributions are given with estimation and some illustrative examples.

Keywords: Lindley distribution, Poisson distribution, Poisson-Lindley distribution, gamma Lindley distribution, maximum-likelihood estimation

Introduction

Statistical distributions are commonly applied to describe real-world phenomena and are most frequently used in different fields such as medicine, finance, biological engineering sciences, and actuarial science. The one-parameter Lindley distribution is used in modeling lifetime data, and appears to perform well. To obtain it, let X be a random variable following the one-parameter distribution with the density function (Lindley, 1958)

$$f(x, \theta) = \begin{cases} \frac{\theta^2 (1+x) e^{-\theta x}}{1+\theta}, & x, \theta > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Sankaran (1970) used (1) assuming that the parameter of a Poisson distribution has Lindley distribution, and it was named the Poisson- Lindley distribution.

Asgharzadeh, Bakouch, and Esmaeili (2013), Ghitany, Al-Mutairi, and Nadarajah (2008), and Ghitany, Atieh, and Nadarajah (2008) studied the new

Razika Grine is a graduate student. Email them at: grinerazika@hotmail.com. Dr. Halim Zeghdoudi is an Associate Professor of Statistics in the Department of Mathematics. Email them at: halim.zeghdoudi@univ-annaba.dz.

distribution bounded to (1) and derived the zero-truncated Poisson-Lindley and Pareto Poisson-Lindley distributions. Sankaran (1970) introduced the discrete Poisson-Lindley distribution by combining the Poisson and Lindley distributions. Mahmoudi and Zakerzadeh (2010) proposed an extended version of the compound Poisson distribution, which was obtained by compounding the Poisson distribution with the generalized Lindley distribution, which was further analyzed by Zakerzadeh and Dolati (2009). Zeghdoudi and Nedjar (2016a) and Shanker and Mishra (2013) introduced the pseudo-Lindley and quasi-Lindley distributions, based on mixtures of gamma (2, θ) and exponential (θ) distributions, where the density function of the random variable X is given by

$$f_{QL}(x; \theta, \alpha) = \begin{cases} \frac{\theta(\alpha + \theta x)e^{-\theta x}}{\alpha + 1}, & x, \theta > 0; \alpha > -1 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Zeghdoudi and Nedjar (2016b, 2017) proposed compound Poisson distributions, named the Poisson Gamma Lindley (PGaL) distribution and Poisson pseudo-Lindley, by compounding Poisson and gamma Lindley (pseudo-Lindley) Nedjar and Zeghdoudi (2016a, 2016b) distributions. The purpose of this study is to introduce a new lifetime distribution by compounding Poisson and quasi-Lindley distributions, which may be useful in modeling lifetime data and biological sciences.

Poisson Quasi-Lindley Distribution

Consider $dF(\lambda) = e^{\lambda\Phi}h(\lambda)B(\Phi)d\lambda$, where $h(\lambda) = \alpha + \theta\lambda$ and $B(\Phi) = -\Phi / (\alpha + 1)$, then the compound Poisson distribution is (Sankaran, 1970)

$$\begin{aligned} P_x(\Phi) &= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} dF(\lambda) \\ &= \frac{B(\Phi)}{x!} \left[\alpha \int_0^\infty e^{(\Phi-1)\lambda} \lambda^x d\lambda + (-\Phi) \int_0^\infty e^{(\Phi-1)\lambda} \lambda^{x+1} d\lambda \right] \\ &= \frac{-\Phi}{\alpha + 1} \left(\frac{\alpha(1-\Phi) - \Phi(x+1)}{(1-\Phi)^{x+2}} \right) \end{aligned} \quad (3)$$

Then, replace Φ with $-\theta$:

$$P_x(\theta) = \frac{\theta}{\alpha+1} \left(\frac{\alpha(1+\theta) + \theta(x+1)}{(1+\theta)^{x+2}} \right) \quad (4)$$

Now, the density function of Poisson quasi-Lindley (PQL) is given by

$$f_{\text{PQL}}(x; \alpha, \theta) = \frac{\theta(\alpha + \theta + \alpha\theta + \theta x)}{(1+\alpha)(1+\theta)^{x+2}}, \quad x = 0, 1, \dots, \theta > 0, \alpha > -1 \quad (5)$$

Remark 1: If $\alpha = \theta$, this distribution is the Poisson-Lindley distribution.

The first and second derivatives of $f_{\text{PQL}}(x)$ are

$$\frac{d}{dx} f_{\text{PQL}}(x) = - \frac{\theta(\theta \ln(\theta+1) - \theta + \alpha \ln(\theta+1) + \theta \alpha \ln(\theta+1) + x \theta \ln(\theta+1))}{(\theta+1)^{x+2}(\alpha+1)} \quad (6)$$

and

$$\begin{aligned} & \frac{d^2}{dx^2} f_{\text{PQL}}(x) \\ &= - \frac{\theta \ln(\theta+1) [\theta \ln(\theta+1) - 2\theta + \alpha \ln(\theta+1) + \theta \alpha \ln(\theta+1) + x \theta \ln(\theta+1)]}{(\theta+1)^{x+2}(\alpha+1)} \end{aligned} \quad (7)$$

When $\frac{d}{dx} f_{\text{PQL}}(x) = 0$, the solution is

$$\hat{x} = \frac{1}{\ln(\theta+1)} - \frac{(\theta + \alpha + \theta \alpha)}{\theta}$$

and

$$\frac{d^2}{dx^2} f_{\text{PQL}}(\hat{x}) = \frac{-\theta^2 \ln(\theta+1)(\theta+1)^{\frac{1}{\theta \ln(\theta+1)} [\theta \ln(\theta+1) - \theta + \alpha \ln(\theta+1) + \theta \ln(\theta+1)] - 2}}{(\alpha+1)} < 0$$

For $\theta, \alpha, \hat{x} > 0$,

$$\hat{x} = \frac{1}{\ln(\theta+1)} - \frac{(\theta + \alpha + \theta\alpha)}{\theta}$$

is the unique critical point at which $f_{PQL}(x; \theta, \alpha)$ is maximum and $f_{PQL}(x)$ is concave. But if $\hat{x} < 0$, the density function $f_{PQL}(x)$ is decreasing in x . Therefore, the mode of PQL is given by

$$\text{Mode}(X) = \begin{cases} \frac{1}{\ln(\theta+1)} - \frac{(\theta + \alpha + \theta\alpha)}{\theta}, & \forall \theta > 0, \alpha > -1 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

The cumulative distribution function (cdf) of the PQL is

$$F_{PQL}(x) = 1 - \frac{\alpha + 2\theta + \alpha\theta + \theta x + 1}{(1 + \alpha)(1 + \theta)^{x+2}}, \quad x = 0, 1, \dots, \theta > 0, \alpha > -1 \quad (9)$$

The plots of density and distribution for some value of α and θ are given in Figures 1 and 2.

Figure 1. Plots of the density function for some parameter values

Figure 2. Plots of the distribution function for some parameter values

Survival and Hazard Rate Function

Let

$$S_{PQL}(x) = 1 - F_{PQL}(x) = \frac{\alpha + 2\theta + \alpha\theta + \theta x + 1}{(1 + \alpha)(1 + \theta)^{x+2}} \quad (10)$$

and

$$h_{PQL}(x) = \frac{f_{PQL}(x)}{1 - F_{PQL}(x)} = \frac{\theta(\alpha + \theta + \alpha\theta + \theta x)}{\alpha + 2\theta + \alpha\theta + \theta x + 1} \quad (11)$$

be the survival and hazard rate function of PQL, respectively.

Proposition 1: Let $h_{PQL}(x)$ be the hazard rate function of X . Then $h_{PQL}(x)$ is increasing.

Proof: According to Glaser (1980) and from the density function of PQL,

$$\begin{aligned}\rho(x) &= \frac{f'_{\text{PQL}}(x)}{f_{\text{PQL}}(x)} \\ &= \frac{\theta \ln(\theta+1) - \theta + \alpha \ln(\theta+1) + \theta \alpha \ln(\theta+1) + x \theta \ln(\theta+1)}{\alpha + \theta + \alpha \theta + \theta x}\end{aligned}\quad (12)$$

it follows that

$$\rho'(x) = \frac{\theta^2}{(\theta + \alpha + x\theta + \theta\alpha)^2} > 0$$

$\forall x, \alpha, \theta$, implying that $h_{\text{PQL}}(x)$ is increasing.

Maximum Likelihood Estimates

Consider the point estimation of the parameters that index the $\text{PQL}(\theta, \alpha)$. Let the log-likelihood function of a single observation (say x_i) for the vector of parameters (θ, α) be written as

$$\begin{aligned}\ln l(x; \alpha, \theta) \\ = \ln \theta + \ln(\alpha + \theta + \alpha \theta + \theta x) - \ln(1 + \alpha) - 2 \ln(1 + \theta) - x \ln(1 + \theta)\end{aligned}\quad (13)$$

The derivatives of $\ln l(x; \theta, \alpha)$ with respect to θ and α are:

$$\frac{\partial \ln l(x; \alpha, \theta)}{\partial \theta} = \frac{1}{\theta} + \left(\frac{1 + \alpha + x}{\alpha + \theta + \alpha \theta + \theta x} \right) - \frac{2 + x}{1 + \theta}\quad (14)$$

$$\frac{\partial \ln l(x; \alpha, \theta)}{\partial \alpha} = \left(\frac{1 + \theta}{\alpha + \theta + \alpha \theta + \theta x} \right) - \frac{1}{1 + \alpha}\quad (15)$$

The maximum likelihood estimators $\hat{\theta}$ of θ and $\hat{\alpha}$ of α are obtained by solving non-linear equations

$$\hat{\theta} = \frac{1}{x}, \quad \hat{\alpha} = \frac{-x}{1+x}\quad (16)$$

and

$$\begin{cases} E(\hat{\theta}) = \frac{\theta(\alpha + \theta + \alpha\theta)}{(1+\alpha)(1+\theta)^3} \\ -1 \leq E(\hat{\alpha}) \leq 0 \end{cases} \quad (17)$$

Proposition 2: Let $X_1, X_2, \dots, X_n$ be n independent random variables from the $PQL(\alpha, \theta)$ distribution. Then the moment generating function (mgf) of $S = \sum_{i=1}^n X_i$ is given by

$$M_S(t) = \frac{(\alpha\theta e^t + \theta^2)^n}{(1+\alpha)^n (e^t + \theta e^t - e^{2t})^n} \quad (18)$$

and

$$M_X(t) = E(e^{tX}) = \frac{\alpha\theta e^t + \theta^2}{(1+\alpha)(e^t + \theta e^t - e^{2t})} \quad (19)$$

Proof: The mgf of X is

$$M_X(t) = E(e^{tX}) = \frac{\alpha\theta e^t + \theta^2}{(1+\alpha)(e^t + \theta e^t - e^{2t})}$$

According to (19) and using the independent random variables $X_1, X_2, \dots, X_n$, the mgf of $S = \sum_{i=1}^n X_i$. Also, successive derivation is used and, by recurrence, find (18).

Corollary 1: Let $X \sim PQL(\theta, \alpha)$. Then the mean and variance for X are

$$E(X) = \frac{2+\alpha}{(1+\alpha)\theta} \quad (20)$$

$$V(X) = \frac{2+4\alpha+\alpha^2+\theta(\alpha+2)(1+\alpha)}{(1+\alpha)^2 \theta^2} \quad (21)$$

Proof: $E(X) = M'_X(t=0)$, $E(X^2) = M''_X(t=0)$, and $V(X) = E(X^2) - E^2(X)$. Then

$$\begin{aligned} E(X) &= \frac{2+\alpha}{(1+\alpha)\theta} \\ E(X^2) &= \frac{6+2\theta+\alpha(\theta+2)}{(1+\alpha)\theta^2} \end{aligned} \quad (22)$$

which achieves the proof.

Moments Estimates

Using the first moment m and second moment m_2 about the PQL distribution, we have

$$\begin{cases} m = \frac{2+\alpha}{(1+\alpha)\theta} \\ m_2 = \frac{6+2\theta+\alpha(2+\theta)}{(1+\alpha)\theta^2} \end{cases} \quad (23)$$

where $m_2 = S^2 + m^2$ and S^2 is the variance. Solve this non-linear system and find the couple (θ, α) , where $(\theta, \alpha) > 0$ for all $S > 0$, $m > 0$. The solving of the non-linear system (23) gives

$$(m_2 - m)\theta^2 - 4m\theta + 2 = 0 \quad \text{and} \quad \alpha = \frac{2 - \theta m}{\theta m - 1} \quad (24)$$

The solution of $(m_2 - m)\theta^2 - 4m\theta + 2 = 0$ is

$$-\frac{1}{m - m_2} \left(2m + \sqrt{2} \sqrt{m - m_2 + 2m^2} \right) \quad \text{if } m - m_2 \neq 0 \quad (25)$$

because

$$\hat{\theta} = -\frac{1}{m - m_2} \left(2m + \sqrt{2} \sqrt{m - m_2 + 2m^2} \right) \quad \text{and} \quad \hat{\alpha} = \frac{2 - \hat{\theta}m}{\hat{\theta}m - 1} \quad (26)$$

The Quantile Function of the Poisson Quasi-Lindley Distribution

Lambert W Function

The Lambert W function is a standard due to its implementation in the computer algebra system Maple in the 1980s (Conte & de Boor, 1980) and, subsequently, Corless, Gonnet, Hare, Jeffrey, and Knuth (1996) provided a comprehensive survey of the history, theory, and applications of this function. The Lambert W function is defined as the solution of the equation:

$$W(z)\exp(W(z)) = z, \quad z \text{ is a complex number} \quad (27)$$

If z is a real number such that $z \geq -1/e$ then $W(z)$ becomes a real function and there are two possible real branches. The real branch taking on values in $(-\infty, -1]$ is called the negative branch and denoted by W_{-1} . The real branch taking on values in $[-1, \infty)$ is called the principal branch and denoted by W_0 . Equation (8) has two possible solutions if $z \in (-1/e, 0]$ and a unique solution if $z \geq 0$. For the results in this note, use the negative branch W_{-1} , which satisfies the following elementary properties: $W_{-1}(-1/e) = -1$, $W_{-1}(z)$ decreasing as z increases to 0, and $W_{-1}(z) \rightarrow -\infty$ as $z \rightarrow 0$.

Lemma 1: Let a, b, c , and d be fixed complex numbers. The solution of the equation $z + abz = c$ with respect to $z \in \mathbb{C}$ is

$$z = c - \frac{1}{d \ln(b)} W(ab^c \ln(b)) \quad (28)$$

For details of the proof, see Jodrá (2010).

The quantile function of X is $Q_X(u) = F_X^{-1}(u)$, $0 < u < 1$. An explicit expression for Q_X in terms of the Lambert W function follows.

Theorem 1: For any $\theta, \alpha > 0$, the quantile function of the PQL distribution X is

$$Q_X(u) = -\frac{\alpha + 2\theta + \alpha\theta + 1}{\theta} - \frac{1}{\ln(1+\theta)} W_{-1}\left(\frac{\ln(1+\theta)}{\theta(1+\theta)^{\frac{\alpha+\alpha\theta+1}{\theta}}}(u-1)\right) \quad (29)$$

where W_{-1} denotes negative branch of Lambert W function.

Proof: For any fixed θ, α , let $u \in (0, 1)$. We have to solve the equation $F_X(x) = u$ with respect to x , for $x > 0$. Solve the following equation; the first quantiles are obtained by substituting $u = 1/4, 1/2, 3/4$ in equation (29)

$$\begin{aligned} Q_1 &= F^{-1}\left(\frac{1}{4}, \theta, \alpha\right) \\ &= -\frac{\alpha + 2\theta + \alpha\theta + 1}{\theta} - \frac{1}{\ln(1+\theta)} W_{-1}\left(\frac{\ln(1+\theta)}{\theta(1+\theta)^{\frac{\alpha+\alpha\theta+1}{\theta}}}\left(\frac{1}{4}-1\right)\right) \end{aligned} \quad (30)$$

$$\begin{aligned} Q_2 &= F^{-1}\left(\frac{1}{2}, \theta, \alpha\right) \\ &= -\frac{\alpha + 2\theta + \alpha\theta + 1}{\theta} - \frac{1}{\ln(1+\theta)} W_{-1}\left(\frac{\ln(1+\theta)}{\theta(1+\theta)^{\frac{\alpha+\alpha\theta+1}{\theta}}}\left(\frac{1}{2}-1\right)\right) \end{aligned} \quad (31)$$

$$\begin{aligned} Q_3 &= F^{-1}\left(\frac{3}{4}, \theta, \alpha\right) \\ &= -\frac{\alpha + 2\theta + \alpha\theta + 1}{\theta} - \frac{1}{\ln(1+\theta)} W_{-1}\left(\frac{\ln(1+\theta)}{\theta(1+\theta)^{\frac{\alpha+\alpha\theta+1}{\theta}}}\left(\frac{3}{4}-1\right)\right) \end{aligned} \quad (32)$$

Illustrative Examples

Example 1

Shown in Table 1 are some quantile of the PQL distribution, which were calculated from the closed-form expression for $Q_X(u)$ given in Theorem 1.

Table 1. Quantile of the PQL distribution

u	$\theta = 0.1, \alpha = 0.1$	$\theta = 0.1, \alpha = 0.5$	$\theta = 3.0, \alpha = 1.0$	$\theta = 5.0, \alpha = 1.0$
0.01	3.28840	7.87570	-0.33136	-0.50238
0.05	4.42420	8.60620	-0.29346	-0.47390
0.10	5.76500	9.54370	-0.24395	-0.43667
0.15	7.06270	10.51300	-0.19181	-0.39743
0.25	9.63630	12.56900	-7.83450×10^{-2}	-0.31190
0.30	10.94900	13.67100	-1.61900×10^{-2}	-0.26497
0.35	12.30000	14.83200	5.02850×10^{-2}	-0.21474
0.40	13.70500	16.06400	0.12176	-0.16067
0.45	15.18000	17.38100	0.19911	-0.10211
0.50	16.74500	18.79900	0.28342	-3.82080×10^{-2}
0.55	18.42500	20.34100	0.37617	3.21560×10^{-2}
0.60	20.25200	22.03700	0.47930	0.11049
0.65	22.26900	23.92800	0.59557	0.19889
0.70	24.53900	26.07600	0.72901	0.30046
0.75	27.15700	28.57300	0.88581	0.41995
0.80	30.28300	31.57700	1.07530	0.56534
0.85	34.21100	35.37800	1.32000	0.75150
0.90	39.59600	40.62500	1.66010	1.01170
0.95	48.50800	49.36800	2.23390	1.45160

Table 2. Mode, mean, and median for PQL

	$\theta = 0.01, \alpha = 0.10$	$\theta = 0.10, \alpha = 0.50$	$\theta = 0.05, \alpha = 1.00$
Q_1	104.660	12.569	30.601
Median = Q_2	172.920	18.799	41.819
Q_3	272.770	28.573	60.008
Mean	89.3990	3.9921	-1.5041
Mode	190.910	16.667	30.000

Displayed in Table 2 are the mode, mean and median for PQL distribution for different choices of parameters θ and α .

ON POISSON QUASI LINDLEY DISTRIBUTION

Simulation

The behavior of the MM estimators are examined for a finite sample size (n). A simulation study consisting of following steps is being carried out for each triplet $(\theta, \alpha; n)$, where $\theta = 0.01, \alpha = 0.1, 0.01, 1$ and for $\alpha = 0.5, \theta = 0.05, 1, 5$, and $n = 10, 30, 50$. The steps are:

- Choose the initial values of θ_0, α_0 for the corresponding elements of the parameter vector $\Theta = (\theta, \alpha)$ to specify PQL distribution;
- Choose sample size n ;
- Generate N independent samples of size n from PQL (θ, α) ;
- Compute the MM estimate Θ_n of Θ_0 for each of the N samples;
- Compute the mean of the obtained estimators over all N samples.

Note the

$$\text{average bias}(\theta) = \frac{1}{N} \sum_{i=1}^n (\Theta_i - \Theta_0) \quad (33)$$

and the average square error

$$\text{MSE}(\theta) = \frac{1}{N} \sum_{i=1}^n (\Theta_i - \Theta_0)^2 \quad (34)$$

Table 3. Average bias of the simulated estimates

n	$\theta = 0.01, \alpha = 0.1$		$\theta = 0.01, \alpha = 0.01$		$\theta = 0.01, \alpha = 1$	
	$\text{bias}(\theta)$	$\text{bias}(\alpha)$	$\text{bias}(\theta)$	$\text{bias}(\alpha)$	$\text{bias}(\theta)$	$\text{bias}(\alpha)$
10	0.000025560	-0.005612900	0.000002500	-0.000505090	0.000264000	-0.088393000
30	0.000008520	-0.001870960	0.000000833	-0.000168363	0.000088000	-0.029464330
50	0.000005112	-0.001122580	0.000000500	-0.000101018	0.000052800	-0.017678600

n	$\theta = 0.05, \alpha = 0.5$		$\theta = 1, \alpha = 0.5$		$\theta = 5, \alpha = 0.5$	
	$\text{bias}(\theta)$	$\text{bias}(\alpha)$	$\text{bias}(\theta)$	$\text{bias}(\alpha)$	$\text{bias}(\theta)$	$\text{bias}(\alpha)$
10	0.000684200	-0.038232000	0.027430000	-0.061022000	0.575600000	-0.111320000
30	0.000298060	-0.012744000	0.009143333	-0.020340660	0.191866600	-0.037106600
50	0.000136840	-0.007646400	0.005486000	-0.012204400	0.115120000	-0.022264000

Shown in Table 3, $\hat{\theta}$ is positively biased with $\text{bias}(\theta) \rightarrow 0$ for $\theta \rightarrow 0$, and $\hat{\alpha}$ is negatively biased with $\text{bias}(\alpha) \rightarrow 0$ for $\alpha \rightarrow 0$. Shown in Table 4, $\text{MSE}(\theta)$ and $\text{MSE}(\alpha) \rightarrow 0$ where $\theta \rightarrow 0$ and $n \rightarrow \infty$.

Example 2

Shown in Table 5 are some distributions of copying groups of random digits with expected frequencies obtained by fitting the Poisson, Poisson-Lindley, and PQL distributions.

Table 4. Average MSE of the simulated estimates

n	$\theta = 0.01, \alpha = 0.1$		$\theta = 0.01, \alpha = 0.01$		$\theta = 0.01, \alpha = 1$	
	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$
10	0.0000000066	0.0003150460	6.3×10^{-11}	0.0000025511	0.0000006970	0.0781332240
30	0.0000000022	0.0001050150	2×10^{-11}	0.0000008504	0.0000002323	0.0260444080
50	0.0000000013	0.0000630090	10^{-11}	0.0000005102	0.0000001394	0.0156266440

n	$\theta = 0.05, \alpha = 0.5$		$\theta = 1, \alpha = 0.5$		$\theta = 5, \alpha = 0.5$	
	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$	$\text{MSE}(\theta)$	$\text{MSE}(\alpha)$
10	0.0000046813	0.0146168500	0.0075240400	0.0372368400	3.3131536000	0.1239214240
30	0.0000015604	0.0048722800	0.0025080160	0.0124228000	1.1043845300	0.0413071410
50	0.0000009363	0.0029233700	0.0015048090	0.0074473600	0.6626307200	0.0247842840

Table 5. Comparison between Poisson, Poisson-Lindley, and Poisson quasi-Lindley distributions

No. of errors per group	Obs. freq.	Poisson	Poisson-Lindley		PQL
	$m = 0.9, n = 2.451$	$\hat{\theta} = 1.08$	$\hat{\theta} = 1.547$	$\hat{\theta} = 1.398, \hat{\beta} = 0.786$	
0	35	25.207	31.856		32.152
1	11	22.686	16.031		15.320
2	8	10.209	7.677		7.602
3	4	3.062	3.557		3.542
4	3	0.689	1.609		1.613
5	1	0.124	0.715		0.749
Total	62	62	62		62
χ^2	-	24.524	3.271		3.216

Conclusion

A new two-parameter distribution is proposed, referred to as the PQL distribution, which contains the Poisson Lindley distribution as special case. Various properties of the distribution are examined including the density function (pdf), cumulative distribution (cdf), survival and hazard rate function, moment generating function (mgf), mean, variance, and some results. Also, maximum likelihood estimates and moment estimates are discussed. The PQL model was fitted to several real data sets to show the potential of the new proposed distribution. The PQL distribution gives a much closer fit than the Poisson and Poisson-Lindley distributions, and thus can be considered as an important tool for modeling lifetime data. This suggests that the new model provides more accurate estimates as well as better fits.

Acknowledgements

The authors acknowledge Editor for the constant encouragement to finalize the paper. Further, the authors acknowledge profound thanks to an anonymous referee for giving critical comments which have immensely improved the presentation of the paper.

References

- Asgharzadeh, A., Bakouch, H. S., & Esmaeili, L. (2013). Pareto Poisson-Lindley distribution and its applications. *Journal of Applied Statistics*, 40(8), 1-18. doi: [10.1080/02664763.2013.793886](https://doi.org/10.1080/02664763.2013.793886)
- Conte, S. D., & de Boor, C. (1980). *Elementary numerical analysis* (3rd ed.). New York, NY: McGraw Hill.
- Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, 5(1), 329-359. doi: [10.1007/bf02124750](https://doi.org/10.1007/bf02124750)
- Ghitany, M. E., Al-Mutairi, D. K., & Nadarajah, S. (2008). Zero-truncated Poisson-Lindley distribution and its applications. *Mathematics and Computers in Simulation*, 79(3), 279-287. doi: [10.1016/j.matcom.2007.11.021](https://doi.org/10.1016/j.matcom.2007.11.021)
- Ghitany, M. E., Atieh, B., & Nadarajah, S. (2008). Lindley distribution and its applications. *Mathematics and Computers in Simulation*, 78(4), 493-506. doi: [10.1016/j.matcom.2007.06.007](https://doi.org/10.1016/j.matcom.2007.06.007)

- Glaser, R. E. (1980). Bathtub and related failure rate characterizations. *Journal of the American Statistical Association*, 75(371), 667-672. doi: 10.2307/2287666
- Jodrá, P. (2010). Computer generation of random variables with Lindley or Poisson-Lindley distribution via the Lambert W function. *Mathematics and Computers in Simulation*, 81(4), 851-859. doi: 10.1016/j.matcom.2010.09.006
- Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1), 102-107. Available from <http://www.jstor.org/stable/2983909>
- Mahmoudi, E. & Zakerzadeh, H. (2010) Generalized Poisson-Lindley distribution. *Communications in Statistics – Theory and Methods*, 39(10), 1785-1798. doi: 10.1080/03610920902898514
- Nedjar, S., & Zeghdoudi, H. (2016a). Gamma Lindley distribution and its application. *Journal of Applied Probability and Statistics*, 11(1), 129-138.
- Nedjar, S., & Zeghdoudi, H. (2016b). On gamma Lindley distribution: Properties and simulations. *Journal of Computational and Applied Mathematics*, 298, 167-174. doi: 10.1016/j.cam.2015.11.047
- Sankaran, M. (1970). 275. Note: The discrete Poisson-Lindley distribution. *Biometrics*, 26(1), 145-149. doi: 10.2307/2529053
- Shanker, R., & Mishra, A. (2013). A quasi Lindley distribution. *African Journal of Mathematics and Computer Science Research*, 6(4), 64-71.
- Zakerzadeh, H., & Dolati, A. (2009). Generalized Lindley distribution. *Journal of Mathematical Extension*, 3(2), 1-17. Retrieved from <http://www.ijmex.com/index.php/ijmex/article/view/45>
- Zeghdoudi, H., & Nedjar, S. (2016a). A pseudo Lindley distribution and its application. *Afrika Statistika*, 11(1), 923-932. doi: 10.16929/as/2016.923.83
- Zeghdoudi, H., & Nedjar, S. (2016b). *On Poisson gamma Lindley distribution: Properties and applications* [Unpublished manuscript].
- Zeghdoudi, H., & Nedjar, S. (2017). On Poisson pseudo Lindley distribution: Properties and applications. *Journal of Probability and Statistical Science*, 15(1), 19-28.

Detection of Outliers in Univariate Circular Data using Robust Circular Distance

Ehab A. Mahmood
University Putra Malaysia
Serdang, Malaysia

Sohel Rana
East West University
Dhaka, Bangladesh

Habshah Midi
University Putra Malaysia
Serdang, Malaysia

Abdul Ghapor Hussin
National Defence University of Malaysia
Kuala Lumpur, Malaysia

A robust statistic to detect single and multi-outliers in univariate circular data is proposed. The performance of the proposed statistic was tested by applying it to a simulation study and to three real data sets, and was demonstrated to be robust.

Keywords: Circular data, outliers, masking, swamping

Introduction

Statistical data can be classified according to their distributional topologies into two sets. First, linear data can be represented on a straight line. Second, circular data can be represented on the circumference of a unit circle. Circular data can be measured either in degrees, when they are distributed in the interval $[0^\circ - 360^\circ)$, or in radians, in the interval $[0 - 2\pi)$. Circular data can arise in contrasting scientific fields such as earth sciences, meteorology, biology, physics, psychology, image analysis, and medicine. The classical statistics that we apply to linear data cannot be used for circular data because of the geometrical properties of the circular data. For example, if we have two circular data points at 100° and 300° , then the arithmetic mean according to the linear measure is equal to 200° . However, the mean direction is equal to 80° according to the geometrical theory of the circle.

A special statistical measure is needed to deal with circular data. Between 5% and 10% of any set of statistical data are surprising points that are often called outliers (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). They may unduly affect the statistical analysis and the final outcomes.

Ehab A. Mahmood is a PhD student in the Department of Mathematics, Faculty of Science. Email them at: eee.mahmood@gmail.com.

There are many methods to detect outliers in linear univariate data. However, little is known regarding detecting outliers in circular data. It can be defined as an observation that is discordant in comparison with the rest of the sample. In linear data, an outlier is an observation that is extreme. However, the outliers in circular data may not have extreme values. For example, consider the following circular data (Collett, 1980):

10,15,20,25,30,350

As linear data, it is obvious that the last observation, 350, is an outlier. However, as circular data, it is clear that 350 is consistent with the other observations. Therefore, the methods used to detect outliers in circular data are different from those used for linear data.

Mardia (1975) suggested a statistic to identify a single outlier in univariate circular data. He considered the observation that is the most influential on the resultant length to be an outlier. Collett (1980) proposed four test statistics, namely L , C , D , and M , to identify a single outlier in univariate circular data. It was found for small samples sizes that it is better to use the C and D statistics. However, no statistic was recommended to detect multiple outliers, and typical methods are only successful in detecting a single outlier. Furthermore, there was no discussion on how to identify an outlier when the sample size is large.

Fisher (1993) summarized three causes of outliers in statistical data: mis-recording, unwitting sampling from another population, and vagaries of sampling resulting in the occasional isolated value. In this identification he used the M statistic, which had already been suggested by Collett (1980), and did not propose a new statistic.

Mardia and Jupp (2000) suggested that circular data could be tested by considering three factors: The first was the mean resultant length. They promoted the use of either the *Mardia* (Mardia, 1975) statistic or the C statistic (Collett, 1980). The second was the likelihood ratio test for slippage in the model. For circular data, they considered either the likelihood ratio test for location slippage in a von Mises distribution (Collett, 1980) or the likelihood ratio test for concentration slippage in a Fisher distribution (Fisher, Lewis, & Willcox, 1981). Their final factor was the exponential distribution. Some tests for this factor had been suggested by Fisher et al. However, Mardia and Jupp did not suggest a new test statistic.

Jammalamadaka and SenGupta (2001) promoted the use of the P - P plot as a simple graphical way of detecting outliers in circular data. Furthermore, they proposed two statistics: The first of these was the locally most powerful invariant

(LMPI) statistic. They used LMPI for the circular data that were applied in the wrapped stable and uniform mixture model WSM. Second, they proposed using a likelihood ratio testing (LRT) approach to identify outliers in circular data. They tested the hypothesis that $\vartheta_1, \vartheta_2, \dots, \vartheta_{i-1}, \vartheta_{i+1}, \dots, \vartheta_n$ follow a von Mises distribution $[vM(\mu_0, k)]$ and that ϑ_i is distributed as $[vM(\mu_1, k)]$, where i is unknown. They applied this test to two cases: first, when μ_0, μ_1 , and k are all known and second, when only k is known. They calculated the power of the procedure and the probability of detecting outliers perfectly by comparing both the LMPI and the LRT approaches with the L statistic (Collett, 1980) and with the statistic of Mardia (1975). They noted that their statistics are better than the other statistics. However, they did not propose a way to test circular data if k is unknown.

Otieno and Anderson-Cook (2005) tested three of the preferred directions, mean direction, median direction (Fisher, 1993), and the Hodges-Lehmann (HL) estimate (Otieno & Anderson-Cook, 2003). They concluded circular HL is a good compromise between circular mean and circular median, like its counterpart for linear data. The HL estimator is less robust to outliers compared to the median, but it is an efficient alternative because it has a smaller circular variance.

Abuzaid, Mohamed, and Hussin (2009) proposed the A statistic to detect an outlier in univariate circular data. This depends on the sum of the circular distances from any point to all other points on the circumference of the unit circle. They depended on calculating both the probability that the contaminant observation was an extreme observation and could be identified as an outlier and the probability of a type II error as a measure for comparing their suggestion with the C , D , and M statistics. However, the probability results for the A statistic are close to the results for the C statistic and they did not test their suggestion for large sample size or apply it to identify multiple outliers.

Abuzaid (2010) used the geometrical properties of the chord of a circle for detecting an outlier in univariate circular data. However, this suggestion detects only a single outlier. Abuzaid, Hussin, Rambli, and Mohamed (2012) then suggested a test statistic to detect outliers in univariate and bivariate circular data. The test statistic was based on the approximate distribution of the circular distances between the sample points. Nonetheless, they did not evaluate their statistic using any statistical measures. Moreover, they suggested a way to identify only a single outlier.

Mohamed, Rambli, Khaliddin, and Ibrahim (2015) proposed a procedure to identify single outliers and patches of outliers in univariate circular data. It is based on spacing theory in circular data. They compared their procedure with the C , D , and A statistics. However, their procedure is difficult, especially if the circular data

have multiple outliers. Furthermore, the rates of swamping, the identification of inliers as outliers (Barnett & Lewis, 1978), are relatively high. For more information, see Beckman and Cook (1983) and Barnett and Lewis (1978), who reviewed the literature on the detection of outliers in various areas of statistical data.

The aim of this study is to investigate the robustness of our proposed statistic to detect outliers in univariate circular data when the data follow a von Mises distribution. The circular distance between any circular data point and the circular median is considered as a statistic test to identify outliers.

The circular median is defined as any angle ϑ such that half of the data points lie in the arc $\vartheta, \vartheta + \pi$ and the majority of the data points are nearer to ϑ than to $\vartheta + \pi$ (Mardia & Jupp, 2000). This will be compared with existing methods to detect a single outlier. Furthermore, the aim is to identify outliers when there is a high level of contamination and with large sample sizes, using various statistical measures to evaluate the procedure.

von Mises Distribution

Let $\vartheta_1, \vartheta_2, \dots, \vartheta_n$ be circular observations following a von Mises distribution with mean direction μ and concentration parameter k , denoted by $[vM(\mu, k)]$. The probability density function of the von Mises distribution is given by Hamelryck, Mardia, and Ferkinghoff-Borg (2012) as follows:

$$g(\vartheta, \mu, k) = \frac{1}{2\pi I_0} e^{k \cos(\vartheta - \mu)} \quad (1)$$

where $0 \leq \mu < 2\pi$, $k \geq 0$, and I_0 denotes the modified Bessel function of the first kind and order 0, which can be defined as follows:

$$I_0(k) = \frac{1}{2\pi} \int_0^{2\pi} e^{k \cos(\vartheta)} d\vartheta$$

If $k = 0$, then the probability density function of the von Mises distribution will be the same as the probability density function of the uniform distribution of circular data, where

$$g(\vartheta, \mu, 0) = \frac{1}{2\pi}$$

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

The mean direction of the circular observations is estimated according to the following formula (Fisher, 1993):

$$\hat{\mu} = \begin{cases} \tan^{-1}(s/c) & \text{if } s > 0, c > 0 \\ \tan^{-1}(s/c) + \pi & \text{if } c < 0 \\ \tan^{-1}(s/c) + 2\pi & \text{if } s < 0, c > 0 \end{cases} \quad (2)$$

where

$$s = \sum_{i=1}^n \sin(\mathcal{G}_i), \quad c = \sum_{i=1}^n \cos(\mathcal{G}_i)$$

The mean resultant length $\bar{R}$ is a measure of the concentration of the circular observations at a specific point of the circumference of the circle. It is calculated using this formula:

$$\bar{R} = \sqrt{\bar{c}^2 + \bar{s}^2} \quad (3)$$

where $0 \leq \bar{R} \leq 1$, $\bar{c} = c / n$, $\bar{s} = s / n$.

$\bar{R} = 0$ is satisfied if and only if the circular data are widely dispersed on the circumference ($\bar{c} = 0$ and $\bar{s} = 0$). $\bar{R} = 1$ is satisfied if and only if the circular data have a high concentration at a specific point ($\bar{c} + \bar{s} = 1$).

The concentration parameter k of the circular observations is estimated according to the following formula (Fisher, 1993):

$$\hat{k} = \begin{cases} 2\bar{R} + \bar{R}^3 + \frac{5}{6} \bar{R}^5 & \text{if } \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \frac{0.43}{(1-\bar{R})} & \text{if } 0.53 \leq \bar{R} < 0.85 \\ (\bar{R}^3 - 4\bar{R}^2 + 3\bar{R})^{-1} & \text{if } \bar{R} \geq 0.85 \end{cases} \quad (4)$$

Methods

Detection of a Single Outlier by Existing Methods

The following methods for the detection of single outlier will be compared in univariate circular data with the suggested procedure.

Mardia Statistic

Mardia (1975) suggested a statistic to identify a single outlier in univariate circular data. The outlier was considered to be the observation that is the most influential on the resultant length. Therefore,

$$Mar = \min \left\{ \frac{n-1-R_{(-i)}}{n-R} \right\} \quad (5)$$

where $R_{(-i)}$ is the resultant length after omitting the i^{th} observation and R is the resultant length for the full data set.

M Statistic

Collett (1980) suggested the M statistic to detect an outlier in univariate circular data. This is given as

$$M = \max \left\{ \frac{R_{(-i)} - R + 1}{n - R} \right\} = \frac{R_k - R + 1}{n - R} \quad (6)$$

where $R_k = \max\{R_{(-i)}\}$.

A Statistic

Abuzaid et al. (2009) used the circular distance between the circular observations $\mathcal{G}_i$ and $\mathcal{G}_j$ as suggested by Rao (1969). This is given as

$$d_{ij} = 1 - \cos(\mathcal{G}_i - \mathcal{G}_j)$$

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

where $d_{ij} \in [0, 2]$. The sum of all circular distances of the ϑ_j to all other observations is given by

$$D_j = \sum_{i=1}^n \left(1 - \cos(\vartheta_i - \vartheta_j)\right), \quad j = 1, 2, \dots, n$$

Abuzaid et al. (2009) argued that if the observation θ_j is an outlier (and so lies far away from the other observations), the value of D_j will increase. Therefore, the A statistic to detect an outlier in the circular univariate data is based on the average circular distance

$$\frac{D_j}{n-1}$$

when omitting the observation θ_j . The statistic is given as follows:

$$A = \max \left\{ \frac{D_j}{2(n-1)} \right\}, \quad j = 1, 2, \dots, n \quad (7)$$

where $A \in [0, 1]$. Jammalamadaka and SenGupta (2001) considered the circular distance between any two circular data points to be the smallest arc between them on the circumference. They calculate the circular distance between ϑ_i and ϑ_j as follows:

$$d_{ij}^* = \pi - \left| \pi - \left| \vartheta_i - \vartheta_j \right| \right| \quad (8)$$

where $d_{ij}^* \in [0, \pi]$.

Abuzaid (2010) used equation (8) to calculate the sum of all circular distances from the observation ϑ_j to all other observations. This is given as

$$D_j^* = \sum_{i=1}^n \left(\pi - \left| \pi - \left| \vartheta_i - \vartheta_j \right| \right| \right), \quad j = 1, 2, \dots, n$$

An alternative statistic is given by

$$A^* = \max \left\{ \frac{D_j^*}{n-1} \right\}, \quad j = 1, 2, \dots, n$$

where $A^* \in [0, \pi]$. Abuzaid expected that the statistic A^* has a similar performance to the A statistic.

Chord Statistic

Abuzaid (2010) used the geometrical properties of the chord of a circle to develop an alternative test to identify an outlier in circular univariate data. A chord is a segment that connects two different points on circumference of the circle. The length of the chord between ϑ_i and ϑ_j can be calculated using the following formula:

$$\text{crd}(d_{ij}) = 2r \sin \frac{d_{ij}^*}{2r}$$

where r is the radius, so in the unit circle $r = 1$, and d_{ij}^* is the smallest arc length between ϑ_i and ϑ_j , which is calculated from equation (8).

In the unit circle, calculating S_j , the sum of the lengths of all the chords passing through observation ϑ_j , was proposed as

$$S_j = \sum_{i=1}^n \text{crd}(d_{ij}) = 2 \sum_{i=1}^n \sin \frac{d_{ij}^*}{2}, \quad j = 1, 2, \dots, n$$

where $0 \leq S_j \leq 2(n-1)$. When $d_{ij}^* = 0$,

$$\sum_{i=1}^n \sin \frac{d_{ij}^*}{2} = 0$$

while

$$\sum_{i=1}^n \sin \frac{d_{ij}^*}{2} = n-1$$

when $d_{ij}^* = \pi$. Therefore, if S_j has the maximum value, this suggests that $\mathcal{G}_j$ is a candidate for the outlier. The chord statistic is given by

$$Chord = \max \left\{ \frac{S_j}{2(n-1)} \right\}, \quad j = 1, 2, \dots, n \quad (9)$$

The Proposed Method

A robust circular distance $RCDu$ statistic is now proposed to identify outliers in the circular data. It depends on two main points: first, the fact that the outliers in circular data may not be extreme values and second, an important property of the von Mises distribution is to be symmetric about the mean direction. However, the circular median is more efficient than the mean direction when the circular data have outliers (Ducharme & Milasevic, 1987). He and Simpson (1992) recommended the circular median is more robust than mean direction when the data do not follow von Mises distribution. Therefore, use the circular distance between any observation and circular median as a statistic to detect single and multi-outliers.

Suppose $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_n$ are circular observations located on the circumference of a unit circle. To apply the proposed procedure, there are several possible ways to calculate the circular distance $dist_{(i)}$ between $\mathcal{G}_i$ and the circular median med because of the circular geometry of the data. The cases are as follows:

- i. If $0 \leq med \leq \pi$

$$dist_{(i)} = \begin{cases} |\mathcal{G}_i - med| & \text{if } |\mathcal{G}_i - med| \leq \pi \\ 2\pi - \mathcal{G}_i + med & \text{if } |\mathcal{G}_i - med| > \pi \end{cases} \quad (10)$$

- ii. If $\pi \leq med \leq 2\pi$

$$dist_{(i)} = \begin{cases} |\mathcal{G}_i - med| & \text{if } |\mathcal{G}_i - med| \leq \pi \\ 2\pi - med + \mathcal{G}_i & \text{if } |\mathcal{G}_i - med| > \pi \end{cases} \quad (11)$$

If $\mathcal{G}_i$ is an outlier then $dist_{(i)}$ is expected to be relatively large. Therefore, the cut-off point is given by

$$RCDu = \max(dist) \quad (12)$$

Table 1. Cut-off points for $RCDu$ statistic with 10%

n	$k=2$	3	4	5	6	7	8	9	10	12	15	20
10	2.550	1.900	1.490	1.280	1.130	1.040	0.974	0.910	0.848	0.775	0.710	0.598
20	2.800	2.190	1.660	1.440	1.270	1.150	1.070	1.000	0.938	0.843	0.756	0.646
30	2.820	2.350	1.760	1.520	1.330	1.780	1.090	1.010	0.974	0.860	0.762	0.662
40	2.850	2.500	1.860	1.560	1.350	1.240	1.130	1.060	1.020	0.911	0.797	0.786
50	3.000	2.610	1.930	1.610	1.400	1.270	1.170	1.090	1.040	0.925	0.818	0.698
60	3.030	2.660	1.990	1.630	1.420	1.300	1.190	1.110	1.080	0.940	0.835	0.715
70	3.060	2.700	2.030	1.660	1.480	1.320	1.210	1.130	1.060	0.960	0.849	0.725
80	3.060	2.750	2.090	1.720	1.500	1.340	1.230	1.140	1.080	0.981	0.858	0.733
90	3.080	2.800	2.120	1.750	1.510	1.350	1.250	1.150	1.100	0.990	0.866	0.740
100	3.080	2.830	2.150	1.750	1.520	1.360	1.260	1.170	1.110	0.995	0.872	0.752
110	3.090	2.860	2.190	1.770	1.550	1.380	1.270	1.180	1.120	0.996	0.886	0.755
120	3.090	2.880	2.230	1.790	1.560	1.390	1.280	1.190	1.130	1.010	0.895	0.759
130	3.100	2.890	2.270	1.810	1.570	1.390	1.300	1.200	1.140	1.030	0.897	0.770
140	3.100	2.900	2.270	1.850	1.580	1.410	1.300	1.210	1.140	1.010	0.905	0.776
150	3.110	2.930	2.310	1.840	1.590	1.420	1.320	1.220	1.150	1.040	0.917	0.782
160	3.110	2.940	2.330	1.870	1.610	1.440	1.320	1.230	1.150	1.040	0.913	0.784
170	3.110	2.960	2.380	1.860	1.620	1.440	1.330	1.240	1.160	1.040	0.917	0.788
180	3.110	2.970	2.420	1.880	1.630	1.450	1.340	1.250	1.170	1.050	0.931	0.792
190	3.110	2.980	2.390	1.920	1.630	1.460	1.350	1.250	1.170	1.060	0.928	0.795
200	3.110	2.990	2.450	1.910	1.640	1.480	1.350	1.260	1.170	1.070	0.932	0.797

Consequently, $\mathcal{G}_i$ is identified as an outlier if $dist_{(i)}$ exceeds the cut-off point. We will depend on a triple measure of robustness to evaluate the proposed method:

- i. Proportion of outliers detected.
- ii. Rate of masking.
- iii. Rate of swamping.

This triple measure of robustness is very popular in the robustness literature for evaluating a particular method. A high proportion of outliers detected, and low masking and swamping rates, are always considered to be good robustness properties for any outlier detection statistic.

Results

The Cut-Off Point for the $RCDu$ Statistic

The $RCDu$ statistic has no simple known distributional form. Therefore, a series of simulation studies of univariate circular data are carried out to find the cut-off point

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

for the $RCDu$ statistic using Monte Carlo methods. The same procedure has been used by Jammalamadaka and SenGupta (2001) and Abuzaid et al. (2009). Twenty different sample sizes of $n = 10, 20, 30, \dots, 200$ and twelve values of concentration parameter $k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20$ are used in these simulation studies. First, generate a set of circular data such that $\vartheta \sim \text{vM}(0, k)$, for each sample size n and concentration parameter k . Then the $RCDu$ statistic is calculated. The process is replicated 5000 times to generate the $RCDu$ statistic for each combination of sample size n and concentration parameter k . Finally, the 10% and 5% upper points of $RCDu$ are tabulated in Tables 1 and 2, respectively. These tabulated values for different sample sizes and concentrations can be used as cut-off points for the proposed statistic. However, it is possible to find the cut-off points for any sample size and concentration parameter. R codes to generate any cut-off points are available from the corresponding author.

Table 2. Cut-off points for $RCDu$ statistic with 5%

n	$k = 2$	3	4	5	6	7	8	9	10	12	15	20
10	2.760	2.200	1.650	1.430	1.160	1.100	1.080	1.020	0.948	0.857	0.770	0.663
20	2.960	2.490	1.850	1.600	1.400	1.250	1.170	1.090	1.020	0.913	0.809	0.701
30	2.980	2.650	1.980	1.670	1.460	1.290	1.180	1.100	1.070	0.934	0.827	0.712
40	3.000	2.760	2.100	1.720	1.460	1.350	1.230	1.140	1.100	0.968	0.984	0.838
50	3.060	2.840	2.170	1.780	1.520	1.380	1.270	1.170	1.110	0.982	0.865	0.744
60	3.080	2.880	2.230	1.780	1.540	1.410	1.270	1.190	1.130	1.010	0.886	0.767
70	3.100	2.920	2.280	1.840	1.600	1.430	1.300	1.210	1.140	1.020	0.901	0.783
80	3.100	2.950	2.390	1.900	1.630	1.450	1.320	1.220	1.160	1.040	0.917	0.776
90	3.110	2.960	2.440	1.930	1.630	1.460	1.340	1.230	1.170	1.040	0.925	0.790
100	3.110	2.990	2.510	1.930	1.640	1.470	1.360	1.240	1.190	1.050	0.933	0.806
110	3.120	3.000	2.440	1.950	1.680	1.480	1.370	1.270	1.190	1.060	0.942	0.797
120	3.120	3.010	2.490	1.970	1.680	1.500	1.380	1.280	1.200	1.090	0.950	0.807
130	3.120	3.020	2.550	1.990	1.690	1.510	1.390	1.290	1.220	1.090	0.952	0.814
140	3.120	3.030	2.540	2.010	1.710	1.540	1.390	1.290	1.210	1.080	0.958	0.818
150	3.130	3.020	2.610	2.030	1.720	1.530	1.410	1.300	1.210	1.090	0.963	0.831
160	3.120	3.040	2.610	2.060	1.740	1.540	1.400	1.310	1.220	1.120	0.970	0.834
170	3.120	3.040	2.660	2.040	1.740	1.560	1.410	1.320	1.230	1.110	0.974	0.838
180	3.130	3.060	2.730	2.060	1.750	1.560	1.430	1.330	1.250	1.120	0.987	0.843
190	3.120	3.060	2.660	2.130	1.750	1.560	1.440	1.330	1.250	1.120	0.986	0.841
200	3.120	3.060	2.740	2.090	1.760	1.570	1.450	1.340	1.250	1.120	0.990	0.845

The Performance of the *RCDu* Statistic

For a Single Outlier

The performance of the *RCDu* statistic compared with existing statistics, *Mardia*, *M*, *A*, and *Chord*, are compared using Monte Carlo simulations. The study parameters were sample sizes $n = 20, 60, 100$, and 150 with six concentration parameters, $k = 2, 3, 5, 6, 8$, and 10 . The data are contaminated with a single outlier ($\mathcal{G}_c$) defined by

$$\mathcal{G}_c = \mathcal{G} + \lambda \pi \bmod(2\pi) \quad (13)$$

where λ is the degree of contamination, ($0 \leq \lambda \leq 1$).

If $\lambda = 0$, there is no contamination at position $[d]$. If $\lambda = 1$, the circular observation is located at the anti-mode of its initial location.

Replicate these processes 3000 times for all combinations of the sample size and concentration parameter with $\lambda = 0.8$. Figure 1 gives the proportions of outliers detected and the rates of masking and swamping for the 10% and 5% of cut-off points for the sample sizes $n = 60$.

It can be seen that, for small values of concentration parameters, the performance of all the methods is relatively low. This is because the circular data will be more spread around the circumference of the circle for low values of the concentration parameter. Therefore, it is very difficult to detect outliers in this case (Collett, 1980). The proportions of outliers detected for the *A* and *Chord* statistics are close to those for the proposed *RCDu* statistic and have the highest proportions of detection outliers. Consequently, the *RCDu*, *A*, and *Chord* statistics have the lowest rates of masking. There is not swamping for all combinations with 10% and 5% cut-off points because the rates of swamping are equal to 0. The results for $n = 20, 100$, and 150 were consistent with the results in Figure 1, so are not shown. Interested readers can request the corresponding author to provide more results.

For Multi-Outliers

In order to test performance of our statistic for different ratios of contamination, three ratios of contamination (5%, 10%, and 20%) were select, replicated 3000 times for the same combinations of sample sizes and concentration parameters with 10% and 5% of cut-off points of the *RCDu* statistic. The results of the triple measure of $n = 60$ are given in Figure 2. The performance of the *RCDu* statistic is relatively

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

low for small values of concentration parameter (for the same reason that we mentioned above). However, the $RCDu$ statistic successfully identifies outliers for $k \geq 5$ for different ratios of contamination. It has the highest proportion of detection of outliers and the lowest rate of masking. There is no swamping for all ratios of contamination for all combinations. The same results for $n = 20, 100$, and 150 were obtained (also available from the corresponding author).

Figure 1. The proportion of a single outlier detected, and rate of masking and swamping, for different statistics with 10% and 5% of cut-off points for $n = 60$

Figure 2. The proportion of outliers detected, and rates of masking and swamping, for 5%, 10%, and 20% of contamination with 10% and 5% of cut-off points for $n = 60$

Illustrative Examples

Example 1: A sample of 14 frogs was collected from the mud flats near Indianola, Mississippi. After 30 hours, the frogs were released and their directions were taken. Abuzaid et al. (2009) tested these circular data and detected that the observation numbered 14 is an outlier. The original circular data were tested; then outliers were inserted to bring the ratio of contamination to 20% in order to test the performance of the statistic with a high ratio of contamination. The cutoff point is

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

approximately 2.68 according to the results in Table 1 (with 10% cut off point and concentration parameter $\hat{k} = 2.18$). The proportion of outliers detected and rate of masking and swamping are tabulated in Table 3 for all methods.

The competing statistics failed to detect outliers with 20% of contamination. The $dist_{(i)}$ statistic is plotted in Figure 3 for the original data set and with contamination 20%, respectively. As noted in Figure 3a, observation 14 is classified as an outlier because $dist_{(14)} = 3.06$ exceeds the cut-off point. Also, in Figure 3b, note $dist_{(12)} = 2.87$, $dist_{(13)} = 2.73$, and $dist_{(14)} = 3.06$ exceed the cut-off point. Therefore, the observations numbered 12, 13, and 14 are classified as outliers.

Table 3. Comparison of the performance of the different statistics (frogs data)

		<i>Mardia</i>	<i>M</i>	<i>A</i>	<i>Chord</i>	<i>RCDu</i>
Original data set	Proportion of outlier	1	1	1	1	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 20%	Proportion of outlier	0	0	0	0	1
	Rate of masking	1	1	1	1	0
	Rate of swamping	0	0	0	0	0

Figure 3. $dist_{(i)}$ statistic for frogs direction data

Table 4. Comparison of the performance of the different statistics (paleocurrent orientations data)

		<i>Mardia</i>	<i>M</i>	<i>A</i>	<i>Chord</i>	<i>RCDu</i>
Original data set	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 5%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 10%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 20%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0

Example 2: Next, consider the data given by Jammalamadaka and SenGupta (2001, p. 238). The sample size (40) represent measurements of the first sample of paleocurrent orientations from three bedded sandstone layers, measured on the Belford Anticline, New South Wales. They detected that observation number 24 is an outlier. In order to test performance of the statistics with different ratios of contamination, we insert outliers to bring the ratio of contamination to 5%, 10%, and 20%. The 10% cut-off point with concentration parameter $\hat{k} = 2$ is 2.85. The proportion of detection of outliers and rate of masking and swamping are tabulated in Table 4.

The other statistics fail to detect outliers with all ratios of contamination. However, the *RCDu* statistic succeeds in identifying all of them without any swamping. The $dist_{(i)}$ is plotted in Figure 4 for the original data set and with 5%, 10%, and 20% contamination.

Note in Figure 4a that $dist_{(24)} = 2.854$ exceeds the cut-off point. Therefore, it is classified as an outlier. This identification coincides with the identification by Jammalamadaka and SenGupta (2001) that this data point is an outlier. In Figure 4b, note that $dist_{(24)} = 2.915$ and $dist_{(40)} = 3.037$ exceed the cut-off point. Therefore, the observations numbered 24 and 40 are classified as outliers. Also, as expected, in Figure 4c, $dist_{(1)} = 3.124$, $dist_{(17)} = 3.072$, $dist_{(24)} = 3.054$, and $dist_{(40)} = 2.932$ exceed the cut-off point. Therefore, they are classified as outliers. The results $dist_{(1)} = 3.089$, $dist_{(9)} = 3.037$, $dist_{(10)} = 2.967$, $dist_{(17)} = 3.142$, $dist_{(24)} = 2.985$, $dist_{(33)} = 3.142$, $dist_{(39)} = 3.124$, and $dist_{(40)} = 2.862$ are greater than the cut-off point, so they are detected as outliers. They are given in Figure 4d.

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

Figure 4. $dist_{(t)}$ statistic for paleocurrent orientations data

Example 3: In the final example, consider the wind direction data with sample size 129 that is considered by Abuzaid (2010, pp. 152-153). The data measurements were recorded over a period of 22.7 days along the Holderness coastline (the North Sea coast of Humberside in the United Kingdom). We insert outliers to bring the ratio of contamination to 5%, 10%, and 20%. The estimated concentration parameter is $\hat{k} = 7$, so the 10% cut-off point is equal to 1.39. The proportion of outliers detected and the rates of masking and swamping are tabulated in Table 5 for all the methods.

Figure 5. $dist_{(i)}$ statistic for circular residuals

To illustrate this, the $dist_{(i)}$ statistic is plotted in Figure 5 for the original data set and with 5%, 10%, and 20% of contamination. All the statistics perform the same for the original data set. However, the M and $Mardia$ statistics fail to detect outliers after contamination. Besides, the A and $Chord$ statistics cannot identify all of the outliers at the 10% and 20% contamination levels. In contrast, the $RCDu$ statistic has the greatest proportion of outliers detected without swamping for all ratios of contamination.

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

Table 5. Comparison of the performance of the different statistics (wind direction data)

		<i>Mardia</i>	<i>M</i>	<i>A</i>	<i>Chord</i>	<i>RCDu</i>
Original data set	Proportion of outlier	1.00	1.00	1.00	1.00	1.00
	Rate of masking	0.00	0.00	0.00	0.00	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 5%	Proportion of outlier	0.14	0.14	1.00	1.00	1.00
	Rate of masking	0.86	0.86	0.00	0.00	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 10%	Proportion of outlier	0.08	0.08	0.92	0.77	1.00
	Rate of masking	0.92	0.92	0.08	0.23	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 20%	Proportion of outlier	0.00	0.00	0.73	0.46	1.00
	Rate of masking	1.00	1.00	0.27	0.54	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00

Conclusion

The *RCDu* statistic was proposed to detect single and multi-outliers in univariate circular data. The proposed statistic is evaluated based on the proportion of outliers detected and the masking and swamping rates. The proposed statistic has the highest proportion of outliers detected and the lowest rates of masking and swamping. Moreover, the proposed *RCDu* statistic is able to detect outliers in data with a high level of contamination. Also, the proposed statistic is successful in detecting outliers in a large data set. Hence, we suggest that the *RCDu* statistic should be used to detect outliers in univariate circular data.

References

- Abuzaid, A. H. (2010). *Some problems of outliers in circular data* (Unpublished doctoral thesis). University of Malaya, Kuala Lumpur, Malaysia.
- Abuzaid, A. H., Hussin, A. G., Rambli, A., & Mohamed, I. (2012). Statistics for a new test of discordance in circular data. *Communications in Statistics – Simulation and Computation*, 41(10), 1882-1890. doi: 10.1080/03610918.2011.624239
- Abuzaid, A. H., Mohamed, I. B., & Hussin, A. G. (2009). A new test of discordancy in circular data. *Communications in Statistics – Simulation and Computation*, 38(4), 682-691. doi: 10.1080/03610910802627048
- Barnett, V. & Lewis, T. (1978). *Outliers in statistical data*. New York, NY: Wiley.

- Beckman, R. J., & Cook, R. D. (1983). Outlier.....s. *Technometrics*, 25(2), 119-149. doi: [10.2307/1268541](https://doi.org/10.2307/1268541)
- Collett, D. (1980). Outliers in circular data. *Applied Statistics*, 29(1), 50-57. doi: [10.2307/2346410](https://doi.org/10.2307/2346410)
- Ducharme, G. R., & Milasevic, P. (1987). Some asymptotic properties of the circular median. *Communications in Statistics – Theory and Methods*, 16(3), 659-664. doi: [10.1080/03610928708829394](https://doi.org/10.1080/03610928708829394)
- Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge, UK: Cambridge University Press. doi: [10.1017/cbo9780511564345](https://doi.org/10.1017/cbo9780511564345)
- Fisher, N. I., Lewis, T., & Willcox, M. E. (1981). Tests of discordancy for samples from Fisher's distribution on the sphere. *Applied Statistics*, 30(3), 230-237. doi: [10.2307/2346346](https://doi.org/10.2307/2346346)
- Hamelryck, T., Mardia, K., & Ferkinghoff-Borg, J. (2012). *Bayesian methods in structural bioinformatics*. Berlin, Germany: Springer-Verlag. doi: [10.1007/978-3-642-27225-7](https://doi.org/10.1007/978-3-642-27225-7)
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. (1986). Robust statistics: The approach based on influence functions. New York, NY: Wiley. doi: [10.1002/9781118186435](https://doi.org/10.1002/9781118186435)
- He, X., & Simpson, D. G. (1992). Robust direction estimation. *The Annals of Statistics*, 20(1), 351-369. doi: [10.1214/aos/1176348526](https://doi.org/10.1214/aos/1176348526)
- Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics*. Singapore: World Scientific Publishing. doi: [10.1142/4031](https://doi.org/10.1142/4031)
- Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3), 349-393. Available from <http://www.jstor.org/stable/2984782>
- Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester, UK: John Wiley & Sons Ltd. doi: [10.1002/9780470316979](https://doi.org/10.1002/9780470316979)
- Mohamed, I. B., Rambli, A., Khaliddin, N., & Ibrahim, A. I. N. (2015). A new discordancy test in circular data using spacings theory. *Communications in Statistics – Simulation and Computation*, 45(8), 2904-2916. doi: [10.1080/03610918.2014.932799](https://doi.org/10.1080/03610918.2014.932799)
- Otieno, B. S., & Anderson-Cook, C. M. (2003). *Hodges-Lehmann estimator of preferred direction for circular* (Technical Report 03-3). Blacksburg, VA: Virginia Tech Department of Statistics.
- Otieno, B. S., & Anderson-Cook, C. M. (2005). Effect of position of an outlier on the influence curve of the measures of preferred direction for circular

DETECTION OF OUTLIERS IN UNIVARIATE CIRCULAR DATA

data. *Journal of Modern Applied Statistical Methods*, 4(1), 81-89. doi:
[10.22237/jmasm/1114906140](https://doi.org/10.22237/jmasm/1114906140)

Rao, J. S. (1969). *Some contributions to the analysis of circular data*
(Unpublished doctoral thesis). Calcutta, India: Indian Statistical Institute.

Around Gamma Lindley Distribution

Hamouda Messaadia
Badji-Mokhtar University
Annaba, Algeria

Halim Zeghdoudi
Badji-Mokhtar University
Annaba, Algeria

Some remarks and correction on a new distribution, Gamma Lindley, of which the Lindley distribution is a particular case, are given pertaining to its parameter space.

Keywords: Lindley distribution, Gamma Lindley distribution

Introduction

A new distribution was proposed by Nedjar and Zeghdoudi (2016a, 2016b) called the Gamma Lindley distribution (GaL). It is a mixture of a gamma $(2, \theta)$ and one-parameter Lindley distribution. There are numerous examples of using a mixture of two known distributions is to generate a new distribution. For example, Shanker, Sharma, and Shanker (2013) used a mixture of exponential (θ) and gamma $(2, \theta)$ to create a two-parameter Lindley distribution. Zakerzadeh and Dolati (2010) used a gamma (α, θ) and gamma $(\alpha + 1, \theta)$ to create a generalized Lindley distribution. Zeghdoudi and Nedjar (2016, 2017) introduced another new distributions, called the pseudo-Lindley distribution, which is based on mixtures of gamma $(2, \theta)$ and exponential (θ) distributions. Here, a mixture of a gamma $(2, \theta)$ and a one-parameter Lindley distribution is used to generate a Gamma Lindley distribution, which is useful in modeling lifetime data and survival analysis and actuarial science.

Nedjar and Zeghdoudi (2016a, 2016b) developed various properties of the distribution, such as the probability density function (pdf), cumulative distribution function (cdf), survival and hazard rate function, moment generating function (mgf), mean, variance, and quantile functions, Lorenz curve, and some results on stochastic orderings. Plots of the pdf and cdf for some parameter values were also given, along with maximum likelihood estimates and moment estimates. However, there the parameter space was incorrect. The density function of the random variable X was given by:

Hamouda Messaadia is a graduate student. Email them at: ha.messadia@gmail.com. Dr. Halim Zeghdoudi is an Associate Professor of Statistics in the Department of Mathematics. Email them at: halim.zeghdoudi@univ-annaba.dz.

AROUND GAMMA LINDLEY DISTRIBUTION

$$f_{\text{GaL}}(x; \theta, \beta) = \theta^2 ((\beta + \beta\theta - \theta)x + 1) \frac{e^{-\theta x}}{\beta(1 + \theta)}$$

Unfortunately, $f_{\text{GaL}}(x; \theta, \beta)$ is not a proper pdf, because each of them can be negative for some values of the parameters $\theta > 0$, $\beta > 0$. For example, see [Figures 1 and 2](#).

To obtain a proper pdf for $f_{\text{GaL}}(x; \theta, \beta)$, modify the parameter space to be $\theta > 0$, $\beta > (\theta / (1 + \theta))$. Now it can be shown that the proper $f_{\text{GaL}}(x; \theta, \beta)$, where $\theta > 0$ and $\beta > (\theta / (1 + \theta))$ are, in fact, general cases of a two-parameter Lindley distribution with pdf

$$f_{\text{TPL}}(x; \theta, \alpha) = \theta^2 (\alpha x + 1) \frac{e^{-\theta x}}{\alpha + \theta}, \quad x, \theta, \alpha > 0$$

Taking $\alpha = \beta + \beta\theta - \theta$ in the pdf of the Gamma Lindley distribution leads to the pdf of the two-parameter Lindley distribution.

Figure 1. $(\theta, \beta) = (4, 0.5)$ with $\beta = 0.5 < (\theta / (1 + \theta)) = (4 / 5)$

Figure 2. $(\theta, \beta) = (1, 0.1)$ with $\beta = 0.1 < (\theta / (1 + \theta)) = (1 / 2)$

Conclusion

This new distribution might attract wider sets of applications in actuarial science, finance, medicine, and engineering. The reliability behavior of the Gamma Lindley distribution allows an improved performance for lifetime data modeling, and the hazard rate function can have various shapes, so this approach is more realistic and provides a greater degree of flexibility. Also, a new version of a compound Poisson distribution named the Poisson gamma Lindley (PGaL) distribution may be obtained by compounding the Poisson and Gamma Lindley distributions, which is applicable to the collective risk model by considering the proposed distribution as primary distribution and exponential and Erlang as secondary distributions.

Acknowledgements

The authors thank the Editor, Pr. Shlomo Sawilowsky, and referees for their useful suggestions which lead to the improvement of the paper.

References

Nedjar, S., & Zeghdoudi, H. (2016a). Gamma Lindley distribution and its application. *Journal of Applied Probability and Statistics*, 11(1), 129-138.

AROUND GAMMA LINDLEY DISTRIBUTION

- Nedjar, S., & Zeghdoudi, H. (2016b). On gamma Lindley distribution: Properties and simulations. *Journal of Computational and Applied Mathematics*, 298, 167-174. doi: [10.1016/j.cam.2015.11.047](https://doi.org/10.1016/j.cam.2015.11.047)
- Shanker, R., Sharma, S., & Shanker, R. (2013). A two-parameter Lindley distribution for modeling waiting and survival times data. *Applied Mathematics*, 4(2), 363-368. doi: [10.4236/am.2013.42056](https://doi.org/10.4236/am.2013.42056)
- Zakerzadeh, H., & Dolati, A. (2009). Generalized Lindley distribution. *Journal of Mathematical Extension*, 3(2), 1-17. Retrieved from <http://www.ijmex.com/index.php/ijmex/article/view/45>
- Zeghdoudi, H., & Nedjar, S. (2016). A pseudo Lindley distribution and its application. *Afrika Statistika*, 11(1), 923-932. doi: [10.16929/as/2016.923.83](https://doi.org/10.16929/as/2016.923.83)
- Zeghdoudi, H., & Nedjar, S. (2017). On Pseudo Lindley distribution: Properties and applications. *New Trends in Mathematical Sciences*, 5(1), 59-65. doi: [10.20852/ntmsci.2017.126](https://doi.org/10.20852/ntmsci.2017.126)

A Double EWMA Control Chart for the Individuals Based on a Linear Prediction

Rafael Perez Abreu
University of Northern Colorado
Greeley, CO

Jay R. Schaffer
University of Northern Colorado
Greeley, CO

Industrial process use single and double Exponential Weighted Moving Average control charts to detect small shifts in it. Occasionally there is a need to detect small trends instead of shifts, but the effectiveness to detect small trends. A new control chart is proposed to detect a small drift.

Keywords: EWMA, DEWMA, control charts, linear drift, forecast, average run length

Introduction

One of the most useful tools to assure the quality of a product or process in manufacture industry are quality control charts. Shewhart (1926) developed the control charts tool to identify when a process was producing a good or a defective product. Today, many control charts have developed to ensure quality through the control of certain characteristics of interest. The main idea of control charts is to detect as soon as possible when this characteristic has changed. In this sense, control charts are designed to detect a shift quickly. Several control charts have been designed to detect small shifts, while others were designed to detect big shifts. In practice, however, we occasionally wish to detect small trends, instead of shifts, in the process; this gradual changes may be due to tool wear or similar causes. Examples of this phenomenon are commonly observed in several manufacturing processes and administrative activities. The effectiveness of these methods to determine small trends in a process has not been thoroughly researched in the current literature. Knoth (2012) reviewed this literature and invited the statistical process community to extend research in this area in order to enhance the knowledge about drift detection.

Rafael Perez Abreu is a doctoral student in the Applied Statistics and Research Methods Department. Email him at: pere4435@bears.unco.edu.

A DEWMA BASED ON LINEAR PREDICTION

A double exponentially weighted moving average (DEWMA) control chart was initially developed by Shamma and Shamma (1992). Zhang and Chen (2005) presented an extension of the exponentially weighted moving average (EWMA) technique to a DEWMA technique. These two DEWMA techniques are the same, and so are their conclusions. Research regarding DEWMA was developed, taking the DEWMA control chart as a reference. For example, Mahmoud and Woodall (2010) conducted a study to compare some characteristics between the EWMA and the DEWMA. Alkahtani (2013) assessed the robustness of DEWMA and EWMA control charts for abnormal processes. Extensions for a multivariate DEWMA control chart case exist. For example, Alkahtani and Schaffer (2012) developed a multivariate DEWMA control chart for detecting shifts in the mean vector of a multivariate normal quality characteristic distribution.

The DEWMA control chart is constructed on the assumption of a data stream of X_i random values following a normal distribution, initially $X_i \sim N(\mu_0, \sigma)$; the DEWMA $S'_i = \lambda S_i + (1 - \lambda)S'_{i-1}$ is then calculated, where $S_i = \lambda X_i + (1 - \lambda)S_{i-1}$ with its corresponding initial values. The control chart is built by plotting the value S'_i with its limits, using k times the variance of S'_i . The DEWMA value is plotted with the upper and lower limits versus i . The DEWMA and EWMA control charts work efficiently to detect small shifts when the mean of the process has changed slightly, and the classical Shewhart control chart works well to detect big shifts (more than twice the standard deviation of the process). What happens, however, if after a period of stability, the process has a permanent small change (a linear trend or drift in the stream of X_i)?

Brown (1962) proposed using the smoothing technique to forecast the demand of goods. These forecast methods are the basis of other more complex forecast methods. Smoothing techniques to produce forecasting are well known in business, where it is essential to predict the demand for goods and services (Hanke & Wichern, 2009). According to the fundamental theorem of exponential smoothing explained by Brown and Meyer (1961), a linear prediction can be forecast by a DEWMA with a linear relationship by the follow equation:

$$F_{i+t} = a_i + b_i t \quad (1)$$

where F_{i+t} is the forecast in the t period,

$$a_i = 2S_i - S'_i \quad (2)$$

and

$$b_i = \frac{\lambda}{1-\lambda}(S_i - S'_i) \quad (3)$$

Details of the development of these equations can be seen in Brown (1962) and Yates (1968). Also, a similar linear prediction equation like the one shown in equation (1) can be built using Holt forecast equations or moving average equations, or other forecast techniques as explained in Hanke and Wichern (2009).

The main idea is to build a three individual control charts under the null hypothesis of the in-control process, assuming the $X_i \sim N(\mu_0, \sigma^2)$. The first control chart is the a_i control chart, where the center line is $E(a_i)$ and the upper and lower limits are given by

$$E(a_i) \pm k\sqrt{\text{Var}(a_i)}$$

This control chart tests the null hypothesis that the forecast level is equal to μ_0 , (i.e. $E(a_i) = \mu_0$) at time t .

The second control chart is for b_i , a control chart for the forecast slope, where the center line of the control chart is $E(b_i) = 0$ and the upper and lower limits can be built as

$$E(b_i) \pm k\sqrt{\text{Var}(b_i)}$$

This control chart tests the null hypothesis $b_i = 0$ at time t (i.e. the forecast linear trend is equal to zero) versus the alternative hypothesis $b_i \neq 0$ (i.e. the slope differs from zero). The main idea is to detect a change when the slope differs from zero.

The third control chart is the sum $F_{i+t} = a_i + b_it$, a control chart for the forecast assuming a linear prediction, where the center line is $E(F_{i+t}) = E(a_i + b_it) = E(a_i) + tE(b_i)$ and the upper and lower limits can be built as

$$E(F_t) \pm k\sqrt{\text{Var}(F_t)}$$

This control chart tests the null hypothesis $F_t = \mu_0$ at time t (i.e. the forecast linear trend is equal to μ_0 , the target value) versus the alternative hypothesis $F_t \neq \mu_0$ (i.e.

the mean level differs from the target). The principal idea is to try and detect a linear drift as soon as it occurs.

The Control Charts to Detect Small Shifts

The EWMA control chart was introduced by Roberts (1959). According to Lucas and Saccucci (1990), the EWMA control chart was a good alternative to the Shewhart control chart when the interest is in detecting small shifts. The EWMA is generally used with individual observations; therefore, this control chart will be discussed when $n = 1$.

The Exponentially Weighted Moving Average Control Chart

For a $X_i \sim N(\mu_0, \sigma)$, $i = 1, 2, \dots, n$, the EWMA control statistic S_i is explained by Montgomery (2007) as:

$$S_i = \lambda X_i + (1 - \lambda) S_{i-1} \quad (4)$$

where $0 < \lambda < 1$ and $S_0 = \mu_0$. It can be shown

$$E(S_i) = \mu_0 \quad (5)$$

and

$$\text{Var}(S_i) = \frac{\lambda}{2 - \lambda} \left[1 - (1 - \lambda)^{2i} \right] \sigma^2 \quad (6)$$

For large values of i , the asymptotic variance becomes

$$\text{Var}_{\text{asym}}(S_i) = \left(\frac{\lambda}{2 - \lambda} \right) \sigma^2 \quad (7)$$

Therefore, the control limits, and center line become

$$\begin{aligned} \text{UCL} &= \mu_0 + k\sigma\sqrt{\frac{\lambda}{2-\lambda}} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - k\sigma\sqrt{\frac{\lambda}{2-\lambda}} \end{aligned}$$

The Double Exponentially Weighted Moving Average Control Chart

For a $X_i \sim N(\mu_0, \sigma)$, $i = 1, 2, \dots, n$, the DEWMA control statistic S'_i was first developed by Shamma and Shamma (1992). It is defined as

$$S'_i = \lambda S_i + (1 - \lambda) S'_{i-1} \quad (8)$$

$$S_i = \lambda X_i + (1 - \lambda) S_{i-1} \quad (9)$$

where $0 < \lambda < 1$ and $S_0 = S'_0 = \mu_0$. It can be shown that

$$E(S'_i) = \mu_0 \quad (10)$$

and

$$\begin{aligned} &\text{Var}(S'_i) \\ &= \lambda^4 \frac{1 + (1 - \lambda)^2 - (1 - \lambda)^{2i} \left((i + 1)^2 - (2i^2 + 2i - 1)(1 - \lambda)^2 + i^2 (1 - \lambda)^4 \right)}{(1 - (1 - \lambda)^2)^3} \sigma^2 \quad (11) \end{aligned}$$

For large values of i , the asymptotic variance becomes

$$\text{Var}_{\text{asym}}(S'_i) = \frac{\lambda(2 - 2\lambda + \lambda^2)}{(2 - \lambda)^3} \sigma^2 \quad (12)$$

Then for large values of i the control limits become

$$\begin{aligned} \text{UCL} &= \mu_0 + k\sigma \sqrt{\frac{\lambda(2-2\lambda+\lambda^2)}{(2-\lambda)^3}} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - k\sigma \sqrt{\frac{\lambda(2-2\lambda+\lambda^2)}{(2-\lambda)^3}} \end{aligned}$$

Mahmoud and Woodall (2010) show how these variances can be obtained.

A Proposed Control Chart to Detect Small Change in the Trends

Double Exponentially Weighted Moving Average Based on a Linear Prediction

A new control charts is now proposed to detect linear trends. The double exponentially weighted moving average based on a linear prediction (DEWMABLP) is constructed assuming a stream of variables $X_i \sim N(\mu_0, \sigma)$, then the DEWMA is $S'_i = \lambda S_i + (1-\lambda)S'_{i-1}$, where $S_i = \lambda X_i + (1-\lambda)S_{i-1}$ and the smooth linear forecast is

$$F_{i+t} = a_i + b_i t \quad (13)$$

where F_{i+t} is the forecast in the t period ahead,

$$a_i = 2S_i - S'_i$$

and

$$b_i = \frac{\lambda}{1-\lambda}(S_i - S'_i)$$

F_t is called the statistic of the DEWMABLP. It is possible to create three control charts: first, a control chart for the intercept a_i that will be similar than the EWMA control chart; second, a control chart for the slope b_i that is used to test if there is a linear drift or trend; and third, a control chart for a linear prediction t

periods ahead of i , F_t , that can be used to test if the statistic one period forecast ahead is or not in statistical control.

An Intercept Prediction DEWMA Control Chart (a_t)

The center line for an a_i control chart is the expected value of a_i . It is

$$\begin{aligned} E(a_i) &= E(2S_i - S'_i) \\ &= E(2S_i) - E(S'_i) \\ &= 2E(S_i) - E(S'_i) \\ &= 2\mu_0 - \mu_0 \\ &= \mu_0 \end{aligned} \tag{14}$$

This can be verified using equations (5) and (10). Using equations (6) and (11), the variance of a_i can be obtained as

$$\text{Var}_{\text{asym}}(a_i) = \text{Var}(2S_i - S'_i) \tag{15}$$

Brown (1962) showed the asymptotic variance for a predict value of a_i is

$$\text{Var}_{\text{asym}}(a_i) = \frac{\lambda(1 + 4(1 - \lambda) + 5(1 - \lambda)^2)}{(1 + (1 - \lambda))^2} \sigma^2 \tag{16}$$

For large values of i the control limits for the a_i control chart become

$$\begin{aligned} \text{UCL} &= \mu_0 + k\sigma \sqrt{\frac{\lambda(1 + 4(1 - \lambda) + 5(1 - \lambda)^2)}{(1 + (1 - \lambda))^2} \sigma^2} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - k\sigma \sqrt{\frac{\lambda(1 + 4(1 - \lambda) + 5(1 - \lambda)^2)}{(1 + (1 - \lambda))^2} \sigma^2} \end{aligned}$$

A Slope Prediction DEWMA Control Chart (b_i)

In a similar manner, the center line for b_i is the expected value of b_i . Using the equations (5) and (10) it can be shown that

$$\begin{aligned} E(b_i) &= E\left(\frac{\lambda}{1-\lambda}(S_i - S'_i)\right) \\ &= \frac{\lambda}{1-\lambda}(E(S_i) - E(S'_i)) \\ &= \frac{\lambda}{1-\lambda}(\mu_0 - \mu_0) \\ &= 0 \end{aligned} \tag{17}$$

The variance of b_i is defined as:

$$\begin{aligned} \text{Var}(b_i) &= \text{Var}\left(\frac{\lambda}{1-\lambda}(S_i - S'_i)\right) \\ &= \left(\frac{\lambda}{1-\lambda}\right)^2 \text{Var}(S_i - S'_i) \end{aligned} \tag{18}$$

Brown (1962) gave the asymptotic variance of b_i for large values of i as

$$\text{Var}_{\text{asym}}(b_i) = \sigma^2 \frac{2\lambda^3}{(1+(1-\lambda))^2} \tag{19}$$

Then, for large values of i the control limits for the b_i chart become

$$\begin{aligned} \text{UCL} &= k\sigma \sqrt{\frac{2\lambda^3}{(1+(1-\lambda))^3}} \\ \text{CL} &= 0 \\ \text{LCL} &= -k\sigma \sqrt{\frac{2\lambda^3}{(1+(1-\lambda))^3}} \end{aligned}$$

A Linear Trend Prediction Double EWMA Control Chart (F_t)

Using equations (14) and (17), it can be shown that the expected value of F_t is

$$\begin{aligned} E(F_{i+t}) &= E(a_i + b_i t) \\ &= E(a_i) + t E(b_i) \\ &= \mu_0 + 0 \\ &= \mu_0 \end{aligned}$$

The variance of F_t is

$$\begin{aligned} \text{Var}(F_{i+t}) &= \text{Var}(a_i + b_i t) \\ &= \text{Var}(a_i) + \text{Var}(b_i t) + 2\text{Cov}(a_i, b_i t) \end{aligned} \quad (20)$$

The covariance term $2\text{Cov}(a_i, b_i t)$ in the previous equation was investigated via simulation to verify the possible independence between a_i and b_i . Simulation for the covariance between a_i and b_i were performed for several values of the smooth parameter λ , considering a process under the in-control null hypothesis. The simulation yielded values very close to zero for the $\text{Cov}(a_i, b_i)$. These results suggest that the covariance $\text{Cov}(a_i, b_i)$ can be considered negligible. Nevertheless, For $t = 1$, Brown (1962) gives the asymptotic covariance of $\text{Cov}(a_i, b_i)$:

$$\text{Cov}(a_i, b_i) = \frac{\lambda^2 (1 + 3(1 - \lambda))}{(1 + (1 - \lambda))^3} \sigma^2 \quad (21)$$

Substituting equations (16), (19), and (21) in equation (20) it is possible to obtain the asymptotic variance of the F_t as

$$\begin{aligned} \text{Var}_{\text{asym}}(F_t) &= \sigma^2 \left(\frac{\lambda (1 + 4(1 - \lambda) + 5(1 - \lambda)^2)}{(1 + (1 - \lambda))^2} + \frac{2\lambda^3}{(1 + (1 - \lambda))^3} + \frac{\lambda^2 (1 + 3(1 - \lambda))}{(1 + (1 - \lambda))^3} \right) \end{aligned} \quad (22)$$

Then, for large values of i the control limits and the center line for the F_t control chart become

UCL

$$= \mu_0 + k\sigma \sqrt{\frac{\lambda(1+4(1-\lambda)+5(1-\lambda)^2)}{(1+(1-\lambda))^2} + \frac{2\lambda^3}{(1+(1-\lambda))^3} + \frac{\lambda^2(1+3(1-\lambda))}{(1+(1-\lambda))^3}}$$

CL = μ_0

LCL

$$= \mu_0 - k\sigma \sqrt{\frac{\lambda(1+4(1-\lambda)+5(1-\lambda)^2)}{(1+(1-\lambda))^2} + \frac{2\lambda^3}{(1+(1-\lambda))^3} + \frac{\lambda^2(1+3(1-\lambda))}{(1+(1-\lambda))^3}}$$

(23)

Design of a Double EWMA Base on Linear Prediction Control Chart F_t

The design parameters for this chart are constructed with k times the multiple of sigma, the standard deviation used in the control limits, and the value of λ , the smooth parameter. It is possible to choose these parameters to give a mean performance of average run length (ARL) under the null hypothesis (H_0), i.e. ARL_0 , for a certain number. For example, an $ARL_0 = 370$ is the equivalent of an ARL of a Shewhart control chart under H_0 for 3σ as its control limits; the DEWMABLP can be designed with $k = 2.16$ and $\lambda = 0.10$ to obtain an $ARL_0 = 373 \approx 370$.

Assessing the Performance of DEWMABLP

In order to assess the performance of this new chart, its performance was compared with the performance of the EWMA, DEWMA, and Shewhart control charts. This comparison was made using the average run length under out-of-control (ARL_1).

Design Parameters for EWMA, DEWMA, and DEWMABLP Control Charts

A Monte Carlo simulation with 10,000 replications with an ARL_0 in-control was fixed approximately to 370 for all control charts under study: EWMA, DEWMA, DEWMABLP, and Shewhart control charts. In order to be fair, all charts were set to an $ARL_0 \approx 370$. Table 1 shows the parameters for λ and k for EWMA, DEWMA, and DEWMABLP control charts to give an $ARL_0 \approx 370$. Also, the standard deviation of ARL_0 is displayed.

Table 1. Average run length under H_0 for several λ and k

Control Chart	λ	k	ALR ₀	sd(ALR ₀)
EWMA	0.01	1.980	372.5	450.1
EWMA	0.05	2.511	371.8	372.1
EWMA	0.10	2.710	369.6	367.2
EWMA	0.15	2.800	371.1	366.0
EWMA	0.20	2.862	370.0	367.4
DEWMA	0.01	1.300	374.6	533.8
DEWMA	0.05	1.920	373.3	382.6
DEWMA	0.10	2.220	368.4	370.5
DEWMA	0.15	2.408	373.6	369.5
DEWMA	0.20	2.530	374.9	380.7
$F_t = a_t + b_t t$	0.01	1.725	375.2	557.0
$F_t = a_t + b_t t$	0.05	2.035	377.5	399.4
$F_t = a_t + b_t t$	0.10	2.160	373.3	386.0
$F_t = a_t + b_t t$	0.15	2.240	379.1	385.4
$F_t = a_t + b_t t$	0.20	2.287	374.5	373.7

Simulations were conducted to compare the performance of the EWMA, DEWMA, DEWMALP, and Shewhart control charts. The ARL under linear drift (ARL₁) for several slopes, that is the out-of-control, were compared between all this control charts. The control chart with lowest ARL₁ is considered the best chart.

The simulation was written in R. A stream of X_t for $t = 1, 2, \dots, 100$ observations were created such that X_t are independent and identically distributed as $X_t \sim N(\mu_0 = 10, \sigma = 1)$, and then another stream X_t was created such that of X_t for $t = 101, \dots, 200$ observations where $X_t \sim N(\mu_0 = 10 + \beta t \sigma, \sigma = 1)$. This procedure was repeated several times using values of 0, 0.025, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.75, and 1 for the slope β .

Results

A summary of the simulation is presented in Tables 2 to 5. Table 2 shows the ARL₁ for these several slopes and $\lambda = 0.20$. In the same manner, Tables 3, 4, and 5 show the ARL₁ for these several slopes and $\lambda = 0.10, 0.05$, and 0.01 , respectively.

When $\lambda = 0.2$, it can be observed in Table 2 that the ARL₁ of the F_t control chart is less than the ARL₁ of the other control charts only when the slope $\beta > 0.20$. For $\lambda = 0.10$ in Table 3, the ARL₁ of the F_t control chart is less than the ARL₁ of the other control charts when the slope $\beta \geq 0.10$. In similar way, for $\lambda = 0.05$ in Table 4, the ARL₁ of the F_t control chart is less than the ARL₁ of the other control charts for slope values $\beta \geq 0.05$. Finally, for $\lambda = 0.01$ in Table 5, the ARL₁ of the

A DEWMA BASED ON LINEAR PREDICTION

F_t control chart is less than the ARL_1 of the other control charts for slope values between $0.025 \leq \beta < 0.200$. For values $\beta \geq 0.200$, the ARL_1 of Shewhart chart has the best performance.

Table 2. Average run length under different slopes, $\lambda = 0.20$

k	3.000	2.862	2.530	2.287
Slope	Shewhart	EWMA	DEWMA	F_t
0.000	370.00	370.00	374.85	374.50
0.025	50.12	30.80	29.41	36.40
0.050	30.64	19.64	19.36	21.97
0.100	18.41	12.59	13.14	13.42
0.150	13.59	9.79	10.62	10.16
0.200	11.00	8.27	9.25	8.41
0.300	8.11	6.49	7.60	6.45
0.400	6.51	5.46	6.63	5.37
0.500	5.52	4.84	6.02	4.71
0.750	4.05	3.84	5.04	3.67
1.000	3.27	3.27	4.43	3.11

Table 3. Average run length under different slopes, $\lambda = 0.10$

k	3.000	2.710	2.220	2.160
Slope	Shewhart	EWMA	DEWMA	F_t
0.000	370.0	369.6	368.4	373.3
0.025	50.15	28.99	29.83	30.78
0.050	30.68	19.12	20.98	19.52
0.100	18.52	12.85	15.27	12.70
0.150	13.67	10.16	12.74	9.91
0.200	11.01	8.62	11.21	8.33
0.300	8.14	6.92	9.47	6.62
0.400	6.56	5.90	8.39	5.62
0.500	5.52	5.21	7.70	4.94
0.750	4.04	4.22	6.57	3.96
1.000	3.29	3.63	5.83	3.40

Table 4. Average run length under different slopes, $\lambda = 0.05$

k	3.000	2.511	1.920	2.035
Slope	Shewhart	EWMA	DEWMA	F_t
0.000	370.00	371.80	373.33	377.51
0.025	50.17	29.19	24.78	29.39
0.050	30.60	19.70	18.69	19.36
0.100	18.55	13.62	18.69	13.18
0.150	13.65	10.83	15.74	10.44
0.200	11.07	9.36	14.23	8.93
0.300	8.13	7.51	12.09	7.14
0.400	6.52	6.46	10.85	6.11
0.500	5.49	5.75	9.93	5.43
0.750	4.06	4.67	8.54	4.40
1.000	3.27	4.02	7.63	3.78

Table 5. Average run length under different slopes, $\lambda = 0.01$

k	3.000	1.980	1.300	1.735
Slope	Shewhart	EWMA	DEWMA	F_t
0.000	370.0	372.5	374.6	375.2
0.025	49.94	33.41	56.08	33.14
0.050	30.79	23.45	43.63	23.09
0.100	18.49	16.56	34.15	16.20
0.150	13.72	13.50	29.56	13.18
0.200	11.09	11.65	26.70	11.37
0.300	8.12	9.45	23.08	9.21
0.400	6.54	8.14	20.76	7.92
0.500	5.51	7.28	19.18	7.09
0.750	4.07	5.96	16.71	5.78
1.000	3.29	5.16	15.11	5.02

Conclusions

The new DEWMABLP control chart works better than the other control charts to detect linear trends in the cases where a small linear trend is present. It works better when slopes are between 0.05 and 0.75 times the standard deviation. The EWMA

control chart also performed well, but with an ARL_1 slightly higher than DEWMABLP. The DEWMA works better for small shifts, but works poorly for linear drifts. Also, it is observed that the performance of the DEWMABLP overcomes the performance of the Shewhart, EWMA, and DEWMA control charts when a linear drift is present and the slope of this linear drift is greater than the parameter λ of the DEWMABLP, EWMA, and DEWMA. It can be concluded that the new DEWMABLP control chart can be used as an alternative when it is suspected that a linear drift can occur in the process after a period of stability. Of course, the DEWMABLP is not designed to detect a shift in the process; therefore, if a shift and a drift are expected at the same time, it should be used in combination with other control charts. This is a similar practice as utilizing both the Shewhart and EWMA control charts with the intention to detect both small and big shifts.

References

- Alkahtani, S. S. (2013). Robustness of DEWMA versus EWMA control charts to non-normal processes. *Journal of Modern Applied Statistical Methods*, 12(1), 148-163. doi: [10.22237/jmasm/1367381820](https://doi.org/10.22237/jmasm/1367381820)
- Alkahtani, S., & Schaffer, J. (2012). A double multivariate exponentially weighted moving average (dMEWMA) control chart for a process location monitoring. *Communications in Statistics – Simulation and Computation*, 41(2), 238-252. doi: [10.1080/03610918.2011.585004](https://doi.org/10.1080/03610918.2011.585004)
- Brown, R. G. (1962). *Smoothing, forecasting and prediction of discrete time series*. New York, NY: Prentice-Hall.
- Brown, R. G., & Meyer, R. F. (1961). The fundamental theorem of exponential smoothing. *Operations Research*, 9(5), 673-685. doi: [10.1287/opre.9.5.673](https://doi.org/10.1287/opre.9.5.673)
- Hanke, J. E., & Wichern, D. (2009). *Business forecasting*. Essex, UK: Pearson.
- Knoth, S. (2012). More on control charting under drift. In: H.-J. Lenz, W. Schmid, & P.-T. Wilrich (Eds.), *Frontiers in statistical quality control* (Vol. 10) (pp. 53-68), Heidelberg, Germany: Physica Verlag. doi: [10.1007/978-3-7908-2846-7_4](https://doi.org/10.1007/978-3-7908-2846-7_4)

Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, 32(1), 1-29. doi: [10.2307/1269835](https://doi.org/10.2307/1269835)

Mahmoud, M. A., & Woodall, W. H. (2010). An evaluation of the double exponentially weighted moving average control chart. *Communications in Statistics – Simulation and Computation*, 39(5), 933-949. doi: [10.1080/03610911003663907](https://doi.org/10.1080/03610911003663907)

Montgomery, D. C. (2007). *Introduction to statistical quality control*. Hoboken, NJ: John Wiley & Sons.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3), 239-250. doi: [10.2307/1266443](https://doi.org/10.2307/1266443)

Shamma, S. E., & Shamma, A. K. (1992). Development and evaluation of control charts using double exponentially weighted moving averages. *International Journal of Quality & Reliability Management*, 9(6), 18. doi: [10.1108/02656719210018570](https://doi.org/10.1108/02656719210018570)

Shewhart, W. A. (1926). Quality control charts. *The Bell System Technical Journal*, 5(4), 593-603. doi: [10.1002/j.1538-7305.1926.tb00125.x](https://doi.org/10.1002/j.1538-7305.1926.tb00125.x)

Yates, B. D. (1968). *A review of some basic exponential smoothing models*. (Master's thesis). Laramie, WY: University of Wyoming.

Zhang, L., & Chen, G. (2005). An extended EWMA mean chart. *Quality Technology & Quantitative Management*, 2(1), 39-52. doi: [10.1080/16843703.2005.11673088](https://doi.org/10.1080/16843703.2005.11673088)

The Impact of Inappropriate Modeling of Cross-Classified Data Structures on Random-Slope Models

Feifei Ye

University of Pittsburgh
Pittsburgh, PA

Laura Daniel

Independent Consultant
Pittsburgh, PA

Previous studies that explored the impact of misspecification of cross-classified data structure as strictly hierarchical are limited to random intercept models. This study examined the effects of misspecification of a two-level, cross-classified, random effect model (CCREM) where both the level-1 intercept and slope were allowed to vary randomly. Results suggest that ignoring one of the crossed factors produced considerably underestimated standard errors for: 1) the regression coefficients of the level-1 predictor; 2) the inappropriately modeled predictor associated with the misspecified crossed factor; and 3) and their interaction. This misspecification also resulted in a significant inflation of the level-1 residual variances and the intercept and slope variance components across the levels of the remaining crossed factor in hierarchical linear model.

Keywords: Hierarchical linear model, cross-classified random effect model, Monte Carlo study

Introduction

Multilevel datasets in educational and social studies may have cross-classified, not purely nested, higher level sampling units, which prevents the use of hierarchical linear model. Examples of such data structure include students cross-classified by school and neighborhood (Raudenbush & Bryk, 2002), such as students in the same school coming from different neighborhoods and students in the same neighborhood going to different schools. Cross-classification also occurs in a longitudinal study when students have different math teachers at different grade levels, such as student math achievement cross-classified by the student and the teacher at a certain grade level.

Feifei Ye is an Assistant Professor of Research Methodology in the Department of Psychology in Education. Email her at: feifeiye.s@gmail.com.

The cross-classified random effects model (CCREM) is used to properly account for the cross-classified data structure. For example, the cross-classification of students in schools and neighborhoods can be modeled by a two-level CCREM, with both schools and neighborhoods considered level-2 units. In the pure hierarchical model in which neighborhoods are nested in schools, a three-level HLM is needed to separate the school (at level-3) and neighborhood (at level-2) effects, while assuming students in the same neighborhood all go to the same school.

With the recent development in computer software that can perform CCREM, a growing number of applied researchers have called upon the method to correctly model cross-classified data structures. A search of recent education research (1994-2014) using ERIC via EBSCO, using the keyword “cross-classified” indicated that 36 peer-reviewed empirical studies have used CCREM. However, given the prevalence of cross-classified data structures in educational research, only 36 studies in the last 20 years seems to be a small number.

In addition to infrequent use of the CCREM, there is also evidence that its use is, sometimes, consciously avoided by social scientists. In a study of neighborhood effects on educational achievement, Ainsworth (2002) removed subjects from the analysis that moved into new neighborhoods during the course of the study. Ma and Wilkins (2002) studied students’ science achievement growth between the 7th and 12th grades using HLM to control for middle school clustering but ignoring high school clustering. More recently, Witherspoon and Ennett (2011) examined rural youths’ developmental trajectories of self-reported grades, affective outcomes, and behavioral educational outcomes from 6th to 12th grade. The authors did not consider the nesting of adolescents within schools for two reasons, the reported complication of students changing schools over the course of the study and the small number of 9 schools. However, they did not use the school membership as a predictor in the model to account for the clustering of students in a school.

There are several suspected reasons that researchers may be hesitant to use CCREMs. First, there may be difficulty in ensuring the model prerequisites have been satisfied. In order for the use of a CCREM to be justified, both of the level-2 units for which a subject is cross-classified have to be randomly selected from larger populations. In the case where each of the cross-classified units are not randomly selected from larger populations, there has been some debate on whether the use of CCREM is appropriate. For example, Teitler and Weiss (2000) used a CCREM to examine the sexual behavior of youths who are cross-classified by census tract and school. Unless both census tract and school were randomly selected, the findings cannot be considered generalizable across all census tracts and schools in the U.S. Second, accurate information may not be readily available regarding

cluster membership in cross-classified collective structures. Third, the existence of CCREM may not be well known, or avoided due to its complexity.

The hesitance to use CCREMs may lead to misspecification of the cross-classified data. It was shown using empirical or simulated datasets misspecifying CCREM may result in misleading conclusions (Fielding, 2002; Goldstein, 1994; Luo & Kwok, 2009; Meyers & Beretvas, 2006). There are two approaches to misspecifying the cross-classified data structure, HLM-delete and HLM-complete (Meyers & Beretvas, 2006). For example, for students who were nested within a cross-classification of middle and high schools, HLM-delete omitted subjects who did not attend the main middle school that fed into their particular high school. This resulted in a strictly hierarchical dataset with students nested in middle schools which were nested in high schools, restricting the generalizability of the findings only to subjects who are not cross-classified (e.g., those who attended the main middle school). HLM-complete utilized all subjects but ignored the middle school clustering. Thus, students were only nested in high schools and the middle schools were not modeled as a separate level.

Meyers and Beretvas (2006) compared the two HLM approaches and CCREM when modelling test scores from students who were cross-classified by middle and high schools using the 1988 National Educational Longitudinal Study (NELS). They found that the fixed parameter estimates and their standard errors, the level-1 residual variance and its standard error were all similar between the HLM and CCREM models. However, the estimated between high school variance differed substantially between the models. The HLM-Delete model had the highest value followed by the HLM-Complete model, with the CCREM presenting the lowest between high school variance. This suggests in the HLM models the between middle school variance was masked to produce an inflated between high school variance.

Meyers and Beretvas (2006) replicated their real data analysis of the 1988 NELS data with a simulation study. Five factors were included in their design: correlation between the residuals of two cross-classified factors, number of feeder middle schools, number of levels of cross-classified units, average middle school size, and intraclass correlation (ICC) values. The CCREM model had students (level-1) nested within a cross-classification of middle and high schools (level-2), while the HLM model had students (level-1) nested within high schools (level-2), ignoring the middle school clustering. Both the CCREM and HLM models included three predictors, a student variable, a middle school variable, and a high school variable. However, while the CCREM modeled both the middle and high school characteristics as level-2 predictors, the HLM purposely modeled the middle school

characteristic which is a level-2 variable on level-1, as a student level characteristic. Results suggested that although the fixed parameter estimates were not affected as a result of misspecification, the standard errors for the middle school predictor that was included on the first level were. They found the relative biases of the standard error estimates under the CCREM model were all acceptable. However, most of the relative biases of the standard error values in the HLM model were intolerably high and negative. Furthermore, they found the between high school variance parameter was overestimated when the model incorrectly ignored the middle school clustering.

Luo and Kwok (2009) extended the above study by examining the impact of misspecification of CCREMs in a three-level model with two random factors crossed at the top level and at the intermediate level respectively. They found that ignoring one of the crossed factors biased the variance component estimates and standard errors of the fixed effects regression coefficients. The variance components of adjacent levels were overestimated and the variance component of the remaining crossed factor was underestimated. Further, misspecification resulted in underestimation of the standard error of the regression coefficient associated with a predictor of the ignored crossed factor and overestimation of the standard error of the regression coefficient of a predictor at a lower level.

Shi, Leite, and Algina (2010) assessed the effect of omitting the random interaction effect in CCREMs on parameter estimate and standard errors. No bias was found for the fixed effects. For random effects, variances at level-2 were affected but not those at level-1.

These three simulation studies examining the misspecification of CCREM focused on the random intercept model where the slope of the student (level-1) predictor remains constant at the cross-classified levels (e.g., middle school and high school). It is common for the effect of student predictors to vary across schools, and researchers are more interested in whether the effect is predicted by the school level predictors (i.e., cross-level interaction). Consequently a comparison of CCREM and HLM models with random slopes and intercepts warrants further investigation.

Studies in which the impact of misspecification of cross-classified datasets was explored can be compared to those exploring the impact of omitting a level in a purely hierarchical model (see, e.g., Moerbeek, 2004), which, similarly, were focused on random intercept models. This may be because analytical results are derivable with a closed-form solution for random intercept models, but not for random slope models (Van Landeghem, De Fraine, & Van Damme, 2005). When analytical results cannot be obtained, simulation studies are needed to evaluate the consequence of model misspecification. Random slope models are of interest in

cross-sectional research (e.g., Swanson & Stevenson, 2002) and longitudinal research. A random slope model is prevalent especially for the growth curve models as the changing slope across time is considered to vary across subjects. In summary, it is important to evaluate how ignoring cross-classified data structure affects the performance of fixed effects and variance components estimators in random slope models, and this study fills the gap. It is hypothesized ignoring cross-classified data structure will influence the standard error of fixed effects related to the misspecified level, and variance components of both random intercepts and slopes.

Methodology

Simulation Design

The design of this simulation mirrored the study of Meyers and Beretvas (2006) with one major difference: the slope of the level-1 predictor was modeled as randomly varying across both the middle and high schools. Four factors were manipulated: correlation between the level-2 residuals (0, 0.40); the intraclass correlation (ICCs) (0.05, 0.15, 0.25); the number of cross-classified units, (i.e., the number of middle schools and high schools, 30 and 50); and the average middle school size (20, 40). This resulted in a total of 24 conditions; for each condition, 2,000 datasets were generated.

Most of the simulation conditions in Meyers and Beretvas (2006) were adopted, except the added conditions with ICC = 0.25 as large ICC is common in with national longitudinal databases in education (Hedges & Hedberg, 2007). This study may be the first in which the impact of ignoring one level in multilevel analysis in random slope models is examined. It was considered important to keep the simulation conditions similar to the previous literature (Meyers & Beretvas, 2006; Luo & Kwok, 2009; Shi et al., 2010) so the results could be directly compared. As found when generating values, the distribution of predictors and the coefficients impacted the bias results. The simulation studies on CCREM, including Luo and Kwok (2009) and Shi et al. (2010) were based on similar parameter values as Meyers and Beretvas.

Unlike Meyers and Beretvas (2006), the number of feeder middle schools into high schools was held constant at two. Refer to their study for a detailed explanation of the way in which middle school students were organized into varying high schools. Meyer and Beretvas found the number of feeder middle schools did not affect outcome biases measures. Shi et al. (2010) came to a similar conclusion in their study of omitting random interaction effect in the cross-classified random

effects models. Different numbers of feeder schools resulted in different degrees of sparseness in the generated cross-classified cells. For example, with 30 high schools and 30 middle schools, there would be a total of 900 cells, and 840 (93.33%) cells in the two-feeder condition and 810 (90%) cells in the three-feeder condition would be empty. The degree of sparseness was found not to influence the bias outcome.

The correlation between the level-2 random effects for the middle and the high schools was set to be either 0 or 0.4. The zero correlation condition was chosen as a baseline model to compare to the 0.4. Meyers and Beretvas (2006) chose 0.4 to mimic the fairly related data structures found in many applied datasets (e.g. students who attend low socioeconomic middle schools are likely to attend low socioeconomic high schools as well).

The number of middle schools and high schools were either both 30 or both 50. The number of students in each middle school was randomly generated, either from a normal distribution with a mean of 20 and a standard deviation of 2 or from a normal distribution with a mean of 40 and a standard deviation of 2.

The three levels of ICCs were 0.05, 0.15, and 0.25. Meyers and Beretvas (2006) chose 0.05 and 0.15 to represent small and moderate ICCs as their examination of applied studies and textbook examples suggested that conditional ICCs ranged from 0.009 to 0.24, ($M = 0.066$, $SD = 0.0682$). The 0.25 was added here to represent large ICCs. Hedges and Hedberg (2007) examined achievement scores in longitudinal surveys with national probability samples and found that the average ICC was about 0.22 for all schools across Grades K-12, higher than the widely-used guidelines of 0.05-0.15.

Model Used to Generate Data

Data were using a two-level cross-classified model where students were cross-classified by middle school and high school and both level-1 intercepts and slopes were set to be random. Following the notation of Raudenbush and Bryk (2002), the level-1 equation of the CCREM model is:

$$Y_{i(jk)} = \pi_{0(jk)} + \pi_{1(jk)}X_{i(jk)} + e_{i(jk)}, \quad e_{i(jk)} \sim N(0, \sigma^2) \quad (1)$$

where $Y_{i(jk)}$ represents a student's achievement score, $\pi_{0(jk)}$ is the adjusted mean of students who had zero values on the student level predictor, $X_{i(jk)}$, and attended the same middle and high school combination. $\pi_{1(jk)}$ is the regression coefficient of the student level predictor. The level-1 residual, $e_{i(jk)}$, is the difference with which the

student's score varies from the cluster's adjusted mean, and has a unity variance. The level-2 equations are:

$$\begin{cases} \pi_{0(jk)} = \gamma_{000} + \gamma_{010}Z_j + \gamma_{020}W_k + b_{0j0} + c_{00k} \\ \pi_{1(jk)} = \gamma_{100} + \gamma_{110}Z_j + \gamma_{120}W_k + b_{1j0} + c_{10k} \end{cases} \quad (2)$$

The intercept model is the same as in Meyers and Beretvas (2006). The level-1 intercept is a function of an intercept, γ_{000} , a high school predictor, W_k , a middle school predictor, Z_j , plus the residual that is decomposed into two separate components for middle and high schools, b_{0j0} and c_{00k} , respectively. However, unlike Meyers and Beretvas, the slope, $\pi_{1(jk)}$, of the level-1 predictor, $X_{i(jk)}$, is now modeled as randomly varying across middle and high schools, reflected in the b_{1j0} and c_{10k} terms. The random slope was also predicted by W_k and Z_j , resulting with two cross-level interaction terms.

The intercept (γ_{000}) was set to be at 100 and the coefficients of the predictors in the model ($\gamma_{100}, \gamma_{010}, \gamma_{020}, \gamma_{120}, \gamma_{110}$) were set to be 0.5 to replicate the study by Meyers and Beretvas (2006). The predictors (W_k, Z_j, X_{jk}) were all generated from normal distributions ($M = 0, SD = 10$). The mean of these distributions were set to be zero because in models with interaction terms, predictors should be centered to avoid multicollinearity. The lower-order effects (X, Z, W) were interpretable as the linear slope of a predictor given the other predictors at zero values. In applied research, researchers often need to make a decision regarding centering the student level predictor, X , as grand mean or group mean centered. Here, setting the mean of X to be zero in each combination of a middle school and a high school will produce similar results while centering with grand or group mean, thereby reducing the possible confounding effects introduced by the different centering approaches.

The CCREM model incorporated five variance components of $b_{0j0}, b_{1j0}, c_{00k}, c_{10k}$, and $e_{i(jk)}$. The residuals, b_{0j0} and b_{1j0} represented the amount of variation among middle schools in their adjusted mean (intercepts) and among the slopes of the student level predictor, after accounting for the middle and high school predictors. Similarly, c_{00k} and c_{10k} were the amount of variation among the high schools in the adjusted mean (intercepts) and the slopes of the student level predictor, after accounting for the middle and high school predictors. The variances of b_{0j0}, c_{00k} , and their covariance depend on the values of ICC and correlation.

Table 1. Covariance matrix of random effects in data generation model for different correlation and ICC values

Correlation	ICC	Generating distribution of random intercepts and slopes			
0.00	0.05	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.0556 & 0 & 0 & 0 \\ 0 & 0.0278 & 0 & 0 \\ 0 & 0 & 0.0556 & 0 \\ 0 & 0 & 0 & 0.0278 \end{pmatrix}$
	0.15	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.2143 & 0 & 0 & 0 \\ 0 & 0.1072 & 0 & 0 \\ 0 & 0 & 0.2143 & 0 \\ 0 & 0 & 0 & 0.1072 \end{pmatrix}$
	0.25	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.5000 & 0 & 0 & 0 \\ 0 & 0.2500 & 0 & 0 \\ 0 & 0 & 0.5000 & 0 \\ 0 & 0 & 0 & 0.2500 \end{pmatrix}$
0.40	0.05	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.0556 & 0 & 0.0222 & 0 \\ 0 & 0.0278 & 0 & 0 \\ 0.0222 & 0 & 0.0556 & 0 \\ 0 & 0 & 0 & 0.0278 \end{pmatrix}$
	0.15	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.2143 & 0 & 0.0857 & 0 \\ 0 & 0.1072 & 0 & 0 \\ 0.0857 & 0 & 0.2143 & 0 \\ 0 & 0 & 0 & 0.1072 \end{pmatrix}$
	0.25	$\begin{pmatrix} b_{0j0} \\ b_{1j0} \\ c_{00k} \\ c_{10k} \end{pmatrix}$	$\sim N$	$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.5000 & 0 & 0.2000 & 0 \\ 0 & 0.2500 & 0 & 0 \\ 0.2000 & 0 & 0.5000 & 0 \\ 0 & 0 & 0 & 0.2500 \end{pmatrix}$

Presented in Table 1 are the generating distribution of the residuals in random intercepts and slopes (b_{0j0} , b_{1j0} , c_{00k} , c_{10k}) by correlation and ICC values. Using τ as the symbol for variance, the ICC is calculated as

$$\frac{\tau_{b_{0j0}}}{\tau_{b_{0j0}} + \tau_{c_{00k}} + \sigma^2}$$

for students attending the same middle school and

$$\frac{\tau_{c_{00k}}}{\tau_{b_{0j0}} + \tau_{c_{00k}} + \sigma^2}$$

for students attending the same high school. With $\tau_{b_{0j0}} = \tau_{c_{00k}}$ and $\sigma^2 = 1$, solve the values of the variances and then covariances (calculated as the square root of the product of corresponding variances and the correlation) for different conditions. Thus, for conditions with zero correlation, the variances of b_{0j0} and c_{00k} were set at 0.0556 for ICC = 0.05, at 0.2143 for ICC = 0.15, and at 0.50 for ICC = 0.25. For conditions with 0.4 correlation, the variances of b_{0j0} and c_{00k} were set with the same values of the corresponding conditions with zero correlation, while the covariance between b_{0j0} and c_{00k} was set at = 0.0222 for ICC = 0.05, 0.0857 for ICC = 0.15, and 0.2 for ICC = 0.25. The variances of b_{1j0} and c_{10k} were set to be half of the variances of b_{0j0} and c_{00k} as the variances of slopes are often smaller than those of intercepts (Raudenbush & Liu, 2001). The covariances between random intercepts and random slopes (b_{0j0} and b_{1j0} ; c_{00k} and c_{10k}) were set to zero. Even though random intercepts and random slopes may be correlated in applied research, it was not considered in order to compare results to previous studies by introducing only variance components of random slopes. Introducing intercept-slope covariance may confound such comparison.

Data Generation and Analysis

Cross-classified datasets were generated using equations (1) and (2). First, we generated 50 (or 30 depending on the number of middle schools) 1×4 vectors from multivariate normal distributions as in Table 1. Combinations of middle schools and high schools were generated following the condition with two middle school feeders in Meyers and Beretvas (2006). The residuals of middle schools were then sorted into an ascending order. For all middle schools except the one with the largest residual, sixty percent of a middle school's students were sent to the high school paired with this middle school, and the rest were sent to the high school paired with the middle school with the next higher residual. For the middle school

with the largest residual, 40% of its students were sent to the high school paired with the middle school with the smallest residual.

Then, 2000 data sets were generated for each of the 24 simulation conditions in SAS 9.1.3. For each of the generated data set, SAS PROC MIXED was used to perform the CCREM analyses and HLM analyses using the full maximum likelihood estimation. The HLM addressed the high school clustering, but ignored the middle school clustering. The HLM included the middle school predictor and its interaction with the student predictor as the student level predictors. The level-1 equation of HLM model is

$$Y_{ij} = \pi_{0j} + \pi_{1j}X_{ij} + \pi_{2j}Z_{ij} + \pi_{3j}X_{ij}Z_{ij} + e_{ij} \quad (3)$$

This first level shows how the student's score, Y_{ij} , is a function of the intercept, π_{0j} , plus a student predictor, X_{ij} , and its weight, π_{1j} , as well as a middle school predictor erroneously included on the first level, Z_{ij} , and its weight, π_{2j} , the interaction between the two first level predictors, $X_{ij}Z_{ij}$, weighted by π_{3j} , plus an error term, e_{ij} , which captures the deviation from the student's score from the high school mean conditioned on the three predictors. The level-2 equations for the HLM are

$$\begin{cases} \pi_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \pi_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \\ \pi_{2j} = \gamma_{20} \\ \pi_{3j} = \gamma_{30} \end{cases} \quad (4)$$

The level-1 intercept, π_{0j} , is predicted by an overall mean for all students conditioned on all predictors, γ_{00} , a high school predictor, W_j , its weight, γ_{01} , plus the error term, u_{0j} , which is the difference from a high school's adjusted mean to the overall mean for all high schools. The slopes, π_{1j} , are predicted by an overall slope, γ_{10} , plus a high school predictor, W_j , and its weight, γ_{11} , plus the error term, u_{1j} . The slope for the middle school predictor, π_{2j} , was held constant, as was the slope of the interaction term between the student and the middle school predictor.

The HLM analyses incorporated variance components of three residuals: u_{0j} , u_{1j} , and e_{ij} . The first residual, u_{0j} represents the amount of variation in the intercepts, or the mean achievement across high schools that remained unexplained after accounting for the predictors at the first and second levels. u_{1j} stands for the amount of residual variation in the slopes of the student level predictor across high schools,

and e_{ij} is the amount of student level variation that exists after controlling for the predictors. For comparison with CCREM estimates, only the variance components of intercepts and slopes between high schools and level-1 residuals were investigated. HLM analysis did not contain the middle school variance component, and thus could not be compared with CCREM on these estimates.

Outcome Measures

The outcome measures included relative biases of parameter estimates and standard error estimates of fixed effects and variance components. The relative bias of parameter estimates, $B(\hat{\theta})$, was calculated using equation (5), where $\bar{\hat{\theta}}_r$ is the mean of the r^{th} parameter estimate across the 2,000 replication and θ_r is the actual value of the r^{th} parameter (Hoogland & Boomsma, 1998).

$$B(\bar{\hat{\theta}}) = \frac{\bar{\hat{\theta}}_r - \theta_r}{\theta_r} \quad (5)$$

The relative bias of the standard errors (SEs), $B(\bar{\hat{S}}_{\hat{\theta}_r})$ was calculated using equation (6), where $\bar{\hat{S}}_{\hat{\theta}_r}$ is the mean standard error across the 2,000 replications and $S_{\hat{\theta}_r}$ is the standard deviation of the parameter estimates (i.e., empirical standard error) (Hoogland & Boomsma, 1998).

$$B(\bar{\hat{S}}_{\hat{\theta}_r}) = \frac{\bar{\hat{S}}_{\hat{\theta}_r} - S_{\hat{\theta}_r}}{S_{\hat{\theta}_r}} \quad (6)$$

Analysis

Descriptive statistics were used to examine whether the bias was acceptable for each condition; that is, whether the relative parameter bias was less than 0.05 and the relative standard error bias was less than 0.1 in absolute value (Hoogland & Boomsma, 1998). When biases were not acceptable in all simulation conditions for a parameter, analyses of variance (ANOVA) were conducted to determine which factor(s) affected the relative biases regarding this parameter. The outcome variables included the relative deviations of parameter estimates, calculated as the differences between the parameter estimates and the population parameters divided

by the population parameters, and the relative deviations of standard error estimates, calculated as the deviations of the sample standard error estimate from the empirical standard error divided by the empirical standard error.

The between-subjects factors in ANOVAs included the correlation between the level-2 residuals, zero and non-zero; the intraclass correlation, 0.05, 0.15, and 0.25; the number of cross-classified units, 30 and 50; and the average middle school size, 20 and 40. The within-subjects factor included the model type (CCREM and HLM). Due to the large number of replications, statistical significance was not examined. Instead, partial eta squared effect sizes, η_p^2 , were computed as a measure of practical significance. Only effects that were practically significant with $\eta_p^2 > 0.01$ were interpreted, which was considered a very conservative cut-off value for an appreciable effect in similar simulation studies (Shi et al., 2010).

Figure 1. Relative standard error bias of coefficients of W (the high school predictor), X (the student level predictor), Z (the middle school predictor), and interactions between X and W and between X and Z

Results

Relative Parameter and Standard Error Bias of Fixed Effects

All relative parameter biases for fixed effect estimates met the criterion of Hoogland and Boomsma (1998) (< 0.05) for acceptable parameter bias amounts. All relative standard error biases for fixed effect estimates, as shown in Figure 1, met Hoogland and Boosma's criterion (< 0.10) for acceptable parameter bias amounts except the HLM estimates of coefficients of X , Z , and XZ .

Coefficient of X , the Student Level Predictor For coefficients of X (γ_{100} for CCREM and γ_{10} for HLM), the CCREM standard error estimates were in the acceptable range for all conditions with biases ranging from -0.07 to 0.01 ($M = -0.03$, $SD = 0.02$). The HLM standard error estimates were unacceptable for all conditions with biases ranging from -0.20 to -0.13 ($M = -0.16$, $SD = 0.02$). Given the substantial bias found in the HLM standard error estimates, a 2 (correlation) $\times$ 3 (ICC) $\times$ 2 (number of cross-classified units) $\times$ 2 (average middle school size) between-subjects ANOVA was conducted on the relative deviations of HLM standard error estimates. Only the main effect of number of middle schools was significant ($\eta_p^2 = 0.012$). The other effects had $\eta_p^2 < 0.004$. Biases in conditions with 30 schools ($M = -0.18$, $SD = 0.15$) were larger in magnitude than those with 50 schools ($M = -0.15$, $SD = 0.09$).

Coefficient of Z , the Middle School Predictor For the coefficient of Z (γ_{010} for CCREM and γ_{20} for HLM), CCREM standard error estimates were acceptable. The relative standard error bias ranged from -0.07 to 0.00 ($M = -0.03$, $SD = 0.02$). However, HLM standard error estimates were negatively biased with a range from -0.37 to -0.15 ($M = -0.25$, $SD = 0.06$). The between-subjects ANOVA conducted on the relative deviations of HLM standard error estimates indicated that the significant effects included the main effects of ICC ($\eta_p^2 = 0.08$) and middle school size ($\eta_p^2 = 0.14$). The other effects had $\eta_p^2 < 0.003$. Pairwise comparison suggested that biases for ICC = 0.25 ($M = -0.29$, $SD = 0.14$) were larger in magnitude than those for ICC = 0.15 ($M = -0.26$, $SD = 0.13$), which in turn were larger than those for ICC = 0.05 ($M = -0.20$, $SD = 0.13$). Biases were larger in magnitude when middle school size was 40 ($M = -0.30$, $SD = 0.12$) than 20 ($M = -0.20$, $SD = 0.14$).

Coefficient of XZ, the Cross-Level Interaction between the Student Level Predictor and the Middle School Predictor For the coefficient of XZ (γ_{110} for CCREM and γ_{30} for HLM), the CCREM standard error estimates were in the acceptable range for all conditions with the biases ranging from -0.07 to -0.01 ($M = -0.03$, $SD = 0.02$). The standard error biases of HLM estimates exceeded the acceptable level for all of the conditions ranging from -0.87 to -0.76 ($M = -0.82$, $SD = 0.04$). A between-subjects ANOVA on the relative deviations of HLM standard error estimates presented similar results as those for the coefficient of Z with the main effects of ICC ($\eta_p^2 = 0.24$) and middle school size ($\eta_p^2 = 0.45$) significant. The other effects had $\eta_p^2 < 0.008$. Pairwise comparison suggested that biases increased in magnitude for larger ICCs ($M = -0.84$, $SD = 0.04$ for ICC = 0.25; $M = -0.83$, $SD = 0.04$ for ICC = 0.15; $M = -0.80$, $SD = 0.05$ for ICC = 0.05). Biases were larger in magnitude when middle school size was 40 ($M = -0.85$, $SD = 0.03$) than 20 ($M = -0.79$, $SD = 0.04$).

In summary, the HLM model resulted with deflated standard errors of regression coefficients related to the student level predictor, the middle school predictor, and their interaction. This is not surprising, because the HLM model ignores the middle school level and disaggregates the middle school predictor at the student level.

Table 2. Relative bias of variance estimates when correlation between residuals was 0

ICC	No. of schools	Middle school size	Variance of intercept across high schools		Variance of slope across high schools		Student-level residual variance	
			CCREM	HLM	CCREM	HLM	CCREM	HLM
0.05	30	20	-0.108	0.232	-0.054	0.432	0.000	1.219
0.05	30	40	-0.103	0.299	-0.057	0.411	0.002	1.260
0.05	50	20	-0.079	0.328	-0.039	0.464	0.001	1.255
0.05	50	40	-0.035	0.412	-0.031	0.471	0.000	1.283
0.15	30	20	-0.088	0.252	-0.055	0.424	0.000	4.711
0.15	30	40	-0.059	0.357	-0.050	0.411	0.000	4.836
0.15	50	20	-0.057	0.326	-0.029	0.466	-0.001	4.825
0.15	50	40	-0.045	0.398	-0.032	0.464	0.000	4.979
0.25	30	20	-0.066	0.267	-0.053	0.423	0.000	11.012
0.25	30	40	-0.055	0.353	-0.056	0.414	0.000	11.384
0.25	50	20	-0.038	0.364	-0.030	0.473	0.000	11.319
0.25	50	40	-0.042	0.399	-0.029	0.460	0.000	11.578

Figure 2. Relative parameter bias of variance estimates of intercepts across high schools for conditions with correlation = 0; the two numbers on the horizontal axis were number of schools / middle school size

Bias of the Random Effects Variance Components for Zero Correlation Conditions

As Meyers and Beretvas (2006) pointed out, when correlation between middle school and high school intercepts residual was not zero, the true value of the variance components was not known because the analysis using CCREM assumed that the correlation of cross-classified units was zero. Thus, only relative parameter bias is reported for the student level residual variance and variance of intercepts and slopes between high schools for conditions when correlation was zero (Table 2). For conditions when correlation was 0.4, compare the variance estimates of CCREM and HLM.

Variance of Level-1 Intercepts across High Schools The relative bias ranged from -0.108 to -0.035 for CCREM estimates and from 0.232 to 0.412 for HLM estimates across the 24 simulation conditions (Figure 2). The CCREM estimates in the conditions with larger sample sizes (50 schools and an average of 40 students per school) were acceptable with relative biases less than 0.05. CCREM estimates in the other conditions were slightly negatively biased. However, HLM estimates were positively biased in all conditions. A mixed effect ANOVA was conducted on the relative deviations of the parameter estimates. Only the main effect of model was significant ($\eta_p^2 = 0.45$) with CCREM estimates having smaller

deviations ($M = -0.06$, $SD = 0.37$) than HLM estimates ($M = 0.33$, $SD = 0.68$). The other effects had $\eta_p^2 < 0.003$.

Figure 3. Relative parameter bias of variance estimates of slopes across high schools for conditions with correlation = 0; the two numbers on the horizontal axis were number of schools / middle school size

Figure 4. Relative parameter bias of level-1 residual variance estimates for conditions with correlation = 0; the two numbers on the horizontal axis were number of schools / middle school size

IMPACT OF INAPPROPRIATE MODELING OF CROSS-CLASSIFIED DATA

Variance of Level-1 Slope across High Schools CCREM estimates of the slope variance across high schools were acceptable with the relative biases ranging from -0.057 to -0.029 (Figure 3). However, HLM estimates were unacceptable with biases ranging from 0.411 to 0.473. A mixed effect ANOVA on the relative deviations of the parameter estimates resulted with only one appreciable effect, the main effect of model ($\eta_p^2 = 0.76$). The other effects had $\eta_p^2 < 0.004$.

Student-Level Residual Variance There was no bias in estimating student level residual variance using CCREM because the biases ranged from -0.0038 to -0.0007. However, the HLM model overestimated the student level residual variance with the bias ranging from 1.219 to 11.578. The mixed ANOVA indicated that the appreciable effects included: interaction effect between model and ICC ($\eta_p^2 = 0.79$), and main effects of model ($\eta_p^2 = 0.88$) and ICC ($\eta_p^2 = 0.79$). As shown in Figure 4, the degree of overestimation of HLM estimates increased with the value of ICC.

In summary, the HLM model inflated the variance components of the level-1 residual, the random intercept and slope across high schools. In other words, while omitting the middle school level, variance across middle schools was imposed at the student and high school level.

Table 3. Mean of random effect variance estimates when correlation between residuals was 0.4

ICC	No. of schools	Middle school size	Var of intercept across high schools		Var of slope across high schools		Student-level residual variance	
			CCREM	HLM	CCREM	HLM	CCREM	HLM
0.05	30	20	0.060	0.090	0.026	0.040	0.996	2.221
0.05	30	40	0.061	0.097	0.026	0.040	0.999	2.254
0.05	50	20	0.064	0.099	0.027	0.041	0.998	2.245
0.05	50	40	0.063	0.102	0.027	0.041	0.998	2.289
0.15	30	20	0.232	0.357	0.101	0.153	0.996	5.750
0.15	30	40	0.225	0.381	0.102	0.153	0.998	5.848
0.15	50	20	0.239	0.382	0.104	0.158	0.997	5.858
0.15	50	40	0.232	0.396	0.104	0.157	0.998	5.950
0.25	30	20	0.524	0.857	0.237	0.357	0.998	11.969
0.25	30	40	0.508	0.891	0.238	0.356	0.999	12.297
0.25	50	20	0.543	0.908	0.242	0.367	0.997	12.269
0.25	50	40	0.526	0.928	0.243	0.367	0.999	12.565

Figure 5. Variance estimates of intercepts across high schools for conditions with correlation = 0.4; the two numbers on the horizontal axis were number of schools / middle school size

Figure 6. Variance estimates of slopes across high schools for conditions with correlation = 0.4; the two numbers on the horizontal axis were number of schools / middle school size

Random Effects Variance Components for 0.4 Correlation Conditions

Compared in Table 3 are CCREM and HLM estimates of the variances of intercepts and slopes across high schools, and the student level residual variance for conditions with 0.4 correlation. The CCREM estimates were close to the generating

values of these variances. For variances of intercepts and slopes (Figures 5 and 6), the HLM estimates were larger than CCREM estimates, and the differences increased with larger ICC values. For student level residual variance (Figure 7), HLM estimates were almost twice as large as the CCREM estimates when $ICC = 0.05$, four times larger than the CCREM estimates when $ICC = 0.15$, and twelve times larger when $ICC = 0.25$. Note the variances of intercepts and slopes over the middle schools, the ignored crossed factor in HLM, were the same as those of the high schools in the data generation. It is evident that the variance contributed by the ignored crossed factor (i.e., the middle school) was imposed on the variance components of the other remaining crossed factor (i.e., the high school) and the student level when using HLM.

Relative Standard Error Bias of Random Effects Variance Components

Presented in Table 4 and Figures 8 and 9 are the relative standard error biases across the 24 conditions of estimates of intercepts and slopes variances. None of the CCREM and HLM standard error estimates were acceptable. For intercepts variances, CCREM standard error estimates had biases ranging from 2.44 to 3.91 ($M = 3.10$, $SD = 0.54$) and HLM standard error estimates had biases ranging from 3.48 to 7.60 ($M = 5.33$, $SD = 1.21$). For slope variances, CCREM standard error estimates had biases ranging from 2.62 to 4.06 ($M = 3.10$, $SD = 0.54$) and HLM standard error estimates had biases ranging from 3.64 to 5.96 ($M = 4.82$, $SD = 0.90$).

Figure 7. Level-1 residual variance estimates for conditions with correlation = 0.4; the two numbers on the horizontal axis were number of schools / middle school size

Figure 8. Relative standard error bias of variance estimates of intercepts across high schools; the two numbers on the horizontal axis were number of schools / middle school size

Figure 9. Relative standard error bias of variance estimates of slopes across high schools; the two numbers on the horizontal axis were number of schools / middle school size

IMPACT OF INAPPROPRIATE MODELING OF CROSS-CLASSIFIED DATA

Table 4. Relative standard error bias of random effect variance estimates

Correlation	ICC	No. of schools	Middle school size	SE of intercept variance across high schools		SE of slope variance across high schools	
				CCREM	HLM	CCREM	HLM
0.0	0.05	30	20	2.92	4.15	2.62	3.64
0.0	0.05	30	40	2.49	3.48	2.79	3.90
0.0	0.05	50	20	3.91	5.64	3.85	5.44
0.0	0.05	50	40	3.58	4.99	3.81	5.50
0.0	0.15	30	20	2.54	4.74	2.75	3.99
0.0	0.15	30	40	2.52	4.14	2.76	3.99
0.0	0.15	50	20	3.64	6.73	3.89	5.71
0.0	0.15	50	40	3.72	6.14	3.91	5.76
0.0	0.25	30	20	2.49	5.00	2.78	4.10
0.0	0.25	30	40	2.62	4.54	2.76	4.09
0.0	0.25	50	20	3.74	7.60	3.83	5.71
0.0	0.25	50	40	3.67	6.48	4.06	5.96
0.4	0.05	30	20	2.83	3.94	2.68	3.80
0.4	0.05	30	40	2.49	3.58	2.77	3.90
0.4	0.05	50	20	3.48	5.00	3.85	5.45
0.4	0.05	50	40	3.56	5.15	3.96	5.67
0.4	0.15	30	20	2.44	4.58	2.66	3.87
0.4	0.15	30	40	2.54	4.48	2.77	4.02
0.4	0.15	50	20	3.50	6.51	3.95	5.83
0.4	0.15	50	40	3.49	6.23	3.86	5.70
0.4	0.25	30	20	2.51	5.28	2.82	4.14
0.4	0.25	30	40	2.53	4.95	2.66	3.93
0.4	0.25	50	20	3.49	7.43	3.96	5.90
0.4	0.25	50	40	3.66	7.28	3.79	5.68

Mixed effect ANOVA was conducted on the relative deviations of standard error estimates for intercepts variances and slope variances. For intercepts variances, the appreciable effects included the interaction effects between model and ICC ($\eta_p^2=0.14$), model and the number of schools ($\eta_p^2=0.06$), model and middle school size ($\eta_p^2=0.01$), and main effects of model ($\eta_p^2=0.63$), ICC ($\eta_p^2=0.05$), the number of schools ($\eta_p^2=0.23$), and middle school size ($\eta_p^2=0.01$). In Figure 9, we could see that HLM estimates were more positively biased than CCREM estimates. The differences in the bias between HLM and CCREM estimates were

larger for larger ICC values, 50 schools (vs. 30), and middle school size of 20 (vs. 40).

For slope variances, the effects that met the criterion for an appreciable effect included: the interaction effects between model and the number of schools ($\eta_p^2 = 0.06$), and main effects of model ($\eta_p^2 = 0.64$), and the number of schools ($\eta_p^2 = 0.29$). Figure 9 shows that HLM estimates were more positively biased than CCREM estimates and such difference was larger for conditions with 50 schools (vs. 30), and middle school size of 20 (vs. 40).

Power of Detecting Nonzero Variance Components

Power was all 100% for hypothesis tests regarding slope variances across high schools for both CCREM and HLM estimates, but not for intercepts variances. Figure 10 presents power levels for hypothesis tests regarding intercepts variances across high schools. When ICC was 0.05, there was no difference in power between CCREM and HLM. For higher ICC values, the CCREM estimates had near one power level while HLM estimates had significantly lower power in conditions with middle school size of 20. The difference in the power was the largest for the conditions with the smallest sample sizes (30 schools with 20 students on average per school).

Figure 10. Power of detecting nonzero intercept variances using CCREM and HLM models; the two numbers on the horizontal axis were number of schools / middle school size

Conclusion

In this study, previous research of ignoring a crossed random factor was extended by examining its effect on CCREM models with both random intercepts and slopes. One kind of model misspecification of CCREM models was considered: misspecifying a cross-classified level-2 unit (e.g., middle school) and imposed its characteristics (e.g., the SES of the middle school) as a predictor at level-1 (e.g., student level). Our results were consistent to previous Monte Carlo research (Meyers & Beretvas, 2006; Luo & Kwok, 2009) in that employing HLM on cross-classified data resulted in 1) unbiased estimates of regression coefficients of all fixed effects; 2) unbiased standard errors of regression coefficient of the high school predictor (the remaining crossed factor); 3) underestimated standard errors of regression coefficient for the middle school predictor (the predictor misspecified as level-1 predictor); 4) overestimated level-1 residual variance; and 5) overestimated variance of random intercepts across high schools.

In the previous studies, a random intercept model was adopted in which the effect of the student level predictor was fixed across different middle and high schools. In this study, this limitation was freed, and the effect of the student level predictor was modeled as random across middle and high schools. In addition, two cross-level interaction effects were modeled in which the effect of the student level predictor was affected by a middle school predictor and a high school predictor. The addition of random slope changed the impact of model misspecification on the fixed effect regression coefficient of the student level predictor found in previous studies. Meyers and Beretvas (2006) found misspecified HLM produced unbiased standard error of the regression coefficient of the student level predictor. However, with the addition of the random slope to model in the current study, it was found this standard error was underestimated and magnitude of underestimation was larger when there were fewer cross-classified units (middle and high schools).

The focus in previous studies on random intercept models omitted the cross-level interaction between the student level predictor and the high school predictor, the interaction between the student level predictor and the middle school predictor, as well as the variance of random slopes of the student level predictor across high schools. In the current study, these omissions were taken into account and it was found model misspecification did not impact the regression coefficient or the standard error of the cross-level interaction between the student level predictor and the high school predictor. The reason may be that clustering of students in high schools was considered in the HLM. However, it was found for the interaction between the student level predictor and the middle school predictor, the standard

error was underestimated using HLM and the magnitude of underestimation was higher for larger ICC values and larger size schools. The dependence of underestimation on the size of the schools was also found by Meyers and Beretvas (2006). This underestimation may be due to the contribution of the per-cluster sample size to the design effect in cluster sampling. According to Snijders and Bosker (1999), the design effect is larger when the sample size per cluster is larger, and thus more bias is introduced when the homogeneity within a cluster is ignored. In addition, the design effect is larger when ICC is larger, thus explaining more biases for larger ICC values.

The results from this study suggest, across all conditions considered, the use of CCREM for cross-classified datasets produced more accurate results when compared to incorrectly using a hierarchical model. It appears that when a predictor is modeled on a level lower than it should be and a level of nesting is ignored, the parameter of the true predictor on that lower level, as well as its interaction with the erroneously modeled predictor, may have biased estimates. This result revealed the insidious effect that ignoring a level of nesting may have on the variables in the model that were actually correctly specified. The downward bias of the standard error for regression coefficients of predictors at level-1, predictors at the ignored level, and their cross-level interaction has important implications for applied researchers who are utilizing an HLM. This underestimation inflates the Type I error rate. Specifically, the degree of underestimation of HLM standard error estimates is the most pronounced for the cross-level interaction. The false identification of significant interaction effects could lead to excessive effort in exploring and explaining such interaction effects. When ICC was not substantial (e.g., at 0.05), the standard error estimates in the misspecified HLM models were still significantly underestimated. This suggests trying to model the cross-classification even when there is concern with the degree of dependence within a crossed factor.

The effect of misspecification on random intercepts variance found in previous studies applies to random slopes variance as well. In the current study, the variances of the level-1 intercepts and slopes across high schools were both overestimated in the misspecified HLM. However, this does not necessarily imply that misspecified HLM models would be more powerful in detecting the nonzero variance components. The standard error estimates of CCREM and HLM are both positively biased while HLM estimates have larger biases. With substantial clustering, CCREM is found to have more power in detecting the nonzero variance components, but HLM has similar power when the number of schools and school sizes are large. Caution is needed in using the standard error of an estimated

variance component to conduct a hypothesis test. The distribution of the estimated variance component converges only slowly to normality as the sample size increases, and thus a Wald z -test on variances is not recommended (Raudenbush & Bryk, 2002). With positively biased standard error estimates, CCREM still produces acceptable power when ICC suggests substantial clustering ($ICC \geq 0.15$). Raudenbush and Bryk (2002) states that the software of HLM provided a better approach to hypothesis testing of variance components, and further studies may be needed in comparing their approach to Wald's z -test in hypothesis testing of variance components.

The biased variance estimates of HLM have serious consequences. If a level-2 crossed factor is ignored, and instead an HLM is implemented, the notion of how much student level variance remains unexplained after accounting for the predictors would be severely inaccurate. Using a CCREM greatly reduces this bias and provides an unbiased estimate of how much variation in the student level remains after controlling for the predictors. Moreover, the overestimation of variances of intercepts and slopes across the remaining crossed factor can lead to incorrect conclusions on the degree of remaining dependence after all predictors are controlled for. This may lead to the identification of a greater number of predictors than actually needed to explain the overestimated variances.

In summary, CCREM should be applied to cross-classified data to avoid inflated Type I error of some fixed effects and overestimated variances of level-1 intercepts and slopes. When CCREM cannot be applied to a cross-classified data due to lack of information of cluster membership, if the conditions are similar to those used in this study, ignoring a crossed factor will not impact the predictors of the correctly modeled crossed factor. Hypothesis tests regarding predictors at the lower level and associated with the ignored crossed factor will have inflated Type I error rates. Intercept and slopes variances will be overestimated.

Limitations and Suggestions for Future Research

The generalizability of results from this study is restricted to the manipulated conditions of the four independent variables in this study and thus is not applicable to all research cases. Indeed, there were several limitations of this study that can be addressed by research methodologists in the future. The principle limitations of this study are contingent on the manipulation of the four factors. Furthermore, as Meyers and Beretvas (2006) purported, more research is needed when the number of cross-classified units (middle and high schools) is not identical. The covariance between random intercepts and random slopes could be nonzero. Last, only partial

cross-classification was considered in this study. As Luo and Kwok (2009) pointed out, cross-classification structure has important impact on the effect of model misspecification of cross-classified data. Different cross-classification structures should be evaluated in the future for random slope models.

References

- Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, 81(1), 117-152. doi: [10.1353/sof.2002.0038](https://doi.org/10.1353/sof.2002.0038)
- Fielding, A. (2002). Teaching groups as foci for evaluating performance in cost-effectiveness of GCE advanced level provision: some practical methodological innovations. *School Effectiveness and School Improvement*, 13(2), 225-246. doi: [10.1076/sesi.13.2.225.3435](https://doi.org/10.1076/sesi.13.2.225.3435)
- Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, 22(3), 364-375. doi: [10.1177/0049124194022003005](https://doi.org/10.1177/0049124194022003005)
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87. doi: [10.3102/0162373707299706](https://doi.org/10.3102/0162373707299706)
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26(3), 329-367. doi: [10.1177/0049124198026003003](https://doi.org/10.1177/0049124198026003003)
- Luo, W., & Kwok, O.-m. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, 44(2), 182-212. doi: [10.1080/00273170902794214](https://doi.org/10.1080/00273170902794214)
- Ma, X. & Wilkins, J. L. M. (2002). The development of science achievement in middle and high school: Individual differences and school effects. *Evaluation Review*, 26(4), 395-417. doi: [10.1177/0193841x02026004003](https://doi.org/10.1177/0193841x02026004003)
- Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, 41(4), 473-497. doi: [10.1207/s15327906mbr4104_3](https://doi.org/10.1207/s15327906mbr4104_3)
- Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, 39(1), 129-149. doi: [10.1207/s15327906mbr3901_5](https://doi.org/10.1207/s15327906mbr3901_5)
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6(4), 387-401. doi: [10.1037/1082-989X.6.4.387](https://doi.org/10.1037/1082-989X.6.4.387)

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modeling. *British Journal of Mathematical and Statistical Psychology*, 63(1), 1-15. doi: [10.1348/000711008X398968](https://doi.org/10.1348/000711008X398968)

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessment. *Educational Evaluation and Policy Analysis*, 24(1), 1-27. doi: [10.3102/01623737024001001](https://doi.org/10.3102/01623737024001001)

Teitler, J. O., & Weiss, C. C. (2000). Effects of neighborhood and school environment on transitions to first sexual intercourse. *Sociology of Education*, 73(2), 112-132. doi: [10.2307/2673240](https://doi.org/10.2307/2673240)

Van Landeghem, G., De Fraine, B., & Van Damme, J. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research*, 40(4), 423-434. doi: [10.1207/s15327906mbr4004_2](https://doi.org/10.1207/s15327906mbr4004_2)

Witherspoon, D., & Ennett, S. (2011). Stability and change in rural youths' educational outcomes through the middle and high school years. *Journal of Youth and Adolescence*, 40(9), 1077-1090. doi: [10.1007/s10964-010-9614-6](https://doi.org/10.1007/s10964-010-9614-6)

Brief Reports

A Remark for the Admissibility of Rao's U-test

Z. D. Bai

Northeast Normal University
 Changchun, Jilin, China

C. R. Rao

Pennsylvania State University
 State College, PA

M. T. Tsai

Academia Sinica
 Nankang, Taipei, Taiwan

Let $\{\mathbf{X}_i; 1 \leq i \leq n\}$ be independent and identically distributed random vectors (i.i.d.r.v.) with a p -variate normal distribution with mean vector $\boldsymbol{\theta}$ and dispersion matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is assumed to be positive definite (p.d.). Partition $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{\theta}_1: p_1 \times 1$, $\boldsymbol{\theta}_2: p_2 \times 1$, $\boldsymbol{\Sigma}_{11}: p_1 \times p_1$, $\boldsymbol{\Sigma}_{22}: p_2 \times p_2$, $p_1 + p_2 = p$, $0 < p_2 < p$. The problem of interest is to test

$$\begin{aligned} H_0 : \boldsymbol{\theta}_1 = \mathbf{0}, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{\Sigma} \text{ unspecified} \\ \text{versus} \\ H_1 : \boldsymbol{\theta}_1 \neq \mathbf{0}, \boldsymbol{\theta}_2 = \mathbf{0}, \boldsymbol{\Sigma} \text{ unspecified.} \end{aligned} \tag{1}$$

For every $n (\geq 2)$, let

$$\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^n \mathbf{X}_i \text{ and } \mathbf{S} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})',$$

and express Hotelling's T^2 -statistic as

$$T^2 = n(n-1) \bar{\mathbf{X}}' \mathbf{S}^{-1} \bar{\mathbf{X}}.$$

Dr. Bai is faculty in the School of Mathematics and Statistics. Email at baizd@nenu.edu.cn. Prof. Rao, Sc.D. (Cantab), FRS, is the Eberly Professor Emeritus of Statistics and the Director of Center for Multivariate Analysis at Penn State, and Research Professor in Biostatistics at the University at Buffalo. Email him at crr1@psu.edu. Dr. Tsai is a research fellow in the Institute of Statistical Science. Email at mttsai@stat.sinica.edu.tw.

Partition $\bar{\mathbf{X}}$ and $\mathbf{S}$ the same as $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, respectively, and define

$$\begin{aligned}\bar{\mathbf{X}}_{1:2} &= \bar{\mathbf{X}}_1 - \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \bar{\mathbf{X}}_2, \\ \mathbf{S}_{11:2} &= \mathbf{S}_{11} - \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \mathbf{S}_{21}.\end{aligned}$$

For the problem (1), Rao (1946, 1949) proposed two test statistics which are of the forms

$$W = \frac{n(n-1) \bar{\mathbf{X}}_{1:2}' \mathbf{S}_{11:2}^{-1} \bar{\mathbf{X}}_{1:2}}{1 + n(n-1) \bar{\mathbf{X}}_2' \mathbf{S}_{22}^{-1} \bar{\mathbf{X}}_2}$$

and

$$U = n(n-1) \bar{\mathbf{X}}_{1:2}' \mathbf{S}_{11:2}^{-1} \bar{\mathbf{X}}_{1:2}$$

respectively.

Tsai (2003) generalized the Stein (1956) approach to show that Rao's U -test is admissible for the problem (1), which seems contradictory to Marden and Perlman (1980) who proved that Rao's U -test is inadmissible for the problem

$$\begin{aligned}H_0 : \boldsymbol{\theta}_1 = \mathbf{0}, \boldsymbol{\Sigma} \text{ unspecified} \\ \text{versus} \\ H_1 : \boldsymbol{\theta}_1 \neq \mathbf{0}, \boldsymbol{\Sigma} \text{ unspecified.}\end{aligned} \tag{2}$$

There is no contradiction between the two results, because the parameter spaces for the two problems are different. Rao's parameter space is $\Theta_R = \{(\boldsymbol{\theta}_1; \mathbf{0}; \boldsymbol{\Sigma})\}$ and MP's is $\Theta_{MP} = \{(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2; \boldsymbol{\Sigma})\}$. Rao's parameter space is smaller than MP's. It is possible that a test is admissible for a smaller parameter space while inadmissible for a larger parameter space. There may exist a test ϕ , such that $\beta_\phi(\boldsymbol{\theta}) \leq \beta_U(\boldsymbol{\theta})$, for all $\boldsymbol{\theta} \in \Theta_R$, $\beta_U(\boldsymbol{\theta}) \leq \beta_\phi(\boldsymbol{\theta})$, for all $\boldsymbol{\theta} \in \Theta_{MP} \setminus \Theta_R$, and $\beta_U(\boldsymbol{\theta}) < \beta_\phi(\boldsymbol{\theta})$, for some $\boldsymbol{\theta} \in \Theta_{MP} \setminus \Theta_R$, where $\beta_\phi(\boldsymbol{\theta})$ denotes the power of the test ϕ at the parameter $\boldsymbol{\theta}$.

Marden and Perlman (1980) proved the admissibility of Hotelling's T^2 -test for the problem (2), while Tsai (2003) proved the inadmissibility of Hotelling's T^2 -test for the problem (1). The inadmissibility result of Tsai (2003) does not contradict to the admissibility result of Marden and Perlman (1980). If ψ is a test

controlling T^2 for the problem (1), that is, $\beta_{T^2}(\theta) \leq \beta_{\psi}(\theta)$, for all $\theta \in \Theta_R$ and $\beta_{T^2}(\theta) < \beta_{\psi}(\theta)$, for some $\theta \in \Theta_R$, we must have $\beta_{T^2}(\theta) > \beta_{\psi}(\theta)$, for some $\theta \in \Theta_{MP} \setminus \Theta_R$. Hence, it does not control T^2 for the problem (2).

Although the results of Tsai (2003) and Marden and Perlman (1980) are not contradictory, the results of Tsai (2003) were obtained for an accurate model of Rao.

Acknowledgments

The first author was partly supported by Grants from NSFC 11571067 and 11471140.

References

- Marden, J. & Perlman, M. D. (1980). Invariant tests for means with covariates. *The Annals of Statistics*, 8(1), 25-63. doi: [10.1214/aos/1176344890](https://doi.org/10.1214/aos/1176344890)
- Rao, C. R. (1946). Tests with discriminant functions in multivariate analysis. *Sankhya*, 7, 407-413.
- Rao, C. R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhya*, 9, 343-366.
- Stein, C. (1956). The admissibility of Hotelling's T^2 -test. *The Annals of Mathematical Statistics*, 27(3), 616-623. doi: [10.1214/aoms/1177728171](https://doi.org/10.1214/aoms/1177728171)
- Tsai, M. T. (2003). *Admissibility of invariant tests for means with covariates* (Technical Report No. 2003-05). Nankang, Taiwan: Institute of Statistical Science, Academia Sinica.

Algorithms and Code

JMASM 46: Algorithm for Comparison of Robust Regression Methods In Multiple Linear Regression By Weighting Least Square Regression (SAS)

Mohamad Shafiq
Universiti Sains Malaysia
Kelantan, Malaysia

Wan Muhamad Amir
Universiti Sains Malaysia
Kelantan, Malaysia

Nur Syabiha Zafakali
Universiti Sains Malaysia
Kelantan, Malaysia

The aim of this study is to compare different robust regression methods in three main models of multiple linear regression and weighting multiple linear regression. An algorithm for weighting multiple linear regression by standard deviation and variance for combining different robust method is given in SAS along with an application.

Keywords: Multiple linear regression, robust regression, M, LTS, S, MM estimation

Introduction

Multiple linear regression (MLR) is a statistical technique for modeling the relationship between one continuous dependent variable from two or more independent variables. A typical data template is compiled in [Table 1](#).

Table 1. Data template for multiple linear regression

i	y_i	x_{i0}	x_{i1}	x_{i2}	..	x_{ip}
1	y_1	1	x_{11}	x_{12}	...	x_{1p}
2	y_2	1	x_{21}	x_{22}	...	x_{2p}
.	.	.	.	.		.
.	.	.	.	.		.
n	y_n	1	x_{n1}	x_{n2}	...	x_{np}

Sources: Ahmad et al., 2016a; 2016b

Mohamad Shafiq is a postgraduate student in the School of Dental Sciences. Email him at shafiqmat786@gmail.com. Dr. Amir is an Associate Professor of Biostatistics. Email him at wmamir@usm.my.

It is used when there are two or more independent variables and a single dependent variable where the equation below shows the model population information:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + L + \beta_k x_{ki} + \varepsilon_i \quad (1)$$

where

β_0 is the intercept parameter, and

$\beta_0, \beta_1, \beta_2, \dots, \beta_{k-1}$ are the parameters associated with $k - 1$ predictor variables.

The dependent variable Y is written as a function of k independent variables, $x_1, x_2, \dots, x_k$. A random error term is added to equation as to make the model more probabilistic rather than deterministic. The value of the coefficient β_i determines the contribution of the independent variables x_i , and β_0 is the y -intercept (Ahmad et al., 2016a; 2016b). The coefficients $\beta_0, \beta_1, \dots, \beta_k$ are usually unknown because they represent population parameters. Below is the data presentation for multiple linear regression. A general linear model in matrix form can be defined by the following vectors and matrices as:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Robust Regression

Robust regression is a method used when the distribution of the residual is not normally distributed and there are some outliers which affect the model (Susanti et al., 2014). It detects the outliers and provides better results (Chen, 2002). A common method of robust regression is the M estimate, introduced by Huber (1973), which is as efficient as Ordinary Least Square (OLS), and is considered the simplest approach. The Least Trimmed Squares (LTS) estimation was introduced by Rousseeuw (1984), and is a high breakdown value method. So, too, is the S estimation, another high breakdown value method with a higher statistical

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

efficiency than LTS estimation (Rousseeuw & Yohai, 1984). The S estimation is used to minimize the dispersion of residuals. The MM estimation, a special type of M estimation introduced by Yohai (1987), combines high breakdown value estimation and efficient estimation. The M estimation has a higher breakdown value and greater statistical efficiency than the S estimation.

Calculation for linear Regression using SAS

```
/* First do a simple linear regression */
proc reg data = temp1;
model y = x;
run;

/* Compute the absolute and squared residuals*/
data temp1.resid;
set temp1.pred;
absresid=abs(residual);
sqresid=residual**2;

/* Run a Regression with the absolute residuals and squared residuals */
/* to get estimated standard deviation and estimated variance */
proc reg data=temp1.resid;
model absresid=x;
output out=temp1.s_weights p=s_hat;

model sqresid=x;
output out=temp1.v_weights p=v_hat;

/* Compute weight using standard deviation */
data temp1.s_weights;
set temp1.s_weights;
s_weight=1/(s_hat**2);
label s_weight = "weights using absolute residuals";

/* Compute weight using variances */
data temp1.v_weights;
set temp1.v_weights;
v_weight=1/v_hat;
```

```

label v_weight = "weights using squared residuals";

/* Run a Weighted Least Square using estimated Standard Deviation */
/* and Variances */
proc reg data=temp1.s_weights;
weight s_weight;
model y = x;
run;

proc reg data=temp1.v_weights;
weight v_weight;
model y = x;
run;

/* Approach the Estimation Method Procedure for Robust Regression */
/* in this case, using the four methods LTS, M, MM and S-estimation */
proc robustreg data = temp1 method =LTS;
model y = x;
run;

```

An Illustration of a Medical Case

A case study of triglycerides will illustrate the different methods for robust regression.

Table 1. Description of the variables

Variables	Code	Description
Triglycerides	Y	Triglycerides level of patients (mg/dl)
Weight	X1	Weight (kg)
Total Cholesterol	X2	Total cholesterol of patients (mg/dl)
Proconvertin	X3	Proconvertin (%)
Glucose	X4	Glucose level of patients (mg/dl)
HDL-Cholesterol	X5	High density lipoprotein cholesterol (mg/dl)
Hip	X6	Hip circumference (cm)
Insulin	X7	Insulin level of patients (IU/ml)
Lipid	X8	Taking lipid lowering medication (0 = no, 1= yes)

Sources: Ahmad & Shafiq, 2013; Ahmad et al., 2014

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Algorithm for Weighting Multiple Linear Model Regression by different Robust Regression Methods

Title 'Alternative Modeling on Weighting Multiple linear regression';

Data Medical;

input Y X1 X2 X3 X4 X5 X6 X7 X8;

Datalines;

168	85.77	209	110	114	37	130.0	17	0
304	58.98	228	111	153	33	105.5	28	1
72	33.56	196	79	101	69	88.5	6	0
119	49.00	281	117	95	38	104.2	10	1
116	38.55	197	99	110	37	92.0	12	0
87	44.91	184	131	100	45	100.5	18	0
136	48.09	170	96	108	37	96.0	13	1
78	69.43	163	89	111	39	103.0	8	0
223	47.63	195	177	112	39	95.0	15	0
200	55.35	218	108	131	31	104.0	33	1
159	59.66	234	112	174	55	114.0	14	0
181	68.97	262	152	108	44	114.5	20	1
134	51.49	178	127	105	51	100.0	21	0
162	39.69	248	135	92	63	93.0	9	1
96	56.58	210	122	105	56	103.4	6	0
117	63.48	252	125	99	70	104.2	10	0
106	66.70	191	103	101	32	103.3	16	0
120	74.19	238	135	142	50	113.5	14	1
119	60.12	169	98	103	33	114.0	13	0
116	36.60	221	113	88	60	94.3	11	1
109	56.40	216	128	90	49	107.1	13	0
105	35.15	157	114	88	35	95.0	12	0
88	50.13	192	120	100	54	100.0	11	0
241	56.49	206	137	148	79	113.0	14	1
175	57.39	164	108	104	42	103.0	15	0
146	43.00	209	116	93	64	97.0	13	0
199	48.04	219	104	158	44	97.0	11	0
85	41.28	171	92	86	64	95.4	5	0
90	65.79	156	80	98	54	98.5	11	1
87	56.90	247	128	95	57	106.3	9	0
103	35.15	257	121	111	69	89.5	13	0

121	55.12	138	108	104	36	109.0	13	0
223	57.17	176	112	121	38	114.0	32	0
76	49.45	174	121	89	47	101.0	8	0
151	44.46	213	93	116	45	99.0	10	1
145	56.94	228	112	99	44	109.0	11	0
196	44.00	193	107	95	31	96.5	12	0
113	53.54	210	125	111	45	105.5	19	0
113	35.83	157	100	92	55	95.0	13	0

;

Run;

ods rtf file='result_ex1.rtf' ;

```
/* This first step is to make the selection of the data that have a
significant impact on triglyceride levels. The next step is to perform
the procedure of modeling linear regression model and run the regression
to get the residuals*/
```

```
proc reg data= Medical;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
output out=work.pred r=residual;
run;
```

```
/* Compute the Absolute and Squared Residuals*/
```

```
data work.resid;
set work.pred;
absresid=abs(residual);
sqresid=residual**2;
```

```
/* Run a Regression Compute the Absolute and Squared Residuals to Get
Estimated Standard Deviation and Variances*/
```

```
proc reg data=work.resid;
model absresid=X1 X2 X3 X4 X5 X6 X7 X8;
output out=work.s_weights p=s_hat;

model sqresid=X1 X2 X3 X4 X5 X6 X7 X8;
output out=work.v_weights p=v_hat;
run;
```

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

```
/* Compute the Weight Using Estimated Standard Deviation and Variances*/
data work.s_weights;
set work.s_weights;
s_weight=1/(s_hat**2);
label s_weight = "weights using absolute residuals";

data work.v_weights;
set work.v_weights;
v_weight=1/v_hat;
label v_weight = "weights using squared residuals";

/* Do a Weighted Least Squares Using the Weight from the Estimated
Standard Deviation*/
proc reg data=work.s_weights;
weight s_weight;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
run;

/* Do a Weighted Least Squares Using the Weight from the Estimated
Variances*/
proc reg data=work.v_weights;
weight v_weight;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
run;

/* Do Robust Regression, a Four Estimation Method to compare which are
LTS, M, MM and S-Estimation For Weighted Least Square using estimated
Standard Deviation*/
proc robustreg method=LTS data=work.s_weights;
weight s_weight;
model Y = X1 X2 X3 X4 X5 X6 X7 X8 / diagnostics leverage;
run;

/* Do a Robust Regression, a Four Estimation Method compare which are
LTS, M, MM and S-Estimation For Weighted Least Square using estimated
Variances*/
proc robustreg method=LTS data=work.v_weights;
weight v_weight;
```

```
model Y = X1 X2 X3 X4 X5 X6 X7 X8 / diagnostics leverage;
run;
```

Results

Compiled in Table 2 are the results from the multiple regression analysis using the original data. Compiled in Table 3 are the results for the weighted least square by standard deviation and weighted least square by variance. The residual plots do not indicate any problem with the model, as can be seen in Figures 1-3. A normal distribution appears to fit the sample data fairly well. The plotted points form a reasonably straight line. In our case, the residual plots bounce randomly around the 0 line (residual vs. predicted value). This supports the reasonable assumption that the relationship is linear.

Table 2. Parameter Estimates for Original Data

Variables	Parameter Estimate	Standard Error	P value
Intercept	-86.56544	102.93662	0.4070
X1	-1.08598	0.95288	0.2634
X2	-0.06448	0.21973	0.7712
X3	0.61857	0.36615	0.1015
X4	1.10882	0.33989	0.0028
X5	-0.52289	0.57119	0.3673
X6	0.81327	1.38022	0.5601
X7	2.77339	1.25026	0.0343
X8	22.40585	14.51449	0.1331

Table 3. Parameter Estimates for Weighted Multiple Linear Regression

Variables	Weighted Least Square MLR (SD)			Weighted Least Square MLR (V)		
	Parameter Estimate	Standard Error	P value	Parameter Estimate	Standard Error	P value
Intercept	-150.25787	90.05385	0.1056	-139.33900	90.60374	0.1353
X1	-1.30694	0.59423	0.0357	-1.19482	0.68833	0.0936
X2	-0.01586	0.17670	0.9291	0.05784	0.19730	0.7716
X3	0.44460	0.35706	0.2227	0.36626	0.44451	0.4169
X4	0.89106	0.38240	0.0267	1.01359	0.37253	0.0111
X5	-0.23352	0.44853	0.6064	-0.24328	0.52342	0.6457
X6	1.74405	1.10677	0.1256	1.35688	1.20057	0.2680
X7	2.81731	1.29607	0.0377	3.17543	1.31793	0.0228
X8	16.87506	10.34963	0.1135	15.78743	12.16151	0.2048

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Figure 1. Fit Diagnostic for y

Figure 2. Fit Diagnostic for y-weighted least square using standard deviation

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Figure 3. Fit Diagnostic for y-weighted least square using variances

Shown in Table 2 are the variables x_4 ($p = 0.0028$) and x_7 ($p = 0.0343$) were statistically significant for the multiple regression analysis. Shown in Table 3 are the variables x_1 ($p = 0.0357$), x_4 ($p = 0.0267$) and x_7 ($p = 0.0377$), which were statistically significant for weighted least square by standard deviation. The weighted least square by variance model shows the variable x_4 ($p = 0.0111$) and x_7 ($p = 0.0028$). RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data, which are to observe how close the data points are to the model predicted values. Lower value of RMSE indicated a better fit. The RMSE for weighted least square by variance (1.08) shows a lower value compared to the weighted least square standard deviation (1.31) and multiple regression (36.4). A higher R-squared value indicated how well the data fit the model and also indicates a better model. The model multiple regression analysis has R-squared of 0.62, weighted standard deviation multiple regression has R-squared of 0.67 and weighted variance multiple regression has R-squared of 0.63.

Shown in Table 4 is a comparison of the models—multiple linear regression (model 1), weighted least square by standard deviation (model 2) and weighted least square by variance (model 3)—using the four different robust methods, which are M estimation, LTS estimation, S estimation and MM estimation. The LTS estimation has high R-squared in three of the models compared to other robust methods. The S estimation also has high R-squared compared to MM and M estimation.

Table 4. Comparison of Model by using different Robust Method

Method	Model 1			Model 2			Model 3		
	Outlier	Leverage	R^2	Outlier	Leverage	R^2	Outlier	Leverage	R^2
<i>M</i>	0.0000	0.2051	0.4662	0.0769	0.2051	0.5761	0.1622	0.1892	0.5090
<i>LTS</i>	0.1282	0.2051	0.7289	0.1282	0.2051	0.7289	0.1351	0.1892	0.7032
<i>S</i>	0.0000	0.2051	0.5230	0.0000	0.2051	0.6079	0.0000	0.1892	0.5232
<i>MM</i>	0.0000	0.2051	0.4602	0.0000	0.2051	0.5843	0.0000	0.1892	0.5214

From Figure 4-6 there is a detection of outlier in observations. They present a regression diagnostics plot (a plot of the standardized residuals of robust regression LTS versus the robust distance). As indicated in Figure 4 and 5, observation 37 is identified as outlier. The observations of 2, 9, 24, and 27 are identified as outlier and leverage. Observations 10, 18 and 33 are identified as leverage point. In Figure 6, observation 35 is identified as outlier, observations 2, 8, 23, and 26 are identified as outlier and leverage, and observations 10, 17 and 27 are identified as leverage. The leverage plots available in SAS software are considered useful and effective in detecting multicollinearity, non-linearity, significance of the slope, and outliers (Lockwood & Mackinnon, 1998).

ALGORITHM FOR COMPARISON OF ROBUST REGRESSION METHODS

Figure 4. Outlier and Leverage Diagnostic for Y using LTS (Model 1)

Figure 5. Outlier and Leverage Diagnostic for Y using LTS (Model 2)

Figure 6. Outlier and Leverage Diagnostic for Y using LTS (Model 3)

Conclusion

SAS code for four different methods of robust regression was considered: M estimation, LTS estimation, S estimation, and MM estimation. They provide a better understanding of the weighted multiple linear regression and different robust method underlying of relative contributions.

Acknowledgements

The authors would like to express their gratitude to Universiti Sains Malaysia for providing research funding (Grant 304/PPSG/61313187, School of Dental Sciences, Universiti Sains Malaysia).

References

Ahmad, W. M. A. W., & Shafiq, M. (2013). High density lipoprotein cholesterol predicts triglycerides level in three distinct phases of blood pressure. *International Journal of Sciences: Basic and Applied Research*, 10(1), 38-46. Retrieved from <http://gssrr.org/index.php?journal=JournalOfBasicAndApplied&page=article&op1=view&path%5B%5D=1111&path%5B%5D=1098>.

Ahmad, W. M. A. W., Shafiq, M., Nurfadhlin, H., & Nor Azlida, A. (2014). A study of triglycerides level in three distinct phases of human blood pressure: A case study from previous projects. *Applied Mathematical Sciences*, 8(46), 2289-2305. doi: [10.12988/ams.2014.42145](https://doi.org/10.12988/ams.2014.42145).

Ahmad, W. M. A. W., Shafiq, M. M. I., Hanafi A. R., Puspa, L., & Nor Azlida, A. (2016a). Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression. *Journal of Modern Applied Statistical Methods*, 15(1), 884-892. doi: [10.22237/jmasm/1462077900](https://doi.org/10.22237/jmasm/1462077900)

Ahmad, W. M. A. W., Arif, M. A., Nor Azlida, A., & Shafiq, M. (2016b). An Alternative Method for Multiple Linear Model Regression Modeling, a Technical Combining of Robust, Bootstrap and Fuzzy Approach. *Journal of Modern Applied Statistical Methods*, 15(2), 743-754. doi: [10.22237/jmasm/1478004120](https://doi.org/10.22237/jmasm/1478004120).

Chen, C. (2002). Robust Regression and Outlier Detection with the ROBUSTREG Procedure. *Proceedings of the 27th SAS Users Group International Conference*. Cary NC: SAS Institute, Inc.

Huber, P.J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5), 799-821. doi: [10.1214/aos/1176342503](https://doi.org/10.1214/aos/1176342503)

Linear Regression. (n.d.) In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Linear_regression#Simple_and_multiple_regression.

Lockwood, C. M. & Mackinnon, D. P. (1998). Bootstrapping the standard 7 error off the mediated effect. *Proceedings of the 23rd Annual Meeting of SAS 8 Users Group International*. Cary, NC: SAS Institute, Inc.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388), 871-880. doi: [10.2307/2288718](https://doi.org/10.2307/2288718)

Rousseeuw, P. J. and Yohai, V. (1984). Robust Regression by Means of S estimators. In J. Franke, W. Härdle, and R. D. Martin, Eds. *Robust and Nonlinear Time Series Analysis (Lecture Notes in Statistics 26)*. New York: Springer Verlag, pp. 256-274. doi: [10.1007/978-1-4615-7821-5_15](https://doi.org/10.1007/978-1-4615-7821-5_15)

Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *Journal of the American Statistical Association*, 88(421), 237-244. doi: [10.1080/01621459.1993.10594315](https://doi.org/10.1080/01621459.1993.10594315)

Susanti, Y., Pratiwi, H., Sulistijowati, S., & Liana, T. (2014). M Estimation, S Estimation, and MM estimation in Robust Regression. *International Journal of Pure and Applied Mathematics*, 91(3), 349-360. doi: [10.12732/ijpam.v91i3.7](https://doi.org/10.12732/ijpam.v91i3.7)

Yohai V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *Annals of Statistics*, 15(2), 642-656. doi: [10.1214/aos/1176350366](https://doi.org/10.1214/aos/1176350366)

JMASM 47: ANOVA_HOV: A SAS Macro for Testing Homogeneity of Variance in One-Factor ANOVA Models (SAS)

Isaac Li
University of South Florida
Tampa, FL

Yi-Hsin Chen
University of South Florida
Tampa, FL

Yan Wang
University of South Florida
Tampa, FL

Patricia Rodríguez de Gil
University of South Florida
Tampa, FL

Thanh Pham
University of South Florida
Tampa, FL

Diep Nguyen
University of South Florida
Tampa, FL

Eun Sook Kim
University of South Florida
Tampa, FL

Jeffrey D. Kromrey
University of South Florida
Tampa, FL

Variance homogeneity (HOV) is a critical assumption for ANOVA whose violation may lead to perturbations in Type I error rates. Minimal consensus exists on selecting an appropriate test. This SAS macro implements 14 different HOV approaches in one-way ANOVA. Examples are given and practical issues discussed.

Keywords: ANOVA, homogeneity, SAS

Introduction

The analysis of variance (ANOVA) is regarded as one of the most powerful and flexible methods for testing hypotheses about population means. When the underlying assumptions are satisfied, ANOVA provides powerful statistical tests, but if they are not satisfied, Type I error rates are inflated and statistical power reduced. A critical assumption of ANOVA is homogeneity of variance (HOV), that is, the compared populations have equal variances, which can be expressed as

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2. \quad (1)$$

Isaac Li is a graduate student of Educational Measurement and Research. Email him at liy1@mail.usf.edu.

Although it was found that moderate deviations from the HOV assumption might not seriously affect the results in ANOVA (Glass, Peckham, & Sanders, 1972), there remains the concern about large violations of HOV that can lead to invalid inferences. In fact, careful attention has been paid to this issue, and numerous procedures have been developed for the purpose of evaluating the HOV assumption although each of these procedures has limitations. For example, the classic F test has been commonly applied, but was found to be sensitive to departures from normality. Similarly, Snedecor and Cochran (1989) discovered that Bartlett's test (Bartlett, 1937) is sensitive to violations of the normality assumption. Therefore, it is advisable to turn to alternative tests that are less dependent on the assumption of normality.

Cases of these alternative approaches which are relatively insensitive to departures from normality include the Levene (Levene, 1960), Brown-Forsythe (BF; Brown & Forsythe, 1974), and O'Brien (OB; O'Brien, 1979, 1981) tests. Simulation studies (Algina, Olejnick, & Ocanto, 1989; Conover, Johnson, & Johnson, 1981; Lee, Katz, & Restori, 2010; Olejnick & Algina, 1987) have revealed differences in performance among these tests. For example, the OB test provided Type I error rates near the nominal alpha in unbalanced samples, but with platykurtic distributions which exhibit lighter tails than a normal distribution, it was more sensitive to variance differences than the BF test. When sample sizes were equal, OB had a power advantage used with platykurtic distributions and had a slight power advantage when used with mesokurtic distributions (i.e., exhibiting same tail weight as a normal distribution) regardless of whether the distributions were skewed or not. On the other hand, the BF test had a power advantage used with leptokurtic distributions (i.e., having heavier tails than a normal distribution) regardless of the skewness. When sample sizes were unequal, results were different because the relative power of these tests depended on the direction of the relationship between the population variances and sample sizes. For example, the power of the OB test was improved when used with skewed-platykurtic and symmetric platykurtic distributions. The power of BF was also enhanced when the relationship between population variance and sample size was direct (i.e. larger samples come from populations with larger variances) and with leptokurtic or mesokurtic distributions. With other distributions, the tests had similar power.

The appropriate test should be selected in the context of their research questions, sample data, and analysis plan. It would be difficult to reach a consensus in answering the question what method should be applied for testing the HOV assumption, because every study is different and no procedure is

superior to the rest in an absolute sense. Furthermore, some testing procedures are less known, or how to implement them, which makes it harder to be selected as a suitable HOV method. The goal of the current study is two-fold: 1) introduce fourteen different approaches that can be used to evaluate the HOV assumption and 2) provide a SAS macro designed to implement these statistical tests.

Statistical Methods for Testing the HOV Assumption

Bartlett Test. Bartlett (1937) proposed a special use of the chi-square test for testing the HOV assumption. The computation of the Bartlett's chi-square is as follows:

$$\chi^2 = \frac{(N-k) \log \left[\frac{\sum_{j=1}^k (n_j - 1) S_j^2}{(N-k)} \right] - \sum_{j=1}^k (n_j - 1) \log(S_j^2)}{1 + \frac{\left(\sum_{j=1}^k \frac{1}{n_j - 1} \right) - \frac{1}{(N-k)}}{3(k-1)}}, \quad (2)$$

where N is total sample size, n_j is the group j sample size, k refers to number of groups, and S_j^2 denotes group j variance. The null hypothesis that the subpopulation variances are equal will be rejected if Bartlett's χ^2 statistic is greater than the critical χ^2 value with degrees of freedom $k - 1$.

Levene's Test (Absolute and Squared). To overcome departures from normality, Levene (1960) proposed to use the absolute values or squared residuals, which transforms a test of variances into a test of means that is relatively robust to the violation of the normality assumption (Sayago & Asuero, 2004). The statistics of the absolute values and squared residuals used for the Levene's test are defined as follows (Katz, Restori, & Lee, 2009):

$$Z_{ij} = |Y_{ij} - \bar{Y}_{.j}| \text{ and } Z_{ij}^2 = (Y_{ij} - \bar{Y}_{.j})^2, \quad (3)$$

where Y_{ij} is the raw score and $\bar{Y}_{.j}$ denotes the mean of the j^{th} group.

The test statistic, W , which is approximately distributed like F , is defined as:

$$W = \frac{(N-k) \sum_{j=1}^k n_j (\bar{Z}_{.j} - \bar{Z}_{..})^2}{(k-1) \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{Z}_{ij} - \bar{Z}_{.j})^2}, \quad (4)$$

where $\bar{Z}_{.j}$ represents the group mean of Z_{ij} and $\bar{Z}_{..}$ is the grand mean. The obtained W value is compared to the F critical value (F_{crit}) with degrees of freedom $(k-1)$ and $(N-k)$ for numerator and denominator, respectively. The null hypothesis that the group variances are equal is rejected if $W > F_{crit}$.

Brown-Forsythe Test. This test follows the idea of Levene's test but uses the group median instead of the group mean in calculation of absolute deviation. This test first transforms the dependent variable values into absolute values of deviations from the group median:

$$z_{ij} = |Y_{ij} - \tilde{Y}_j| \quad (5)$$

where $\tilde{Y}_j$ is the median of group j and z_{ij} refers to the transformed value of Y_{ij} .

The obtained statistic W is computed using the same formula as that in the Levene's Test presented above. The critical values of W are obtained from the F table with $(k-1)$ and $(N-k)$ degrees of freedom. If the p -value $< \alpha$, H_0 is rejected and we conclude that the variances are not equal. The Brown-Forsythe test is more robust than Levene's test when the population distribution is skewed (Brown & Forsythe, 1974).

O'Brien's Test. O'Brien (1979) proposed a method that transforms original scores so they would represent sample variances. The transformed scores are used for ANOVA or the Welch test as a dependent variable. The transformation he proposed is the weighted average of a modified Levene's squared difference that is computed as

$$r_{ij}(w) = \frac{(w + n_j - 2)n_j(Y_{ij} - \bar{Y}_{.j})^2 - wS_j^2(n_j - 1)}{(n_j - 1)(n_j - 2)}, \quad (6)$$

where S_j^2 represents the within-group unbiased estimate of variance for sample j and w ($0 \leq w \leq 1$) denotes the weighting factor.

The mean of the transformed values per group, $\bar{r}_j = \frac{\sum r_{ij}}{n_j} = S_j^2$, will equal the variance computed for that group. The weighted average, $r_{ij}(w)$ is a modification of Levene's squared difference from the group mean ($w = 0$), and a jackknife pseudo value of $S_j^2(w = 1)$. O'Brien (1981) suggested setting $w = .5$ as default. Then the F statistic computed on r_{ij} in the standard ANOVA procedure would be the test statistic for O'Brien's test.

Bootstrap Brown-Forsythe Test. Boos and Brownie (2004) recommended a bootstrap approach for comparing variances where power outweighs computational simplicity. The test they investigated in earlier work (Boos & Brownie, 1989) focused on drawing bootstrap samples from

$$\bar{S} = \{Y_{ij} - \bar{Y}_{.j}\}, \quad (7)$$

which will simply be the residuals from the original sample. Instead of drawing a separate bootstrap sample from each of the groups, bootstrapping is performed by pooling the residuals across groups. In each bootstrap sample, a test statistic for variances is computed and the p -value for the bootstrap test is obtained as the proportion of bootstrap samples with a statistic's value that is greater than that observed in the original data. The authors recommended using the median version of Levene's test statistic (that is, the Brown-Forsythe statistic), then obtaining the p -value via the bootstrap rather than the F distribution. This recommendation was also made by Lim and Loh (1996): their simulation study found that the bootstrapped version of Levene's test provided more power than the F distribution version.

Ramsey Conditional Test: Brown-Forsythe or O'Brien. Ramsey (1994) proposed a conditional procedure based on the Brown-Forsythe method and the O'Brien method. He suggested the appropriate test between the two methods should be selected conditional on a test of kurtosis. To decide which procedure to use, the kurtosis for each of k groups is computed by using Pearson's traditional measure:

$$b_2 = \frac{m_4}{m_2^2}, \quad (8)$$

where $m_r = \Sigma(Y_{ij} - \bar{Y}_{.j})^r / n_j$ and m_2 is the second moment about the mean or the biased sample variance. Thus, the formula for kurtosis of group j , b_{2j} , is

$$b_{2j} = \frac{\frac{\sum (Y_{ij} - \bar{Y}_{.j})^4}{n_j}}{\left[\frac{\sum (Y_{ij} - \bar{Y}_{.j})^2}{n_j} \right]^2}. \quad (9)$$

The kurtosis value for each group is then compared to critical values obtained from a table provided by Ramsey and Ramsey (1993). A score of -1 , 0 , or 1 is recorded depending on the test being significantly platykurtic, nonsignificant, or significantly leptokurtic, respectively. These scores of kurtosis were added across groups as a total score, S , which was used to identify the population as platykurtic if $S \leq -1$, mesokurtic if $S = 0$, or leptokurtic if $S \geq 1$. The O'Brien method will be applied if the data are platykurtic and the Brown-Forsythe method will be used if the data are mesokurtic or leptokurtic.

Cochran's Test (Arithmetic and Harmonic). Cochran's C (Cochran, 1941) is essentially an outlier test that defines a test statistic as the ratio of the largest group variance to the sum of the sample variances:

$$C = \frac{S_{\max}^2}{\sum S_j^2} \quad (10)$$

Critical values of Cochran's C can be obtained from the central F distribution:

$$\text{Critical } C = \frac{1}{1 + \frac{k-1}{F_{\frac{\alpha}{k}, n-1, (k-1)(n-1)}}}, \quad (11)$$

where n is the number of observations in each group (for the balanced design) and F refers to the critical value of F at α/k with $df = n - 1$, $(k - 1)(n - 1)$. For an unbalanced design one could either use an arithmetic average of n or the harmonic

mean of n to calculate degrees of freedom in the numerator. If the obtained value of C exceeds the critical value, the null hypothesis of variance homogeneity is rejected.

G Test. As an alternative to Cochran's C , 't Lam (2010) suggested using the G test which is a ratio of the product of the largest variance and its degrees of freedom to the sum of the products of each variance and its degrees of freedom:

$$G = \frac{v_{\max} S_{\max}^2}{\sum v_j S_j^2} \quad (12)$$

Critical values of the G test can be obtained from the central F distribution:

$$\text{Critical } G = \frac{1}{1 + \frac{v_{\text{pool}} / (v_{\max} - 1)}{F_{\frac{\alpha}{k}, \bar{n}_j - 1, (k-1)(\bar{n}_j - 1)}}}, \quad (13)$$

where v_{pool} is the pooled degrees of freedom; $v_{\max}$ denotes the degrees of freedom for the group with the largest variance; $\bar{n}_j$ is the mean number of observations in each group; $F_{\alpha/k}$ denotes the critical value of F at α/k with $df = \bar{n}_j - 1, (k-1)(\bar{n}_j - 1)$.

The Maximum F-ratio Test ($F_{\max}$) (Arithmetic and Harmonic). Hartley (1950) developed the $F_{\max}$ test (a.k.a. Hartley's test) for comparing three or more population variances. Hartley's test requires independent random samples of the same size from normally distributed populations (Ott & Longnecker, 2010). The $F_{\max}$ test is computed by taking the ratio of the largest group variance ($S_{\max}^2$) to the smallest group variance ($S_{\min}^2$) and its formula is presented as follows:

$$F_{\max} = \frac{S_{\max}^2}{S_{\min}^2}. \quad (14)$$

The value of F_{max} is compared to a critical value from the table containing the sampling distribution of Hartley's test. A F_{max} value less than the critical value provides evidence that the groups have similar or equal variances. Similar to the Cochran's C test for an unbalanced design, an arithmetic average of the sample size or a harmonic mean can be used to calculate degrees of freedom in the numerator.

Z-Variance Test.

Overall and Woodward (1974) proposed the Z-variance test for the HOV assumption. This test is an approximate test based on Fisher and Yates (1963), which transforms the chi-square statistics with large degrees of freedom (df) into standard normal deviate z -scores. Fisher and Yates' (1963) formula provided a z -score transformation for chi-square statistics based on large df :

$$z = \sqrt{2\chi^2} - \sqrt{2(df)-1}. \quad (15)$$

Because sample variance estimates have a chi-square distribution,

$$\chi_{n-1}^2 = \frac{(n-1)S^2}{\sigma^2}, \quad (16)$$

where S^2 is the sample variance estimate; σ^2 is the true population variance. A z -score transformation of within-cell variance estimates that appears adequate for samples as small as $n = 10$ or larger is then computed as

$$Z_j = \sqrt{\frac{c(n_j-1)S_j^2}{MS_w}} - \sqrt{c(n_j-1)-1}, \quad (17)$$

where $c = 2 + 1/n_j$, n_j is the sample size for the j^{th} group; S_j^2 denotes unbiased variance estimate for the j th group, and MS_w denotes the pooled within-group mean square (or MS_E as the pooled within-group error variance).

The F -ratio with $k-1$ and ∞ degrees of freedom for testing the homogeneity of variance of k samples variance is given as

$$F = \frac{\sum_{j=1}^k Z_j^2}{k-1}. \quad (18)$$

Note the division of the sum of squared z-scores by $k - 1$, as the estimate of σ^2 . The right-hand side of the equation can be conceived as a between groups sum of squares divided by $df = k - 1$. Overall and Woodward (1974) found that the Z-variance test performed very well with normally distributed data but produced too many Type I errors when samples were from leptokurtic or skewed distributions.

Modified Z-variance Test. To improve the performance of the Z-variance test when sample distributions are leptokurtic or skewed, Overall and Woodward (1976) proposed the modified Z-variance approach to testing the HOV. Based on their examination of the robustness of this modification to the z-variance against four other tests, in which the authors investigated a c value based on sample size, skewness and kurtosis, they determined c to be a scaling coefficient that affects the variability of the Z_j values. The formula for c is,

$$c = 2.0 \left(\frac{2.9 + .2 / n_j}{K} \right)^{1.6(n_j - 1.8K + 14.7)/n_j}, \quad (19)$$

where n_j is the sample size of the j^{th} group and K refers to the mean of the kurtosis indices from all groups. Overall and Woodward's kurtosis index is the 4th power of the Z-scores within each group divided by $n_j - 2$ degrees of freedom,

$$K = \frac{\sum Z_{ij}^4}{n_j - 2}, \quad (20)$$

where

$$Z_{ij} = \frac{Y_{ij} - \bar{Y}_{.j}}{\sqrt{\frac{n_j - 1}{n_j} S_j^2}}. \quad (21)$$

Methodology

ANOVA_HOV: A SAS Macro

ANOVA_HOV is a macro is written to implement the 14 procedures to test the homogeneity of variance assumption in one-way ANOVA models. Before the macro is defined, two datasets of critical values are created for the F_{max} test and Ramsey's conditional test (see [Appendix A](#) for SAS code). Critical values at alpha levels of .05 and .01 for the F_{max} test are included. To use the ANOVA_HOV macro researchers need to input their data set into SAS, identify variables, and then call the macro. The macro will run each of the fourteen HOV tests using the data set and input variables.

Macro input parameters. Use of the macro requires specification of four arguments for the macro call, which are briefly described below.

- data.* Specifies the name of the input data set.
- iv.* Specifies the name of the independent variable.
- dv.* Specifies the name of the dependent variable.
- nboots.* Specifies the number of bootstrap samples used in the Bootstrap Brown-Forsythe Test. A large number of replications are conducive to consistency in estimation.

Results

To illustrate the use of the macro, two examples are provided. Data in the examples were downloaded from the database of the Integrated Postsecondary Education Data System (IPEDS), a system of interrelated surveys conducted annually by the U.S. Department of Education's National Center for Education Statistics (NCES) ([NCES, 2014](#)). IPEDS gathers information from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. The latest wave of data available was collected in 2014 and only excerpts of the data were involved in the examples below.

For Example 1, the dependent variable was enrollment of Hispanic students and the HOV was tested on the independent variable, institutions of different categories. The four categories of institutions ("cat") were 1) degree-granting with primarily baccalaureate degrees or above; 2) degree-granting but less than 50% above bachelor's level; 3) degree-granting at levels of Associate's and

ANOVA_HOV: A SAS MACRO

certificates; and 4) non-degree-granting. The number of enrolled Hispanic students per institution (“sumef”) ranged from 0 to several thousands. To evaluate the macro in a situation of equal group sizes, a random sample of 10 institutions were drawn from each category to create a sample of 40 out of a total of over 7,000 institutions in the data set. The syntax below shows how the data were read into SAS to be analyzed.

```
data enrfrn;
  input cat sumef @@;
  datalines;
2 94 2 3 2 278 2 19 2 2 2 3 2 18 2 10 2 1 2 50
3 8 3 2 3 12 3 19 3 79 3 8 3 20 3 23 3 15 3 1
4 2582 4 983 4 743 4 2 4 4 4 19 4 13 4 339 4 197 4 4
5 10 5 20 5 1 5 153 5 1 5 60 5 1 5 3 5 21 5 4
;;;
```

The statement below invokes the macro by calling the variables.

```
%hov (data = enrfrn, iv = cat, dv = sumef, nboots = 5000);
```

Independent Variable:	cat	
N of Groups:	4	
Dependent Variable:	sumef	
Total N of Observations:	40	
	Obtained	
Test	Value	p
Brown and Forsythe	3.5791	.0231
Levene's (absolute values)	9.3372	.0001
Levene's (squared values)	1.9722	.1356
O'Brien	1.747	.1748
Bartlett	102.025	0
Bootstrap Brown-Forsythe	3.5791	.3402
Ramsey's Conditional Test	3.5791	.0231
Z-variance	20.0296	0
MZ-variance test	3.5384	.014
Cochran's C test (with arithmetic mean)	0.9848	p < .10
Cochran's C test (with harmonic mean)	0.9848	p < .10
G test	0.9848	p < .10
F-max test (with arithmetic mean)	1312.1533	p < .01
F-max test (with harmonic mean)	1312.1533	p < .01

Figure 1. Output for Example 1.

The output from the macro run is shown in Figure 1. The top paragraph gives a summary of the data while the rest of output includes the name of the

HOV tests, the obtained values for each test, and their associated p -values. Note that only comparisons with critical values are reported for the bottom five tests because their p -values are not calculated.

In the second example, the ANOVA_HOV macro was applied to a large data set from IPEDS with unequal group sizes. The independent variable was a classification of how the institution is operated (“control”): 1) public institution, 2) private not-for-private institution, and 3) private for-profit institution. Both of the first two categories included more than 1,000 institutions whereas there were fewer than 200 in the third group. The dependent variable in the analysis was total salaries and wages expenditures as part of the institution’s library budget (“lsalwag”), which has a mean of \$1,143,651. The data file in analysis contains nearly 3,000 observations and is named “libs1”. The macro was invoked using the statement below and the number of bootstraps was set at 5,000 once again.

```
%hov (data = libs1, iv = control, dv = lsalwag, nboots = 5000);
```

The output of this analysis is provided in [Figure 2](#).

Independent Variable:	control	
N of Groups:	3	
Dependent Variable:	lsalwag	
Total N of Observations:	2964	
Test	Obtained Value	p
Brown and Forsythe	12.0649	0
Levene's (absolute values)	20.569	0
Levene's (squared values)	1.0383	.3542
O'Brien	1.0376	.3544
Bartlett	713.5098	0
Bootstrap Brown-Forsythe	12.0649	.1590
Ramsey's Conditional Test	12.0649	0
Z-variance	163.616	0
MZ-variance test	24.4983	0
Cochran's C test (with arithmetic mean)	0.5688	p < .10
Cochran's C test (with harmonic mean)	0.5688	p < .10
G test	0.5317	p < .10
F-max test (with arithmetic mean)	158.3482	p < .01
F-max test (with harmonic mean)	158.3482	p < .01

Figure 2. Output for Example 2.

A Simulation Study

To investigate the accuracy and precision of the fourteen approaches for testing the homogeneity of variance assumption in one-way ANOVA models, a simulation study (Wang, Rodríguez de Gil, Chen, Kromrey, Kim, Pham, Nguyen,

& Romano, 2017) was conducted employing six design factors: number of groups, average number of observations per group, sample size pattern, variance pattern, maximum group variance ratio, and population distributions measured by skewness and kurtosis. The performance of the fourteen methods was examined at six nominal alpha levels (.01, .05, .10, .15, .20, and .25). For each condition, 5,000 samples were generated and Type I error and statistical power examined. Bradley's (1978) liberal criterion for robustness was set at $\pm 0.5\alpha$ around the nominal alpha and eta-square effect sizes were estimated to explore the significant impacts of the research design factors on the variability in the Type I error estimates. Lastly, statistical power was estimated only for conditions in which Type I error was adequately controlled.

Of the approaches, Levene's test with the squared residuals, Brown-Forsythe, O'Brien, Ramsey's conditional procedure, and Bootstrap Brown-Forsythe tests were the five tests that maintained adequate Type I error control and were superior to other tests. Meanwhile, the power of these five tests was found to be acceptable and the power differences between them were subtle. O'Brien's test had slightly less power than the other four tests and the Bootstrap Brown-Forsythe test tended to have slightly greater power.

Population shape and average group size were two design factors that affected Type I error rates. Specifically, Ramsey and O'Brien were the only two tests that maintained adequate Type I error control across all the population shapes investigated. Among the five best tests, Brown-Forsythe and Bootstrap Brown-Forsythe tests had conservative Type I error rates if the distribution shape was extremely leptokurtic. In contrast, Levene's test with the squared residuals tended to have liberal Type I error rates when the shape is skewed. With normal populations, 12 out of 14 approaches maintained adequate Type I error control. Bartlett, Cochran with the harmonic mean, and Z-variance tests maintained Type I error control extremely well. Levene's test with squared values, F_{max} with the harmonic mean, and Z-variance had larger power. Z-variance test seemed to be the best choice for a normal distribution. Average group size had significant impact on Type I error control for four of the five best performers including Ramsey, Brown-Forsythe, Bootstrap Brown-Forsythe, and Levene's test with squared values. Increasing the sample size of a group improved Type I error control; on the other hand, when the average per-group sample size falls to five, only O'Brien's test maintained adequate Type I error control, which improves for all tests as the nominal alpha level grew.

Conclusion

When examining the differences between two or more group means, ANOVA is among the most commonly used procedures. The assumptions of variance homogeneity and normality continue to be critical; in particular, violations of HOV may impact Type I error rates. This macro provides the researcher with the ability to apply fourteen different evaluation measures mainly for two purposes: first, for testing substantive hypotheses about population variances (e.g., do educational enrichment programs increase heterogeneity of student achievement scores?), and secondly, for assessing the tenability of homogeneity of variance in consideration of a subsequent test of mean differences. Different nominal alpha levels may be indicated for these uses.

HOV tests must be selected with care because it is difficult to assess population distribution shape based on samples, more so with small samples, which might result in severe consequences of non-normality for most of these tests. It is recommended that when average group is fewer than 10, the O'Brien test should be used because it gives a better Type I error control; for average group sizes between 10 and 20, the Ramsey test provides the most power; as for large groups with more than 20 observations, the Bootstrap Brown-Forsythe test will supply the most power.

References

- Algina, J., Olejnik, S., & Ocanto, R. (1989). Type I error rates and power estimates for selected two-sample tests of scale. *Journal of Educational and Behavioral Statistics*, 14(4), 373-384. doi: [10.3102/10769986014004373](https://doi.org/10.3102/10769986014004373).
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 160(901), 268-282. doi: [10.1098/rspa.1937.0109](https://doi.org/10.1098/rspa.1937.0109)
- Boos, D. D., & Brownie, C. (1989). Bootstraps methods for testing homogeneity of variances. *Technometrics*, 31(1), 69-82. doi: [10.1080/00401706.1989.10488477](https://doi.org/10.1080/00401706.1989.10488477).
- Boos, D. D., & Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statistical Science*, 19(4), 571-578. doi: [10.1214/088342304000000503](https://doi.org/10.1214/088342304000000503)
- Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x).

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346), 364-367. doi: [10.1080/01621459.1974.10482955](https://doi.org/10.1080/01621459.1974.10482955).

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Human Genetics*, 11(1), 47-52. doi: [10.1111/j.1469-1809.1941.tb02271.x](https://doi.org/10.1111/j.1469-1809.1941.tb02271.x).

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. doi: [10.1080/00401706.1981.10487680](https://doi.org/10.1080/00401706.1981.10487680).

Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological agricultural and medical research*. New York: Hafner.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, 42(3), 237-288. doi: [10.3102/00346543042003237](https://doi.org/10.3102/00346543042003237)

Hartley, H. O. (1950). The maximum f-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, 37(3/4), 308-312. doi: [10.2307/2332383](https://doi.org/10.2307/2332383).

Katz, G. S., Restori, A. F., & Lee, H. B. (2009). A Monte Carlo study comparing the Levene test to other homogeneity of variance tests. *North American Journal of Psychology*, 11, 511-522.

't Lam, R. U. E. (2010). Scrutiny of variance results for outliers: Cochran's test optimized. *Analytica Chimica Acta*, 659(1-2), 68-84. doi: [10.1016/j.aca.2009.11.032](https://doi.org/10.1016/j.aca.2009.11.032)

Lee, H. B., Katz, G. S., & Restori, A. F. (2010). A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics*, 6(3), 359-366. doi: [10.3844/jmssp.2010.359.366](https://doi.org/10.3844/jmssp.2010.359.366)

Levene, H. (1960). Robust tests for the equality of variance. In I. Olkin, Ed. *Contributions to Probability and Statistics*, pp. 278-292. Palo Alto, CA: Stanford Univ. Press.

Lim, T. S., & Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis*, 22(3), 287-301. doi: [10.1016/0167-9473\(95\)00054-2](https://doi.org/10.1016/0167-9473(95)00054-2)

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 74(368), 877-880. doi: [10.1080/01621459.1979.10481047](https://doi.org/10.1080/01621459.1979.10481047)

- O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, 89(3), 570-574. doi: 10.1037/0033-2909.89.3.570
- Olejnik, S. F., & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics*, 12, 45-61. doi: 10.2307/1164627
- Ott, L. R., & Longnecker, M. (2011). *An introduction to statistical methods and data analysis* (6th Ed.). Belmont, CA: Brooks/Cole, Cengage Learning.
- Overall, J. E., & Woodward, J. A. (1974). A simple tests for heterogeneity of variance in complex factorial designs. *Psychometrika*, 39(3), 311-318. doi: 10.1007/bf02291705.
- Overall, J. E., & Woodward, J. A. (1976). *A robust and powerful test for heterogeneity of variance*. Galveston, TX: University of Texas Medical Branch Psychometric Laboratory.
- Ramsey, P. H., & Ramsey, P. P. (1993). Updated version of the critical values of the standardized fourth moment. *Journal of Statistical Computation and Simulation*, 44(3-4), 231-241. doi: 10.1080/00949659308811460
- Ramsey, P. H. (1994). Testing variances in psychological and educational research. *Journal of Educational & Behavioral Statistics*, 19(1), 23-42. doi: 10.3102/10769986019001023.
- Sayago, A., & Asuero, A. G. (2004). Fitting straight lines with replicated observations by linear regression: Part II. Testing for homogeneity of Variances. *Clinical Reviews in Analytical Chemistry*, 34(3-4), 133-146). doi: 10.1080/10408340490888599.
- Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2014). *The Integrated postsecondary education data system (IPEDS)*. [Data file and code book]. Retrieved from <https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx>.
- Wang, Y., Rodriguez de Gil, P., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Pham, T., Nguyen, D., & Romano, J. L. (2017). Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement* , 77(2), 305-329. doi: 10.1177/0013164416645162.

Appendix A: The SAS Macro ANOVO_HOV

```

data test;
  input group Y;
  datalines;
1      49.66
1      47.86
1      51.61
2      49.68
2      51.02
2      50.02
3      50.45
3      49.09
3      45.73
4      46.39
4      49.25
4      48.39
4      51.99
4      47.84
4      57.99
4      53.28
4      48.08
4      48.10
4      53.48
4      49.98
;;;
run;

* Critical values for the Fmax test;
data cv05a;
  input nuval c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12;
  cards;
2 39.0 87.5 142 202 266 333 403 475 550 626 704
3 15.4 27.8 39.2 50.7 62.0 72.9 83.5 93.9 104 114 124
4 9.60 15.5 20.6 25.2 29.5 33.6 37.5 41.1 44.6 48.0 51.4
5 7.15 10.8 13.7 16.3 18.7 20.8 22.9 24.7 26.5 28.2 29.9
6 5.82 8.38 10.4 12.1 13.7 15.0 16.3 17.5 18.6 19.7 20.7

```

```

7 4.99 6.94 8.44 9.70 10.8 11.8 12.7 13.5 14.3 15.1 15.8
8 4.43 6.00 7.18 8.12 9.03 9.78 10.5 11.1 11.7 12.2 12.7
9 4.03 5.34 6.31 7.11 7.80 8.41 8.95 9.45 9.91 10.3 10.7
10 3.72 4.85 5.67 6.34 6.92 7.42 7.87 8.28 8.66 9.01 9.34
12 3.28 4.16 4.79 5.30 5.72 6.09 6.42 6.72 7.00 7.25 7.48
15 2.86 3.54 4.01 4.37 4.68 4.95 5.19 5.40 5.59 5.77 5.93
20 2.46 2.95 3.29 3.54 3.76 3.94 4.10 4.24 4.37 4.49 4.59
30 2.07 2.40 2.61 2.78 2.91 3.02 3.12 3.21 3.29 3.36 3.39
60 1.67 1.85 1.96 2.04 2.11 2.17 2.22 2.26 2.30 2.33 2.36

```

```
;
```

```
data cv01a;
```

```
Input nuval c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12;
```

```
Cards;
```

```

2 199 448 729 1036 1362 1705 2063 2432 2813 3204 3605
3 47.5 85 120 151 184 216 249 281 310 337 361
4 23.2 37 49 59 69 79 89 97 106 113 120
5 14.9 22 28 33 38 42 46 50 54 57 60
6 11.1 15.5 19.1 22 25 27 30 32 34 36 37
7 8.89 12.1 14.5 16.5 18.4 20 22 23 24 26 27
8 7.50 9.9 11.7 13.2 14.5 15.8 16.6 17.9 18.9 19.8 21
9 6.54 8.5 9.9 11.1 12.1 13.1 13.9 14.7 15.3 16.0 16.6
10 5.85 7.4 8.6 9.6 10.4 11.1 11.8 12.4 12.9 13.4 13.9
12 4.91 6.1 6.9 7.6 8.2 8.7 9.1 9.5 9.9 10.2 10.6
15 4.07 4.9 5.5 6.0 6.4 6.7 7.1 7.3 7.5 7.8 8.0
20 3.32 3.8 4.3 4.6 4.9 5.1 5.3 5.5 5.6 5.8 5.9
30 2.63 3.0 3.3 3.4 3.6 3.7 3.8 3.9 4.0 4.1 4.2
60 1.96 2.2 2.3 2.4 2.4 2.5 2.5 2.6 2.6 2.7 2.7

```

```
;
```

```
data cv01;
```

```
Set cv01a;
```

```
Array cc[11] c2 - c12;
```

```
Do kval = 2 to 12;
```

```
    Cv01 = cc[kval-1];
```

```
    Output;
```

```
End;
```

```
Keep nuval kval cv01;
```

```
Run;
```

ANOVA_HOV: A SAS MACRO

```
Data cv05;
  Set cv05a;
  Array cc[11] c2 - c12;
  Do kval = 2 to 12;
    Cv05 = cc[kval-1];
    Output;
  End;
Keep nuval kval cv05;
Run;

proc sort data=cv01; by nuval kval;
proc sort data=cv05; by nuval kval;

data fmaxcv;
  merge cv05 cv01;
  by nuval kval;
run;

* Critical values for Ramsey's conditional test;
data cvw2;
  input cs lvalue05 uvalue05;
  cards;
4 -1.22 0.83
5 -1.14 1.50
6 -1.19 1.91
7 -1.23 2.15
8 -1.20 2.28
9 -1.20 2.33
10 -1.19 2.40
11 -1.19 2.39
12 -1.16 2.40
13 -1.17 2.34
14 -1.15 2.32
15 -1.13 2.31
16 -1.13 2.27
17 -1.11 2.23
18 -1.10 2.22
```

```

19 -1.09 2.19
20 -1.08 2.16
;
run;

options mprint minoperator noquotelenmax ;

%macro ANOVA_HOV (data=, iv=, dv=, nboots=);

* Prepare tables;
proc sort data = &data;
    by &iv;
proc means noprint data = &data;
    class &iv;
    var &dv;
    output out = cells mean = cellmn n = cs median = m50 kurt=cg2
var=vargrp;
data grand (drop=&iv _:) cells;
    set cells;
    if _type_ = 0 then output grand;
    else output cells;

data _null_;
    set grand;
    call symputx('mean',cellmn);
    call symputx('variance',vargrp);

proc means noprint data = cells;
    var cs;
    output out = total n = howmany mean=avg_cs max=max_cs min=min_cs
sum=sum_cs; run;
data _null_;
    set total;
    call symputx('n_groups',howmany);
    call symputx('avg_cs',avg_cs);
    call symputx('sum_cs',sum_cs);
run;

```

ANOVA_HOV: A SAS MACRO

```
%put total is &sum_cs;

data head;
  format labl $45. val $15. pr $15.;
  labl = "Independent Variable:"; val = "&iv"; pr = " "; output;
  labl = "N of Groups:"; val = put("&n_groups", $10.) ; pr = " ";
output;
  labl = "Dependent Variable:"; val = "&dv"; pr = " "; output;
  labl = "Total N of Observations:"; val = put("&sum_cs", $10.); pr =
" "; output;
  labl = " "; val = " "; pr = " "; output;
  labl = " "; val = "Obtained"; pr = " "; output;
  labl = "Test"; val = "Value"; pr = "p"; output;
run;

%do nn = 1 %to &n_groups;
  data _null_;
    set cells;
    if &nn = _n_;
      call symput("size&nn",cs);
%end;
run;

* Bartlett, BF, Leveneabs, Levenesquare, Obrien;
ods listing close;

proc glm data=&data ;
  class &iv;
  model &dv = &iv / ss3 ;
  means &iv / hovtest=bartlett;
  means &iv / hovtest=bf;
  means &iv / hovtest=levене(type=abs);
  means &iv / hovtest=levене(type=square);
  means &iv / hovtest=obrien (w=.50);
  ods output overallanova=anova1 hovftest = hov bartlett = bart (keep=df
chisq probchisq);
run;
```



```

data hoveff (keep= df fvalue probf rename=(df=df_b fvalue=value
probf=p)) hoverr (rename=(df=df_w) keep=df);
  set hov;
  if source = 'Error' then output hoverr;
  else output hoveff;
run;

```

```

data hovtest;
  length labl $45. ;
  merge hoveff hoverr;
  n_groups = &n_groups;
  avg_cs = &avg_cs;
  if _n_ = 1 then do;
    labl = 'Brown and Forsythe';
  end;
  else if _n_ = 2 then do;
    labl = "Levene's (absolute values)";
  end;
  if _n_ = 3 then do;
    labl = "Levene's (squared values)";
  end;
  else if _n_ = 4 then do;
    labl = "O'Brien";
  end;
run;

```

```

data bart;
  length labl $45.;
  set bart;
  labl = 'Bartlett';
  rename df = df_b chisq = value probchisq = p;
run;

```

```

data sastests (keep= labl val pr);
  set hovtest bart ;
  val = input(round(value,.0001), $15.);
  pr = input(round(p,.0001), $15.);
run;

```

ANOVA_HOV: A SAS MACRO

```
* Bootstrap BF;
data obtained; *Retrieve sample BF value from previous PROC GLM ODS
table;
    set hov;
    if method = 'BF' and source ne 'Error';
    fo = fvalue;
    keep fo;
proc sort data = cells; by &iv;
data deviates;
    merge &data cells;
    by &iv;
    dev = abs(&dv - m50);
run;

data outboot(drop=__i);
    do Replicate = 1 to &nboots;
        do __i = 1 to &sum_cs;
            p = int(1 + &sum_cs*(ranuni(0)));
            set deviates point=p;
            output;
        end;
    end;
    stop;
run;
data outboot;
    set outboot;
    by replicate;
    if first.replicate then obsnum = 0;
    obsnum + 1;
run;

data outboot;
    set outboot;
    * Assign group membership to observations in bootstrap samples,
maintaining group sizes in original sample;
    cumulate = 0;
    %do i = 1 %to &n_groups;
```

```

        cell&i = &&size&i;
        if obsnum gt cumulate then group = &i;
        cumulate = cumulate + cell&i;
    %end;
run;

proc glm data=outboot outstat=stats;
    class &iv;
    model dev = &iv /ss3;
    means &iv / hovtest=bf;
    ods output hovftest = hov_boot;
    by replicate;
run;
data hov_boot;
    set hov_boot;
    if upcase(source) = upcase("&iv");
    f_boot = fvalue;
    keep f_boot;
ods listing;
data stats;
    if _n_ = 1 then set obtained;
    retain fo;
    set hov_boot;
    if fo >= f_boot then count+1;
    p=strip(put(1-count/&nboots,8.4));
    run;
data bf_final (keep = labl val pr);
    length labl $ 45;
    set stats;
    if _n_ = &nboots;
    labl = 'Bootstrap Brown-Forsythe';
    val = input(round(fo,.0001), $15.);
    pr = input(p, $15.);
run;

*Ramsey conditional H0V test;
proc sort data=cells ; by vargrp;

```

ANOVA_HOV: A SAS MACRO

```

data kur1; *Division by zero here;
  set cells;
  b2 = 3*(cs-1)/(cs+1) + (cs-2)*(cs-3)/((cs+1)*(cs-1))*cg2;
  mnb2 = 3*(cs-1)/(cs+1);
  varb2 = 24*cs*(cs-2)*(cs-3)/((cs+1)*(cs+1)*(cs+3)*(cs+5));
  x = (b2 - mnb2)/sqrt(varb2);
  moment=6*(cs*cs- 5*cs+2)/((cs+7)*(cs+9))*sqrt(6*(cs+3)*(cs+5)/(cs*(cs-
2)*(cs-3)));
  a = 6 + 8/moment*(2/moment+sqrt(1+4/(moment**2)));
  z_b2 = (1-2/(9*a))-((1-2/a)/(1+x*sqrt(2/(a-4))))**(1/3))/sqrt(2/(9*a));
  pbz_b2 = 2*(1-probnorm(abs(z_b2)));
  w2 = (cs+1)*b2/(cs-1) - 3;
run;

proc sort data = kur1; by cs;
proc sort data = cvw2; by cs;
data kur2;
  merge cvw2 kur1;
  by cs;
data kur3;
  set kur2;
  if b2 ne .;
  s = 0;
  if (cs <= 20) and (w2 < 0) and (w2 < lvalue05) then s = -1;
  else if (cs <= 20) and (w2 > 0) and (w2 > uvalue05) then s = 1;
  else if (cs > 20) and (b2 < 0) and (pbz_b2 < 0.05) then s = -1;
  else if (cs > 20) and (b2 > 0) and (pbz_b2 < 0.05) then s = 1;
run;

proc means data = kur3 noprint; * Only one obs in kur3;
  output out=kur4 sum(s) = sums;
run;

data kur5;
  if _n_ = 1 then set kur4;
  retain sums;
  set sastests;

```

```

        if (labl = 'Brown and Forsythe') or (labl = "O'Brien");
run;

data ramseytest (keep = labl val pr);
    length labl $ 45;
    set kur5;
    if sums < 0 and labl = 'Brown and Forsythe' then delete;
    if sums > -1 and labl = "O'Brien" then delete;
    labl = "Ramsey's Conditional Test";
run;

*Cochran;
data ordermax;
    set cells;
    nu = cs - 1;
    nu_x_var = nu*vargrp;
    nu_pooled = &sum_cs - &n_groups;

proc univariate data = ordermax noprint;
    var nu_x_var;
    output out= get_g_den sum = s_den;
run;

proc sort data = ordermax;
    by descending vargrp descending nu;
data _null_;
    set ordermax;
    if _n_ = 1 then do;
        call symput('nu_j',nu);;
        call symput('nu_pooled',nu_pooled);;
    end;
run;

proc univariate data = ordermax noprint;
    var vargrp nu_x_var;
    output out = largevar max=big_ol_var big_nu_var min=sml_ol_var
min_nu_var sum =sumvars g_gen;

```

ANOVA_HOV: A SAS MACRO

```
run;

data getarray;
  set cells;
  obs = _n_;
  array nn[100] n1-n100;
  retain n1-n100;
  nn[obs] = cs;
  keep n1-n100;
run;

data calc_harmonica;
  set getarray;
  har_avesize= harmean(of n1-n100);
  if _n_ = &n_groups;
run;
proc means data=calc_harmonica noprint;
  var har_avesize;
  output out=calc_harm mean=;
run;

data cochran_c_calc;
  set largevar ;
  g_test = (big_ol_var*&nu_j)/g_gen;
  cochran_c = big_ol_var/sumvars;
run;

data cochran_c_crit;
  merge largevar calc_harm (keep=har_avesize) ;

* Computation of critical values for Cochran C test of variances using
arthaver;

* +-----+
  Researcher inputs these three values
* +-----+;
  k = &n_groups; * Number of Groups;
  n1 = &avg_cs; * N for arthm_ave;
```

```

n2= har_avesize; *N for harmonic average;

* +-----+
Critical value calculations for arithmetic
average and harmonic mean
* +-----+;

dfna = n1 - 1;
dfda = (n1 - 1)*(k - 1);
dfnh = n2 - 1;
dfdh = (n2 - 1)*(k - 1);
crit_f_a01 = finv(1-.01/k,dfna,dfda);
crit_f_a05 = finv(1-.05/k,dfna,dfda);
crit_f_a10 = finv(1-.10/k,dfna,dfda);
crit_f_h01 = finv(1-.01/k,dfnh,dfdh);
crit_f_h05 = finv(1-.05/k,dfnh,dfdh);
crit_f_h10 = finv(1-.10/k,dfnh,dfdh);

cochran_crit__arith_ave01 = 1/(1 + ((k-1)/crit_f_a01));
cochran_crit__arith_ave05 = 1/(1 + ((k-1)/crit_f_a05));
cochran_crit__arith_ave10 = 1/(1 + ((k-1)/crit_f_a10));
cochran_crit__har_ave01 = 1/(1 + ((k-1)/crit_f_h01));
cochran_crit__har_ave05 = 1/(1 + ((k-1)/crit_f_h05));
cochran_crit__har_ave10 = 1/(1 + ((k-1)/crit_f_h10));

crit_F_g01 = finv(1-.01/k,&nu_j,&nu_pooled - &nu_j);
crit_F_g05 = finv(1-.05/k,&nu_j,&nu_pooled - &nu_j);
crit_F_g10 = finv(1-.10/k,&nu_j,&nu_pooled - &nu_j);
g_test_crit01 = 1/(1 + ((&nu_pooled/&nu_j - 1)/crit_f_g01));
g_test_crit05 = 1/(1 + ((&nu_pooled/&nu_j - 1)/crit_f_g05));
g_test_crit10 = 1/(1 + ((&nu_pooled/&nu_j - 1)/crit_f_g10));

run;

data cca (keep=lab1 cochran_c cochran_crit__arith_ave01
cochran_crit__arith_ave05 cochran_crit__arith_ave10
rename=(cochran_c=value
cochran_crit__arith_ave01=p01 cochran_crit__arith_ave05=p05
cochran_crit__arith_ave10=p10))

```

ANOVA_HOV: A SAS MACRO

```

        cch      (keep=labl      cochran_c      cochran_crit__har_ave01
cochran_crit__har_ave05 cochran_crit__har_ave10
                rename=(cochran_c=value      cochran_crit__har_ave01=p01
cochran_crit__har_ave05=p05 cochran_crit__har_ave10=p10))
        ccg (keep=labl g_test g_test_crit01 g_test_crit05 g_test_crit10
                rename=(g_test=value      g_test_crit01=p01
g_test_crit05=p05 g_test_crit10=p10));
        length labl $ 45;
        merge cochran_c_calc cochran_c_crit;
run;

data cctest (keep = labl val pr);
    set cca cch ccg;
    length pr $ 15;
    if _n_ = 1 then labl = "Cochran's C test (with arithmetic mean)";
    else if _n_ = 2 then labl = "Cochran's C test (with harmonic mean)";
    else if _n_ = 3 then labl = "G test";
    if value gt p01 then pr = "p < .01";
    if value gt p05 then pr = "p < .05";
    if value gt p10 then pr = "p < .10";
    else if value le p10 then pr = "p > .10";
    val = input(strip(put(value, 8.4)), $15.);
run;

* Fmax with arithmetic mean;
data fmax_cal;
    set largevar;
    fmax = big_ol_var/sml_ol_var;
    k = &n_groups; * Number of Groups;
    kval = k;
    n1 = &avg_cs;
    nu = round(n1 - 1);
    if nu < 11 then nuval = nu;
    else if nu = 11 then nuval = 10;
    else if 11 < nu < 15 then nuval = 12;
    else if 15 <= nu < 20 then nuval = 15;
    else if 20 <= nu < 30 then nuval = 20;
    else if 30 <= nu < 60 then nuval = 30;

```



```

    else if 60 <= nu then nuval = 60;
run;
proc sort data=fmax_cal; by nuval kval;
proc sort data=fmaxcv; by nuval kval;
data fmaxcv1;
    merge fmaxcv fmax_cal (in=a);
    by nuval kval;
    if a;run;
data fmax;
    set fmaxcv1;
    length lab1 $ 45 pr $ 15 ;
    lab1 = 'F-max test (with arithmetic mean)';
    val = input(round(fmax,.0001), $15.);

    if fmax gt cv01 then pr = "p < .01";
    else if fmax gt cv05 then pr = "p < .05";
    else if fmax lt cv05 then pr = "p > .05";
    keep lab1 val pr;
run;

* fmax with harmonic mean;
data fmax_cal_har;
    merge largevar calc_harm (keep=har_avesize);
    fmax = big_ol_var/sml_ol_var;
    k = &n_groups; * Number of Groups;
    kval = k;
    n1 = har_avesize; * group size (harmonic mean);
    nu = round(n1 - 1);

    if nu < 11 then nuval = nu;
    else if nu = 11 then nuval = 10;
    else if 11 < nu < 15 then nuval = 12;
    else if 15 <= nu < 20 then nuval = 15;
    else if 20 <= nu < 30 then nuval = 20;
    else if 30 <= nu < 60 then nuval = 30;
    else if 60 <= nu then nuval = 60;
run;

```

ANOVA_HOV: A SAS MACRO

```

proc sort data=fmax_cal_har; by nuval kval;
proc sort data=fmaxcv; by nuval kval;
data fmaxcv1_har;
  merge fmaxcv fmax_cal_har (in=a);
  by nuval kval;
  if a;
data fmax_har ;
  set fmaxcv1_har;
  length lab1 $ 45 pr $ 15 ;
  lab1 = 'F-max test (with harmonic mean)';
  val = input(round(fmax,.0001), $15.);

  if fmax gt cv01 then pr = "p < .01";
  else if fmax gt cv05 then pr = "p < .05";
  else if fmax lt cv05 then pr = "p > .05";
  keep lab1 val pr;
run;

* Z variance;

data _null_;
  set anova1;
  if source = 'Error' then call symput('mse',ms);
  if source = 'Corrected Total' then call symput('totN',DF+1);
run;

data z2;
  set cells;
  c= 2 + (1/cs);
  z=sqrt((c*(cs-1)*vargrp)/&mse)-sqrt(c*(cs-1)-(c/2));
  z_squared=z**2;
  keep z_squared;
run;
proc means noprint data = z2 sum;
  var z_squared;
  output out = z2a sum = sum_z2;

```

```

data z2b (keep = lab1 val pr);
  set z2a;
  length lab1 $ 45 ;
  lab1 = "Z-variance";
  value = sum_z2/(&n_groups-1);
  df_b = &n_groups-1;
  df_w = &totn - &n_groups;
  p= 1-probchi(sum_z2,df_b);
  val = input(round(value,.0001), $15.);
  pr = input(round(p,.0001), $15.);
run;

* Modified Z variance;

proc means noprint data=&data;
  var &dv;
  by &iv;
  output out=tt var=var_group n=n_group mean=mean_group;

data ones (drop=_type_ _freq_);
  merge &data tt;
  by &iv;

* calc kurtosis;
data temp;
  set ones;
  by &iv;
  z=(&dv - mean_group) / sqrt ((n_group - 1) * var_group / n_group);
  z_4power=z**4;
  if first.&iv then sumz4 = 0;
  sumz4 + z_4power ;
  if last.&iv then output;
  keep &iv var_group n_group mean_group sumz4;

data temp ;
  set temp;
  K_i=sumz4/(n_group - 2);
proc means noprint data = temp;

```

ANOVA_HOV: A SAS MACRO

```

var K_i;
output out = temp2 mean = K;

data temp3;
  if _N_ = 1 then set temp2;
  retain K;
  set temp;
  c= 2*((2.9+0.2/n_group)/K)**(1.6*(n_group-1.8*K+14.7)/n_group);
  z=sqrt( (c*(n_group-1)*var_group) / &mse ) - sqrt (c*(n_group-1)-
(c/2) );
  z_squared=z**2;
  keep z_squared;

proc means noprint data = temp3 sum;
  var z_squared;
  output out = z2x sum = sum_z2;

data z2y (keep = lab1 val pr);
  set z2x;
  length lab1 $ 45 ;
  lab1 = "MZ-variance test";
  value = sum_z2/(&n_groups-1);
  df_b = &n_groups-1;
  p = 1 - probchi(sum_z2,df_b);
  val = input(round(value,.0001), $15.);
  pr = input(round(p,.0001), $15.);
run;

* Results;
data toprint;
  set head sastests bf_final ramseytest z2b z2y cctest fmax fmax_har;
  check_p = substr(pr,1,2);
  if check_p = '0.' then do;
    substr(pr,1,2) = ' .';
  end;
  pr = LEFT(pr);
  drop check_p;
run;

```

```
proc print data = toprint noobs label;  
  title 'Tests of Homogeneity of Variance';  
  label lab1 = '00'x val = '00'x pr = '00'x;  
run;  
  
%mend ANOVA_HOV;
```

JMASM 48: The Pearson Product-Moment Correlation Coefficient and Adjustment Indices: The Fisher Approximate Unbiased Estimator and the Olkin-Pratt Adjustment (SPSS)

David A. Walker
Northern Illinois University
DeKalb, IL

This syntax program is intended to provide an application, not readily available, for users in SPSS who are interested in the Pearson product–moment correlation coefficient (r) and r biased adjustment indices such as the Fisher Approximate Unbiased estimator and the Olkin and Pratt adjustment.

Keywords: Pearson product-moment correlation coefficient, SPSS, syntax

Introduction

The purpose for this computational program is to provide an application not readily available for users in the frequently employed Statistical Package for the Social Sciences (SPSS) software who are interested in the Pearson product-moment correlation coefficient (r) and r biased adjustment indices. The intent is that this program may assist users whose research importance is predicated on concepts such as point estimate bias or accuracy of point estimates to infer applicable and more robust suggestions about their data principally in a small sample size situation.

Correlation Coefficient

The Pearson product-moment correlation coefficient is employed extensively in social science research (Smithson, 2000) as a correlational technique between two variables (X and Y) and also in concurrence with numerous univariate and

David Walker is a Professor of Educational Research and Assessment. Email him at: dawalker@niu.edu.

multivariate methods “...to analyze the underlying relationship between the variables of interest prior to or following the main analysis” (Padilla & Veprinsky, 2014, p. 824). To be sure, there are alternative correlational methods that have been proposed to estimate the population correlation, ρ , (Donner & Rosner, 1980; Hotelling, 1953; Olkin, 1967), but Pearson’s r appears to be the most frequently-applied statistic in this milieu.

Within the correlation coefficient’s bivariate relationship between X and Y , it is assumed that this pairing has a linear relationship and both X and Y have a normal distribution (Olkin & Pratt, 1958), where “...observations follow a bivariate normal distribution with means (μ_{xi}, μ_{yi}) , standard deviations $(\sigma_{xi}, \sigma_{yi})$...” (Donner & Rosner, 1980, p. 69). The sample correlation coefficient can be represented as

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{\left[n(\sum x^2) - (\sum x)^2 \right] \left[n(\sum y^2) - (\sum y)^2 \right]}} \quad (1)$$

where n = number of x, y pairs, xy = product of xy , and $\sum xy$ = sum the product.

Fisher (1915, 1921, 1924) found that the correlation coefficient was comprised of an asymmetrical distribution, which also influenced this index’s standard error, causing r to be a biased estimator of ρ under normal distribution conditions particularly with small sample sizes (i.e., for Fisher, “small” $N = 18$). Zimmerman, Zumbo, and Williams (2003) pointed out that the notion of “bias” in this situation is derived specifically from the sample mean associated with the r metric. Additionally, Zimmerman et al. noted that, practically,

This discrepancy [bias] may not be crucial if one is simply investigating whether or not a correlation exists. However, if one is concerned with an accurate estimate of the magnitude of a non-zero correlation in test and measurement procedures, then the discrepancy may be of concern. (p. 134)

Bishara and Hittner (2015) established that the threshold for a “small” sample size was $N < 20$, where “... the absolute bias becomes negligible (less than .01) for a sample size greater than 20” (p. 786); noting that the bias decreased as the N increased. Further, Zimmerman et al. (2003) determined that the extent of the aforementioned estimation issue, where r could underestimate ρ by “...as much as .03 or .04 under some realistic conditions...” (p. 134). They also noted that r

could achieve a positive bias as high as 0.05 under non-normal distribution conditions.

***r*-Based Bias Adjustments**

To correct for the inherent bias affiliated with *r*, Fisher (1915) proposed the Fisher Approximate Unbiased (*r*_{FAU}) estimator, which assumes bivariate normality, and can be characterized as

$$r \left[1 + \frac{(1-r^2)}{2n} \right] \quad (2)$$

where *r* = sample correlation coefficient. Additionally, Olkin and Pratt (1958), also assuming bivariate normality, suggested a second unbiased adjustment to *r* (*r*_{OP}), which can be represented as

$$r \left[1 + \frac{(1-r^2)}{2(n-3)} \right] \quad (3)$$

Through a simulation study, Zimmerman et al. (2003) reported that the *r*_{OP} and the *r*_{FAU} adjustments were effectively the same when *N* ≥ 20, but when *N* < 20, *r*_{OP} corrected bias more precisely than *r*_{FAU}. This finding was also corroborated in a simulation conducted by Walker (2016). Gorsuch and Lehmann (2010) supported the use of these *r*-based bias adjustments, though Bishara and Hittner (2015) were more cautious of their use in the presence of non-normal conditions.

Data and Programs

The example used here is comprised of a small sample, where *N* = 16, and are labor statistic data derived from Longley (1967). The full data set consists of seven economic-based variables measured from 1947 to 1962. The sample correlation is between the *Y* variable, the total derived employment, and an *X* variable, the number of people unemployed. As seen below, the user would enter in the program the sample correlation coefficient (*r*) and the sample size (*N*) in the space between BEGIN DATA and END DATA.

DAVID A. WALKER

```
*****
Copyright David A. Walker, 2016
Contact dawalker@niu.edu
Northern Illinois University, 325 Graham, DeKalb, IL 60115
      **APA 6th Edition Citation**
Walker, D. A. (2016). r and r biased adjustment indices [Computer program].
DeKalb, IL: Author.
*****.
```

```
DATA LIST LIST /r (F8.3) N (F8.0).
```

```
*****
NOTE: Between BEGIN DATA and END DATA, put the Pearson's Correlation
Coefficient (r) and the sample size (N)
*****.
```

```
BEGIN DATA
```

```
.502 16
```

```
END DATA.
```

```
COMPUTE rFAU = ((1+(1-r**2)/(2*N))*r).
```

```
COMPUTE rOP = ((1+(1-r**2)/(2*(N-3)))*r).
```

```
COMPUTE FISHERZ = .5*LN((1+r)/(1-r)).
```

```
COMPUTE t = r*SQRT(N-2)/SQRT(1-r**2).
```

```
COMPUTE p1 = CDF.T(t,N-2).
```

```
COMPUTE p = (1-p1)*2.
```

```
COMPUTE Power = (1-CDFNORM(1.96-ABS(FISHERZ*SQRT(N-3)))).
```

```
COMPUTE r2 = r**2.
```

```
FORMAT rFAU TO r2 (F9.3).
```

```
VARIABLE LABELS r 'Pearson Correlation Coefficient r'/r2 'Variance
Explained by the Relationship r2'/ Power 'Post-Hoc Power'/p 'p-value'/rFAU
'Fisher Approximately Unbiased (rFAU) r'/rOP 'Olkin & Pratt (rOP) Adjusted
r'/.

```

```
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
```

```
  /VARIABLES= r p r2 Power
```

```
  /TITLE "r Effect Size and Power".
```

```
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
```

```
  /VARIABLES= r rFAU rOP
```

```
  /TITLE "r and r Bias Adjustments".
```

r AND r BIASED ADJUSTMENT INDICES

Table 1. r , effect size, and power

Pearson correlation coefficient (r)	p -value	Variance explained by the relationship (r^2)	Post-hoc power
0.502	0.048	0.252	0.512

Table 2. Estimates for r and r bias adjustments

Pearson correlation coefficient (r)	Fisher approximately unbiased r (r_{FAU})	Olkin & Pratt adjusted r (r_{OP})
0.502	0.514	0.516

Results

After implementation of the program, the results in Table 1 from this example display the sample-based correlation coefficient (0.502) along with its subsequent p -value (0.048); denoting statistical significance at the 0.05 level (note: $p = 0.000$ from the program would default to < 0.001). Additionally, the matrix generated an r^2 effect size (note: applicable when statistical significance is realized) that indicated a substantial amount of the variance, or about 25%, was explained in the bivariate relationship between X and Y . Also, the model's overall post-hoc power value, which was based on alpha established at 0.05 and the sample size of 16, was expectedly not robust at 0.512, where power ≥ 0.80 is desired in social science research (Nunnally, 1978).

The results from Table 2 exhibit the correlation coefficient and the r_{FAU} and the r_{OP} bias adjustments. As would be expected, the bias-adjusted indices r_{FAU} (0.514) and r_{OP} (0.516) were very comparable, but noticeably higher in value than r (0.502) (i.e., $>$ the aforementioned threshold of 0.01 or +0.012 and +0.014, respectively).

Conclusion

Given the information derived from the tables, such as the probable point estimate bias, the program affords users with more accurate estimates, which may provide a study with added robust inferences about the data (i.e., particularly with a small sample size). As noted by Zimmerman et al. (2003) concerning the utility of applying an r -based adjustment, "...if one is troubled by the slight bias in the correlation coefficient for normal populations, it is clear that it can be largely eliminated by the Fisher approximate unbiased estimator or by the Olkin and Pratt estimator" (p. 155).

References

- Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement*, 75(5), 785-804. doi: [10.1177/0013164414557639](https://doi.org/10.1177/0013164414557639)
- Donner, A., & Rosner, B. (1980). On inferences concerning a common correlation coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(1), 69-76. doi: [10.2307/2346412](https://doi.org/10.2307/2346412)
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507-521. doi: [10.2307/2331838](https://doi.org/10.2307/2331838)
- Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3-32.
- Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, 3, 329-332.
- Gorsuch, R. L., & Lehmann, C. S. (2010). Correlation coefficients: Mean bias and confidence interval distortions. *Journal of Methods and Measurement in the Social Sciences*, 1(2), 52-65. doi: [10.2458/v1i2.114](https://doi.org/10.2458/v1i2.114)
- Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2), 193-232. Available from <http://www.jstor.org/stable/2983768>
- Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the viewpoint of the user. *Journal of the American Statistical Association*, 62(319), 819-841. doi: [10.1080/01621459.1967.10500896](https://doi.org/10.1080/01621459.1967.10500896)
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Olkin, I. (1967). Correlations revisited. In J. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102-128). Chicago, IL: Rand McNally.
- Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, 29(1), 201-211. doi: [10.1214/aoms/1177706717](https://doi.org/10.1214/aoms/1177706717)
- Padilla, M. A., & Veprinsky, A. (2014). Bootstrapped deattenuated correlation: Nonnormal distributions. *Educational and Psychological Measurement*, 74(5), 823-830. doi: [10.1177/0013164414531780](https://doi.org/10.1177/0013164414531780)

r AND r BIASED ADJUSTMENT INDICES

Smithson, M. J. (2000). *Statistics with confidence: An introduction for psychologists*. Thousand Oaks, CA: Sage Publications.

Walker, D. A. (2016). *Fisher's approximately unbiased r and Olkin and Pratt's adjusted r for samples < 30* [Computer program].

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica*, 24(1), 133-158. Retrieved from <https://www.uv.es/psicologica/articulos1.03/9.ZUMBO.pdf>

JMASM 49: A Compilation of Some Popular Goodness of Fit Tests for Normal Distribution: Their Algorithms and MATLAB Codes (MATLAB)

Metin Öner
Manisa Celal Bayar University
Manisa, Turkey

İpek Deveci Kocakoç
Dokuz Eylül University
İzmir, Turkey

The main purpose of this study is to review calculation algorithms for some of the most common non-parametric and omnibus tests for normality, and to provide them as a compiled MATLAB function. All tests are coded to provide p -values for those normality tests, and the proposed function gives the results as an output table.

Keywords: Normality test, non-parametric test, MATLAB function, p -value

Introduction

One of the most important assumptions for parametric statistical methods is that the sample data come from a normally-distributed population. As this assumption holds, t -tests, variance analysis, factor analysis, and many more methods gain power. Normality of error terms is one of the most important assumptions for regression analysis. For many statistical analyses, a normality test is one of the most important and necessary things to do along with the examination of distributional features of data with descriptive statistics, outlier detection, and heteroscedasticity tests.

Hypotheses for fitness testing if sample data or variable(s) follow a normal distribution (i.e. goodness-of-fit testing) are as follows:

$$H_0: x \in N(\mu, \sigma)$$

$$H_1: x \notin N(\mu, \sigma)$$

Most popular normality tests can be listed as follows:

Metin Öner is an Associate Professor in the School of Applied Sciences. Email him at: metin.oner@cbu.edu.tr.

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

- (a) Empirical Distribution Function (EDF) tests:
 - Kolmogorov-Smirnov test
 - Limiting form (KS-Lim)
 - Stephens Method (KS-S)
 - Marsaglia Method (KS-M)
 - Lilliefors test (KS-L)
 - Anderson-Darling (AD) test
 - Cramer-Von Mises (CvM) test
- (b) Tests based on Regression and Correlation:
 - Shapiro-Wilk (SW) test
 - Shapiro-Francia (SF) test
- (c) Moment $(\sqrt{b_1}, b_2)$ tests:
 - Jarque-Bera (JB) test
 - D'Agostino and Pearson (DAP) test

Shown in Table 1, normality tests covered by some commercial and non-commercial statistical packages/software. Only some common tests are computed by these packages, and researchers are obliged to use these ready-made tests, or code them in any language. Some tests are not covered by any of the packages, and most packages do not give any hints on which test to use under which assumptions.

Therefore, the purpose of this article is to review the most common algorithms used for the computation of p -values for normality testing and compile them in one MATLAB routine for researchers who want to examine their data via different normality tests.

Table 1. Normality tests covered by some commercial and non-commercial statistical software

Package	EDF Based						Reg-Corr		Moment	
	KS-Lim	KS-S	KS-M	KS-L	AD	CvM	SW	SF	JB	DAP
Minitab	X				X					
NCSS	X				X					
SAS	X						X			
SPSS	X			X			X			
Stata							X	X		X
Statistica	X			X			X			
XL-Stat				X	X		X		X	
R Project				X	X	X	X	X		X
Analyse-it	X				X		X			

Empirical Distribution Function Tests

Let $y_1, y_2, \dots, y_n$ be ordered values of $x_1, x_2, \dots, x_n$ sample values. If i denotes the frequency of y_k in k^{th} order, the empirical distribution function is defined as the following step function:

$$F_n(y) = \begin{cases} 0, & y < y_1 \\ i/n, & y_k \leq y < y_{k+1}, \quad k = 1, 2, \dots, n-1 \\ 1, & y \geq y_n \end{cases} \quad (1)$$

The following EDF tests are based on this function. Computation procedures are given here for these tests. For further information about the formulas and the interpretation of EDF statistics, see Hollander and Wolfe (1999) and Gibbons and Chakraborti (1992). For details about the k -sample analogs of the Kolmogorov-Smirnov and Cramer-von Mises statistics used by NPAR1WAY, see Kiefer (1959).

Kolmogorov-Smirnov Normality Test

The Kolmogorov-Smirnov (KS) test statistic is computed with the help of the D_n statistic, which is defined as follows:

$$D_n = \sup_x |F_n(x) - F_0(x)| \quad (2)$$

where “sup” in equation (2) denotes the supremum, that is the maximum of the values in a given interval.

The D_n test statistic will be the greatest vertical distance between $F(x)$ and $F_0(x)$ (Kolmogorov, 1933):

$$D_n = \max \left\{ |D_n^-|, |D_n^+| \right\} \quad (3)$$

where $D_n^- = F(x_{k-1}) - F_0(x_k)$ and $D_n^+ = F(x_k) - F_0(x_k)$. The KSz statistic is computed as follows:

$$\text{KSz} = \sqrt{n} D_n \quad (4)$$

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

The stepwise procedure given below covers the mutual steps required for calculation of p -values for four KS-type tests. The first eight steps are also identical for the AD and CvM tests. The rest of the calculation steps are given for each EDF test.

- Step 1: Enter sample $\mathbf{x}$ vector
- Step 2: $y_i = \text{sort}(\mathbf{x})$
- Step 3: $i = 1, 2, \dots, n$
- Step 4: $F_n(\mathbf{y}) = i / n$
- Step 5: $\text{mean}(\mathbf{y}) = \Sigma y_i / n$
- Step 6: $\sigma(\mathbf{y}) = \sqrt{\Sigma (y_i - \text{mean}(\mathbf{y}))^2 / n - 1}$
- Step 7: $z_i = (y_i - \text{mean}(\mathbf{y})) / \sigma(\mathbf{y})$
- Step 8: $F_0(\mathbf{y}) = u_i = \phi(z_i)$
- Step 9: $D_{\text{plus}} = \text{abs}(F_0(\mathbf{y}) - i / n)$
- Step 10: $D_{\text{minus}} = \text{abs}(F_0(\mathbf{y}) - (i - 1) / n)$
- Step 11: $D_n = \max(D_{\text{plus}}, D_{\text{minus}})$
- Step 12: $\text{KSz} = \sqrt{n} D_n$

Limiting Form

Most statistical packages use this method to calculate the KSz test statistic. In this method, if $n \rightarrow \infty$ the distribution of the KSz statistic ($\sqrt{n} D_n$) is asymptotically Kolmogorov distributed. This statistic has the following formula ([Facchinetti, 2009](#)):

$$\lim_{n \rightarrow \infty} \Pr(\sqrt{n} D_n \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2} \quad (5)$$

This method is suitable for cases where the sample size is large and the distribution parameters are known. However, it is being used in cases where n is small and parameters are not known. The p -value for this test can be calculated by the following step:

$$\text{Step 13: } p\text{-value} = 1 - \sum_{k=-20}^{20} (-1)^k \exp\left(-2k^2 \left(\sqrt{n} D_n\right)^2\right)$$

Marsaglia Method

This method was introduced by Marsaglia, Tsang, and Wang (2003). $\Pr(D_n \leq d)$ is calculated by this formula:

$$\Pr(D_n \leq d) = \frac{n!}{n^n} t_{kk} \quad (6)$$

Here, t_{kk} is the $(k, k)^{\text{th}}$ element of the matrix $\mathbf{H}^n$, $\mathbf{H}$ is an $m \times m$ matrix, $m = 2k - 1$, $d = (k - h) / n$, with k (a positive integer), and $0 \leq h < 1$.

Although this method has a complicated algorithm, it provides 13-15 digit accuracy for n ranging from 2 to at least 16000 for one-tailed p -value calculation.

Step 13: Calculate k , m , and h values:

$k = \text{Round up}(nD_n)$

$m = 2k - 1$

$h = k - nD_n$

Step 14: Get the $\mathbf{H}$ matrix using the following procedure:

Get the first column of the $\mathbf{H}$ matrix except the element Hmatrix ($m, 1$).

Loop $i = 1: m - 1$

Hmatrix ($i, 1$) = $(1 - h_i) / i!$

End

Get the m^{th} row of the $\mathbf{H}$ matrix except the element Hmatrix ($m, 1$).

Hmatrix (m, c) = Hmatrix ($r, 1$)^T, $r = 1, \dots, m - 1$, and $c = m - r + 1$

Get the other elements of the $\mathbf{H}$ matrix.

Loop $i = 1: m - 1$

Loop $j = 2: m$

If $i - j + 1 \geq 0$

Hmatrix (i, j) = $1 / (i - j + 1)!$

Else

Hmatrix (i, j) = 0

End

End

End

Get the element Hmatrix ($m, 1$).

If $h \leq 0.5$

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

$$\text{Hmatrix}(m, 1) = (1 - 2h^m) / m!$$

Else

$$\text{Hmatrix}(m, 1) = (1 - 2h^m + \max(0, 2h - 1)^m) / m!$$

Step 15: Calculate the p -value

$$p\text{-value} = \Pr(D_n > d) = 1 - \Pr(D_n \leq d) = 1 - \frac{n!}{n^n} \mathbf{H}^n(k, k)$$

Stephens' Method

Stephens' method uses a D^* test statistic, which is revised by using D_n , to test normality for cases where parameters are not known. The test statistic D^* is calculated based on n via the following equation:

$$D^* = D_n \left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}} \right) \quad (7)$$

Stephens (1986, p. 123) obtained upper tail critical D^* values by Monte Carlo simulations and tabulated them. Here, the p -value is calculated by linear interpolation based on the values of Stephens' table and the calculation steps are as follows:

Step 13: Calculate the modified D^* statistic from equation (7).

Step 14: Calculate the p -value via linear interpolation by using Stephens' critical value table:

p -value =

$$\left\{ \begin{array}{ll} \text{report } p\text{-value} > 0.15 & D^* < 0.775 \\ 0.15 + (D^* - 0.775) \left[(0.10 - 0.15) / (0.819 - 0.775) \right] & 0.775 \leq D^* < 0.819 \\ 0.10 + (D^* - 0.819) \left[(0.05 - 0.10) / (0.895 - 0.819) \right] & 0.819 \leq D^* < 0.895 \\ 0.05 + (D^* - 0.895) \left[(0.025 - 0.05) / (0.995 - 0.895) \right] & 0.895 \leq D^* < 0.995 \\ 0.025 + (D^* - 0.995) \left[(0.01 - 0.025) / (1.035 - 0.995) \right] & 0.995 \leq D^* < 1.035 \\ \text{report } p\text{-value} < 0.01 & D^* \geq 1.035 \end{array} \right.$$

Lilliefors Test for Normality

The Lilliefors (LF) test was presented as a correction of the Kolmogorov-Smirnov test by Lilliefors (1967). Dellal and Wilkinson (1986) provided a numerical approximation to calculate p -values for this method. The LF test is an extension of the Kolmogorov-Smirnov test to the case where the parameters of the hypothesized normal distribution are unknown and estimated by a sample data set. If the mean and variance parameters of the hypothesized normal distribution are not known, the LF test or Stephens' method should be used instead of the Kolmogorov-Smirnov test. The LF test as a correction of the Kolmogorov-Smirnov test should not be confused with the original Kolmogorov-Smirnov test. This study adopts the algorithm used by the statistical software SPSS to calculate the p -value of the LF test, which is based on the use of the critical value table and formulation of Dellal and Wilkinson. The algorithm is given below:

Step 13: Find the critical values $D20$ and $D15$ corresponding to n from the table given by Dellal and Wilkinson (1986). This table is included in the MATLAB codes in the Appendix. $D20$ and $D15$ are the critical values corresponding to $\alpha = 0.20$ and $\alpha = 0.15$ for the LF normality test, respectively. If n lies between two lower and upper n values, find the critical values $D20$ and $D15$ by linear interpolation. The equations of linear interpolation for the critical values $D20$ and $D15$ are given below:

$$D20 = D20_{\text{lower}} + (n - n_{\text{lower}}) \frac{D20_{\text{upper}} - D20_{\text{lower}}}{n_{\text{upper}} - n_{\text{lower}}}$$

$$D15 = D15_{\text{lower}} + (n - n_{\text{lower}}) \frac{D15_{\text{upper}} - D15_{\text{lower}}}{n_{\text{upper}} - n_{\text{lower}}}$$

Step 14: Find the values of a_1 , b_1 , c_1 , a_2 , b_2 , and c_2 via the following equations:

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

$$\begin{aligned}
 a_1 &= -7.01256(n + 2.78019) \\
 b_1 &= 2.99587\sqrt{n + 2.78019} \\
 c_1 &= 2.1804661 + 0.974598/\sqrt{n} + 1.67997/n \\
 a_2 &= -7.90289126054n^{0.98} \\
 b_2 &= 3.180370175721n^{0.49} \\
 c_2 &= 2.2947256
 \end{aligned}$$

Step 15: Calculate the critical value $D10$ corresponding to $\alpha = 0.10$.

If $n \leq 100$, $a = a_1$, $b = b_1$, $c = c_1$, with

$$D10 = \frac{-b_1 - \sqrt{b_1^2 - 4a_1c_1}}{2a_1}$$

If $n > 100$, $a = a_2$, $b = b_2$, $c = c_2$, with

$$D10 = \frac{-b_2 - \sqrt{b_2^2 - 4a_2c_2}}{2a_2}$$

Step 16: Calculate the p -value according to the following formula:

$$p\text{-value} = \begin{cases} 0.10 & \text{if } D_n = D10 \\ \exp(aD_n^2 + bD_n + c - 2.3025851) & \text{if } D_n > D10 \\ 0.15 + (D_n - D15)[(0.10 - 0.15)/(D10 - D15)] & \text{if } D_n \geq D15 \\ 0.20 + (D_n - D20)[(0.15 - 0.20)/(D15 - D20)] & \text{if } D_n \geq D20 \\ \text{report } p\text{-value} > 0.20 & \text{if } D_n \leq D10 \end{cases}$$

Anderson-Darling Normality Test

The Anderson-Darling (AD) statistic, A^2 , was derived from following integral function:

$$A^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \psi(x) dF(x) \quad (8)$$

where $\psi(x)$ is the weight function of the squared difference. The AD test statistic turns into following formula by taking the weight function to be $\psi(x) = [F(x)(1 - F(x))]^{-1}$, where $F(x)$ is the underlying theoretical cumulative distribution (Anderson & Darling, 1952, 1954).

$$A^2 = -n - (1/n) \sum_{i=1}^n (2i-1) [\ln(u_i) + \ln(1-u_{n-i+1})] \quad (9)$$

where $u_i = \varphi(z_i) = \varphi((y_i - \bar{x})/\hat{\sigma})$ is the cumulative probability of the standard normal distribution. Mean and variance are unknown and estimated from the sample.

Stephens (1986) modified the A^2 test statistic to obtain critical values for different sample sizes:

$$A^{*2} = A^2 (1.0 + 0.75/n + 2.25/n^2) \quad (10)$$

The algorithm for calculating the p -value is given below for Anderson-Darling test (Stephens, 1986, p. 127).

- Step 9: Loop $i = 1: n$
 Sum = Sum + $(2i - 1)(\ln u_i + \ln(1 - u_i))$
 End
- Step 10: Calculate the test statistic.
 $A = -n - (1/n) * \text{Sum}$
- Step 11: Calculate the modified test statistic.
 $A^* = A(1 + 0.75/n + 2.25/n^2)$
- Step 12: Calculate the p -value.

$$p\text{-value} = \begin{cases} 1 - \exp(-13.436 + 101.14A^* - 223.73A^{*2}) & \text{if } A^* \leq 0.2 \\ 1 - \exp(-8.318 + 42.796A^* - 59.938A^{*2}) & \text{if } 0.2 < A^* \leq 0.34 \\ \exp(0.9177 - 4.279A^* - 1.38A^{*2}) & \text{if } 0.34 < A^* \leq 0.60 \\ \exp(1.2937 - 5.709A^* + 0.0186A^{*2}) & \text{if } 0.60 < A^* \leq 153.467 \\ 0 & \text{if } A^* > 153.467 \end{cases}$$

Cramer-von Mises Test

Csörgö and Faraway (1996) noted the Cramer-von Mises (CvM) test was independently presented by Cramer (1928) and von Mises (1931). The CvM test is derived from the same integral function as the AD statistic:

$$W^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F(x)]^2 \psi(x) dF(x) \quad (11)$$

where $\psi(x)$ is a weight function of squared differences. When $\psi(x) = 1$, the statistic is referred to as the CvM statistic W^2 and, when $\psi(x) = [F(x)(1 - F(x))]^{-1}$, the statistic is referred to as the AD statistic A^2 .

The CvM test statistic W^2 can be written explicitly as

$$W^2 = \frac{1}{12n} + \sum_{i=1}^n \left[F_0(y_i) - \frac{2i-1}{2n} \right]^2 \quad (12)$$

where $F_0(y_i)$ is the cumulative distribution function of the specified distribution and the y_i are the sorted values of the x_i data set (Scott & Stewart, 2011).

Stephens (1986, p.127) provided the calculation procedure for a modified statistic $W^{*2} = W^2(1 + 0.5 / n)$ as follows:

- Step 9: Loop $i = 1: n$
 Sum = Sum + $(u_i - (2i - 1) / 2n)^2$
 End
- Step 10: Calculate the test statistic.
 $W = (1 / 12n) + \text{Sum}$
- Step 11: Calculate the modified test statistic.
 $W^* = W(1 + 0.5 / n)$
- Step 12: Calculate the p -value

$$p\text{-value} = \begin{cases} 1 - \exp(13.953 + 775.5W^* - 12542.61W^{*2}) & \text{if } W^* < 0.0275 \\ 1 - \exp(-5.903 + 179.546W^* - 1515.29W^{*2}) & \text{if } 0.0275 \leq W^* < 0.051 \\ \exp(0.886 - 31.62W^* + 10.897W^{*2}) & \text{if } 0.051 \leq W^* < 0.092 \\ 1.111 - 34.242W^* + 12.832W^{*2} & \text{if } W^* \geq 0.093 \end{cases}$$

Regression and Correlation Based Tests

Shapiro-Wilk Test

Let $y_1, y_2, \dots, y_i, \dots, y_n$ be ordered values of n independent and identically distributed random samples $(x_1, x_2, \dots, x_i, \dots, x_n)$ coming from a population with unknown mean $\mu \in \Re$ and unknown $\sigma > 0$. The Shapiro-Wilk statistic W for testing normality is then defined as

$$W = \frac{\left(\sum_{i=1}^n a_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (13)$$

where a_i are elements of the vector

$$\mathbf{a}' = \frac{\mathbf{m}'\mathbf{V}^{-1}}{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}$$

where $\mathbf{m}' = (m_1, m_2, \dots, m_n)$ is the vector of expected values of normal order statistics and $\mathbf{V} = [\text{cov}(y_i, y_j)]$ is the covariance matrix of order statistics (Shapiro & Wilk, 1965).

Royston (1982) provided an approximation method for the calculation of the p -value for a normalized W statistic. This method calculates the p -value of the test as the upper tail of the standard normal distribution. The algorithm is given below:

Step 1: Sort the sample observations in ascending order, i.e. $\vec{\mathbf{x}} = (x_1, \dots, x_n)$ into $\vec{\mathbf{y}} = (y_1, \dots, y_n)$.

Step 2: Calculate Blom scores $(\tilde{\mathbf{m}})$ (Solomon & Sawilowsky, 2009).

$$\tilde{m}_i = \Phi^{-1}\left(\frac{i - 0.375}{n + 0.25}\right)$$

Step 3: $m = \tilde{\mathbf{m}}^T \tilde{\mathbf{m}} = \sum_{i=1}^n \tilde{m}_i^2$

Step 4: Calculate the coefficients a_i from the following equations:

For $i = 1, 2, n - 1, n$, and with $u = 1/\sqrt{n}$:

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

$$\begin{aligned}
 a_n &= -2.70606u^5 + 4.43469u^4 - 2.07119u^3 - 0.14798u^2 + 0.22116u + \tilde{m}_n / \sqrt{m} \\
 a_{n-1} &= -3.58263u^5 + 5.68263u^4 - 1.75246u^3 - 0.29376u^2 + 0.04298u + \tilde{m}_n / \sqrt{m} \\
 a_1 &= -a_n \\
 a_2 &= -a_{n-1}
 \end{aligned}$$

For $3 \leq i \leq n-2$:

$$a_i = \tilde{m}_i / \sqrt{\varepsilon}, \text{ where } \varepsilon = \frac{m - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2}{1 - 2a_n^2 - 2a_{n-1}^2}$$

Step 5: Calculate the Shapiro-Wilk test statistic W from equation (13).

Step 6: Normalize the test statistic W :

$$Z = \frac{\ln(1-W) - \mu}{\sigma}$$

where, for $x = \ln n$,

$$\begin{aligned}
 \mu &= -1.5861 - 0.31082x - 0.083751x^2 + 0.0038915x^3 \\
 \sigma &= \exp(-0.4803 - 0.082676x + 0.0030302x^2)
 \end{aligned}$$

Step 7: Calculate the p -value as the upper tail from the standard normal distribution:

$$p\text{-value} = \Pr(Z \geq z) = 1 - \phi(|z|)$$

Shapiro Francia Test

The Shapiro Francia test statistic W' is given by

$$W' = \frac{\left(\sum_{i=1}^n b_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^n \tilde{m}_i X_{(i)}\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n \tilde{m}_i^2} \quad (14)$$

where $y_1 \leq \dots \leq y_n$ are ordered statistics, the $\tilde{m}_i$ are Blom scores, and

$$\mathbf{b} = (b_1, \dots, b_i, \dots, b_n) = \frac{\tilde{\mathbf{m}}^T}{\sqrt{(\tilde{\mathbf{m}}^T \tilde{\mathbf{m}})}} = \frac{\tilde{\mathbf{m}}^T}{\sqrt{\sum_{i=1}^n \tilde{m}_i^2}}$$

which is called the Shapiro Francia statistic for normality tests (Shapiro & Francia, 1972). If the sample data are leptokurtic, the Shapiro-Francia test is recommended; whereas for platykurtic data, the Shapiro-Wilk test is preferred.

Royston (1993) proposed an approximation for the Shapiro-Francia test to calculate the p -value. Mbah and Paothong (2015) use Royston's approximation algorithm for p -value calculation when they compared the Shapiro-Francia test with other tests. The algorithm for Royston's approximation (for sample sizes $5 \leq n \leq 5000$) is given below:

- Step 1: Sort the sample observations in ascending order, i.e. $\vec{\mathbf{x}} = (x_1, \dots, x_n)$ into $\vec{\mathbf{y}} = (y_1, \dots, y_n)$.
- Step 2: Calculate the Shapiro Francia test statistic W' from equation (14).
- Step 3: Normalize the test statistic W' .

$$Z = \frac{\ln(1 - W') - \mu}{\sigma}$$

where, for $u = \ln(n)$, $v = \ln(u)$,

$$\begin{aligned}\mu &= -1.2725 + 1.0521(v - u) \\ \sigma &= 1.0308 + 0.26758(v + 2/u)\end{aligned}$$

- Step 4: Calculate the p -value as the upper tail from the standard normal distribution:

$$p\text{-value} = \Pr(Z \geq z) = 1 - \phi(|z|)$$

Moment Tests

Jarque-Bera Test

The Jarque-Bera (JB) test is a goodness of fit measure calculated from sample kurtosis and skewness (Jarque & Bera, 1987). The normal distribution has a skewness coefficient of zero and a kurtosis of three. The test statistic JB is then defined by:

$$JB = n \left[\frac{(\sqrt{b_1})^2}{6} - \frac{(b_2 - 3)^2}{24} \right] \quad (15)$$

where the sample skewness is $\sqrt{b_1} = \mu_3 / \mu_2^{3/2}$ and the sample kurtosis is $b_2 = \mu_4 / \mu_2^2$, where μ_2, μ_3 , and μ_4 are the second, third, and fourth central moments, respectively. The j^{th} moment is calculated by

$$\mu_j = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^j, \quad j = 2, 3, 4 \quad (16)$$

The JB statistic has an asymptotic chi-square distribution with two degrees of freedom and H_0 should be rejected at a significance level α if $JB \geq \chi_\alpha^2(2)$.

The algorithm for calculating the JB test statistic is as follows:

- Step 1: The sample vector $\mathbf{x}$ is entered.
- Step 2: $\text{mean}(\mathbf{x}) = \sum x_i / n$.
- Step 3: Second moment $\mu_2 = \sum_{i=1}^n (x_i - \text{mean}(\mathbf{x}))^2 / n$.
- Step 4: Third moment $\mu_3 = \sum_{i=1}^n (x_i - \text{mean}(\mathbf{x}))^3 / n$.
- Step 5: Fourth moment $\mu_4 = \sum_{i=1}^n (x_i - \text{mean}(\mathbf{x}))^4 / n$.
- Step 6: Skewness $\sqrt{b_1} = \mu_3 / \mu_2^{3/2}$.
- Step 7: Kurtosis $b_2 = \mu_4 / \mu_2^2$.
- Step 8: Calculate the JB test statistic from equation (16).
- Step 9: Calculate the p -value from the chi-square distribution with two degrees of freedom.

D'Agostino and Pearson Test

The D'Agostino and Pearson (DAP) test aggregates the skewness and kurtosis tests. The test statistic is defined by

$$DAP = Z^2(\sqrt{b_1}) + Z^2(b_2) \quad (17)$$

The skewness test statistic $Z(\sqrt{b_1})$ and kurtosis test statistic $Z(b_2)$ are approximately normally distributed, and the DAP test statistic has an asymptotic chi-square distribution with two degrees of freedom (D'Agostino & Pearson, 1973). The algorithm for calculating the p -value of DAP test is given below:

Step 1: Compute the sample skewness $\sqrt{b_1} = \mu_3 / \mu_2^{3/2}$.

Step 2: Compute the following values:

$$\begin{aligned} Y &= \sqrt{b_1} \sqrt{\frac{(n+1)(n+3)}{6(n-2)}} \\ \beta_2(\sqrt{b_1}) &= \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} \\ W^2 &= -1 + [2\beta_2(\sqrt{b_1}) - 1]^{1/2} \\ \alpha &= [2/(W^2 - 1)]^{1/2} \end{aligned}$$

Step 3: Compute the skewness test statistic $Z(\sqrt{b_1})$:

$$Z(\sqrt{b_1}) = \frac{1}{\sqrt{\ln W}} \ln \left[\frac{Y}{\alpha} + \left[\left(\frac{Y}{\alpha} \right)^2 + 1 \right]^{1/2} \right]$$

Step 4: Compute the sample kurtosis $b_2 = \mu_4 / \mu_2^2$.

Step 5: Compute the following values:

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

$$E(b_2) = \frac{3(n-1)}{(n+1)}$$

$$\text{Var}(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

$$X = \frac{b_2 - E(b_2)}{\text{Var}(b_2)}$$

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)} \left[\frac{6(n+3)(n+5)}{n(n-2)(n-3)} \right]^{1/2}$$

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left[\frac{2}{\sqrt{\beta_1(b_2)}} + \left(1 + \frac{4}{\beta_1(b_2)} \right)^{1/2} \right]$$

Step 6: Compute the kurtosis test statistic $Z(b_2)$.

$$Z(b_2) = \frac{1}{\sqrt{2/(9A)}} \left[\left(1 - \frac{2}{9A} \right) - \left(\frac{1 - 2/A}{1 + X\sqrt{2/(A-4)}} \right)^{1/3} \right]$$

Step 7: Compute the DAP test statistic from equation (17).

$$\text{DAP} = Z^2(\sqrt{b_1}) + Z^2(b_2)$$

Step 8: Calculate the p -value from the chi-square distribution with two degrees of freedom.

Codes, Execution, and Output

All algorithms for these ten normality tests are coded in the Matlab2015 environment and presented as a function (normalitytest.m). Data should be a $1 \times n$ row vector (in “ $\mathbf{x} = [\dots]$ ” format) and entered as a variable in the workspace. The function gives a display of results, as well as a 10×3 matrix named “Results,” including test statistics in the first column, p -values in the second, and test result in the last. The code file is available both as an Appendix and as an .m file in the MathWorks File Exchange under the name “normality test package” (Öner & Deveci Kocakoç, 2016).

Here, utilization of the function is shown on two sample data sets of 20 data. The first data set is normally distributed:

```
x1=[66 53 154 76 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];
```

After the command “Results=normalitytest(x1),” the output in Figure 1 can be obtained. By using the common threshold of $\alpha = 0.05$, data is found to be normally distributed by all tests since all p -values are above 0.05.

Command Window			
>> x1=[66 53 154 76 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];			
>> Results=normalitytest(x1)			
Test Name	Test Statistic	p-value	Normality (1:Normal,0:Not Normal)
KS Limiting Form	0.7188	0.6797	1
KS Stephens Modification	0.7478	0.1500	1
KS Marsaglia Method	0.7188	0.6229	1
KS Lilliefors Modification	0.1607	0.1876	1
Anderson-Darling Test	0.5951	0.1204	1
Cramer- Von Mises Test	0.0956	0.1295	1
Shapiro-Wilk Test	0.9186	0.0930	1
Shapiro-Francia Test	0.9206	0.0935	1
Jarque-Bera Test	2.9892	0.2243	1
DAgostino & Pearson Test	4.6285	0.0988	1

Figure 1. Output for the first data set

Command Window			
>> x2=[66 253 154 276 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];			
>> Results=normalitytest(x2)			
Test Name	Test Statistic	p-value	Normality (1:Normal,0:Not Normal)
KS Limiting Form	1.1035	0.1750	1
KS Stephens Modification	1.1479	0.0100	0
KS Marsaglia Method	1.1035	0.1473	1
KS Lilliefors Modification	0.2468	0.0024	0
Anderson-Darling Test	1.6833	0.0006	0
Cramer- Von Mises Test	0.2751	0.0007	0
Shapiro-Wilk Test	0.7758	0.0004	0
Shapiro-Francia Test	0.7719	0.0007	0
Jarque-Bera Test	13.1210	0.0014	0
DAgostino & Pearson Test	14.7783	0.0006	0

Figure 2. Output for the second data set.

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

The second data set is disturbed by changing some of the values in the first data set:

```
x2=[66 253 154 276 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];
```

After the command “Results=normalitytest(x2),” the output in Figure 2 can be obtained. By using the common threshold of $\alpha = 0.05$, data is found to be normally distributed by only two of the tests. The KS test is known to have some problems with outliers and small sample sizes (Steinskog, Tjøstheim, Kvamstø, 2007). The Stephens’ and Lilliefors modifications seem to overcome these problems, while the limiting form and Marsaglia method do not. Since it is not in the scope of this article to discuss the powers of the tests, the rest of the interpretation is left to the reader.

The function also gives these results as a matrix to use for other purposes. By this function, a gap in statistical computing can be filled for users of MATLAB as well as for researchers who would like to know the calculation details of these test statistics. This MATLAB function can be used to compute all ten test statistics and to have the results of normality tests with just one command. The next step after this study is to build an API with a user-friendly interface to perform the tests without the need of any MATLAB knowledge.

References

- Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2), 193-212. doi: [10.1214/aoms/1177729437](https://doi.org/10.1214/aoms/1177729437)
- Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769. doi: [10.1080/01621459.1954.10501232](https://doi.org/10.1080/01621459.1954.10501232)
- Csörgö, S., & Faraway, J. J. (1996). The exact and asymptotic distributions of Cramer-von Mises statistics. *Journal of Royal Statistical Society. Series B (Methodological)*, 58(1), 221-234. Available from <http://www.jstor.org/stable/2346175>
- D’Agostino, R. B., & Pearson, E. S. (1973). Tests for departures from normality. Empirical results for the distribution of b^2 and $\sqrt{b^1}$. *Biometrika*, 60(3), 613-622. doi: [10.1093/biomet/60.3.613](https://doi.org/10.1093/biomet/60.3.613)

Dellal, G. E., & Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *The American Statistician*, 40(4), 294-296. doi: [10.1080/00031305.1986.10475419](https://doi.org/10.1080/00031305.1986.10475419)

Facchinetti, S. (2009). A procedure to find exact critical values of Kolmogorov-Smirnov test. *Statistica Applicata – Italian Journal of Applied Statistics*, 21(3-4), 337-359. Available from http://sa-ijas.stat.unipd.it/sites/sa-ijas.stat.unipd.it/files/IJAS_3-4_2009_07_Facchinetti.pdf

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2), 163-172. doi: [10.2307/1403192](https://doi.org/10.2307/1403192)

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione [On the empirical determination of a distribution]. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83-91.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399-402. doi: [10.1080/01621459.1967.10482916](https://doi.org/10.1080/01621459.1967.10482916)

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18). doi: [10.18637/jss.v008.i18](https://doi.org/10.18637/jss.v008.i18)

Mbah, A. K., & Paothong, A. (2015). Shapiro-Francia test compared to other normality test using expected *p*-value. *Journal of Statistical Computation and Simulation*, 85(15), 3002-3016. doi: [10.1080/00949655.2014.947986](https://doi.org/10.1080/00949655.2014.947986)

Öner, M., Deveci Kocakoç, I. (2016). *Normality test package* [MATLAB function]. Retrieved from <https://www.mathworks.com/matlabcentral/fileexchange/60147-normality-test-package>

Royston, P. (1982). An extension of Shapiro and Wilk's *W* test for normality to large samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2), 115-124. doi: [10.2307/2347973](https://doi.org/10.2307/2347973)

Royston, P. (1993). A pocket-calculator algorithm for the Shapiro-Francia test for non-normality: An application to medicine. *Statistics in Medicine*, 12(2), 181-184. doi: [10.1002/sim.4780120209](https://doi.org/10.1002/sim.4780120209)

Scott, W. F., & Stewart, B. (2011). Tables for the Lilliefors and modified Cramer-von Mises tests of normality. *Communications in Statistics – Theory and Methods*, 40(4), 726-730. doi: [10.1080/03610920903453467](https://doi.org/10.1080/03610920903453467)

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337), 215-216. doi: [10.1080/01621459.1972.10481232](https://doi.org/10.1080/01621459.1972.10481232)

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591-611. doi: [10.1093/biomet/52.3-4.591](https://doi.org/10.1093/biomet/52.3-4.591)

Solomon, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, 8(2), 448-462. doi: [10.22237/jmasm/1257034080](https://doi.org/10.22237/jmasm/1257034080)

Steinskog, D. J., Tjøstheim, D. B., & Kvamstø, N. G. (2007). A cautionary note on the use of the Kolmogorov-Smirnov test for normality. *Monthly Weather Review*, 135(3), 1151-1157. doi: [10.1175/mwr3326.1](https://doi.org/10.1175/mwr3326.1)

Stephens, M. A. (1986). Tests based on EDF statistics. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 97-194). New York, NY: Marcel Dekker.

Appendix: MATLAB Function (normalitytest.m)

```
function Results=normalitytest(x)

% Enter the data as a row vector in the workspace

%example1: Normally distributed data
%x1=[66 53 154 76 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];

%example1: Disturbed data
%x2=[66 253 154 276 73 118 106 69 87 84 41 33 78 56 55 35 44 135 75 58];

% Alpha value can be changed as required
alpha=0.05;

% KOLMOGOROV-SMIRNOV TEST- LIMITING FORM

n=length(x);
i=1:n;
y=sort(x);
fx=normcdf(zscore(y));
dplus=max(abs(fx-i/n));
dminus=max(abs(fx-(i-1)/n));
Dn=max(dplus,dminus);
KSz=sqrt(n)*Dn;
s=-20:1:20;
a=(-1).^s.*exp(-2*(s.*KSz).^2);
pvalue=1-sum(a);
Results(1,1)=KSz;
Results(1,2)=pvalue;

% KOLMOGOROV-SMIRNOV TEST - STEPHENS MODIFICATION

dKSz=Dn*(sqrt(n)-0.01+0.85/sqrt(n));

if dKSz<0.775
    pvalue=0.15;
```

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

```
elseif dKSz<0.819
    pvalue=((0.10-0.15)/(0.819-0.775))*(dKSz-0.775)+0.15;
elseif dKSz<0.895
    pvalue=((0.05-0.10)/(0.895-0.819))*(dKSz-0.819)+0.10;
elseif dKSz<0.995
    pvalue=((0.025-0.05)/(0.995-0.895))*(dKSz-0.895)+0.05;
elseif dKSz<1.035
    pvalue=((0.01-0.025)/(1.035-0.995))*(dKSz-0.995)+0.025;
else
    pvalue=0.01;
end
Results(2,1)=dKSz;
Results(2,2)=pvalue;

% KOLMOGOROV-SMIRNOV TEST - MARSAGLIA METHOD

k=ceil(n*Dn);
m=2*k-1;
h=k-n*Dn;

Hmatrix=zeros(m,m);

for i=1:m-1
    for j=2:m
        if i-j+1>=0
            Hmatrix(i,j)=1/factorial(i-j+1);
        else
            Hmatrix(i,j)=0;
        end
    end
end

for i=1:m-1
    Hmatrix(i,1)=(1-h^i)/factorial(i);
end

Hmatrix(m,:)=fliplr(Hmatrix(:,1)');
```

```

if h<=0.5
Hmatrix(m,1)=(1 - 2*h^m)/factorial(m);
else
Hmatrix(m,1)=(1 - 2*h^m + max(0,2*h-1)^m)/factorial(m);
end
    lmax = max(eig(Hmatrix));
    Hmatrix = (Hmatrix./lmax)^n;
    pvalue = (1 - exp(gamma1n(n+1) + n*log(lmax) - n*log(n)) *
Hmatrix(k,k));
Results(3,1)=KSz;
Results(3,2)=pvalue;

% KOLMOGOROV-SMIRNOV TEST - LILLIEFORS MODIFICATION

% P = [n D20 D15]
P=[5 0.289 0.303;
    6 0.269 0.281;
    7 0.252 0.264;
    8 0.239 0.250;
    9 0.227 0.238;
   10 0.217 0.228;
   11 0.208 0.218;
   12 0.200 0.210;
   13 0.193 0.202;
   14 0.187 0.196;
   15 0.181 0.190;
   16 0.176 0.184;
   17 0.171 0.179;
   18 0.167 0.175;
   19 0.163 0.170;
   20 0.159 0.166;
   25 0.143 0.150;
   30 0.131 0.138;
   40 0.115 0.120;
  100 0.074 0.077;
  400 0.037 0.039;
  900 0.025 0.026];

```

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

```
aaa=P(:,1)';
subind=max(find(aaa<n));
upind=subind+1;
xxx=P(subind:upind,:);

if aaa(upind)==n
    D20=xxx(2,2);
    D15=xxx(2,3);
else
    D20=xxx(1,2)+(n-aaa(subind))*((xxx(2,2)-xxx(1,2))/(xxx(2,1)-
xxx(1,1)));
    D15=xxx(1,3)+(n-aaa(subind))*((xxx(2,3)-xxx(1,3))/(xxx(2,1)-
xxx(1,1)));
end

a1=-7.01256*(n+2.78019);
b1=2.99587*sqrt(n+2.78019);
c1=2.1804661+0.974598/sqrt(n)+1.67997/n;

a2=-7.90289126054*(n^0.98);
b2=3.180370175721*(n^0.49);
c2=2.2947256;

if n>100
    D10=(-b2-sqrt(b2^2-4*a2*c2))/(2*a2);
    a=a2;
    b=b2;
    c=c2;
else
    D10=(-b1-sqrt(b1^2-4*a1*c1))/(2*a1);
    a=a1;
    b=b1;
    c=c1;
end

if Dn==D10
    pvalue=0.10;
elseif Dn>D10
```

```

        pvalue=exp(a*Dn^2+b*Dn+c-2.3025851);
elseif Dn>=D15
    pvalue=((0.10-0.15)/(D10-D15))*(Dn-D15)+0.15;
elseif Dn>=D20
    pvalue=((0.15-0.20)/(D15-D20))*(Dn-D20)+0.20;
else
    pvalue=0.20;
end
Results(4,1)=Dn;
Results(4,2)=pvalue;

% ANDERSON-DARLING TEST

adj=1+0.75/n+2.25/(n^2);
i=1:n;
ui=normcdf(zscore(y),0,1);
oneminusui=sort(1-ui);
lastt=(2*i-1).*(log(ui)+log(oneminusui));
asquare=-n-(1/n)*sum(lastt);
AD=asquare*adj;

if AD<=0.2
    pvalue=1-exp(-13.436+101.14*AD-223.73*AD^2);
elseif AD<=0.34
    pvalue=1-exp(-8.318+42.796*AD-59.938*AD^2);
elseif AD<=0.6
    pvalue=exp(0.9177-4.279*AD-1.38*AD^2);
elseif AD<=153.467
    pvalue=exp(1.2937*AD-5.709*AD+0.0186*AD^2);
else
    pvalue=0;
end
Results(5,1)=AD;
Results(5,2)=pvalue;

% CRAMER-VON MISES TEST

adj=1+0.5/n;

```

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

```

i=1:n;
fx=normcdf(zscore(y),0,1);
gx=(fx-((2*i-1)/(2*n))).^2;
CvMteststat=(1/(12*n))+sum(gx);
AdjCvM=CvMteststat*adj;

if AdjCvM<0.0275
    pvalue=1-exp(-13.953+775.5*AdjCvM-12542.61*(AdjCvM^2));
elseif AdjCvM<0.051
    pvalue=1-exp(-5.903+179.546*AdjCvM-1515.29*(AdjCvM^2));
elseif AdjCvM<0.092
    pvalue=exp(0.886-31.62*AdjCvM+10.897*(AdjCvM^2));
elseif AdjCvM>=0.093
    pvalue=exp(1.111-34.242*AdjCvM+12.832*(AdjCvM^2));
end
Results(6,1)=AdjCvM;
Results(6,2)=pvalue;

% SHAPIRO-WILK TEST

a=[];
i=1:n;
mi=norminv((i-0.375)/(n+0.25));
u=1/sqrt(n);
m=mi.^2;

a(n)=-2.706056*(u^5)+4.434685*(u^4)-2.07119*(u^3)-
0.147981*(u^2)+0.221157*u+mi(n)/sqrt(sum(m));
a(n-1)=-3.58263*(u^5)+5.682633*(u^4)-1.752461*(u^3)-
0.293762*(u^2)+0.042981*u+mi(n-1)/sqrt(sum(m));
a(1)=-a(n);
a(2)=-a(n-1);
eps=(sum(m)-2*(mi(n)^2)-2*(mi(n-1)^2))/(1-2*(a(n)^2)-2*(a(n-1)^2));
a(3:n-2)=mi(3:n-2)./sqrt(eps);
ax=a.*y;
KT=sum((x-mean(x)).^2);
b=sum(ax)^2;
SWtest=b/KT;

```

```
mu=0.0038915*(log(n)^3)-0.083751*(log(n)^2)-0.31082*log(n)-1.5861;
sigma=exp(0.0030302*(log(n)^2)-0.082676*log(n)-0.4803);
z=(log(1-SWtest)-mu)/sigma;
pvalue=1-normcdf(z,0,1);
Results(7,1)=SWtest;
Results(7,2)=pvalue;
```

% SHAPIRO-FRANCIA TEST

```
mi=norminv((i-0.375)/(n+0.25));
micarp=sqrt(mi*mi');
weig=mi./micarp;
pay=sum(y.*weig)^2;
payda=sum((y-mean(y)).^2);
SFteststa=pay/payda;

u1=log(log(n))-log(n);
u2=log(log(n))+2/log(n);
mu=-1.2725+1.0521*u1;
sigma=1.0308-0.26758*u2;

zet=(log(1-SFteststa)-mu)/sigma;
pvalue=1-normcdf(zet,0,1);
Results(8,1)=SFteststa;
Results(8,2)=pvalue;
```

% JARQUE-BERA TEST

```
E=skewness(y);
B=kurtosis(y);
JBtest=n*((E^2)/6+((B-3)^2)/24);
pvalue=1-chi2cdf(JBtest,2);
Results(9,1)=JBtest;
Results(9,2)=pvalue;
```

% D'AGOSTINO-PEARSON TEST

```
beta2=(3*(n^2+27*n-70)*(n+1)*(n+3))/((n-2)*(n+5)*(n+7)*(n+9));
```

SOME POPULAR GOODNESS OF FIT TESTS FOR NORMALITY

```

wsquare=-1+sqrt(2*(beta2-1));
delta=1/sqrt(log(sqrt(wsquare)));
alfa=sqrt(2/(wsquare-1));

expectedb2=(3*(n-1))/(n+1);
varb2=(24*n*(n-2)*(n-3))/(((n+1)^2)*(n+3)*(n+5));
sqrtbeta=((6*(n^2-5*n+2))/((n+7)*(n+9)))*sqrt((6*(n+3)*(n+5))/(n*(n-2)*(n-3)));
A=6+(8/sqrtbeta)*(2/sqrtbeta+sqrt(1+4/(sqrtbeta^2)));

squarerootb=skewness(y);
Y=squarerootb*sqrt(((n+1)*(n+3))/(6*(n-2)));
zsqrbtbtest=delta*log(Y/alfa+sqrt((Y/alfa)^2+1));

b2=kurtosis(y);
zet=(b2-expectedb2)/sqrt(varb2);
ztestb2=((1-2/(9*A))-((1-2/A)/(1+zet*sqrt(2/(A-4))))^(1/3))/sqrt(2/(9*A));

DAPtest=zsqrbtbtest^2+ztestb2^2;

pvalue=1-chi2cdf(DAPtest,2);
Results(10,1)=DAPtest;
Results(10,2)=pvalue;

% Compare p-value to alpha
for i=1:10
    if Results(i,2)>alpha
        Results(i,3)=1;
    else
        Results(i,3)=0;
    end
end

% Output display

disp(' ')

```



```

disp('Test Name           Test Statistic   p-value   Normality
(1:Normal,0:Not Normal)')
disp('-----')
disp('-----')
fprintf('KS Limiting Form           %6.4f \t    %6.4f           %1.0f
\r',KSz,Results(1,2),Results(1,3))
fprintf('KS Stephens Modification   %6.4f \t    %6.4f           %1.0f
\r',dKSz,Results(2,2),Results(2,3))
fprintf('KS Marsaglia Method         %6.4f \t    %6.4f           %1.0f
\r',KSz,Results(3,2),Results(3,3))
fprintf('KS Lilliefors Modification   %6.4f \t    %6.4f           %1.0f
\r',Dn,Results(4,2),Results(4,3))
fprintf('Anderson-Darling Test        %6.4f \t    %6.4f           %1.0f
\r',AD,Results(5,2),Results(5,3))
fprintf('Cramer-Von Mises Test         %6.4f \t    %6.4f           %1.0f
\r',AdjCvM,Results(6,2),Results(6,3))
fprintf('Shapiro-Wilk Test              %6.4f \t    %6.4f           %1.0f
\r',SWtest,Results(7,2),Results(7,3))
fprintf('Shapiro-Francia Test           %6.4f \t    %6.4f           %1.0f
\r',SFteststa,Results(8,2),Results(8,3))
fprintf('Jarque-Bera Test                %6.4f \t    %6.4f           %1.0f
\r',JBtest,Results(9,2),Results(9,3))
fprintf('DAgostino & Pearson Test        %6.4f \t    %6.4f           %1.0f
\r',DAPtest,Results(10,2),Results(10,3))

```

JMASM 50: A Web-based Shiny Application for Conducting a Two Dependent Samples Maximum Test (R)

Saverpierre Maggio
Wayne State University
Detroit, MI

Gokul Bhandari
University of Windsor
Windsor, ONT, CAN

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

A web-based Shiny application written in R statistical language was developed and deployed online to calculate a new two dependent samples maximum test as presented in Maggio and Sawilowsky (2014b). The maximum test allows researchers to conduct both the dependent samples *t*-test and Wilcoxon signed-ranks tests on same data without raising concerns associated with Type I error inflation and choice of statistical tests (Maggio and Sawilowsky, 2014a). The maximum test in R statistical language provides a friendly user interface.

Keywords: Maximum test, Wilcoxon signed-ranks test, Type I error inflation, R programming, Shiny application, Dependent samples *t*-test

Introduction

The dependent samples *t*-test is used if data are known or expect to be normally distributed. If not, and it is assumed that the treatment alternative is a shift in means then the Wilcoxon Signed-Ranks (WSR) test is to be used (Blair & Higgins, 1985; Bridge & Sawilowsky, 1999; Gerke & Randles, 2010; Wiederman & Alexandrowicz, 2011). Researchers that conduct both tests on the same data unavoidably increase the experiment-wise Type I error rate.

A new maximum test excel calculator was introduced by Maggio and Sawilowsky (2014b). It is based on the maximum of the dependent samples *t*-test and the Wilcoxon Signed-Ranks (WSR) test. The critical value was previously obtained from the joint sampling distribution for the two tests, thereby eliminating concern for Type I error inflation (Maggio & Sawilowsky, 2014a) when

Dr. Maggio is an Adjunct Faculty member in the Department of Evaluation and Research, College of Education. Email him at: spmaggio@uwindsor.ca. Dr. Bhandari is Associate Professor of Information Systems at Odette School of Business. Dr. Sawilowsky is Professor in the Department of Evaluation and Research, College of Education, and the Editor of this Journal. Email him at shlomo@wayne.edu.

conducting both tests. This new test also eliminates choice of one test over another thus allowing multiple tests on same data either serially or in parallel.

Purpose

The purpose of this article is to provide a web-based Shiny application written in R that calculates the two dependent samples maximum test as in Maggio and Sawilowsky (2014a). This application allows for column names (headers) and separators such as commas, semicolons, or tabs. The locations of the data columns for dependent samples can also be indicated by providing an appropriate column number for Series 1 and Series 2 as shown in Figure 1.

Methodology

The application is available online at <https://datawise.shinyapps.io/maxtest/>. A screenshot of the application in R located in Figures 1 and 2.

Input

The process of obtaining the maximum test p -values, critical values and a determination of whether or not to “reject” or “fail to reject” a null hypothesis is as follows:

1. As shown in Figure 1, upload a comma separated values (CSV) file by clicking on the “Choose File” button.
2. Select the significant levels such as 0.05 or 0.01
3. Select one-tail test or two-tail test.

Conclusion

The application reports in the second tab “Data” displays the uploaded data for visual confirmation. In the third tab “ t -test” reports the output for the t -test (statistic, p -value, and degrees of freedom). The fourth tab reports the output of the “Wilcoxon Signed Rank Test” (statistic and p -value). The fifth tab “Maximum Test Output” reports the maximum of the two tests, critical values of the dependent samples t -test and WSR test, maximum test critical value, effective

significance level of the tests, and the decision whether to reject the null or fail to reject the null.

The maximum test can be used in lieu of choice between the dependent samples t -test and the WSR where “both the classical parametric and non-parametric tests can be safely conducted on the same data, with the maximum of the two refereed to the new table of critical values that are designed to maintain the Type I error rate to nominal α while guaranteeing the maximum power of the two tests” (Maggio & Sawilowsky, 2014b, p.4).

Critical values for a two tailed test and an example to follow was provided for in Maggio and Sawilowsky (2014a). An easy to use calculator in excel was provided in Maggio and Sawilowsky (2014b).

The screenshot shows the 'Maximum Test' web-based application interface. The title 'Maximum Test' is at the top left. Below it is a section for uploading data: 'Upload your CSV file containing the data' with a 'Choose File' button and 'No file chosen' text. To the right of this section are five tabs: 'About Maximum Test', 'Data', 't-Test', 'Wilcoxon Signed Rank Test', and 'Maximum Test Output'. The 'Data' tab is currently selected. Below the upload section, there are several input fields and options: a checked 'Header' checkbox, a 'Separator' section with radio buttons for 'Comma' (selected), 'Semicolon', and 'Tab', two input fields for 'Column number for Series 1' (value 1) and 'Column number for Series 2' (value 2), a 'Significance Level' dropdown menu set to '0.05', and a 'Select test type' section with radio buttons for 'One tail' and 'Two tail' (selected).

Figure 1. Interface of the web-based application

About Maximum Test	Data	t-Test	Wilcoxon Signed Rank Test	Maximum Test Output
Outputs	Values			
Larger of critical values from t and WSR tests	7260			
Max Test Critical Value	2.42716			
Effective significance Level	0.025			
Decision	Reject the null			

Figure 2. Example of an out from the Maximum test

References

- Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's sign-ranks test under various population shapes. *Psychological Bulletin*, 97(1), 119-128. doi: [10.1037/0033-2909.97.1.119](https://doi.org/10.1037/0033-2909.97.1.119)
- Bridge, P. K. & Sawilowsky, S. (1999). Increasing physician's awareness of the impact of statistical tests on research outcomes: Investigating the comparative power of the Wilcoxon rank-sum test and independent samples t-test to violations from normality. *Journal of Clinical Epidemiology*, 52(3), 229-235. doi: [10.1016/s0895-4356\(98\)00168-1](https://doi.org/10.1016/s0895-4356(98)00168-1)
- Gerke, T. A. & Randles, H. (2010). A method for resolving ties in asymptotic relative efficiency. *Statistics and Probability Letters*, 80(13), 1065-1069. doi: [10.1016/j.spl.2010.02.021](https://doi.org/10.1016/j.spl.2010.02.021)
- Maggio, S. & Sawilowsky, S. (2014a). A new maximum test via the dependent samples t-test and the Wilcoxon sign rank test. *Applied Mathematics*, 5(10), 110-114. doi: [10.4236/am.2014.51013](https://doi.org/10.4236/am.2014.51013).
- Maggio, S. & Sawilowsky, S. (2014b). JMASM 33: A Two Dependent Samples Maximum Test Calculator: Excel. *Journal of Modern Applied Statistical Methods*, 13(1), 514-518. doi: [10.22237/jmasm/1398918660](https://doi.org/10.22237/jmasm/1398918660)
- Wiederman, W. T., & Alexandrowicz, R. W. (2011). A modified normal scores test for paired data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 7(1), 25-38. doi: [10.1027/1614-2241/a000020](https://doi.org/10.1027/1614-2241/a000020)

Emerging Scholars

Inferential Procedures for Log Logistic Distribution with Doubly Interval Censored Data

Yue Fang Loh

Universiti Putra Malaysia
Seri Kembangan, Malaysia

Jayanthi Arasan

Universiti Putra Malaysia
Seri Kembangan, Malaysia

Habshah Midi

Universiti Putra Malaysia
Seri Kembangan, Malaysia

M. R. Abu Bakar

Universiti Putra Malaysia
Seri Kembangan, Malaysia

The log logistic model with doubly interval censored data is examined. Three methods of constructing confidence interval estimates for the parameter of the model were compared and discussed. The results of the coverage probability study indicated that the Wald outperformed the likelihood ratio and jackknife inferential procedures.

Keywords: doubly interval censored, jackknife, likelihood ratio, log logistic, Wald

Introduction

Doubly interval censored (DIC) data is a type of interval censored (IC) data, which often arises in disease progression studies where the survival time of interest is the elapsed time between two related events that are possibly IC (De Gruttola & Lagakos, 1989; Sun, 2004). Let A and B denote the times of the occurrences of the two events with $A \leq B$ and the survival time, $Y = B - A$. The observations in Y are DIC when A and B are observed in an interval form $A \in (A_L, A_R]$ and $B \in (B_L, B_R]$ respectively with $A_L \leq A_R$ and $B_L \leq B_R$.

A well-known example of DIC data in real life can be seen in acquired immune deficiency syndrome (AIDS) cohort studies where the A and B represent the human immunodeficiency virus (HIV) infection and AIDS diagnosis time respectively, and Y is the AIDS incubation time. The HIV infection time is often determined through periodic blood tests for which it is only known to occur between the last negative test and the first positive test and therefore observations are commonly interval censored. Also, observations on the diagnosis of AIDS could be either right censored (RC) or IC due to, for example, the end of the study

Yue Fang Loh is a PhD student in the Department of Mathematics. Email at yuefangloh@yahoo.com.

and the periodic follow up nature of the study design, thus yielding DIC data on Y (De Gruttola & Lagakos, 1989; Kim, et al., 1993).

Statistical analysis of DIC data was first discussed by De Gruttola & Lagakos (1989) via nonparametric approach to obtain the maximum likelihood estimator of the joint distribution of HIV infection time and AIDS incubation time without truncated data. Since then, many researchers extend the statistical analysis of DIC data, especially in the context of AIDS, to include truncation effect and covariates information in nonparametric and semiparametric approaches. Authors who have contributed include Bacchetti (1990); Bacchetti & Jewell (1991); Kim, et al. (1993); Jewell (1994); Jewell et al. (1994); Gómez & Lagakos (1994); Sun (1995, 1997); Tu (1995); Gómez & Calle (1999); Goggins, et al. (1999); Sun, et al. (1999); Fang & Sun (2001); Pan (2001); and Lim, et al. (2002). The Bayesian approach has gained some attention in analysis of DIC data in recent years for severe acute respiratory syndrome (SARS) disease incubation time (McBryde, et al., 2006) and time to caries development in children (Komárek, et al., 2005; Komárek & Lesaffre, 2006, 2008; Jara, et al., 2010).

Brookmeyer & Goedart (1989) proposed a two-stage parametric regression model for jointly estimating the effects of covariates on risk of HIV infection as well as risk of progression to AIDS disease once infected. They assumed the HIV infection time, A , follows the piecewise exponential distribution and the onset of AIDS disease, B , follows the Weibull distribution. The likelihood function was presented and maximum likelihood estimates (MLEs) were obtained via Newton Raphson iterative procedure. They considered special cases of DIC data where A could be only IC and B could be only RC or observed exactly (OE). The proposed model was later adapted by Darby, et al. (1990) and fitted to data on the development of AIDS in hemophiliacs in the United Kingdom who are seropositive for HIV.

Reich, et al. (2009) studied two procedures for estimating the incubation time distribution. The first procedure defined the likelihood function with DIC data scheme and obtained the MLEs parametrically. They proposed the following likelihood function and obtained the MLE of parameter γ affecting Y , while parameter λ affecting A is assumed to be known,

$$L(g; l) = \sum_{i=1}^n \left\{ \int_{a_{L_i}}^{a_{R_i}} \int_{b_{L_i}}^{b_{R_i}} f_A(a) f_T(b-a) db da \right\}^{d_{DC_i}} \times \left\{ S_T(t_{L_i}) - S_T(t_{R_i}) \right\}^{d_{IC_i}} f_T(t_i)^{d_{OE_i}}. \quad (1)$$

The variables δ_{DC_i} , δ_{IC_i} , and δ_{OE_i} serve as indicators to identify whether the i^{th} subject is DIC, IC or OE. The second procedure involves a data reduction technique to reduce the DIC data to IC data and obtain the MLEs parametrically. They assumed A follows the uniform distribution and Y follows the log normal distribution.

Kiani & Arasan (2012) proposed a parametric model for analyzing DIC data by assuming that both A and Y follow the exponential distribution. Following Kiani & Arasan, proposed here is a parametric model that could be used to analyze DIC data. It is assumed that the first event time A is uniformly distributed and the survival time Y follows a special case of the log logistic distribution with $\gamma = 1$. We assume independent censoring for both A and Y (Oller, et al., 2004) and independence between A and Y , which are classical assumptions for the treatment of DIC survival times. All simulation studies were performed using the R programming language (R Core Team, 2015).

The Model

Let the survival time of interest Y be a non-negative continuous random variable with density function $f_Y(y)$ whereas $f_A(a)$ and $f_B(b)$ denote the density function of the times to the occurrences of the first event A and second event B respectively. Following Reich, et al. (2009), the distribution of b could be obtained if a is given and $f_Y(y)$ is known. Thus,

$$f_{B|A}(b|a) = f_Y(b - a|a). \quad (2)$$

Thus, the joint density function of A and B would be,

$$f_{A,B}(a,b) = f_{B|A}(b|a)f_A(a) = f_Y(b - a|a)f_A(a) = f_Y(b - a)f_A(a) \quad (3)$$

where $Y = B - A$ and A is assumed to be independent of Y . Therefore, the likelihood for a DIC data is as follows,

$$L = \int_{a_L}^{a_R} \int_{b_L}^{b_R} f_{A,B}(a,b) db da = \int_{a_L}^{a_R} \int_{b_L}^{b_R} f_Y(b - a)f_A(a) db da \quad (4)$$

The distributional assumptions on both A and Y allow us to construct the likelihood function of all data. Here, we assume $A \sim U(u_L, u_R)$ and Y follows the

LOG LOGISTIC MODEL WITH DOUBLY INTERVAL CENSORED DATA

log logistic distribution with scale parameter $-\infty < \lambda < \infty$ and known shape parameter $\gamma = 1$. The density function of A is given by

$$f_A(a) = \frac{1}{u_R - u_L}, \quad (5)$$

and the survival function is

$$S_A(a) = \frac{u_R - a}{u_R - u_L}. \quad (6)$$

Similarly, the density and survival function of Y are given respectively as follows:

$$f_Y(y) = \frac{e^y}{(1 + e^y)^2}, \quad (7)$$

$$S_Y(y) = \frac{1}{1 + e^y}. \quad (8)$$

DIC data include IC and RC lifetime data as special cases (Kalbfleisch & Prentice, 2002; Sun, 1998), therefore a comprehensive likelihood function containing all contributions with respect to each type of data need to be defined. For the i^{th} subject, in cases where both A and B are IC, Y is DIC and the likelihood contribution is

$$\begin{aligned} L_{1_i}(I) &= \int_{a_{L_i}}^{a_{R_i}} \int_{b_{L_i}}^{b_{R_i}} f_Y(b - a) f_A(a) db da \\ &= \frac{1}{e^y(u_R - u_L)} \log \left[\frac{\left\{ 1 + e^y(b_{R_i} - a_{R_i}) \right\} \left\{ 1 + e^y(b_{L_i} - a_{L_i}) \right\}}{\left\{ 1 + e^y(b_{R_i} - a_{L_i}) \right\} \left\{ 1 + e^y(b_{L_i} - a_{R_i}) \right\}} \right]. \end{aligned} \quad (9)$$

In cases where A is IC and B is RC, the likelihood contribution is

$$L_{2_i}(\lambda) = \int_{a_{L_i}}^{a_{R_i}} \int_{b_{L_i}}^{\infty} f_Y(b-a) f_A(a) db da = \frac{1}{e^{\lambda}(u_R - u_L)} \log \left[\frac{1 + e^{\lambda}(b_{L_i} - a_{L_i})}{1 + e^{\lambda}(b_{L_i} - a_{R_i})} \right]. \quad (10)$$

In cases where either A or B is OE while the other is IC, Y becomes IC and the interval $(y_{L_i}, y_{R_i}]$ is equal to $(b_i - a_{R_i}, b_i - a_{L_i}]$ when A is IC and $(b_{L_i} - a_i, b_{R_i} - a_i]$ when B is IC. The likelihood contribution is

$$L_{3_i}(\lambda) = \int_{y_{L_i}}^{y_{R_i}} f_Y(y) dy = S_Y(y_{L_i}) - S_Y(y_{R_i}) = \frac{e^{\lambda}(y_{R_i} - y_{L_i})}{(1 + e^{\lambda}y_{L_i})(1 + e^{\lambda}y_{R_i})}. \quad (11)$$

In cases where A is OE and B is RC, Y becomes RC and $y_{D_i} = b_{L_i} - a_i$, the likelihood contribution is

$$L_{4_i}(\lambda) = S_Y(y_{D_i}) = \frac{1}{1 + e^{\lambda}y_{D_i}}. \quad (12)$$

In cases where both A and B are OE, Y becomes OE and $y_i = b_i - a_i$, the likelihood contribution is

$$L_{5_i}(\lambda) = f_Y(y_i) = \frac{e^{\lambda}}{(1 + e^{\lambda}y_i)^2}. \quad (13)$$

The censoring indicators for the i^{th} subject are defined as follows,

$$\begin{aligned} d_{DC_i} &= 1 \text{ if } Y \text{ is DIC, 0 otherwise;} \\ d_{IR_i} &= 1 \text{ if } A \text{ is IC and } B \text{ is RC, 0 otherwise;} \\ d_{IC_i} &= 1 \text{ if } Y \text{ is IC, 0 otherwise;} \\ d_{RC_i} &= 1 \text{ if } Y \text{ is RC, 0 otherwise;} \\ d_{OE_i} &= 1 \text{ if } Y \text{ is OE, 0 otherwise;} \end{aligned} \quad (14)$$

where $\delta_{OE_i} = 1 - (\delta_{DC_i} + \delta_{IR_i} + \delta_{IC_i} + \delta_{RC_i})$. Following that, the likelihood function for the full sample can be written as

$$\begin{aligned}
 L(\lambda) = & \prod_{i=1}^n \left(\frac{1}{e^{\lambda}(u_R - u_L)} \log \left[\frac{\left\{ 1 + e^{\lambda}(b_{R_i} - a_{R_i}) \right\} \left\{ 1 + e^{\lambda}(b_{L_i} - a_{L_i}) \right\}}{\left\{ 1 + e^{\lambda}(b_{R_i} - a_{L_i}) \right\} \left\{ 1 + e^{\lambda}(b_{L_i} - a_{R_i}) \right\}} \right] \right)^{d_{DC_i}} \\
 & \times \left[\frac{1}{e^{\lambda}(u_R - u_L)} \times \log \left\{ \frac{1 + e^{\lambda}(b_{L_i} - a_{L_i})}{1 + e^{\lambda}(b_{L_i} - a_{R_i})} \right\} \right]^{d_{IR_i}} \\
 & \times \left\{ \frac{e^{\lambda}(y_{R_i} - y_{L_i})}{(1 + e^{\lambda}y_{L_i})(1 + e^{\lambda}y_{R_i})} \right\}^{d_{IC_i}} \times \left(\frac{1}{1 + e^{\lambda}y_{D_i}} \right)^{d_{RC_i}} \\
 & \times \left\{ \frac{e^{\lambda}}{(1 + e^{\lambda}y_i)^2} \right\}^{d_{OE_i}}, \quad (15)
 \end{aligned}$$

and the log likelihood function is

$$\begin{aligned}
 (\lambda) = & \sum_{i=1}^n \left\{ \begin{aligned} & \delta_{DC_i} \left(-\lambda - \log(u_R - u_L) + \log \left[\frac{\log \left\{ 1 + e^{\lambda}(b_{R_i} - a_{R_i}) \right\}}{1 + e^{\lambda}(b_{L_i} - a_{L_i})} + \log \left\{ 1 + e^{\lambda}(b_{L_i} - a_{R_i}) \right\} \right. \right. \\ & \left. \left. - \log \left\{ 1 + e^{\lambda}(b_{R_i} - a_{L_i}) \right\} - \log \left\{ 1 + e^{\lambda}(b_{L_i} - a_{R_i}) \right\} \right] \right) \\ & + \delta_{IR_i} \left(-\lambda - \log(u_R - u_L) + \log \left[\frac{\log \left\{ 1 + e^{\lambda}(b_{L_i} - a_{L_i}) \right\}}{1 + e^{\lambda}(b_{L_i} - a_{R_i})} \right] \right) \\ & + \delta_{IC_i} \left\{ \lambda + \log(y_{R_i} - y_{L_i}) - \log(1 + e^{\lambda}y_{L_i}) - \log(1 + e^{\lambda}y_{R_i}) \right\} \\ & - \delta_{RC_i} \log(1 + e^{\lambda}y_{D_i}) + \delta_{OE_i} \left\{ \lambda - 2 \log(1 + e^{\lambda}y_i) \right\} \end{aligned} \right\} \quad (16)
 \end{aligned}$$

Let

$$\begin{aligned}
 A_{1i} &= 1 + e' \left(b_{R_i} - a_{R_i} \right), & A_{9i} &= \frac{e' \left(b_{R_i} - a_{R_i} \right)}{1 + e' \left(b_{R_i} - a_{R_i} \right)}, \\
 A_{2i} &= 1 + e' \left(b_{L_i} - a_{L_i} \right), & A_{10i} &= \frac{e' \left(b_{L_i} - a_{L_i} \right)}{1 + e' \left(b_{L_i} - a_{L_i} \right)}, \\
 A_{3i} &= 1 + e' \left(b_{R_i} - a_{L_i} \right), & A_{11i} &= \frac{e' \left(b_{R_i} - a_{L_i} \right)}{1 + e' \left(b_{R_i} - a_{L_i} \right)}, \\
 A_{4i} &= 1 + e' \left(b_{L_i} - a_{R_i} \right), & A_{12i} &= \frac{e' \left(b_{L_i} - a_{R_i} \right)}{1 + e' \left(b_{L_i} - a_{R_i} \right)}, \\
 A_{5i} &= 1 + e' y_{L_i}, & A_{13i} &= \frac{e' y_{L_i}}{1 + e' y_{L_i}}, \\
 A_{6i} &= 1 + e' y_{R_i}, & A_{14i} &= \frac{e' y_{R_i}}{1 + e' y_{R_i}}, \\
 A_{7i} &= 1 + e' y_{D_i}, & A_{15i} &= \frac{e' y_{D_i}}{1 + e' y_{D_i}}, \\
 A_{8i} &= 1 + e' y_i, & A_{16i} &= \frac{e' y_i}{1 + e' y_i},
 \end{aligned} \tag{17}$$

The first and second partial derivatives of the log likelihood function are given as follows,

$$\frac{\partial \ell(I)}{\partial I} = \sum_{i=1}^n \left[\begin{aligned} & d_{DC_i} \left\{ -1 + \left(\log \frac{A_{1i} A_{2i}}{A_{3i} A_{4i}} \right)^{-1} \left(A_{9i} + A_{10i} - A_{11i} - A_{12i} \right) \right\} \\ & + d_{IR_i} \left\{ -1 + \left(\log \frac{A_{2i}}{A_{4i}} \right)^{-1} \left(A_{10i} - A_{12i} \right) \right\} \\ & + d_{IC_i} \left(1 - A_{13i} - A_{14i} \right) - d_{RC_i} A_{15i} + d_{OE_i} \left(1 - 2A_{16i} \right) \end{aligned} \right], \tag{18}$$

$$\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = \sum_{i=1}^n \left[\delta_{DC_i} \left(\log \frac{A_{1i} A_{2i}}{A_{3i} A_{4i}} \right)^{-2} \left\{ \left(\log \frac{A_{1i} A_{2i}}{A_{3i} A_{4i}} \right) \left(\frac{A_{9i}}{A_{1i}} + \frac{A_{10i}}{A_{2i}} - \frac{A_{11i}}{A_{3i}} - \frac{A_{12i}}{A_{4i}} \right) \right. \right. \\ \left. \left. - (A_{9i} + A_{10i} - A_{11i} - A_{12i})^2 \right\} \right. \\ \left. + \delta_{IR_i} \left(\log \frac{A_{2i}}{A_{4i}} \right)^{-2} \left\{ \left(\log \frac{A_{2i}}{A_{4i}} \right) \left(\frac{A_{10i}}{A_{2i}} - \frac{A_{12i}}{A_{4i}} \right) - (A_{10i} - A_{12i})^2 \right\} \right. \\ \left. - \delta_{IC_i} \left(\frac{A_{13i}}{A_{5i}} + \frac{A_{14i}}{A_{6i}} \right) - \delta_{RC_i} \left(\frac{A_{15i}}{A_{7i}} \right) - 2\delta_{OE_i} \left(\frac{A_{16i}}{A_{8i}} \right) \right]. \quad (19)$$

The observed information matrix $i(\hat{\lambda})$ which can be obtained from the second partial derivatives of the log likelihood function evaluated at $\hat{\lambda}$ provides us with the estimate of the variance,

$$\widehat{\text{var}}(\hat{\lambda}) = \left\{ i(\hat{\lambda}) \right\}^{-1}. \quad (20)$$

The MLE of the parameter in this paper is obtained by solving the likelihood function using Newton Raphson iterative procedure, which was implemented using `maxLik` package (Henningson & Toomet, 2011) in the R programming language.

Simulation Study

A simulation study using $N = 1000$ samples, each with sample sizes $n = 30, 50, 100, 150, 200, 250$ and 300 was conducted to examine how well the estimation procedure works for the model. The $A \sim U(0,16)$ and Y is assumed to follow the log logistic distribution (special case, $\gamma = 1$) with parameter λ . The value of -4.3 was chosen as the true parameter value of λ to simulate the survival times that mimic those seen in lung cancer data (Prentice, 1973).

DIC data mostly arise in epidemiology studies with periodic follow-ups of subjects. It is common for a subject to miss some scheduled follow up appointments. Therefore, each subject will have two sequences of time, potential inspection times and actual inspection times. Assuming all subject with the same sequence of potential inspection $\mathbf{PT} = (pt_1, pt_2, \dots, pt_g)$, two study period, 48 and

60 months is considered and the follow ups are scheduled to be conducted on monthly basis, therefore $g = 48$ and 60 . The subject will turn up for inspection at each of the pt_j with attendance probability q where $0 \leq q \leq 1$ and $j = 1, 2, \dots, g$. Therefore, each subject will have their own sequence of actual inspection times $\mathbf{AT}_i = (at_{i1}, at_{i2}, \dots, at_{ih_i})$ where $0 \leq h_i \leq g$ which is simulated from the Bernoulli distribution with attendance probabilities $q = 1, 0.8$ and 0.6 . It is assumed that all subjects were inspected from the beginning of the study and therefore $at_{i1} = pt_1$ and have been event free at time origin, $y = 0$.

For each subject in a sample, two random numbers u_{1i} and u_{2i} are generated from $U(0,1)$ to produce a_i and y_i where

$$a_i = u_R - (u_R - u_L)u_{1i}, \quad (21)$$

and

$$y_i = e^{-\lambda} \left(\frac{1}{u_{2i}} - 1 \right). \quad (22)$$

Then b_i is calculated from $y_i + a_i$. Following that, the intervals $(a_{L_i}, a_{R_i}]$ and $(b_{L_i}, b_{R_i}]$ are obtained for a_i and b_i respectively. The a_{L_i} will be the largest element of $\mathbf{AT}_i$ which is less than a_i , and a_{R_i} will be the smallest element of $\mathbf{AT}_i$ which is greater than a_i . Similarly, the b_{L_i} will be the largest element of $\mathbf{AT}_i$ which is less than b_i , and b_{R_i} will be the smallest element of $\mathbf{AT}_i$ which is greater than b_i . If $b_i > at_{ih_i}$, then B is RC with $(b_{L_i}, b_{R_i}] = (at_{ih_i}, \infty)$.

In order to randomly select some subjects that are OE on A or B , two time-windows are defined. The time-window for OE on A is $[G_{1i}, G_{2i}] = [a_{L_i} + (a_{R_i} - a_{L_i})u_{3i} - \varepsilon, a_{L_i} + (a_{R_i} - a_{L_i})u_{3i} + \varepsilon]$, and for OE on B is $[G_{3i}, G_{4i}] = [b_{L_i} + (b_{R_i} - b_{L_i})u_{4i} - \varepsilon, b_{L_i} + (b_{R_i} - b_{L_i})u_{4i} + \varepsilon]$ where $\varepsilon = 0.25$ and u_{3i} and u_{4i} are random numbers generated from $U(0,1)$. In cases where a_i and b_i fall in the same interval, these observations are discarded and two new values of a_i and y_i are generated to calculate b_i . This simulation procedure may yield five possible types of data where $0 < a_{L_i} < a_{R_i} \leq b_{L_i} < b_{R_i} < \infty$,

1. $a_{L_i} < a_i \leq a_{R_i}$ and $b_{L_i} < b_i \leq a_{R_i}$ then Y is DIC;
2. $a_{L_i} < a_i \leq a_{R_i}$ and $b_{L_i} < b_i < \infty$ then A is IC, B is RC;

LOG LOGISTIC MODEL WITH DOUBLY INTERVAL CENSORED DATA

- 3a. $a_{L_i} < a_i \leq a_{R_i}$ and $G_{3i} \leq b_i \leq G_{4i}$ then Y is IC;
- 3b. $G_{1i} \leq a_i \leq G_{2i}$ and $b_{L_i} < b_i \leq b_{R_i}$ then Y is IC;
4. $G_{1i} \leq a_i \leq G_{2i}$ and $b_{L_i} < b_i < \infty$ then Y is RC;
5. $G_{1i} \leq a_i \leq G_{2i}$ and $G_{3i} \leq b_i \leq G_{4i}$ then Y is OE.

In Table 1, the proportion of different types of data in each setting indicated.

Table 1. Average percentage of different types of data for the model at 60 and 48 months study periods.

Attendance probability	Study period = 60			Study period = 48		
	1	0.8	0.6	1	0.8	0.6
Y is DIC (%)	12.78	16.64	20.80	10.80	13.91	17.36
A is IC, B is RC (%)	33.43	38.34	43.53	36.80	42.36	48.26
Y is IC (%)	20.02	18.56	16.00	17.01	15.68	13.40
Y is RC (%)	26.02	21.33	16.59	28.75	23.63	18.38
Y is OE (%)	7.75	5.13	3.08	6.65	4.42	2.60

Simulation results

The simulation study was conducted to examine the bias, standard error (SE) and root mean square error (RMSE) of the estimate at different study periods, attendance probabilities and sample sizes.

From Table 1, more DIC data were generated at 60 months study period as compared to 48 months study period. This is due to the fact that chances of observing the event of interest either exactly or in an interval are higher for longer study period. Forty-eight months study period produced more B that is RC. Higher attendance probability produces more uncensored data and shorter width of interval for IC data.

Given in Table 2 are the bias, SE and RMSE of $\hat{\lambda}$ at various sample sizes, n attendance probabilities, q and study periods, g . The values of bias, SE and RMSE for $\hat{\lambda}$ decrease with an increase in n , q and g . The trend indicates that smaller censoring proportion in data, smaller sample, and shorter study period yield estimates that are less efficient and rather inaccurate.

Table 2. Bias, SE and RMSE of $\hat{\lambda}$ for the model at 60 and 48 months study period

q	n	Study period = 60			Study period = 48		
		Bias	SE	RMSE	Bias	SE	RMSE
1	30	-0.0642	0.3633	0.3689	-0.0426	0.3921	0.3944
	50	-0.0543	0.2783	0.2836	-0.0384	0.3000	0.3024
	100	-0.0349	0.1992	0.2022	-0.0393	0.2129	0.2165
	150	-0.0297	0.1655	0.1682	-0.0355	0.1694	0.1731
	200	-0.0286	0.1400	0.1429	-0.0280	0.1413	0.1441
	250	-0.0289	0.1248	0.1281	-0.0289	0.1293	0.1325
	300	-0.0234	0.1121	0.1145	-0.0288	0.1189	0.1223
0.8	30	-0.0703	0.3589	0.3657	-0.0746	0.3880	0.3951
	50	-0.0587	0.2793	0.2854	-0.0542	0.2898	0.2948
	100	-0.0426	0.1918	0.1964	-0.0520	0.2165	0.2227
	150	-0.0351	0.1588	0.1626	-0.0459	0.1720	0.1780
	200	-0.0461	0.1338	0.1415	-0.0431	0.1399	0.1464
	250	-0.0387	0.1179	0.1241	-0.0415	0.1254	0.1321
	300	-0.0354	0.1120	0.1175	-0.0473	0.1167	0.1259
0.6	30	-0.0641	0.3595	0.3652	-0.0975	0.3945	0.4063
	50	-0.0607	0.2747	0.2813	-0.0780	0.2970	0.3070
	100	-0.0614	0.1961	0.2055	-0.0770	0.2057	0.2196
	150	-0.0635	0.1594	0.1715	-0.0689	0.1724	0.1856
	200	-0.0634	0.1347	0.1488	-0.0708	0.1488	0.1648
	250	-0.0623	0.1223	0.1372	-0.0663	0.1273	0.1435
	300	-0.0562	0.1105	0.1240	-0.0663	0.1155	0.1332

Confidence interval estimation

The performance of three CI estimates when applied to the parameter of the proposed model is compared. The first method is based on the asymptotic normality of the MLE or Wald, followed by likelihood ratio and finally the jackknife CI estimate (see Arasan & Lunn, 2009).

Wald confidence interval estimates

Let $\hat{\lambda}$ be the MLE of parameter λ . Cox & Hinkley (1974) showed under mild regularity conditions, $\hat{\lambda}$ is asymptotically normally distributed with mean λ and variance $I(\lambda)^{-1}$ where $I(\lambda)$ is the Fisher information matrix evaluated at λ . The matrix $I(\lambda)$ can be estimated by the observed information matrix evaluated at the MLE, $i(\hat{\lambda})$. The estimate of $\text{var}(\hat{\lambda})$ can be obtained from the inverse of $i(\hat{\lambda})$. If

$z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution, then the $100(1 - \alpha)\%$ confidence interval for λ could be expressed as

$$\hat{\lambda} - z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\lambda})} < \lambda < \hat{\lambda} + z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\lambda})}. \quad (23)$$

Likelihood ratio confidence interval estimates

For a parameter of interest, λ , the likelihood ratio statistic for testing $H_0: \lambda = \lambda_0$ versus $H_1: \lambda \neq \lambda_0$ is given as

$$\psi = -2 \left\{ \ell(\lambda_0) - \ell(\hat{\lambda}) \right\}, \quad (24)$$

where ℓ denote the log likelihood function, λ_0 maximizes $\ell(\lambda_0)$ under H_0 or restricted model and $\hat{\lambda}$ is the MLE of λ . For large sample sizes, ψ is approximately $\chi^2_{(1,1-\alpha)}$. A $100(1 - \alpha)\%$ CI of λ is constructed by finding two values of $\hat{\lambda}$ where we fail to reject H_0 at α significance level which satisfy $\ell(\lambda_0) = \ell(\hat{\lambda}) - \frac{1}{2} \chi^2_{(1,1-\alpha)}$ with $\hat{\lambda}_L < \hat{\lambda}$ and $\hat{\lambda}_R > \hat{\lambda}$.

Jackknife confidence interval estimates

The jackknife is a resampling technique where each subsample removes one observation from the original sample (Efron & Tibshirani, 1993). For a sample $\mathbf{y} = (y_1, y_2, \dots, y_n)$, the i^{th} jackknife sample will be $\mathbf{y}_{(i)} = (y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ for $i = 1, 2, \dots, n$. Let $\hat{\lambda}$ be the MLE for parameter λ , then $\hat{\lambda}_{(i)}$ will be the MLE of $\hat{\lambda}$ obtained from the i^{th} jackknife sample. The jackknife estimate of the parameter λ and jackknife estimate of standard error is then calculated by using

$$\hat{\lambda}_{jack} = \hat{\lambda} - (n-1) \left\{ \hat{\lambda}_{(\cdot)} - \hat{\lambda} \right\}, \quad (25)$$

$$\widehat{se}(\hat{\lambda})_{jack} = \left[\frac{n-1}{n} \sum_{i=1}^n \left\{ \hat{\lambda}_{(i)} - \hat{\lambda}_{(\cdot)} \right\}^2 \right]^{\frac{1}{2}}, \quad (26)$$

where $\hat{\lambda}_{(\cdot)} = \sum_{i=1}^n \frac{\hat{\lambda}_{(i)}}{n}$.

If $t_{(1-\alpha/2, n-1)}$ is the $1 - \alpha/2$ quantile of the student's t distribution at $n - 1$ degrees of freedom, then the $100(1 - \alpha)\%$ jackknife confidence interval for λ could be expressed as

$$\hat{\lambda}_{jack} - t_{(1-\alpha/2, n-1)} \widehat{se}(\hat{\lambda})_{jack} < \lambda < \hat{\lambda}_{jack} + t_{(1-\alpha/2, n-1)} \widehat{se}(\hat{\lambda})_{jack}. \quad (27)$$

Coverage probability study

A coverage probability study was conducted using $N = 1500$ samples, each with sample sizes, $n = 30, 50, 100, 150, 200, 250$ and 300 to compare the performance of the CI estimates at different sample sizes, attendance probabilities and study periods. Other assumptions of the coverage probability study are similar to what was discussed in the simulation study.

The coverage probability error of a CI is the probability that the interval does not contains the true value of the parameter and should preferably be equal or close to the nominal error probability, α . Two nominal error probabilities were chosen as 0.05 and 0.1 . The left and right error probabilities were estimated and the total error probability was calculated. Following Arasan & Lunn (2009) and Kiani & Arasan (2013), the estimated left (right) error probability was obtained by summing up the numbers for the left (right) endpoint which was more (less) than the true parameter value divided by the total number of samples, N . The estimated total error probability was calculated by summing up the number of times in which an interval did not contain the true parameter value divided by N .

The estimated error probabilities for Wald, likelihood ratio and jackknife intervals are given in Equations (28), (29) and (30) respectively as follows,

$$\begin{aligned} \text{left} &= \# \left\{ \hat{\lambda} - z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\lambda})} > \lambda \right\} / 1500, \\ \text{right} &= \# \left\{ \hat{\lambda} + z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\lambda})} > \lambda \right\} / 1500, \end{aligned} \quad (28)$$

$$\begin{aligned} \text{left} &= \# \left\{ \psi > \chi^2_{(1, \alpha)} \text{ and } \hat{\lambda} > \lambda \right\} / 1500, \\ \text{right} &= \# \left\{ \psi > \chi^2_{(1, \alpha)} \text{ and } \hat{\lambda} < \lambda \right\} / 1500, \end{aligned} \quad (29)$$

$$\begin{aligned} \text{left} &= \# \left\{ \hat{\lambda}_{jack} - t_{(1-\alpha/2, n-1)} \widehat{se}(\hat{\lambda})_{jack} > \lambda \right\} / 1500, \\ \text{right} &= \# \left\{ \hat{\lambda}_{jack} - t_{(1-\alpha/2, n-1)} \widehat{se}(\hat{\lambda})_{jack} < \lambda \right\} / 1500. \end{aligned} \quad (30)$$

Following Doganaksoy & Schmee (1993), the interval is called anticonservative if the total error probability is more than $\alpha + 2.58\text{se}(\hat{\alpha})$. If the total error probability is less than $\alpha - 2.58\text{se}(\hat{\alpha})$, the interval is called conservative. The interval is called symmetric when the larger of the left or right error probability is less than 1.5 times the smaller one.

The overall performances of these CI estimates methods was evaluated based on the total numbers of anticonservative (C-), conservative (C) and asymmetrical (S-) intervals. Also, the behavior of the methods at different nominal error probabilities, sample sizes, study periods and attendance probabilities are of interest.

Coverage probability results

Summarized in Table 3 are the results obtained from the coverage probability study. Given in Tables 4 and 5 are the estimated error probabilities in detail. Figures 1 and 2 provide a graphical view of the estimated left and right error probabilities.

From Tables 4 and 5, the estimated total error probabilities of all CI estimates methods are close to the nominal error probabilities, however, most of the intervals produced are highly asymmetric, regardless of the nominal level, study period, attendance probability and sample size. Both Wald and likelihood ratio methods did not produce any conservative interval, however, the jackknife method produced some conservative intervals when sample sizes were small, $n \leq 50$. The likelihood ratio method produced more anticonservative intervals than the Wald and jackknife methods. All CI estimates methods perform poorly when $q = 0.6$. The numbers of anticonservative, conservative and asymmetrical intervals produced by all CI estimates methods are smaller at higher level of α . Also, all CI estimates methods perform slightly better at $g = 48$.

Overall, the Wald method is better than likelihood ratio and jackknife methods in constructing confidence interval for the parameter of the proposed model as it produced the least number of anticonservative and asymmetrical intervals in addition to not producing any conservative interval. From Figures 1 and 2, we can observe that all CI estimate methods work very well when $q = 1$

regardless of the nominal levels and study periods. However, they start to perform poorly when $q < 1$ especially at $q = 0.6$ by deviating far from the nominal error probability as n increases.

Table 3. Summary of the performance of Wald, likelihood ratio and jackknife methods (C- = anticonservative; C = conservative; S- = asymmetrical)

	q	Wald			LR			Jackknife		
		C-	C	S-	C-	C	S-	C-	C	S-
$\alpha = 0.05,$ $g = 60$	1.0	0	0	5	1	0	7	0	1	6
	0.8	0	0	6	0	0	7	0	2	6
	0.6	2	0	6	4	0	7	3	1	6
$\alpha = 0.05,$ $g = 48$	1.0	0	0	5	1	0	6	0	1	5
	0.8	0	0	6	0	0	7	0	2	5
	0.6	3	0	7	3	0	7	2	2	6
$\alpha = 0.1,$ $g = 60$	1.0	0	0	5	0	0	5	0	1	6
	0.8	0	0	6	0	0	7	0	1	6
	0.6	1	0	7	3	0	7	2	1	5
$\alpha = 0.1,$ $g = 48$	1.0	0	0	5	0	0	5	0	1	5
	0.8	0	0	5	0	0	7	0	2	5
	0.6	3	0	7	3	0	7	3	0	7

LOG LOGISTIC MODEL WITH DOUBLY INTERVAL CENSORED DATA

Table 4. Estimated error probabilities of Wald, likelihood ratio and jackknife methods for the model when $\alpha = 0.05$ (C- = anticonservative; C = conservative)

	<i>n</i>	Wald			Likelihood Ratio			Jackknife		
		Left Error	Right Error	Total Error	Left Error	Right Error	Total Error	Left Error	Right Error	Total Error
<i>q</i> = 1, <i>g</i> = 60	30	0.0193	0.0220	0.0413	0.0167	0.0300	0.0467	0.0187	0.0053	0.0240 ^C
	50	0.0247	0.0333	0.0580	0.0227	0.0427	0.0653 ^{C-}	0.0253	0.0200	0.0453
	100	0.0167	0.0307	0.0473	0.0153	0.0360	0.0513	0.0173	0.0267	0.0440
	150	0.0180	0.0353	0.0533	0.0167	0.0393	0.0560	0.0193	0.0313	0.0507
	200	0.0167	0.0360	0.0527	0.0160	0.0380	0.0540	0.0193	0.0353	0.0547
	250	0.0160	0.0340	0.0500	0.0153	0.0353	0.0507	0.0173	0.0333	0.0507
<i>q</i> = 0.8, <i>g</i> = 60	300	0.0133	0.0313	0.0447	0.0127	0.0320	0.0447	0.0140	0.0280	0.0420
	30	0.0167	0.0227	0.0393	0.0153	0.0307	0.0460	0.0173	0.0080	0.0253 ^C
	50	0.0147	0.0360	0.0507	0.0133	0.0433	0.0567	0.0133	0.0213	0.0347 ^C
	100	0.0127	0.0287	0.0413	0.0113	0.0327	0.0440	0.0167	0.0253	0.0420
	150	0.0160	0.0287	0.0447	0.0153	0.0340	0.0493	0.0180	0.0253	0.0433
	200	0.0127	0.0367	0.0493	0.0120	0.0413	0.0533	0.0107	0.0380	0.0487
<i>q</i> = 0.6, <i>g</i> = 60	250	0.0127	0.0300	0.0427	0.0120	0.0333	0.0453	0.0120	0.0293	0.0413
	300	0.0060	0.0467	0.0527	0.0060	0.0487	0.0547	0.0067	0.0440	0.0507
	30	0.0180	0.0193	0.0373	0.0153	0.0333	0.0487	0.0193	0.0093	0.0287 ^C
	50	0.0160	0.0253	0.0413	0.0160	0.0313	0.0473	0.0200	0.0160	0.0360
	100	0.0160	0.0440	0.0600	0.0147	0.0507	0.0653 ^{C-}	0.0160	0.0387	0.0547
	150	0.0113	0.0460	0.0573	0.0100	0.0493	0.0593	0.0133	0.0447	0.0580
<i>q</i> = 1, <i>g</i> = 48	200	0.0080	0.0560	0.0640	0.0073	0.0607	0.0680 ^{C-}	0.0087	0.0527	0.0613 ^{C-}
	250	0.0073	0.0660	0.0733 ^{C-}	0.0067	0.0700	0.0767 ^{C-}	0.0067	0.0627	0.0693 ^{C-}
	300	0.0060	0.0593	0.0653 ^{C-}	0.0060	0.0653	0.0713 ^{C-}	0.0067	0.0593	0.0660 ^{C-}
	30	0.0253	0.0180	0.0433	0.0213	0.0293	0.0507	0.0207	0.0007	0.0213 ^C
	50	0.0240	0.0287	0.0527	0.0193	0.0340	0.0533	0.0227	0.0167	0.0393
	100	0.0247	0.0380	0.0627	0.0233	0.0420	0.0653 ^{C-}	0.0260	0.0280	0.0540
<i>q</i> = 0.8, <i>g</i> = 48	150	0.0133	0.0347	0.0480	0.0133	0.0373	0.0507	0.0147	0.0333	0.0480
	200	0.0140	0.0300	0.0440	0.0127	0.0333	0.0460	0.0153	0.0267	0.0420
	250	0.0147	0.0413	0.0560	0.0140	0.0433	0.0573	0.0153	0.0360	0.0513
	300	0.0120	0.0373	0.0493	0.0107	0.0420	0.0527	0.0127	0.0373	0.0500
	30	0.0207	0.0160	0.0367	0.0193	0.0300	0.0493	0.0200	0.0007	0.0207 ^C
	50	0.0133	0.0287	0.0420	0.0113	0.0340	0.0453	0.0147	0.0153	0.0300 ^C
<i>q</i> = 0.6, <i>g</i> = 48	100	0.0207	0.0367	0.0573	0.0187	0.0433	0.0620	0.0227	0.0287	0.0513
	150	0.0140	0.0387	0.0527	0.0127	0.0453	0.0580	0.0173	0.0373	0.0547
	200	0.0067	0.0360	0.0427	0.0047	0.0393	0.0440	0.0120	0.0327	0.0447
	250	0.0100	0.0407	0.0507	0.0100	0.0440	0.0540	0.0107	0.0367	0.0473
	300	0.0100	0.0440	0.0540	0.0100	0.0473	0.0573	0.0107	0.0433	0.0540
	30	0.0120	0.0267	0.0387	0.0120	0.0460	0.0580	0.0173	0.0013	0.0187 ^C
<i>q</i> = 0.6, <i>g</i> = 48	50	0.0120	0.0347	0.0467	0.0093	0.0460	0.0553	0.0147	0.0160	0.0307 ^C
	100	0.0147	0.0367	0.0513	0.0120	0.0433	0.0553	0.0180	0.0307	0.0487
	150	0.0087	0.0493	0.0580	0.0073	0.0560	0.0633	0.0107	0.0447	0.0553
	200	0.0073	0.0593	0.0667 ^{C-}	0.0060	0.0640	0.0700 ^{C-}	0.0073	0.0547	0.0620
	250	0.0067	0.0633	0.0700 ^{C-}	0.0060	0.0687	0.0747 ^{C-}	0.0067	0.0620	0.0687 ^{C-}
	300	0.0080	0.0660	0.0740 ^{C-}	0.0080	0.0740	0.0820 ^{C-}	0.0087	0.0673	0.0760 ^{C-}

Table 5. Estimated error probabilities of Wald, likelihood ratio and jackknife methods for the model when $\alpha = 0.1$ (C- = anticonservative; C = conservative)

	<i>n</i>	Wald			Likelihood Ratio			Jackknife		
		Left Error	Right Error	Total Error	Left Error	Right Error	Total Error	Left Error	Right Error	Total Error
<i>q</i> = 1, <i>g</i> = 60	30	0.0427	0.0493	0.0920	0.0400	0.0593	0.0993	0.0473	0.0280	0.0753 ^C
	50	0.0433	0.0567	0.1000	0.0427	0.0640	0.1067	0.0507	0.0493	0.1000
	100	0.0333	0.0653	0.0987	0.0327	0.0700	0.1027	0.0373	0.0580	0.0953
	150	0.0327	0.0667	0.0993	0.0320	0.0687	0.1007	0.0353	0.0647	0.1000
	200	0.0387	0.0707	0.1093	0.0360	0.0720	0.1080	0.0427	0.0653	0.1080
	250	0.0333	0.0640	0.0973	0.0327	0.0687	0.1013	0.0347	0.0633	0.0980
	300	0.0327	0.0727	0.1053	0.0313	0.0760	0.1073	0.0353	0.0693	0.1047
<i>q</i> = 0.8, <i>g</i> = 60	30	0.0407	0.0500	0.0907	0.0387	0.0587	0.0973	0.0413	0.0267	0.0680 ^C
	50	0.0347	0.0613	0.0960	0.0287	0.0680	0.0967	0.0440	0.0547	0.0987
	100	0.0307	0.0633	0.0940	0.0293	0.0693	0.0987	0.0353	0.0560	0.0913
	150	0.0253	0.0680	0.0933	0.0240	0.0727	0.0967	0.0287	0.0633	0.0920
	200	0.0273	0.0793	0.1067	0.0253	0.0827	0.1080	0.0293	0.0740	0.1033
	250	0.0240	0.0707	0.0947	0.0233	0.0753	0.0987	0.0280	0.0687	0.0967
	300	0.0220	0.0833	0.1053	0.0220	0.0880	0.1100	0.0233	0.0833	0.1067
<i>q</i> = 0.6, <i>g</i> = 60	30	0.0360	0.0540	0.0900	0.0347	0.0660	0.1007	0.0353	0.0287	0.0640 ^C
	50	0.0353	0.0613	0.0967	0.0347	0.0660	0.1007	0.0413	0.0440	0.0853
	100	0.0273	0.0787	0.1060	0.0267	0.0873	0.1140	0.0327	0.0733	0.1060
	150	0.0247	0.0867	0.1113	0.0240	0.0920	0.1160	0.0267	0.0807	0.1073
	200	0.0187	0.1033	0.1220	0.0173	0.1067	0.1240 ^{C-}	0.0193	0.1020	0.1213 ^{C-}
	250	0.0133	0.1053	0.1187	0.0120	0.1080	0.1200 ^{C-}	0.0133	0.1033	0.1167
	300	0.0133	0.1133	0.1267 ^{C-}	0.0127	0.1227	0.1353 ^{C-}	0.0167	0.1167	0.1333 ^{C-}
<i>q</i> = 1, <i>g</i> = 48	30	0.0433	0.0440	0.0873	0.0393	0.0553	0.0947	0.0427	0.0160	0.0587 ^C
	50	0.0453	0.0507	0.0960	0.0440	0.0560	0.1000	0.0500	0.0360	0.0860
	100	0.0413	0.0740	0.1153	0.0380	0.0807	0.1187	0.0433	0.0607	0.1040
	150	0.0313	0.0700	0.1013	0.0313	0.0753	0.1067	0.0340	0.0600	0.0940
	200	0.0293	0.0600	0.0893	0.0267	0.0647	0.0913	0.0307	0.0567	0.0873
	250	0.0320	0.0767	0.1087	0.0300	0.0827	0.1127	0.0353	0.0653	0.1007
	300	0.0273	0.0707	0.0980	0.0267	0.0693	0.0960	0.0287	0.0680	0.0967
<i>q</i> = 0.8, <i>g</i> = 48	30	0.0413	0.0467	0.0880	0.0387	0.0613	0.1000	0.0433	0.0127	0.0560 ^C
	50	0.0360	0.0513	0.0873	0.0320	0.0640	0.0960	0.0380	0.0347	0.0727 ^C
	100	0.0373	0.0653	0.1027	0.0367	0.0740	0.1107	0.0407	0.0560	0.0967
	150	0.0280	0.0740	0.1020	0.0273	0.0827	0.1100	0.0313	0.0680	0.0993
	200	0.0247	0.0780	0.1027	0.0220	0.0873	0.1093	0.0253	0.0687	0.0940
	250	0.0227	0.0767	0.0993	0.0213	0.0807	0.1020	0.0240	0.0753	0.0993
	300	0.0227	0.0840	0.1067	0.0220	0.0873	0.1093	0.0260	0.0807	0.1067
<i>q</i> = 0.6, <i>g</i> = 48	30	0.0293	0.0640	0.0933	0.0267	0.0753	0.1020	0.0353	0.0307	0.0660 ^C
	50	0.0253	0.0673	0.0927	0.0233	0.0787	0.1020	0.0293	0.0513	0.0807
	100	0.0307	0.0827	0.1133	0.0273	0.0880	0.1153	0.0327	0.0653	0.0980
	150	0.0207	0.0913	0.1120	0.0193	0.0987	0.1180	0.0207	0.0833	0.1040
	200	0.0207	0.1093	0.1300 ^{C-}	0.0180	0.1153	0.1333 ^{C-}	0.0267	0.1047	0.1313 ^{C-}
	250	0.0173	0.1127	0.1300 ^{C-}	0.0160	0.1227	0.1387 ^{C-}	0.0187	0.1087	0.1273 ^{C-}
	300	0.0153	0.1220	0.1373 ^{C-}	0.0147	0.1273	0.1420 ^{C-}	0.0153	0.1167	0.1320 ^{C-}

LOG LOGISTIC MODEL WITH DOUBLY INTERVAL CENSORED DATA

Figure 1. Estimated error probabilities of interval estimates methods when $g = 60$

Figure 2. Estimated error probabilities of interval estimates methods when $g = 48$

Conclusion

The estimation procedure worked well for the log logistic distribution with doubly interval censored data where values of bias, standard error and root mean square error are all reasonably low. The Wald confidence interval estimates performed better than the likelihood ratio and jackknife confidence interval when dealing with doubly interval censored data. The jackknife method required more computational effort than the other two. The finite-difference gradient and Hessian which are included in the `maxLik` package in R programming language could not be applied as the derivatives become unreliable due to the complexity of the model.

Acknowledgements

We gratefully acknowledge financial support from the Ministry of Education Malaysia. The research leading to these results has received funding from the Fundamental Research Grant Scheme (FRGS 2014) under vote no. 5524673.

References

- Arasan, J., & Lunn, M. (2009). Survival model of a parallel system with dependent failures and time varying covariates. *Journal of Statistical Planning and Inference*, 139(3), 944–951. doi: [10.1016/j.jspi.2008.06.007](https://doi.org/10.1016/j.jspi.2008.06.007)
- Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association*, 85(412), 1002–1008. [10.2307/2289596](https://doi.org/10.2307/2289596)
- Bacchetti, P., & Jewell, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics*, 47(3), 947–960. doi: [10.2307/2532651](https://doi.org/10.2307/2532651)
- Brookmeyer, R., & Goedart, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*, 45(1), 325–335. doi: [10.2307/2532057](https://doi.org/10.2307/2532057)
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. CRC Press. doi: [10.1007/978-1-4899-2887-0](https://doi.org/10.1007/978-1-4899-2887-0)
- Darby, S. C., Doll, R., Thakrar, B., Rizza, C. R., & Cox, D. R. (1990). Time from infection with HIV to onset of AIDS in patients with haemophilia in the UK. *Statistics in Medicine*, 9(6), 681–689. doi: [10.1002/sim.4780090615](https://doi.org/10.1002/sim.4780090615)

- De Gruttola, V., & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, 45(1), 1–11. doi: 10.2307/2532030
- Doganaksoy, N., & Schmee, J. (1993). Comparisons of approximate confidence intervals for distributions used in life-data analysis. *Technometrics*, 35(2), 175–184. doi: 10.2307/1269662
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap (Monographs on Statistics and Applied Probability, no. 57)*. London: Chapman and Hall.
- Fang, H. B., & Sun, J. (2001). Consistency of nonparametric maximum likelihood estimation of a distribution function based on doubly interval-censored failure time data. *Statistics & Probability Letters*, 55(3), 311–318. doi: 10.1016/s0167-7152(01)00160-2
- Goggins, W. B., Finkelstein, D. M., & Zaslavsky, A. M. (1999). Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. *Biometrics*, 55(2), 445–451. doi: 10.1111/j.0006-341x.1999.00445.x
- Gómez, G., & Calle, M. L. (1999). Non-parametric estimation with doubly censored data. *Journal of Applied Statistics*, 26(1), 45–58. doi: 10.1080/02664769922647
- Gómez, G., & Lagakos, S. W. (1994). Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*, 50(1), 204–212. doi: 10.2307/2533210
- Henningson, A., & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), 443–458. doi: 10.1007/s00180-010-0217-1
- Jara, A., Lesaffre, E., De Iorio, M., & Quintana, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *The Annals of Applied Statistics*, 4(4), 2126–2149. doi: 10.1214/10-aoas368
- Jewell, N. P. (1994). Non-parametric estimation and doubly-censored data: General ideas and applications to AIDS. *Statistics in Medicine*, 13(19-20), 2081–2095. doi: 10.1002/sim.4780131917
- Jewell, N. P., Malani, H. M., & Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association*, 89(425), 7–18. doi: 10.2307/2291196

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. (2nd Ed.). NY: John Wiley & Sons. doi: [10.1002/9781118032985](https://doi.org/10.1002/9781118032985)

Kiani, K., & Arasan, J. (2012). Interval estimations for parameters of exponential model with doubly interval-censored survival time data. *Proceedings of 2nd Regional Conference on Applied and Engineering Mathematics (RCAEM-II) 2012*, pp. 653–660. Arau, Malaysia: Universiti Malaysia Perlis.

Kiani, K., & Arasan, J. (2013). Gompertz model with time-dependent covariate in the presence of interval-, right-and left-censored data. *Journal of Statistical Computation and Simulation*, 83(8), 1472–1490. doi: [10.1080/00949655.2012.662979](https://doi.org/10.1080/00949655.2012.662979)

Kim, M. Y., De Gruttola, V. G., & Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, 49(1), 13–22. doi: [10.2307/2532598](https://doi.org/10.2307/2532598)

Komárek, A., & Lesaffre, E. (2006). Bayesian semi-parametric accelerated failure time model for paired doubly interval-censored data. *Statistical Modelling*, 6(1), 3–22. doi: [10.1191/1471082x06st107oa](https://doi.org/10.1191/1471082x06st107oa)

Komárek, A., & Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, 103(482), 523–533. doi: [10.1198/016214507000000563](https://doi.org/10.1198/016214507000000563)

Komárek, A., Lesaffre, E., Härkänen, Tomni, Declerck, D., & Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics*, 6(1), 145–155. doi: [10.1093/biostatistics/kxh023](https://doi.org/10.1093/biostatistics/kxh023)

Lim, H., Sun, J., & Matthews, D. E. (2002). Maximum likelihood estimation of a survival function with a change point for truncated and interval-censored data. *Statistics in Medicine*, 21(5), 743–752. doi: [10.1002/sim.986](https://doi.org/10.1002/sim.986)

McBryde, E. S., Gibson, G., Pettitt, A. N., Zhang, Y., Zhao, B., & McElwain, D. L. S. (2006). Bayesian modelling of an epidemic of severe acute respiratory syndrome. *Bulletin of Mathematical Biology*, 68(4), 889–917. doi: [10.1007/s11538-005-9005-4](https://doi.org/10.1007/s11538-005-9005-4)

Oller, R., Gómez, G., & Calle, M. L. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *Canadian Journal of Statistics*, 32(3), 315–326. doi: [10.2307/3315932](https://doi.org/10.2307/3315932)

- Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics*, 57(4), 1245–1250. doi: [10.1111/j.0006-341x.2001.01245.x](https://doi.org/10.1111/j.0006-341x.2001.01245.x)
- Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika*, 60(2), 279–288. doi: [10.1093/biomet/60.2.279](https://doi.org/10.1093/biomet/60.2.279)
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Reich, N. G., Lessler, J., Cummings, D. A. T., & Brookmeyer, R. (2009). Estimating incubation period distributions with coarse data. *Statistics in Medicine*, 28(22), 2769–2784. doi: [10.1002/sim.3659](https://doi.org/10.1002/sim.3659)
- Sun, J. (1995). Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics*, 51(3), 1096–1104. doi: [10.2307/2533008](https://doi.org/10.2307/2533008)
- Sun, J. (1997). Self-consistency estimation of distributions based on truncated and doubly censored survival data with applications to AIDS cohort studies. *Lifetime Data Analysis*, 3(4), 305–313. doi: [10.1023/A:1009609227969](https://doi.org/10.1023/A:1009609227969)
- Sun, J. (1998). Interval censoring. In *Encyclopedia of biostatistics*. (pp. 2090–2095). New York : John Wiley. doi: [10.1002/0470011815.b2a11039](https://doi.org/10.1002/0470011815.b2a11039)
- Sun, J. (2004). Statistical analysis of doubly interval-censored failure time data. *Handbook of Statistics: Advances in Survival Analysis*, 23, 105–122. doi: [10.1016/s0169-7161\(03\)23006-6](https://doi.org/10.1016/s0169-7161(03)23006-6)
- Sun, J., Liao, Q., & Pagano, M. (1999). Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics*, 55(3), 909–914. doi: [10.1111/j.0006-341x.1999.00909.x](https://doi.org/10.1111/j.0006-341x.1999.00909.x)
- Tu, X. M. (1995). Nonparametric estimation of survival distributions with censored initiating time, and censored and truncated terminating time: Application to transfusion data for acquired immune deficiency syndrome. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 44(1), 3–16. doi: [10.2307/2986191](https://doi.org/10.2307/2986191)

Study Evaluating the Alterations Caused in an Exploratory Factor Analysis when Multivariate Normal Data is Dichotomized

Rosilei S. Novak
Federal University of Paraná
Curitiba, Brazil

Jair M. Marques
Federal University of Paraná
Curitiba, Brazil

The relationships resulting from the dichotomization of multivariate normal data is a question that causes concern when using exploratory factor analysis. The relationships in an exploratory factor analysis are examined when multivariate normal data, generated by Monte Carlo methods, is dichotomized.

Keywords: Exploratory factor analysis, normal multivariate data, dichotomized data, Monte Carlo simulation

Introduction

The dichotomization of multivariate normal data is widely used when working with exploratory factor analysis (EFA). The alteration on the variables facilitates the representation, reduces field expenses, and assists on the interpretations of the results. However, this process can lead to information loss from the real data.

The phi correlation coefficient was used for dichotomized data, since the objective was to analyze the impact of the substitution of the tetrachoric correlation coefficient by the phi correlation coefficient. In situations where the tetrachoric correlation matrices are singular, they are not appropriate for factor analysis (Embreson & Reise, 2013, p. 37).

Distortion is frequently verified on original data when data are dichotomized. MacCallum, Zhang, Preacher, and Rucker (2002) presented a practical analysis of dichotomization, illustrating with numerical examples the consequences caused on the original variables. Kubinger (2003) highlighted a problem in psychological

Rosilei Souza Novak is a PhD in the Department of Numerical Methods in Engineering. Email her at: rosileisouzanovak@gmail.com. Jair Mendes Marques is an Associate Professor of Numerical Methods in Engineering. Email him at: jair.m.marques@gmail.com.

studies, where hundreds of tests were developed based on factor analysis using dichotomic variables, leading to compromised results. Fedorov, Mannino, and Zhang (2008) stated dichotomization is a transformation of a continuous result into a binary result. This uncommon approach is prejudicial to hypothesis tests and statistical estimates. Their work was based on Fisher's approach, showing that this transformations leads to a great loss of information when data present normal distribution. In terms of information, this loss affects at least $1 - 2 / \pi$ (or 36%) of the original data. Pearson and Mundform (2010) considered the distortion of original data when it is dichotomized, because the effects caused by this transformation is unknown.

The magnitude of the real loss caused by dichotomization on real data is still unknown in empirical studies. The aim in this study is to shed light on this question. To do so, MATLAB functions were developed to generate random multivariate normal samples using the Monte Carlo simulation method. Those samples were then dichotomized and factor analysis was performed on each normal sample and its corresponding dichotomized sample. Finally, significance tests were performed to compare means and variances between multivariate normal and dichotomized samples.

Methods

This research was conducted with the aid of MATLAB R12 software. A total of 12,600 random multivariate normal samples were generated by the Monte Carlo simulation method. Afterwards, those samples were dichotomized. The 12,600 samples were generated considering the variation of number of variables (14 cases), as shown in Table 1, and to each of these cases the sample size have varied in 9 different situations (2, 3, 4, 5, 10, 20, 30, 40, and 50 times the number of variables). Once each sample was generated 100 times, the final result was $14 \times 9 \times 100 = 12,600$ samples.

Shown in Table 1 are the simulations, where the vectors represent the number of variables per factor. For instance, the vector [3 2] represents 2 factors with 5 variables (3 variables on the first factor and 2 variables on the second factor). The criterion used to select the number of variables was a minimum of 5 and a maximum of 50, and the criterion used to select the number of factors was a minimum of 2 and a maximum of 10.

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

Table 1. Total classification of number of variables per factor

Simulation	Variables	Variables per factor
1	5	[3 2]
2	6	[3 3]
3	7	[4 3]
4	8	[5 3]
5	9	[4 3 2]
6	10	[5 3 2]
7	15	[7 4 2 2]
8	20	[10 5 3 2]
9	25	[10 5 5 3 2]
10	30	[10 8 5 3 2 2]
11	35	[15 5 5 3 3 2 2]
12	40	[15 5 5 5 3 3 2 2]
13	45	[15 5 5 5 4 4 3 2 2]
14	50	[14 6 5 5 4 4 3 3 3 3]

In order to perform this study, two functions were created in MATLAB R12: *Matrizc1* and *Simula1*. The first, *Matrizc1*, generates an iteration according to sample size, number of variables involved, and number of factors, based on 100 random multivariate normal samples. Those samples were then dichotomized. The dichotomization of the multivariate normal samples was performed considering three conditions: $P(z \leq z_c) = 0.25$ (1st dichotomization point), $P(z \leq z_c) = 0.50$ (2nd dichotomization point) and $P(z \leq z_c) = 0.75$ (3rd dichotomization point).

Only multivariate normal samples with the following requirements were considered: Phi correlation matrix, with $MSA > 0.5$ and communalities ≥ 0.7 . The samples which did not matched the established requirements were discarded and replaced, until the total amount of 100 samples was reached.

The second function, *Simula1*, performed factor analysis to each of the 100 samples individually (multivariate normal samples and its corresponding dichotomized samples), obtaining (to each of the 100 samples generated) the MSA (mean sample adequacy) for the percentage of variance explained by the first factor, by the total variance and by the communalities (evaluation of the common proportion of variance of each variable shared with common factors).

The factor analysis was performed using the main components method. The number of factors was determined using the Kaiser criterion. The rotation method used was Varimax orthogonal. Factorial loads were not considered, once many oscillations occurred among samples, disallowing comparisons.

After obtaining the factor analysis results, the same function, Simula1, performed significance tests to compare means and variances of the samples. The following statistical significance tests were performed: Student's t -test (comparing means of multivariate normal and dichotomized data, variance explained by the first factor and total variance explained), T^2 Hotelling test (comparing mean vectors of multivariate normal data and dichotomized data from the communalities), Snedecor f -test (comparing the variances of MSA multivariate normal data and dichotomized data, variance explained by the first factor and total variance explained), with the objective of determining the adequate Student's t -test. Finally, the multivariate chi-square test (comparing the covariance matrix of communalities vectors between multivariate normal and dichotomized data) with the objective of determining the adequate T^2 Hotelling test.

All tests were applied considering a significance level of 0.05. Once the tests are all bilateral, the significant results present $p < 0.025$.

In a summarized manner, the methodology was developed in sequence, according to the following stages:

- Generate 100 multivariate normal samples;
- Generate dichotomized samples corresponding to the multivariate normal samples;
- Perform factor analysis on each of the generated samples (normal and dichotomized);
- Calculate, for the 100 multivariate normal and dichotomized samples, the MSA means, variance explained by the first factor, total variance explained and the vectors of the communalities means;
- Perform statistical tests comparing the results obtained through factor analysis of the multivariate normal and dichotomized data.

This article does not present an extensive list of all the simulations conducted. Instead, a representative group was selected, shown in Table 2. The simulations chosen are 1 and 14 with sizes 2, 5, 20, and 50 times the number of variables.

Table 2. Reduced classification of number of variables per factor

Simulation	Variables	Variables per factor
1	5	[3 2]
14	50	[14 6 5 5 4 4 3 3 3 3]

Results

The results of the study are presented in tables, showing the results of the MSA, the proportion of variance explained by the first factor, and the total proportion of variance related to the four types of sample simulations: one small sample (size equal to 2 times the number of variables), two intermediate samples (sizes equal to 5 and 20 times the number of variables), and one large sample (size equal to 50 times the number of variables), in three different dichotomization points. Tables showing the communalities results also are presented, describing only one type of sample simulation, considering sample sizes equal to the cases already seen (2, 5, 20, and 50 times the number of variables) and three different points of dichotomization. The tables referring to the communalities results are extensive, as they show mean vectors. Since no relevant oscillations occurred among simulations, only the tables referring to the first sample simulation are presented, relating the group behavior.

The tables show sample size, means or mean vectors, and p -values (resulting from significance tests comparing means or MSA mean vectors, proportion of variance explained by the first factor, total proportion of variance and the communalities between multivariate normal data and dichotomized data). The tables do not show variances or covariance matrices and p -values (resulting from the significance tests comparing variances or covariance matrices), even though the tests performed assisted in the selection of the adequate mean and mean vector tests.

Results Obtained for the MSA

Tables 3 and 4 relate the sample size with its corresponding means, and the results from the test of mean difference (pMc) of the MSA from the multivariate normal samples and its corresponding dichotomized samples, in three points of dichotomization. It can be observed in Table 3 the results of the MSA for the sample [3 2] (2 factors and 5 variables). The differences identified between the MSA means from normal and dichotomized data were always significant.

The means from dichotomized data were always larger than the means from normal data, except for the sample with size 250 (1st, 2nd, and 3rd points of dichotomization), sample classified as large. There is no great influence from the points of dichotomization on the results.

Table 3. Means and tests of significance of the MSA for samples with 5 variables and 2 factors – vector [3 2]

Size (n)	Sample	1 st point of dichotomization		2 nd point of dichotomization		3 rd point of dichotomization	
		Mean	pMc	Mean	pMc	Mean	pMc
10	Normal	0.7974	0.00000	0.5587	0.00000	0.5603	0.00000
	Dichotomized	0.8368		0.6192		0.6042	
25	Normal	0.5492	0.00000	0.5600	0.00000	0.5697	0.00015
	Dichotomized	0.5893		0.5968		0.5961	
100	Normal	0.5397	0.00000	0.5480	0.00000	0.5440	0.00000
	Dichotomized	0.5816		0.5920		0.5782	
250	Normal	0.6607	0.00000	0.6545	0.00000	0.6543	0.00000
	Dichotomized	0.6398		0.6431		0.6384	

Table 4. Means and tests of significance of the MSA for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 st point of dichotomization		2 nd point of dichotomization		3 rd point of dichotomization	
		Mean	pMc	Mean	pMc	Mean	pMc
100	Normal	0.6471	0.00000	0.6466	0.00000	0.6435	0.00000
	Dichotomized	0.6673		0.6876		0.6704	
250	Normal	0.6978	0.00000	0.6986	0.00000	0.6999	0.00000
	Dichotomized	0.8064		0.8175		0.8061	
1000	Normal	0.7128	0.00000	0.7125	0.00000	0.7129	0.00000
	Dichotomized	0.8596		0.8655		0.8587	
2500	Normal	0.7141	0.00000	0.7147	0.00000	0.7145	0.00000
	Dichotomized	0.8688		0.8740		0.8689	

Table 4 shows the MSA results for the sample [14 6 5 5 4 4 3 3 3 3], 10 factors and 50 variables. The differences between the MSA means from normal data and dichotomized data were always significant. The MSA was always higher for dichotomized data. The three points of dichotomization do not have influence on the results.

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

Table 5. Means and tests of significance of the MSA for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 st point of dichotomization			2 nd point of dichotomization			3 rd point of dichotomization		
		Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N
10	Normal	56.1392	0.00000	0.8541	55.1227	0.02110	0.8856	55.8176	0.00000	0.8440
	Dichotomized	47.9528			48.8185			47.1141		
25	Normal	53.0267	0.00000	0.8676	53.0125	0.00000	0.8747	53.3529	0.00000	0.8582
	Dichotomized	46.0085			46.3744			45.7901		
100	Normal	52.1435	0.00000	0.8454	52.1435	0.00000	0.8454	52.2820	0.00000	0.8351
	Dichotomized	44.0870			44.0870			43.6632		
250	Normal	52.3523	0.00000	0.8265	52.2140	0.00000	0.8438	52.2068	0.00000	0.8285
	Dichotomized	43.2731			44.0626			43.2547		

Table 6. Means and tests of significance of the variance explained by the first factor for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 st point of dichotomization			2 nd point of dichotomization			3 rd point of dichotomization		
		Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N	Mean	pC1c	Prop. D/N
100	Normal	23.2264	0.00000	0.7476	23.2459	0.00000	0.7760	23.1972	0.00000	0.7486
	Dichotomized	17.3648			18.0395			17.3660		
250	Normal	23.1705	0.00000	0.7505	23.2400	0.00000	0.7739	23.1826	0.00000	0.7528
	Dichotomized	17.3907			17.9855			17.4532		
1000	Normal	23.1376	0.00000	0.7530	23.1392	0.00000	0.7770	23.1397	0.00000	0.7511
	Dichotomized	17.4239			17.9811			17.3806		
2500	Normal	23.0981	0.00000	0.7522	23.1010	0.00000	0.7738	23.1173	0.00000	0.7511
	Dichotomized	17.3765			17.8774			17.3650		

Results Obtained for the Proportion of Variance Explained by the First Factor

Tables 5 and 6 show, for the three points of dichotomization, the means, the p -values of the test of mean difference (pC1c) for the results obtained from the proportion of variance explained by the first factor, and the proportions of the means of the variance explained by the first factor of the dichotomized samples, in comparison to the means of normal samples (D/N).

Table 5 shows the results from the test of mean difference for the proportion of variance explained by the first factor (pC1c) for the samples [3 2], with 5 variables and 2 factors. It can be noted that all differences are significant.

The results are always larger for normal samples. The variance explained by the first factor of the multivariate normal samples was always higher than 52%, and the variance explained by the first factor of the dichotomized samples always lower than 49%. It can be observed that the means of the dichotomized samples correspond, at least, to 82.65% of the mean of the multivariate normal samples ($n = 250$, 1st point of dichotomization), and a maximum of 88.56% ($n = 10$, 2nd point of dichotomization). The points of dichotomization on the comparison results show similar results.

Table 6 shows the results from the test of mean difference for the proportion of variance explained by the first factor (pC1c) for the samples [14 6 5 5 4 4 3 3 3 3], with 50 variables and 10 factors. The differences were all significant, with results always larger for normal samples.

The variance explained by the first factor of the multivariate normal samples was always higher than 23%, and the variance explained by the first factor of the dichotomized samples always lower than 19%. The means of the dichotomized samples corresponds to a minimum of 74.76% of the mean from the multivariate normal sample ($n = 100$, 1st point of dichotomization), and a maximum of 77.70% ($n = 1000$, 2nd point of dichotomization). The three points of dichotomization do not have influence on the results.

Results Obtained for the Total Variance Explained

Tables 7 and 8 show, for the three points of dichotomization, the means of multivariate normal samples and dichotomized samples, p -values from the test of mean difference for the total variance explained (pCc), and the proportions of the means of the total variance explained of dichotomized samples, in comparison to the means of normal samples (D/N). Table 7 shows the results from the test of mean

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

difference for the total variance explained (pCc) for samples with 5 variables and 2 factors [3 2]. All the differences are significant, with results always larger for normal samples. The total variance explained of the factors of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 84%. The means from dichotomized samples correspond to a minimum of 83.87% of the mean from the multivariate normal samples ($n = 250$, 1st point of dichotomization), and a maximum of 89.57% ($n = 10$, 2nd point of dichotomization). The points of dichotomization do not have influence on the results.

Table 8 shows the results from the test of mean difference for the total variance explained (pCc) for samples with 50 variables and 10 factors [14 6 5 5 4 4 3 3 3 3]. All the differences are significant, with results always larger for normal samples.

The total variance explained of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 74%. The means from dichotomized samples correspond to a minimum of 79.72% of the mean from the multivariate normal sample ($n = 2500$, 3rd point of dichotomization), and a maximum of 84.25% ($n = 100$, 2nd point of dichotomization). Results do not vary among the three points of dichotomization.

Table 9 shows the results from the test of mean difference for the total variance explained (pCc) for samples with 50 variables and 10 factors [14 6 5 5 4 4 3 3 3 3]. All differences were significant, with results always larger for normal samples. The total variance explained of the multivariate normal samples was always higher than 86%, and the total variance explained of the dichotomized samples always lower than 74%. The means from dichotomized samples correspond to a minimum of 79.72% of the mean from the multivariate normal sample ($n = 2500$, 3rd point of dichotomization), and a maximum of 84.25% ($n = 100$, 2nd point of dichotomization). The points of dichotomization do not have influence on the results.

Results for the Communalities

Table 9 shows the comparisons of the communalities from multivariate normal and dichotomized data, only for the first sample simulation and the second dichotomization point. Only one point of dichotomization is presented (50/50), which is most widely used by researchers, since the dichotomization do not have influence on the results.

Table 7. Means and tests of significance of the total variance explained for samples with 5 variables and 2 factors – vector [3 2]

Size (n)	Sample	1 st point of dichotomization			2 nd point of dichotomization			3 rd point of dichotomization		
		Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N
10	Normal	93.6434	0.00000	0.8706	93.3969	0.00000	0.8957	94.3784	0.00000	0.8710
	Dichotomized	81.5291			83.6629			82.2129		
25	Normal	89.9031	0.00000	0.8800	89.5597	0.00000	0.8869	89.9636	0.00000	0.8728
	Dichotomized	79.1224			79.4313			78.5291		
100	Normal	88.6243	0.00000	0.8615	89.1904	0.00000	0.8744	88.8288	0.00000	0.8529
	Dichotomized	76.3543			77.9938			75.7635		
250	Normal	87.0965	0.00000	0.8387	86.9444	0.00000	0.8563	86.9663	0.00000	0.8412
	Dichotomized	73.0546			74.4515			73.1579		

Table 8. Means and tests of significance of the total variance explained for samples with 50 variables and 10 factors – vector [14 6 5 5 4 4 3 3 3 3]

Size (n)	Sample	1 st point of dichotomization			2 nd point of dichotomization			3 rd point of dichotomization		
		Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N	Mean	pCc	Prop. D/N
100	Normal	87.7839	0.00000	0.8273	87.8335	0.02110	0.8425	87.7222	0.00000	0.8254
	Dichotomized	72.6315			73.9998			72.4105		
250	Normal	86.9701	0.00000	0.8077	87.0215	0.00000	0.8262	86.9936	0.00000	0.8076
	Dichotomized	70.2458			71.8986			70.2610		
1000	Normal	86.5961	0.00000	0.7988	86.5931	0.00000	0.8186	86.5737	0.00000	0.7985
	Dichotomized	69.1756			70.8880			69.1296		
2500	Normal	86.4823	0.00000	0.7973	86.4811	0.00000	0.8166	86.4973	0.00000	0.7972
	Dichotomized	68.9531			70.6282			68.9640		

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

Table 9. Mean vectors and tests of significance of the communalities for samples with 5 variables and 2 factors – vector [3 2] – 2nd point of dichotomization

Size (<i>n</i>)	Sample	Mean vector	pHc	Prop. D/N
10	Normal	[0.9279, 0.8841, 0.9872, 0.9338, 0.9368]	0.00000	0.8796
	Dichotomized	[0.8329, 0.7777, 0.8912, 0.8428, 0.8386]		0.9027
25	Normal	[0.7928, 0.9015, 0.9070, 0.9078, 0.9689]	0.00000	0.8289
	Dichotomized	[0.6572, 0.8065, 0.8097, 0.8152, 0.8830]		0.9066
100	Normal	[0.7740, 0.8928, 0.9129, 0.9113, 0.9685]	0.00000	0.7870
	Dichotomized	[0.6092, 0.8031, 0.8079, 0.8060, 0.8735]		0.9019
250	Normal	[0.8684, 0.8758, 0.8631, 0.9100, 0.8299]	0.00000	0.8199
	Dichotomized	[0.7580, 0.7437, 0.7555, 0.7849, 0.6805]		0.8753

This table shows mean vectors for the communalities of normal and dichotomized data referring to samples [3 2], with 5 variables and 3 factors, its respective p-values for the results from the test of mean difference for the mean vectors (pHc) and the minimum and maximum proportions given by the mean vectors of the communalities of the dichotomized samples in comparison to normal samples (D/N). The first value corresponds to the minimum proportion and the second to the maximum proportion. The table shows significant differences among the mean vectors, with communalities results always larger for normal samples.

The mean vectors from the dichotomized samples correspond to a minimum of 78.70% of the mean from the multivariate normal samples ($n = 100$), and a maximum of 90.66% ($n = 25$).

According to the analyses performed, the results do not show great variation for the three points of dichotomization.

Conclusion

For the cases studied, the following conclusions can be drawn on the relationships resulting from an EFA between multivariate normal and dichotomized data:

- 1) For the MSA there is no regularity of values for normal data and its corresponding dichotomized data. The results suggest that, with the increase of factors and number of variables, the MSA for dichotomized data presents values higher than the values for normal

data. The differences among the MSA means, with few exceptions on small samples, were always significant.

- 2) For the variance explained by the first factor, the total variance, and the communalities, the differences among the mean values for the normal and dichotomized data were always significant, and the values for normal data were always higher in comparison with the values for dichotomized data. Therefore, normal data always explains dichotomized data more efficiently.
- 3) With regard to the points of dichotomization, in the acquisition of dichotomized data, it can be concluded that its results are very similar, having no influence on the analyses performed.

According to the results obtained for the MSA on the 378 simulations performed (number of variables $14 \times$ samples sizes $9 \times$ dichotomization $3 = 378$), it can be verified that, for the cases involving 2 or 3 factors (simulations 1, 2, 3, 4, 5, and 6), the comparison between the MSA means of multivariate normal data and its respective dichotomized data have not presented significant differences in 6 (8%) cases. In 78 (48%) cases the MSA mean was significantly higher for multivariate normal data and in other 78 (48%) cases it was higher for dichotomized data. In the cases involving 4 to 10 factors and 8 to 50 variables (simulations 7 to 14), were verified 6 (3%) cases where the difference between the MSA means of multivariate normal data and its respective dichotomized data have not presented significant differences. In 21 (10%) cases the MSA mean was significantly higher for multivariate normal data and in the other 189 (87%) cases it was higher for dichotomized data.

In the cases involving 2 or 3 factors, it was verified that the differences were not influenced by the sample size, and in cases with 4 to 10 factors, MSA mean was higher for multivariate normal data only in small samples, with 2, 3 or 4 times the number of variables. Therefore, it can be concluded specially cases where the factor number is higher than 3 and the sample size corresponds to at least 5 times the number of variables, resulted in a higher MSA mean for dichotomized data. These results show that, in this situation, dichotomized data are adequate for the application of factor analysis.

According to the results obtained for the test of mean difference of the variance explained by the first factor between multivariate normal samples and its corresponding dichotomized samples, it was verified that all the 378 cases studied showed significant differences, with multivariate normal data means always higher. For samples with 2 or 3 factors (5 to 10 variables), the minimum ratio between D/N

EFFECTS OF THE DICHOTOMIZATION OF MULTIVARIATE DATA

(proportion of the dichotomized data mean to the multivariate normal mean) was 77.72% and the maximum 91.42%, as for samples with 4 to 10 factors (15 to 20 variables), the ratio was of 64.02% and 80.91% respectively. Therefore, for smaller numbers of factors (2 or 3) the loss of explanation by the first factor when data is dichotomized is less intense than in cases involving higher numbers of factors (4 to 10)

The test of mean difference for the total variance explained of the 378 cases studied have presented significant differences, with means always higher for multivariate normal data. For samples with 2 or 3 factors (5 to 10 variables) the D/N ratio ranged from 81.24% to 89.98%, and for samples with 4 to 10 factors (15 to 50 variables) ranged from 78.87% to 87.16% respectively. Therefore, similarly to the case of the variance explained by the first factor, the total variance explained also presents better results for smaller numbers of factors and variables.

The comparisons between mean vectors of the communalities of multivariate normal samples and the corresponding mean vectors of the communalities of dichotomized samples resulted in significant differences for the 378 cases studied. The components of these vectors were always higher for multivariate normal data. The D/N ratio between components for samples with 2 or 3 factors (5 to 10 variables) ranged from 71.07% to 93.67%, and for samples with 4 to 10 factors (15 to 50 variables) ranged from 55.79% to 94.54%. Therefore, it can be concluded for samples with smaller numbers of factors, the communalities results for dichotomized data, in relation to multivariate normal data, presents better results.

The substitution of multivariate normal data by its corresponding dichotomized data, using the phi correlation coefficient to calculate the correlation matrix, as an alternative to the tetrachoric correlation coefficient (since it is not possible to use this coefficient), will be always viable within the conditions analyzed for the MSA, variance explained by the first factor, total variance explained and communalities.

References

- Embreson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. New York, NY: Psychology Press.
- Fedorov, V., Mannino, F., & Zhang, R. (2008). Consequences of dichotomization. *Pharmaceutical Statistics*, 8(1), 50-61. doi: [10.1002/pst.331](https://doi.org/10.1002/pst.331)

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 45(1). Retrieved from http://www.pabst-publishers.de/psychology-science/1-2003/abstract_06.html

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19-40. doi: [10.1037/1082-989X.7.1.19](https://doi.org/10.1037/1082-989X.7.1.19)

Pearson, R. H., & Mundform, D. J. (2010). Recommended sample size for conducting exploratory factor analysis on dichotomous data. *Journal of Modern Applied Statistical Methods*, 9(2), 359-368. doi: [10.22237/jmasm/1288584240](https://doi.org/10.22237/jmasm/1288584240)

Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique

Oyebayo Ridwan Olaniran
Universiti Tun Hussein Onn Malaysia
Muar, Johor, Malaysia

Waheed Babatunde Yahya
University of Ilorin
Ilorin, Kwara State, Nigeria

The most important ingredient in Bayesian analysis is prior or prior distribution. A new prior determination method was developed under the framework of parametric empirical Bayes using bootstrap technique. By way of example, Bayesian estimations of the parameters of a normal distribution with unknown mean and unknown variance conditions were considered, as well as its application in comparing the means of two independent normal samples with several scenarios. A Monte Carlo study was conducted to illustrate the proposed procedure in estimation and hypothesis testing. Results from Monte Carlo studies showed that the bootstrap prior proposed is more efficient than the existing method for determining priors and also better than the frequentist methods reviewed.

Keywords: Prior, conjugacy, bootstrapping, hypothesis testing, Monte Carlo studies

Introduction

Bayesian statistics have several advantages over the traditional classical (frequentist) statistics ranging from proffering solution to problems related to estimation, testing hypotheses, or estimating confidence regions for unknown parameters. The reason is by use of Bayes' theorem probability density functions are obtained for the unknown parameters. These density functions allow for the estimation of unknown parameters, the testing of hypotheses, and the computation of confidence regions often referred to as the credible interval. Therefore, application of Bayesian statistics has been spreading (Koch, 2007). The process of inductive learning via Bayes' rule is referred to as Bayesian inference (Hoff, 2009).

The Bayesian inference utilizes the posterior distribution $p(\theta | y)$ which describes our belief that θ is the true value, having observed dataset y . The posterior

Oyebayo Ridwan Olaniran is a PhD student and Graduate Research Assistant in the Department of Mathematics and Statistics, Faculty of Applied Science and Technology. Email him at: rid4stat@yahoo.com. Waheed Babatunde Yahya is an Associate Professor in the Department of Statistics, Faculty of Physical Sciences. Email them at dr.yah2009@gmail.com.

distribution is obtained from the prior distribution and sampling model via Bayes' rule:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{\int_{\theta} p(y | \theta)p(\theta) d\theta} \quad (1)$$

The expression given by (1) above is the general Bayes theorem for inference and is the basis for making inferences from a Bayesian perspective in terms of estimation, hypothesis testing, and obtaining credible intervals, as well as making direct probability statements about the quantities in which we are interested (Spiegelhalter, Abrams, & Myles, 2004).

The Bayes theorem is commonly written in its proportional form as $p(\theta | y) \propto p(y | \theta)p(\theta)$. Bayesian inference is based on the posterior distribution, which is the conditional distribution of the parameters or unobserved covariates given the observed data. The posterior distribution summarizes all the information about the parameters and covariates. For example, the mean, median, or mode of the posterior distribution could be used as point estimators. Bayesian inference for θ is then based on the posterior distribution $p(y | \theta)$. For example, a Bayesian estimator of θ is the posterior mean

$$\hat{\theta} = E(\theta | y) = \int_{\theta} \theta p(\theta | y) d\theta$$

A Bayesian analogue to a confidence interval is the credible interval, which is a region with probability $1 - \alpha$ under the posterior distribution. Choices of prior distributions are important. In fact, much of the controversy regarding Bayesian methods revolves around the prior distributions (Wu, 2010).

Priors and Prior Distributions

Priors are carriers of information that is coherently incorporated via Bayes' theorem to the inference. At the same time, parameters are unobservable, and prior specification is subjective in nature. There are two different schools of thought to be considered when choosing priors in Bayesian analysis (Rouder, Speckman, Sun, & Morey, 2009). The first is the subjective Bayes school, which believe that priors should reflect the analyst's *a priori* beliefs about parameters. Usually, these beliefs are informed by the theoretical and experimental context. The second is the objective Bayes school, in which the priors are meant to reflect as few assumptions

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

as possible. Bayes himself proposes a class of uniform prior for the parameter p of a binomial distribution communicated to the Royal Statistical Society by Price in 1763 (Bayes & Price, 1763).

Laplace in the early 1770s extended this prior belief in his principle of insufficient reason and termed it as flat prior. Efron (2012) claimed that the Bayesian/frequentist controversy centers on the use of Bayes' rule in the absence of genuine prior experience. Due to the parameter invariant problem involved when using the uniform prior, Jeffreys (1949) proposed another class of prior that is invariant to parameter transformation (Lesaffre & Lawson, 2013). Both the uniform flat prior and Jeffrey prior are usually referred to as objective or non-informative prior.

In the search of genuine or informative prior, Raifa and Schlaifer proposed the conjugate prior in 1961 (as reported in Bolstad, 2004). The conjugate prior ensures that the posterior distribution class is the same as the prior distribution. Conjugacy is mathematically convenient in that the posterior distribution follows a known parametric form (Gelman et al., 2013). If information is available that contradicts a conjugate parametric family, then it may be necessary to use a more realistic but often inconvenient prior distribution. A conjugate prior can be made informative or non-informative depending on the parameter value assumed. Yahya, Olaniran, and Ige (2014) claimed that the conjugate prior approach needs to be updated when the genuine prior parameter is not available. Using a conjugate prior does not necessarily guarantee an adequate posterior unless the parameter of the prior distribution is correctly specified. The adverse effect of incorrect prior specification is when the prior information did not agree with the data information which might lead to incorrect estimation or inference about the unknown parameter. Solving this problem led to the proposition of empirical Bayes in the early 1950s by Robbins as reported in Robbins (1956), Martiz (1970), Efron and Morris (1973, 1975, 1976), Morris (1983), Casella (1985), Bishop (2005), and recently in Efron (2012, 2013, 2014), Lee (2012), and Lesaffre and Lawson (2013).

Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data (Lee, 2012). This method is often perceived in two forms: the parametric (known functional form) and non-parametric (unknown functional form). Parametric empirical Bayes usually involve the use of conjugate prior with the prior parameters estimated from the data using the maximum likelihood estimation (MLE) method or method of moment (MM) (Lee, 2012). Efron (2014) reported the use of empirical Bayes methods is increasing, although still suffers from an uncertain theoretical basis, enjoying neither the safe haven of Bayes' theorem nor the steady support of frequentist optimality. Their

rationale is often reduced to inserting more or less obvious estimates into familiar Bayesian formulas. This conceals the essential empirical Bayes task: learning an appropriate prior distribution from ongoing statistical experience, rather than knowing it by assumption.

Efficient learning requires both Bayesian and frequentist modeling strategies. Bayesian statistics with well-known distributions are often smooth and easy with the use of conjugate priors with adequate prior parameter specification using subjective or empirical Bayes method. However, in most real life situations, it is often difficult to describe using existing or known functional form. Posterior distributions under this situation are often estimated using Monte Carlo integration or methods popularly referred to as Markov chain Monte Carlo (MCMC) methods. MCMC methods ranges from the Gibbs sampler (Casella & George, 1992), expectation maximization (EM) algorithm, to the Metropolis-Hasting (MH) algorithm (Lee, 2012).

Currently, the focus is on updating the parametric empirical Bayes procedure using bootstrapping resampling procedures. The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data (Efron & Tibshirani, 1993). It amounts to treating the data as if they were the population for the purpose of evaluating the distribution of interest. Under mild regularity conditions, the bootstrap yields an approximation to the distribution of an estimator or test statistic that is at least as accurate as the approximation obtained from first-order asymptotic theory. Thus, the bootstrap provides a way to substitute computation for mathematical analysis if calculating the asymptotic distribution of an estimator or statistic is difficult.

Therefore, the aim of this study is to develop efficient alternative methods for determining the priors within the Bayesian framework using bootstrapping techniques. The usefulness of the proposed method in classical hypothesis testing is demonstrated for comparing two population means from two independent samples. The efficiencies of the proposed methods shall be determined and compared with some of the existing frequentists and Bayesian test methods under different parameters combinations.

Methodology

Consider random sample $y_1, y_2, \dots, y_n$ from $N(\mu, \sigma^2)$. The density function of y is given as

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

$$p(y | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right] \quad (2)$$

The problem is how to effectively estimate the location and scale parameters μ and σ^2 , respectively. The Bayes estimation procedures for μ and σ^2 require estimation of the posterior distribution of μ and σ^2 given y . The posterior density following Bayes' theorem is

$$p(\mu, \sigma^2 | y) = \frac{p(\mu, \sigma^2) p(y | \mu, \sigma^2)}{\int_{\mu} \int_{\sigma^2} p(\mu, \sigma^2) p(y | \mu, \sigma^2) d\sigma^2 d\mu} \quad (3)$$

Bolstad (2004), Murphy, (2007), and Lesaffre and Lawson (2013), among others, used the Normal-Gamma $NG(\mu_0, n_0, \alpha_0, \beta_0)$ natural conjugate prior for μ and $\lambda = \sigma^2$, given as

$$p(\mu, \lambda) = \frac{[\lambda n_0]^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda n_0}{2}(\mu - \mu_0)^2\right] \times \frac{\beta_0^{\alpha_0} \lambda^{\alpha_0-1} \exp(-\lambda \beta_0)}{\Gamma \alpha_0}$$

To characterize the information from the data $D = y_1, y_2, \dots, y_n$, define the likelihood function $L(D | \mu, \sigma^2)$:

$$\begin{aligned} L(D | \mu, \lambda) &= \prod_{i=1}^n \frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(y_i - \mu)^2\right] \\ &= \left(\frac{\lambda^{\frac{1}{2}}}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{\lambda}{2} \sum_{i=1}^n (y_i - \mu)^2\right] \end{aligned}$$

The posterior distribution is of the form

$$p(\mu, \lambda | y) = \frac{p(\mu, \lambda) L(D | \mu, \lambda)}{\int_{\mu} \int_{\lambda} p(\mu, \lambda) L(D | \mu, \lambda) d\lambda d\mu}$$

Murphy (2007) gave the solution of the posterior density which is also Normal-Gamma, i.e. NG($\mu_n, n_n, \alpha_n, \beta_n$) where

$$\mu_n = \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \quad (4)$$

$$n_n = n_0 + n \quad (5)$$

$$\alpha_n = \alpha_0 + \frac{n}{2} \quad (6)$$

$$\beta_n = \beta_0 + \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{n_0 n (\bar{y} - \mu_0)^2}{2(n_0 + n)} \quad (7)$$

and

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Thus, the Bayes estimate of μ is

$$\hat{\mu} = \mu_n$$

and, from (4),

$$\begin{aligned} \hat{\mu} &= \frac{n_0\mu_0 + n\bar{y}}{n_0 + n} \\ &= \frac{n\bar{y}}{n_0 + n} + \frac{n_0\mu_0}{n_0 + n} \end{aligned} \quad (8)$$

Let

$$w = \frac{n}{n_0 + n}$$

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

Then

$$\begin{aligned}\hat{\mu} &= w\bar{y} + (1-w)\mu_0 \\ &= w(\bar{y} - \mu_0) + \mu_0\end{aligned}$$

Similarly, the Bayes estimate of σ^2 is determined by

$$\hat{\sigma}^2 = \frac{\beta_n}{\alpha_n}$$

and from (6) and (7)

$$\hat{\sigma}^2 = \frac{\beta_0 + \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{n_0 n (\bar{y} - \mu_0)^2}{2(n_0 + n)}}{\alpha_0 + \frac{n}{2}} \quad (9)$$

The empirical Bayes version of the above estimate involves estimating the prior parameters $\pi = (\mu_0, n_0, \alpha_0, \beta_0)$ from the data. Thus the empirical Bayes estimate of μ and σ^2 are

$$\hat{\mu}_{EB} = \frac{\hat{n}_0 \hat{\mu}_0 + n \bar{y}}{\hat{n}_0 + n} \quad (10)$$

and

$$\hat{\sigma}_{EB}^2 = \frac{\hat{\beta}_0 + \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\hat{n}_0 n (\bar{y} - \hat{\mu}_0)^2}{2(\hat{n}_0 + n)}}{\hat{\alpha}_0 + \frac{n}{2}} \quad (11)$$

The proposed bootstrap Bayesian version of the estimate of μ and σ^2 involves the following steps:

1. Generation of bootstrap samples from the original data a desired number of times B ,

2. Estimating the hyperparameters (prior parameters) each time the samples are generated using the maximum likelihood (ML) method,
3. Updating the posterior estimates using the hyperparameters in step 2 above using (8) and (9), and
4. Then obtaining the proposed bootstrap empirical Bayesian estimates $\hat{\mu}_{BT}$ and $\hat{\sigma}_{BT}^2$ using

$$\hat{\mu}_{BT} = \frac{1}{B} \sum_{j=1}^B \hat{\mu}_{EBj}$$

$$\hat{\sigma}_{BT}^2 = \frac{1}{B} \sum_{j=1}^B \hat{\sigma}_{EBj}^2$$

The $\hat{\mu}_{BT}$ proposed here has good statistical properties in terms of bias and mean square error (MSE).

To evaluate bias,

$$\begin{aligned} \hat{\mu}_{BT} &= \frac{1}{B} \sum_{j=1}^B \hat{\mu}_{EBj} \\ &= \frac{1}{B} \sum_{j=1}^B \frac{\hat{n}_{0j} \hat{\mu}_{0j} + n \bar{y}}{\hat{n}_{0j} + n} \\ &= \frac{1}{B} \sum_{j=1}^B \left[\frac{n \bar{y}}{\hat{n}_{0j} + n} + \frac{\hat{n}_{0j} \hat{\mu}_{0j}}{\hat{n}_{0j} + n} \right] \end{aligned}$$

Fixing $\hat{n}_{0j} = B$ and $\hat{\mu}_{0j} = \bar{y}_{bj}$, where $\bar{y}_{bj}$ is the j^{th} ML estimate based on the j^{th} y_b bootstrap sample drawn, i.e.

$$\bar{y}_{bj} = \frac{\sum_{i=1}^n y_{ib}}{n}$$

then

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

$$\begin{aligned}
 \hat{\mu}_{BT} &= \frac{1}{B} \sum_{j=1}^B \left[\frac{n\bar{y}}{B+n} + \frac{B\bar{y}_{bj}}{B+n} \right] \\
 &= \frac{1}{B(B+n)} \sum_{j=1}^B [n\bar{y} + B\bar{y}_{bj}] \\
 &= \frac{n\bar{y}}{(B+n)} + \frac{\sum_{j=1}^B \bar{y}_{bj}}{(B+n)}
 \end{aligned}$$

Hence

$$\begin{aligned}
 E[\hat{\mu}_{BT}] &= E \left[\frac{n\bar{y}}{(B+n)} + \frac{\sum_{j=1}^B \bar{y}_{bj}}{(B+n)} \right] \\
 &= \frac{nE[\bar{y}]}{(B+n)} + \frac{\sum_{j=1}^B E[\bar{y}_{bj}]}{(B+n)}
 \end{aligned}$$

Because $\bar{y}$ and $\bar{y}_{bj}$ are known unbiased estimates of μ ,

$$\begin{aligned}
 E[\hat{\mu}_{BT}] &= \frac{n\mu}{(B+n)} + \frac{\sum_{j=1}^B \mu}{(B+n)} \\
 &= \frac{1}{(B+n)} [n\mu + B\mu] \\
 &= \mu
 \end{aligned}$$

Therefore, $\hat{\mu}_{BT}$ is unbiased for estimating μ .

Also, the MSE is the combination of square of bias and variance of the estimate, then following from the above derivation the MSE is just the variance of the estimate. Thus

$$\text{var}[\hat{\mu}_{BT}] = \text{var} \left[\frac{n\bar{y}}{(B+n)} + \frac{\sum_{j=1}^B \bar{y}_{bj}}{(B+n)} \right]$$

$$\begin{aligned}
 &= \frac{n^2 \text{var}[\bar{y}]}{(B+n)^2} + \frac{\sum_{j=1}^B \text{var}[\bar{y}_{bj}]}{(B+n)^2} \\
 &= \frac{n^2 \frac{\sigma^2}{n}}{(B+n)^2} + \frac{\sum_{j=1}^B \frac{\sigma^2}{n}}{(B+n)^2} \\
 &= \left[\frac{n^2 + B}{(B+n)^2} \right] \frac{\sigma^2}{n}
 \end{aligned}$$

Hence, it can be show that the limiting form of

$$\left[\frac{n^2 + B}{(B+n)^2} \right]$$

is 0 by applying L'Hôpital's rule ([Weisstein, 2003](#)):

$$\begin{aligned}
 \lim_{B \rightarrow \infty} \left[\frac{n^2 + B}{(B+n)^2} \right] &= \lim_{B \rightarrow \infty} \left[\frac{\frac{d[n^2 + B]}{dB}}{\frac{d(B+n)^2}{dB}} \right] \\
 &= \lim_{B \rightarrow \infty} \left[\frac{1}{2(B+n)} \right] \\
 &\rightarrow \frac{1}{\infty}
 \end{aligned}$$

Hence

$$\lim_{B \rightarrow \infty} \left[\frac{n^2 + B}{(B+n)^2} \right] = 0$$

The above derivation implies that at a fixed sample size n , the

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

$$\lim_{B \rightarrow \infty} \text{var}[\hat{\mu}_{\text{BT}}] = 0$$

This affirms that the experimenter can control the stability of the estimator by increasing the number of bootstrap sample B . In addition,

$$\text{var}[\hat{\mu}_{\text{BT}}] = \left[\frac{n^2 + B}{(B + n)^2} \right] \frac{\sigma^2}{n} < \text{var}[\hat{\mu}_{\text{ML}}] = \frac{\sigma^2}{n}$$

which implies that the proposed estimator is more efficient than the frequentist ML estimator. This comparison is reasonable because they are both unbiased. Also within the Bayesian realm, it could be observe that the proposed estimator is also more efficient since it minimizes the MSE in terms of bias and variance reduction. The traditional Bayesian estimator minimizes the MSE by reducing the variance alone.

Application to Two-Sample Hypothesis Testing

Consider the situation in which we have independent samples from two normal distributions

$$\begin{aligned} x_1, x_2, \dots, x_{n_1} &\sim N(\mu_1, \sigma_1^2) \\ y_1, y_2, \dots, y_{n_2} &\sim N(\mu_2, \sigma_2^2) \end{aligned}$$

which are independent of each other, and the quantity of interest really is the posterior distribution of

$$\delta = \mu_1 - \mu_2$$

The hypothesis of interest under this scenario might be of the form

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2 \quad (12)$$

or similarly in terms of δ

$$H_0 : \delta = 0 \text{ against } H_1 : \delta \neq 0 \quad (13)$$

Testing the above hypotheses in (12) and (13) using the Bayesian method requires computing $p(\delta | D)$ defined as the posterior distribution of δ given data D (Lee, 2012). The posterior probability of the null hypothesis H_0 can then be estimated using

$$p(H_0 : \delta \leq 0 | D) = \int_{-\infty}^0 p(\delta | D) d\delta \quad (14)$$

If this probability is less than the chosen α , reject the null hypothesis and conclude that H_1 holds. But (14) will fail if the null hypothesis is simple as in the case of (12) and (13) above (because the probability of a specific point on a continuous interval is 0). Bolstad (2004) and Lee (2012) suggested the use of credible interval under this condition. The credible interval for a specified significance level α for parameter δ is

$$p(a \leq \delta \leq b | D) = \int_a^b p(\delta | D) d\delta = 1 - \alpha$$

On construction of the credible interval $[a, b]$ using a specified significance level α , δ is said to be credible if it lies within the interval $[a, b]$, and thus H_0 holds; otherwise, H_1 holds. The bootstrap Bayesian estimates can be used here to determine the posterior density or posterior samples of δ by using the formulae above. The bootstrap Bayesian estimate of parameter δ is $\hat{\delta}_{BT} = \hat{\mu}_{1BT} - \hat{\mu}_{2BT}$. The posterior density of δ using the bootstrap Bayesian approach will likely approach the Gaussian distribution if one is to follow the central limit theorem (since the bootstrap prior distribution used is the sampling distribution of means which is Gaussian). Thus the posterior density of δ under this assumption is $N(\hat{\mu}_{1BT} - \hat{\mu}_{2BT}, \text{var}[\hat{\mu}_{1BT}] + \text{var}[\hat{\mu}_{2BT}])$. This then implies we can construct a frequentist-like test-statistic for the unknown parameter δ as

$$Z_{BT} = \frac{\delta - \hat{\delta}_{BT}}{\sqrt{\text{var}[\hat{\mu}_{1BT}] + \text{var}[\hat{\mu}_{2BT}]}} \sim N(0,1) | H_0 \quad (15)$$

In another parlance, Lee (2012) claimed that to correct for small sample bias and unequal variance bias, the Student's t -distribution with ν degree of freedom would provide a better approximation than the Gaussian distribution. Thus a modification to (15) above is

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

$$t_{BT} = \frac{\delta - \hat{\delta}_{BT}}{\sqrt{\text{var}[\hat{\mu}_{1BT}] + \text{var}[\hat{\mu}_{2BT}]}} \sim t_v \mid H_0 \quad (16)$$

The parameter v of the Student's t -distribution used here indicates the effective sample size to be used for the hypothesis testing. Hence, using $v = \min(n_1, n_2)$ is proposed here.

It is pertinent to note that the above equations (15) and (16) are approximate distributions of $\hat{\delta}_{BT}$. The hypothesis can be tested directly by computing $p(\delta_{BT} \mid D)$ using the difference of the posteriors (generated using bootstrap priors) for the parameters μ_1 and μ_2 . In this regard, the posterior probability of the null hypothesis H_0 can then be estimated using

$$p(H_0 : \delta = 0 \mid D) \approx 2 \left[\min \left(\int_{-\infty}^0 p(\delta_{BT} \mid D) d\delta, \int_0^{\infty} p(\delta_{BT} \mid D) d\delta \right) \right] \quad (17)$$

In the same parlance, posterior probability of the null hypothesis H_0 for (15) and (16) are, respectively,

$$p(H_0 : \delta = 0 \mid D) \approx 2 \left[\min(\Phi(z), 1 - \Phi(z)) \right] \quad (18)$$

and

$$p(H_0 : \delta = 0 \mid D) \approx 2 \left[\min(\psi_v(z), 1 - \psi_v(z)) \right] \quad (19)$$

where $\Phi(z)$ is the cumulative distribution of the standard normal variate z and ψ_v is the cumulative distribution of the Student's t -distribution with mean 0, variance 1, and degrees of freedom v .

The above procedures in (15), (16), and (17) will be evaluated to ascertain which to recommend under specific situation. Consideration of the Bayesian MCMC approach to estimation and testing of equality in means for two groups proposed by Kruschke (2011, 2013) and Kruschke, Aguinis, and Joo (2012) was also achieved. This approach is already implemented in the R statistical package via the package BEST (Kruschke & Meredith, 2014). As a standard check, two frequentist procedures were also considered. The frequentist procedures considered are the pooled variance t -test and unequal variance Welch test (Montgomery & Runger, 2003).

Simulation

To illustrate the proposed bootstrap empirical Bayesian procedures in estimation and hypothesis testing, two Monte Carlo samples were generated from univariate normal distributions with the following mean structures: $\mu_1 = 10$ and $\mu_2 = \mu_1 + \delta$, where $\delta = 0, 1, 2$. The cases of equal and unequal variances were considered with equal variance case define as $\sigma_1^2 = \sigma_2^2 = 4$ and unequal variance case define as $\sigma_1^2 = 4$ and $\sigma_2^2 = 16$. Under equal sample condition, five sample size, $n_1 = n_2 = 5, 10, 20, 30$, and 50 , were used representing sample ranges of extreme low to large sample. Similarly for unequal sample condition, three sample structures were considered, namely $n_1 = 5, n_2 = 10$; $n_1 = 10, n_2 = 30$; and $n_1 = 20, n_2 = 80$. The bootstrap size (B) and number of iterations used were fixed at 1000 .

Results

The empirical type-I error rate (false positive rate) and power (true positive rate) were computed using the frequentist and Bayesian procedures discussed. The role of sample size cannot be overemphasized in estimation and hypothesis testing, therefore more emphasis will be laid on the sample size regarding the results obtained from various procedures used here.

The validity of test procedures can be assessed using the empirical type-I error rate which is the probability that a test function wrongly rejects the null hypothesis when it is true. A test procedure yielding a false positive that is close to the nominal level is often regarded as having been valid. In light of this, the first situation, or Case I in Table 1, considered the common assumptions (equal sample and homoscedasticity) involved while comparing two normal samples. The frequentist traditional pooled t -test produced on average (over all sample sizes) false positive rates that are relatively close to the nominal (0.05) level.

However, this result was not the best if the comparison is made over all the test procedures employed in this paper. For instance, within the Bayesian test procedures, the proposed t_{BT} test procedures produced false positive rates that are closer to the nominal (0.05) level at all the sample sizes considered. Therefore, it can be re-affirmed that the traditional pooled t -test is valid but the proposed t_{BT} is more valid as it yielded an overall average of the empirical type-I error (0.051) that is relatively closer to the nominal 5% level set for the test than the overall average of 0.046 provided by the pooled t -test. The performances of all the tests considered

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

as described above are clearly depicted in the various graphical plots provided in Figure 1 for better understanding.

Table 1. Proportion of empirical type-I error (false positive) rate based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 0$ for the various methods under equal and unequal variance conditions

Sample sizes	Frequentist		Bayesian			
	Welch	Pooled	Existing	Proposed		
			MCMC	BT	Z _{BT}	t _{BT}
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.040	0.051	0.007	0.145	0.119	0.051
$n_1 = n_2 = 10$	0.050	0.052	0.032	0.106	0.080	0.050
$n_1 = n_2 = 20$	0.040	0.042	0.036	0.077	0.062	0.048
$n_1 = n_2 = 30$	0.045	0.046	0.045	0.080	0.066	0.051
$n_1 = n_2 = 50$	0.041	0.041	0.039	0.069	0.067	0.057
Average	0.043	0.046	0.032	0.095	0.079	0.051
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.046	0.054	0.013	0.156	0.137	0.056
$n_1 = n_2 = 10$	0.057	0.059	0.045	0.105	0.098	0.058
$n_1 = n_2 = 20$	0.032	0.035	0.028	0.065	0.057	0.039
$n_1 = n_2 = 30$	0.050	0.052	0.046	0.067	0.063	0.055
$n_1 = n_2 = 50$	0.040	0.040	0.041	0.066	0.058	0.051
Average	0.045	0.048	0.035	0.092	0.083	0.052
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.041	0.044	0.011	0.109	0.106	0.034
$n_1 = 10, n_2 = 30$	0.062	0.053	0.041	0.093	0.088	0.058
$n_1 = 20, n_2 = 80$	0.051	0.056	0.051	0.069	0.069	0.055
Average	0.051	0.051	0.034	0.090	0.088	0.049
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.040	0.019	0.018	0.095	0.084	0.029
$n_1 = 10, n_2 = 30$	0.049	0.006	0.034	0.075	0.069	0.040
$n_1 = 20, n_2 = 80$	0.048	0.005	0.045	0.064	0.062	0.053
Average	0.046	0.010	0.032	0.078	0.072	0.041
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.066	0.120	0.024	0.136	0.128	0.073
$n_1 = 10, n_2 = 30$	0.042	0.163	0.037	0.097	0.094	0.050
$n_1 = 20, n_2 = 80$	0.044	0.159	0.040	0.062	0.063	0.052
Average	0.051	0.147	0.034	0.098	0.095	0.058

Table 2. Proportion of power (true positive rate) based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 1$ for the various methods under equal and unequal variance conditions

Sample sizes	Frequentist		Existing	Bayesian		
	Welch	Pooled	MCMC	Proposed		
				BT	Z _{BT}	t _{BT}
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.087	0.094	0.020	0.245	0.202	0.093
$n_1 = n_2 = 10$	0.171	0.176	0.126	0.262	0.257	0.178
$n_1 = n_2 = 20$	0.327	0.328	0.296	0.389	0.374	0.339
$n_1 = n_2 = 30$	0.485	0.485	0.458	0.531	0.532	0.504
$n_1 = n_2 = 50$	0.690	0.691	0.691	0.740	0.741	0.730
Average	0.352	0.355	0.318	0.433	0.421	0.369
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.075	0.084	0.015	0.196	0.167	0.085
$n_1 = n_2 = 10$	0.100	0.105	0.076	0.157	0.144	0.106
$n_1 = n_2 = 20$	0.162	0.168	0.145	0.200	0.202	0.175
$n_1 = n_2 = 30$	0.225	0.232	0.215	0.265	0.255	0.238
$n_1 = n_2 = 50$	0.353	0.358	0.348	0.416	0.414	0.396
Average	0.183	0.189	0.160	0.247	0.236	0.200
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.114	0.133	0.043	0.227	0.221	0.102
$n_1 = 10, n_2 = 30$	0.253	0.262	0.205	0.333	0.338	0.256
$n_1 = 20, n_2 = 80$	0.505	0.496	0.480	0.570	0.569	0.521
Average	0.291	0.297	0.243	0.377	0.376	0.293
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.074	0.034	0.028	0.147	0.137	0.050
$n_1 = 10, n_2 = 30$	0.166	0.040	0.132	0.211	0.219	0.145
$n_1 = 20, n_2 = 80$	0.355	0.093	0.331	0.409	0.411	0.366
Average	0.198	0.056	0.164	0.256	0.256	0.187
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.065	0.159	0.036	0.174	0.165	0.082
$n_1 = 10, n_2 = 30$	0.117	0.260	0.098	0.176	0.172	0.133
$n_1 = 20, n_2 = 80$	0.177	0.399	0.170	0.218	0.218	0.192
Average	0.120	0.273	0.101	0.189	0.185	0.136

The second scenario, Case II in Table 1, is the case where the frequentist Welch t -test has been established to be better. Here the equal sample sizes, but with heteroscedastic situation, was considered. As expected, the Welch t -test yielded

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

false positive rates that are relatively closer to the nominal (0.05) level. The result of the proposed Bayesian t_{BT} test is not worst-off here as it equally produced false positive rates that are quite close to the 5% nominal level and competes favorably with the results of the Welch test.

Moving to unequal sample sizes and heteroscedastic situations (Cases IV and V), similar results as observed with equal sample and unequal variance situations were observed. To assess the usability of the test procedures, the true positive (power) as assessment criteria was employed. The most powerful test procedures under the varying scenarios is the BT method which is the Bayesian method using the direct bootstrap distribution as can be observed from the various results in Tables 2 and 3 under various parameters combinations. The powers of this test method appreciated better as the values of the effect size, δ increases.

Figure 1. Plots of false positive rate and power (true positive rate) for various scenarios and sample sizes

Table 3. Proportion of power (true positive rate) based on 1000 simulations at varying sample sizes (n_1, n_2) and $\delta = 2$ for the various methods under equal and unequal variance conditions

Sample sizes	Frequentist		Existing	Bayesian		
	Welch	Pooled	MCMC	BT	Z _{BT}	t _{BT}
Case I: Equal sample sizes and equal variance						
$n_1 = n_2 = 5$	0.254	0.279	0.078	0.485	0.468	0.282
$n_1 = n_2 = 10$	0.557	0.564	0.447	0.651	0.657	0.569
$n_1 = n_2 = 20$	0.869	0.871	0.846	0.889	0.891	0.874
$n_1 = n_2 = 30$	0.962	0.962	0.957	0.975	0.980	0.969
$n_1 = n_2 = 50$	0.999	0.999	0.997	0.999	0.999	0.999
Average	0.728	0.735	0.665	0.800	0.799	0.739
Case II: Equal sample sizes and unequal variance						
$n_1 = n_2 = 5$	0.135	0.154	0.049	0.297	0.274	0.152
$n_1 = n_2 = 10$	0.255	0.267	0.193	0.365	0.359	0.269
$n_1 = n_2 = 20$	0.480	0.488	0.433	0.560	0.540	0.494
$n_1 = n_2 = 30$	0.662	0.670	0.648	0.718	0.714	0.689
$n_1 = n_2 = 50$	0.880	0.882	0.872	0.907	0.908	0.900
Average	0.482	0.492	0.439	0.569	0.559	0.501
Case III: Unequal sample sizes and equal variance						
$n_1 = 5, n_2 = 10$	0.368	0.423	0.172	0.557	0.547	0.355
$n_1 = 10, n_2 = 30$	0.718	0.769	0.663	0.813	0.810	0.727
$n_1 = 20, n_2 = 80$	0.969	0.976	0.963	0.983	0.984	0.975
Average	0.685	0.723	0.599	0.784	0.780	0.686
Case IV: Unequal sample sizes and unequal variance with large variance in large sample direction						
$n_1 = 5, n_2 = 10$	0.205	0.101	0.094	0.332	0.352	0.161
$n_1 = 10, n_2 = 30$	0.506	0.250	0.444	0.576	0.572	0.475
$n_1 = 20, n_2 = 80$	0.869	0.598	0.849	0.903	0.897	0.879
Average	0.527	0.316	0.462	0.604	0.607	0.505
Case V: Unequal sample sizes and unequal variance with large variance in small sample direction						
$n_1 = 5, n_2 = 10$	0.131	0.276	0.061	0.290	0.279	0.157
$n_1 = 10, n_2 = 30$	0.280	0.530	0.248	0.410	0.402	0.316
$n_1 = 20, n_2 = 80$	0.536	0.812	0.521	0.611	0.613	0.573
Average	0.316	0.539	0.277	0.437	0.431	0.349

Conclusion

Efficient Bayesian methods were developed for testing equality of two population means from two independent samples. Among all the tests methods considered, it

can be concluded that the most suitable test method is the proposed Bayesian t_{BT} method giving its high level of validity as demonstrated by various results obtained.

References

- Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53, 370-418. doi: [10.1098/rstl.1763.0053](https://doi.org/10.1098/rstl.1763.0053)
- Bishop, C. M. (2005). *Neural networks for pattern recognition*. New York, NY: Oxford University Press
- Bolstad, W. M. (2004). *Introduction to Bayesian statistics*. Hoboken, NJ: John Wiley & Sons. doi: [10.1002/047172212x](https://doi.org/10.1002/047172212x)
- Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2), 83-87. doi: [10.1080/00031305.1985.10479400](https://doi.org/10.1080/00031305.1985.10479400)
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3), 167-174. doi: [10.1080/00031305.1992.10475878](https://doi.org/10.1080/00031305.1992.10475878)
- Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6(4), 1971-1997. doi: [10.1214/12-aos571](https://doi.org/10.1214/12-aos571)
- Efron, B. (2013). *Empirical Bayes modeling, computation, and accuracy* [Unpublished manuscript]. Stanford, CA: Department of Statistics, Stanford University.
- Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science*, 29(2), 285-301. doi: [10.1214/13-sts455](https://doi.org/10.1214/13-sts455)
- Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors – An empirical Bayes approach. *Journal of the American Statistical Association*, 68(341), 117-130. doi: [10.1080/01621459.1973.10481350](https://doi.org/10.1080/01621459.1973.10481350)
- Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311-319. doi: [10.1080/01621459.1975.10479864](https://doi.org/10.1080/01621459.1975.10479864)
- Efron, B., & Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics*, 4(1), 22-32. doi: [10.1214/aos/1176343345](https://doi.org/10.1214/aos/1176343345)
- Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York, NY: Springer Science+Business Media, LLC. doi: [10.1007/978-0-387-92407-6](https://doi.org/10.1007/978-0-387-92407-6)
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 186(1007), 453-461. doi: [10.1098/rspa.1946.0056](https://doi.org/10.1098/rspa.1946.0056)
- Koch, K.-R. (2007). *Introduction to Bayesian statistics*. New York, NY: Springer Science+Business Media, LLC. doi: [10.1007/978-3-540-72726-2](https://doi.org/10.1007/978-3-540-72726-2)
- Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Amsterdam, the Netherlands: Elsevier.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, 142(2), 573-603. doi: [10.1037/a0029146](https://doi.org/10.1037/a0029146)
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, 15(4), 722-752. doi: [10.1177/1094428112457829](https://doi.org/10.1177/1094428112457829)
- Kruschke, J. K., & Meredith, M. (2014). *BEST: Bayesian estimation supersedes the t-test* [R package, version 0.2.2]. Retrieved from <http://CRAN.R-project.org/package=BEST>
- Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Hoboken, NJ: John Wiley & Sons.
- Lesaffre, E., & Lawson, A. B. (2013). *Bayesian biostatistics*. Hoboken, NJ: John Wiley & Sons. doi: [10.1002/9781119942412](https://doi.org/10.1002/9781119942412)
- Maritz, J. (1970). *Empirical Bayes methods*. London, UK: Methuen.
- Montgomery, D. C., & Runger, G. C. (2007). *Applied statistics and probability for engineers* (4th ed.). Hoboken, NJ: Wiley.
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381), 47-65. doi: [10.1080/01621459.1983.10477920](https://doi.org/10.1080/01621459.1983.10477920)
- Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution* [Unpublished manuscript] Vancouver, BC: University of British Columbia Department of Computer Science. Retrieved from <https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf>
- Robbins, H. (1956). An empirical Bayes approach to statistics. In J. Neyman (Ed.), *Proceedings of the third Berkeley Symposium on Mathematical Statistics*

BAYESIAN HYPOTHESIS TESTING USING BOOTSTRAP PRIOR

and Probability. Volume 1. Contributions to the theory of statistics (pp. 157-163). Berkeley, CA: University of California Press.

Rouder, J. N., Speckman, P. L., Sun, D., & Morey, R. D. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225-237. doi: [10.3758/pbr.16.2.225](https://doi.org/10.3758/pbr.16.2.225)

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: Wiley. doi: [10.1002/0470092602](https://doi.org/10.1002/0470092602)

Weisstein, E. W. (2003). L'Hospital's rule. Retrieved from <http://mathworld.wolfram.com/LHospitalsRule.html>

Wu, L. (2010). *Mixed effects models for complex data*. Boca Raton, FL: Chapman and Hall/CRC. doi: [10.1201/9781420074086](https://doi.org/10.1201/9781420074086)

Yahya, W. B., Olaniran, O. R., & Ige, S. O. (2014). On Bayesian conjugate normal linear regression and ordinary least square regression methods: A Monte Carlo study. *Ilorin Journal of Science*, 1(1). 216-227. Retrieved from <http://www.unilorin.edu.ng/ejournals/index.php/ILJS/article/view/1611>

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, <http://digitalcommons.wayne.edu/jmasm>. Three areas are appropriate for *JMASM*:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board. *JMASM* charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. All manuscripts are checked for plagiarism, including self-plagiarism (text recycling), with a zero-tolerance policy. This includes text-to-text or concept-to-concept duplication, figures, tables, and equations. At any time (including after publication), the discovery of a violation of this policy will result in immediate withdrawal by the editor.
2. *JMASM* uses a modified American Psychological Association style guideline.
3. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
4. Manuscripts should be prepared in Microsoft Word (.doc or .docx) only (Wordperfect and .rtf formats may be acceptable – please inquire). Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are NOT acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.

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 submission form requires an Abstract with a 50 word maximum, and a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left justified, indent optional.
 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 the spacebar twice after a period.
-

Journal of Modern Applied Statistical Methods

ISSN: 1538-9472

<http://digitalcommons.wayne.edu/jmasm>

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

Copyrights, Attribution and Usage Policies

Copyright ©2017 JMASM, Inc. *JMASM* retains the copyright for this work for the entire usual period, but grants assignors the right, after one year from the date of publication, to republish the work in whole or in part anywhere and in any format, provided reference is given to the original publication in *JMASM* (see website for further details). Readers may freely access journal content at <http://digitalcommons.wayne.edu/jmasm>.

To Advertisers

Advertisements are accepted at the discretion of the editor. Send requests for advertising information to ea@jmasm.com.

WAYNE STATE
UNIVERSITY
LIBRARY SYSTEM