DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 33

11-1-2007

# Vol. 6, No. 2 (Full Issue)

JMASM Editors

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

Recommended Citation

Editors, JMASM (2007) "Vol. 6, No. 2 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 6: Iss. 2, Article 33.
Available at: http://digitalcommons.wayne.edu/jmasm/vol6/iss2/33

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 administrator of DigitalCommons@WayneState.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-669
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Journal Of Modern Applied Statistical Methods

Shlomo S. Sawilowsky
*Editor*
College of Education
Wayne State University

Harvey Keselman
*Associate Editor*
Department of Psychology
University of Manitoba

Bruno D. Zumbo
*Associate Editor*
Measurement, Evaluation, & Research Methodology
University of British Columbia

Vance W. Berger
*Assistant Editor*
Biometry Research Group
National Cancer Institute

John L. Cuzzocrea
*Assistant Editor*
Educational Research
University of Akron

Todd C. Headrick
*Assistant Editor*
Educational Psychology and Special Education
Southern Illinois University-Carbondale

Alan Klockars
*Assistant Editor*
Educational Psychology
University of Washington

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-669

Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Editorial Board

Subhash Chandra Bagui
Department of Mathematics & Statistics
University of West Florida

J. Jackson Barnette
School of Public Health
University of Alabama at Birmingham

Vincent A. R. Camara
Department of Mathematics
University of South Florida

Ling Chen
Department of Statistics
Florida International University

Christopher W. Chiu
Test Development & Psychometric Rsch
Law School Admission Council, PA

Jai Won Choi
National Center for Health Statistics
Hyattsville, MD

Rahul Dhanda
Forest Pharmaceuticals
New York, NY

John N. Dyer
Dept. of Information System & Logistics
Georgia Southern University

Matthew E. Elam
Dept. of Industrial Engineering
University of Alabama

Mohammed A. El-Saidi
Accounting, Finance, Economics &
Statistics, Ferris State University

Felix Famoye
Department of Mathematics
Central Michigan University

Barbara Foster
Academic Computing Services, UT
Southwestern Medical Center, Dallas

Shiva Gautam
Department of Preventive Medicine
Vanderbilt University

Dominique Haughton
Mathematical Sciences Department
Bentley College

Scott L. Hershberger
Department of Psychology
California State University, Long Beach

Joseph Hilbe
Departments of Statistics/ Sociology
Arizona State University

Sin–Ho Jung
Dept. of Biostatistics & Bioinformatics
Duke University

Jong-Min Kim
Statistics, Division of Science & Math
University of Minnesota

Harry Khamis
Statistical Consulting Center
Wright State University

Kallappa M. Koti
Food and Drug Administration
Rockville, MD

Tomasz J. Kozubowski
Department of Mathematics
University of Nevada

Kwan R. Lee
GlaxoSmithKline Pharmaceuticals
Collegeville, PA

Hee-Jeong Lim
Dept. of Math & Computer Science
Northern Kentucky University

Balgobin Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute

J. Sunil Rao
Dept. of Epidemiology & Biostatistics
Case Western Reserve University

Karan P. Singh
University of North Texas Health
Science Center, Fort Worth

Jianguo (Tony) Sun
Department of Statistics
University of Missouri, Columbia

Joshua M. Tebbs
Department of Statistics
Kansas State University

Dimitrios D. Thomakos
Department of Economics
Florida International University

Justin Tobias
Department of Economics
University of California-Irvine

Dawn M. VanLeeuwen
Agricultural & Extension Education
New Mexico State University

David Walker
Educational Tech, Rsrch, & Assessment
Northern Illinois University

J. J. Wang
Dept. of Advanced Educational Studies
California State University, Bakersfield

Dongfeng Wu
Dept. of Mathematics & Statistics
Mississippi State University

Chengjie Xiong
Division of Biostatistics
Washington University in St. Louis

Andrei Yakovlev
Biostatistics and Computational Biology
University of Rochester

Heping Zhang
Dept. of Epidemiology & Public Health
Yale University

INTERNATIONAL

Mohammed Ageel
Dept. of Mathematics, & Graduate School
King Khalid University, Saudi Arabia

Mohammad Fraiwan Al-Saleh
Department of Statistics
Yarmouk University, Irbid-Jordan

Keumhee Chough (K.C.) Carriere
Mathematical & Statistical Sciences
University of Alberta, Canada

Michael B. C. Khoo
Mathematical Sciences
Universiti Sains, Malaysia

Debasis Kundu
Department of Mathematics
Indian Institute of Technology, India

Christos Koukouvinos
Department of Mathematics
National Technical University, Greece

Lisa M. Lix
Dept. of Community Health Sciences
University of Manitoba, Canada

Takis Papaioannou
Statistics and Insurance Science
University of Piraeus, Greece

Nasrollah Saebi
Computing, Information Systems & Math
Kingston University, UK

Keming Yu
Department of Statistics
University of Plymouth, UK

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-669
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Journal Of Modern Applied Statistical Methods

*Invited Articles*


| | | |
|---|---|---|
| 355 – 360 | **Philip H. Ramsey**, **Patricia P. Ramsey** | Optimal Trimming and Outlier Elimination |
| 361 – 366 | **Rand R. Wilcox** | An Omnibus Test When Using A Regression Estimator With Multiple Predictors |
| 367 – 379 | **Bradley E. Huitema**, **Joseph W. McKean**, **Sean Laraway** | Time-Series Intervention Analysis Using ITSACORR: Fatal Flaw |


*Regular Articles*


| | | |
|---|---|---|
| 380 – 398 | **Lisa M. Lix**, **Anita M. Lloyd** | A Comparison of Procedures for the Analysis of Multivariate Repeated Measurements |
| 399 – 412 | **Scott J. Richter**, **Melinda H. McCann** | Multiple Comparison of Medians Using Permutation Tests |
| 413 – 420 | **Jennifer E. V. Lloyd**, **Bruno D. Zumbo** | The Non-Parametric Difference Score: A Workable Solution for Analyzing Two-Wave Change When The Measures Themselves Change Across Waves |
| 421 – 442 | **S. Jonathan Mends-cole** | Probability Coverage and Interval Length for Welch's and Yuen's Techniques: Shift in Location, Change in Scale, and (Un)Equal Sizes |
| 443 – 455 | **Michèle Weber** | The Effect of Different Degrees of Freedom of the Chi-square Distribution on the Statistical Power of the t, Permutation t, and Wilcoxon Tests |
| 456 – 468 | **Vic Hasselblad**, **Yuliya Lokhnygina** | Tests for 2 × 2 Tables in Clinical Trials |
| 469 – 475 | **D. B. Stark**, **J. F. Reed III** | Sensitivity Curves for Asymmetric Trimming Hinge Estimators |
| 476 – 486 | **Panagiotis Mantalos**, **Ghazi ShukurCausality**, **Pär Sjölander** | The Effect of GARCH (1,1) on the Granger Test in Stable VAR Models |
| 487 – 491 | **Leader Navaei** | Large Deviations Techniques for Error Exponents to Multiple Hypotheses LAO Testing |

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-669
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00


| | | |
|---|---|---|
| 492 – 502 | **Mathachan Pathiyil**, **E. S. Jeevanand** | Semi Parametric Estimation of Some Reliability Measures of Geometric Distribution |
| 503 – 516 | **Mohammad F. Al-Saleh**, **Hani M. Samawi** | Inference on Overlapping Coefficients in Two Exponential Populations |
| 517 – 529 | **Rudy A. Gideon** | The Correlation Coefficients |
| 530 – 536 | **L. W. Huson** | Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data |
| 537 – 543 | **Joseph L. Balloun**, **Hilton Barrett** | From Information Lost to Knowledge Gained: The Benefits of Analyzing All the Research Evidence |
| 544 – 550 | **Kussiy K. Alyass** | Global Measure of the Deviation of a Wavelet Density Estimator |
| 551 – 560 | **James D. Stamey**, **Thomas L. Bratcher**, **Dean M. Young** | Bayesian Subset Selection of Binomial Parameters Using Possibly Misclassified Data |
| 561 – 572 | **M. A. Islam**, **R. I. Chowdhury**, **K. P. Singh** | Covariate Dependent Markov Models for Analysis of Repeated Binary Outcomes |
| 573 – 588 | **Michaela N. Gelin**, **Bruno D. Zumbo** | Operating Characteristics of the DIF MIMIC Approach Using Jöreskog's Covariance Matrix with ML and WLS Estimation for Short Scales |
| 589 – 595 | **Ayman Baklizi** | A Simple Method for Finding Empirical Likelihood Type Intervals for the ROC Curve |
| 596 – 607 | **Michael B. C. Khoo** | A Modified Control Chart for Samples Drawn From Finite Populations |
| 608 – 618 | **Sandra Hall**, **Matthew S. Mayo**, **Xu-Feng Niu** **James C. Walker** | Generalized Linear Mixed-Effects Models for the Analysis of Odor Detection Data |
| 619 – 629 | **Shou Hsing Shih**, **Chris P. Tsokos** | A Weighted Moving Average Process for Forecasting |
| 630 – 644 | **Adriana Pérez** | Longitudinal Evaluation of Estimates in an Establishment Survey After Ration Imputation |

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-669
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

*Brief Reports*

645 – 648 **W. J. Hurley** A Note on Probability Trees

*Early Scholars*

649 – 656 **Ganesh Dutta**, **Premadhis Das**, **Nripes Kumar Mandal** Optimum Choice of Covariates for a Series Of SBIBDS Obtained Through Projective Geometry

657 – 666 **Anwar Hassan**, **Sheikh Bilal Ahmad** A New Generalization of Negative Polya-Eggenberger Distribution and its Applications

*Letters to the Editor*

667 **Ian R. White** Letters to the Editor

668 – 669 **Kung-Jong Lui** Reply to Ian R. White

*JMASM* is an independent print and electronic journal (http://tbf.coe.wayne.edu/jmasm), publishing (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.

Editorial Assistant: **Jonathan Lent**
Internet Sponsor: **Paula C. Wood**, Dean, College of Education, Wayne State University

| | | |
|---|---|---|
| Cushing-Malloy, Inc. | (888) 295-7244 toll-free (Phone) | Sales & Information: |
| Internet: www.cushing-malloy.com | (734) 663-5731 (Fax) | skehoe@cushing-malloy.com |

# The easy way to find open access journals

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

www.doaj.org

The Directory of Open Access Journals covers free, full text, quality controlled scientific and scholarly journals. It aims to cover all subjects and languages.

## Aims

- Increase visibility of open access journals
- Simplify use
- Promote increased usage leading to higher impact

## Scope

The Directory aims to be comprehensive and cover all open access scientific and scholarly journals that use a quality control system to guarantee the content. All subject areas and languages will be covered.

## In DOAJ browse by subject

Agriculture and Food Sciences
Biology and Life Sciences
Chemistry
General Works
History and Archaeology
Law and Political Science
Philosophy and Religion
Social Sciences
Arts and Architecture
Business and Economics
Earth and Environmental Sciences
Health Sciences
Languages and Literatures
**Mathematics and statistics**
Physics and Astronomy
Technology and Engineering

*Contact*
**Lotte Jørgensen**, Project Coordinator
Lund University Libraries, Head Office
E-mail: lotte.jorgensen@lub.lu.se
Tel: +46 46 222 34 31

Funded by
www.soros.org

Hosted by
LUND UNIVERSITY
www.lu.se

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-360

Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

*Invited Articles*

# Optimal Trimming and Outlier Elimination

Philip H. Ramsey
Queens College of CUNY, Flushing

Patricia P. Ramsey
Fordham University

Five data sets with known true values are used to determine the optimal number of pairs that should be trimmed in order to produce the minimum relative error. The optimal trimming in the five data sets is found to be 1%, 5%, 7%, 10% and 28%. The 28% rate is shown to be an outlier among the five data sets. Results of four data sets are used to establish cutoff values for outlier detection in two robust methods of outlier detection.

Key words: Median absolute deviation, Box-and-whisker plot, MAD statistic.

## Introduction

Outliers have been considered a serious problem for the application of many statistical procedures, especially when assuming an underlying normal distribution. Barnett and Lewis (1978) provided a detailed treatment of outliers and a number of procedures for outlier detection. Barnett and Lewis state, "We shall define an outlier in a set of data to be an observation (or set of observations) which appears to be inconsistent with the remainder of that set of data" (p. 4). Similar definitions have been provided by others (Everitt, 2002; Marriott, 1990).

The presence of outliers has been shown to seriously bias traditional statistical procedures (Wilcox, 2001). Symmetric trimming of a data set by removing a specified percentage of data points from each tail of a distribution is a simple method of removing outliers. A 10% trim would remove the top and bottom 10% of the data. In general, $100\alpha$% trimming of a sample of size N would remove $[100\alpha N]$ from the top and bottom of the N ordered observations where [ ] implies the greatest lower integer.

Trimming the data biases the standard deviation of a data set but that problem can be overcome (Wilcox, 2001). However, the number

Philip H. Ramsey is Professor of Psychology. E-mail him at Philip.Ramsey@qc.cuny.edu. This research was supported in part by a PSC-CUNY grant. Patricia P. Ramsey is Professor of Management Systems. Email her at ramseyphd@fordham.edu.

of pairs trimmed (i.e. the value of $\alpha$) must be determined. Wilcox as argued for $\alpha = .20$. Some researchers may find eliminating 40% of the data to be excessive. Some others may even resist any trimming unless outlier detection can be objectively confirmed. Trimming has been found to be beneficial in testing differences in means (Kowalchuk, Keselman, Wilcox, & Algina, 2006; Lix & Keselman, 1998).

## Methodology

One of the simplest methods for evaluating an observation as a possible outlier would be to divide the deviation from the mean by the standard deviation. The problem is that an outlier biases the standard deviation upward thus reducing the ratio and making the observation appear less extreme. This "masking" effect is particularly strong when more than one outlier is present (Barnett & Lewis, 1978; Wilcox, 2001).

If a set of N observations, $X_1, \ldots, X_N$, is placed in order by size, the set can be identified by the order statistics, $X_{(1)}, \ldots, X_{(N)}$. If N is odd, the median, M, becomes the middle value, $X_{\left(\frac{N+1}{2}\right)}$.

If N is even, $\underline{M}$ becomes the midpoint of the middle two values, $\left\{ X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N+2}{2}\right)} \right\} / 2$. The median of the absolute deviations from the median (MAD), can be taken as a measure of variability. In particular, MAD/.6745 can be taken as an estimate of the population standard deviation, $\sigma$, in a normal distribution. Dividing an observation's absolute deviation from M by MAD/.6745, defines the MAD statistic which can be taken as an estimated value in a standard normal deviate (Wilcox, 2001. p. 36). Wilcox suggests that a ratio exceeding 2.0 identifies the observation as an outlier. The use of the MAD statistic removes the problem of masking. However, the criterion value, 2.0, may be too small, identifying too many observations as outliers. For example, if one is drawing random samples from a perfectly, normally distributed population then the probability of a standard normal deviate exceeding 2.0 is .0455. A sample of size, N = 100, could be expected to have four or five observations identified as outliers (i.e. 4 or 5 false positives).

Another approach using the median, M, can be raced back to Tukey's (1977) box-and-whisker plots. For N even, the ordered values, $X_{(i)}$, are divided into the top and bottom half. The median of the bottom half is $Q_1$, the first quartile of the original data. The median of the top half is $Q_3$, the third quartile of the original data. For N odd, the ordered values, $X_{(i)}$, are again divided into the top and bottom halves but the middle value (i.e. M) is included in both the top and bottom half. The values of $Q_1$ and $Q_3$ are again taken as the medians of the respective subgroups.

The interquartile range, IR, is $Q_3 - Q_1$. Any observation $X_i$ exceeding $Q_3 + \underline{m}IR$ (with $\underline{m}$ usually taken to be 1.5), is identified as an outlier. Likewise, any observation $X_i$ less than $Q_1 - \underline{m}IR$, is identified as an outlier. In sampling from a normal distribution, the probability of obtaining a single observation outside this interval (with $\underline{m} = 1.5$) would be .0070. In a sample of size, N = 100, one should expect only about one such observation identified as an outlier (i.e. one false positive). The multiplier, $\underline{m}$, could be increased to reduce the number of false positives but how high should it be and what balance should be set between false positives and false negatives?

Some authors have presented illustrative data sets when defining outliers. Everitt (2002, p. 274) identified the value 198 as an outlier in the data set {125. 128, 130, 131, 198}. For that data set, $\underline{M} = 130$ and MAD = 2. The MAD statistic for the observation, 198, would be 22.5 and well above the 2.0 cutoff value. If Everitt's data set were to be taken as a defining criterion for an outlier then the MAD statistic would need to exceed 22.5. The values $Q_3 = 131$ and IR = 3 would require an IR multiplier of m = 22.4 to match the value 198. It is unlikely that Everitt or any other author intended to use a data set to define a cutoff point for an outlier but Everitt is using a much more extreme example than has been recommended for outlier detection.

## Results

Stigler (1977) reported 24 data sets that may be of use in the present investigation. Most of the

data sets were subsets of larger sets. Each data set contained observations of 18th and 19th century investigations of physical phenomena for which nearly exact values are now known. Such data sets make it possible to compare statistical estimates to 'true values' in real data. Data Sets 1 to 8 all estimated the parallax of the sun with a 'true value' of 8.798. The 158 values were combined and designated Data Set 25 for the present investigation. Data Set 17 included 23 observations from Michelson's 1882 data estimating the velocity of light with a 'true value' of 710.5. Data Set 23 included 66 observations from Newcomb's measurements of the passage of light with a 'true value' of 33.02. Data Set 19 included 29 observations from Cavendish's 1798 determinations of the density of the earth with a 'true value' of 5.517. Data Set 24 included 100 observations from Michelson's 1879 estimation of the velocity of light in air with a 'true value' of 734.5. These five data sets include all of the data reported by Stigler.

Stigler (1977) reported trimming at 10%, 15%, and 25%. Stigler included eight other robust estimators for a total of 11. For each data set the 11 estimators were used to estimate the true value. The mean absolute deviation of the 11 estimators from the true value was designated $s_j$ for data set j. For a given data set j, each of the eleven estimators had a relative error computed as the deviation of the estimated value and true value then divided by $s_j$. These relative errors were one criterion used to compare the 11 estimators.

The five data sets selected for the present investigation were used to evaluate various degrees of trimming. The present approach is to remove one observation from each end of the ordered data set and calculate the relative error just as was done by Stigler. Additional pairs were removed until the minimum relative error was determined. The minimum relative error satisfied two objectives. First, it established an ideal degree of trimming for each data set. Second, it provided an estimator of an outlier detection criterion. That is, if outliers are responsible for poor estimation then the point at which estimation is best might be taken as the point at which an outlier or multiple outliers have been eliminated.

Table 1 presents all 23 observations for Data Set 17 and the analysis needed for outlier detection. The largest observation, 1051, produces a MAD statistic of 4.061 as the most extreme of the 23 observations. Table 2 presents the relative errors (REs) for the mean, trimmed means eliminating one to five pairs of observations, and the median. The minimum RE is .8418 and occurs with a single pair of means removed or 5% trimming.

From Table 1 the largest and smallest observations, 1051 and 573, are considered to be potential outliers. Their elimination produces the minimum RE. The criterion for MAD statistic must be less than 4.061 in order to ensure that this most extreme pair is rejected. However, if the criterion is less than 2.874 then a second pair of means would be trimmed. The midpoint, 3.468, of 4.062 and 2.874 could be taken as the best estimate for outlier detection for Data Set 17 to reject one and only one pair of means.

The interquartile range, IR, in Table 1 is IR = 803 – 703.5 = 99.5. The maximum IR multiplier, $\underline{m}$, to ensure that either $Q_3 + \underline{m}IR$ or $Q_1 - \underline{m}IR$ will lead to the rejection of the most extreme pair, 1051 and 573, is 2.5. Similarly, the minimum value of m to prevent the detection of a second pair of means is 1.261. The midpoint is $\underline{m} = (1.261 + 2.5)/2 = 1.88$.

Applying the same analysis as was applied to Data Set 17 to the other four data sets produces the results summarized in Table 3. Averages are calculated for four data sets (17, 19, 23 & 25). Data Set 24 is separated and appears to be a possible outlier among the five data sets. The averages of the four relevant data sets are shown and the midpoints of maximum and minimum averages are presented as well. The value of 3.5 for MAD statistic cutoffs is well above the 2.0 value suggested by Wilcox. The 2.0 value for $\underline{m}$, the IR multiplier, is well above the original value of 1.5.

The optimal trimming percentages of the five data sets are 1, 5, 7, 10, and 28. The MAD statistic for the value 28 is 4.72. That exceeds the original 2.0 criterion as well as the 3.5 criterion derived from the other four data sets. The cutoff point for the IR multiplier, $\underline{m} = 2.0$, would be $Q_3 + 2.0IR = 10 + 2.0(5) = 20.0$. The 28% trimming of Data Set 24 is well above this 20.0 cutoff.

Table 1. Analysis of Data Set 17, Michelson's 1882 Data Estimating the Velocity of Light with a 'true value' of 710.5

| | X | Order Sequence | D = \|X-M\| | Ordered D Values | D(.6745/MAD) |
|---|---|---|---|---|---|
| | 1051 | 1 | 277 | 277 | 4.061 |
| | 883 | 2 | 109 | 201 | 1.598 |
| | 851 | 3 | 77 | 196 | 1.129 |
| | 820 | 4 | 46 | 175 | 0.674 |
| | 816 | 5 | 42 | 163 | 0.616 |
| $Q_3 = 803$ | 809 | 6 | 35 | 109 | 0.513 |
| | 797 | 7 | 23 | 92 | 0.337 |
| | 796 | 8 | 22 | 78 | 0.323 |
| | 796 | 9 | 22 | 77 | 0.323 |
| | 781 | 10 | 7 | 63 | 0.103 |
| | 778 | 11 | 4 | 51 | 0.059 |
| M = 774 | 774 | 12 | 0 | **46** | 0.000 |
| | 772 | 11 | 2 | 42 | 0.029 |
| | 748 | 10 | 26 | 35 | 0.381 |
| | 748 | 9 | 26 | 26 | 0.381 |
| | 723 | 8 | 51 | 26 | 0.748 |
| | 711 | 7 | 63 | 23 | 0.924 |
| $Q_1 = 703.5$ | 696 | 6 | 78 | 22 | 1.144 |
| | 682 | 5 | 92 | 22 | 1.349 |
| | 611 | 4 | 163 | 7 | 2.390 |
| | 599 | 3 | 175 | 4 | 2.566 |
| | 578 | 2 | 196 | 2 | 2.874 |
| | 573 | 1 | 201 | 0 | 2.947 |

MAD = 46

Table 2. Trimmed Means and Relative Errors (REs) for Data Set 17 with $s_j = 48$ with RE Calculated for the Mean and Up to Five Pairs of Values Trimmed. Optimal trimming occurs at 5% with RE = 0.8418.

| | Value | RE | Trimming |
|---|---|---|---|
| Mean = | 756.217 | .9524 | 0% |
| Mean – 1 = | 750.905 | **0.8418** | 5% |
| Mean – 2 = | 753.053 | 0.8865 | 10% |
| Mean – 3 = | 756.353 | 0.9553 | 15% |
| Mean – 4 = | 761.800 | 1.0688 | 20% |
| Mean – 5 = | 763.769 | 1.1098 | 25% |
| Median = | 774 | 1.3229 | |

Table 3. Maximum and Minimum Values Needed for the Optimal Trimming

| | DS25 | DS 17 | DS23 | DS19 | Ave. | DS24 |
|---|---|---|---|---|---|---|
| Opt. Trim | 1% | 5% | 7% | 10% | 5.75% | 28% |
| MAD-MAX | 7.05 | 4.062 | 2.474 | 1.64 | 3.805 | 0.5395 |
| Midpoint | 6.695 | 3.468 | 2.2485 | 1.58 | 3.5 | 0.5395 |
| MAD-MIN | 6.34 | 2.874 | 2.023 | 1.52 | 3.189 | 0.5395 |
| IR-MAX | 5.479 | 2.5 | 1.143 | 0.646 | 2.439 | -0.0556 |
| Midpoint | 4.6175 | 1.8805 | 0.9285 | 0.549 | 2.0 | -0.11 |
| IR-MIN | 3.756 | 1.261 | 0.714 | 0.452 | 1.547 | -0.1667 |

The cutoff from the original, $\underline{m} = 1.5$, would be $Q_3 + 1.5IR = 10 + 1.5(5) = 17.5$. Of course, 28exceeds this more conservative value of 17.5.

## Conclusion

In sampling from a standard normal distribution the probability of exceeding a value of 3.5 is approximately .0005. Even in a sample of size, N = 1000, a single, false-positive indication of an outlier would not be expected. Again sampling from a standard normal distribution the probability of identifying an outlier with the $\underline{m} = 2.0$ multiplier for IR would be approximately 0.0008. In that case a sample of size, N = 1000, might be expected to produce one, false-positive observation.

As a final point, note that Data Set 24 does suggest that trimming even in excess of 20% may sometimes be justified. However, to the extent that present results are applicable, trimming by no more than 10% is more likely to be optimal.

## References

Barnett, V. and Lewis, T. (1978), *Outliers in statistical data*. New York: Wiley.

Everitt, B. S. (2002), *The Cambridge dictionary of statistics* (2nd ed.), Cambridge, UK: Cambridge University Press.

Kowalchuk, R. K., Keselman, H. J., Wilcox, R. R., & Algina, J. (2006). Multiple comparison procedures, trimmed means and transformed statistics, *Journal of Modern Applied Statistical Methods*, *5*, 44-65.

Lix, L. M. & Keselman, H. J. (1998). To trim or not to trim: Tests of mean equality under heteroscadasticity and non normality, *Educational and Psychological Measurement*, *58*, 409-429 (Errata -58,853).

Marriott, F. H. C. (1990), *A dictionary of statistical terms* (5th ed.). New York: Longman Scientific & Technical.

Stigler, S. M. (1977), "Do robust estimators work with real data?" *The Annals of Statistics*, *5*, 1055-1098.

Tukey, J. W. (1977), *Exploratory data analysis*, Reading, MA: Addison-Wesley.

Wilcox, R. R. (2001), *Fundamentals of modern statistical methods*, New Haven, CT: Springer.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 361-366
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# An Omnibus Test When Using A Regression Estimator With Multiple Predictors

Rand R. Wilcox
University of South Carolina

In quantile regression, the goal is to estimate the $\gamma$ quantile of $Y$ given values for $p$ predictors. Methods for making inferences about the individual slope parameters have been proposed, some of which have been found to perform very well in simulations. But for an omnibus test that all slope parameters are zero, it appears that little is known about how best to proceed. For the special case $\gamma$=.5, a drop-in-dispersion test has been recommended, but it requires a large sample size to control the probability of a Type I error and it assumes that the usual error term is homoscedastic. The article suggests an alternative method that performs well in simulations, it allows heteroscedasticity, and it can be used when $\gamma \neq .5$.

Key words: Robust regression, tests of independence, bootstrap methods.

## Introduction

Consider the random variables $X_1,\ldots,X_p,Y$ having some unknown (p+1)-variate distribution and let $Y_\gamma$ be the conditional $\gamma$ quantile of $Y$ given $X_1,\ldots,X_p$. When using the Koenker and Bassett (1978) quantile regression method, the goal is to estimate $Y_\gamma$ assuming that

$$Y_\gamma = \alpha_\gamma + \beta_{1\gamma}X_1 + \ldots + \beta_{p\gamma}X_\gamma \qquad (1)$$

where the unknown parameters $\beta_{1\gamma},\ldots,\beta_{p\gamma}$ and $\alpha_\gamma$ are estimated based on the random sample $(X_{i1},\ldots,X_{ip},Y_i)$, $i=1,\ldots,n$. The special case $\gamma$=.5 corresponds to what is called the least absolute value regression estimator, meaning that the estimates of the parameters are chosen so as to minimize the sum of the absolute values of the residuals. This special case predates ordinary least squares by about a half century and offers protection against the deleterious effects of outliers among the Y values. As is probably evident, choices for $\gamma$ other than .5 can be revealing and help add perspective on the association among the variables under study.

As a simple example, consider data from a study conducted by Williams, Stanchina,

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. He is the author of seven textbooks on statistics, the most recent of which is *Introduction to Robust Estimation and Hypothesis Testing* (2005, 2nd ed., San Diego, CA: Academic Press).

Table. 1 Values for $d_0$ and $d_1$

| p | $d_0$ | | | | $d_1$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=.1 | $\alpha$=.05 | $\alpha$=.025 | $\alpha$=.01 | $\alpha$=.1 | $\alpha$=.05 | $\alpha$=.025 | $\alpha$=.01 |
| 2 | .2179 | .1203 | .0588 | .0430 | -.00196 | -.00117 | -00056 | -.00055 |
| 3 | .2814 | .1840 | .1143 | .0364 | -.00300 | -.00223 | -.00149 | -.00044 |
| 4 | .4478 | .3356 | .2624 | .1546 | -.00580 | -.00476 | -.00396 | -.00240 |
| 5 | .6373 | .4250 | .3097 | .1590 | -.00896 | -.00630 | -.00474 | -.00248 |
| 6 | .7699 | .5648 | .4111 | .2734 | -.01120 | -.00858 | -.00640 | -.00439 |

Bezdjian, Skrok, Raine and Baker (2005). A portion of the study dealt with the association between a so-called Q score resulting from the Porteus maze test, which is used to evaluate intelligence and executive functioning, and how this Q score is related to a measure of delinquency. Figure 1 shows a scatterplot of the data. The sample size is n=943. Also shown are the regression lines corresponding to $\gamma$=.5, .8 and .9. As is evident, based on the typical response, as measured by the median or even the .8 quantile, there is little or no indication of an association. (The p-value when $\gamma$=.8 is approximately .36.) But for $\gamma$=.9, the regression line has a positive slope that is significantly different from zero at the .05 level. (For another recent illustration of the practical value of quantile regression methods, see Angrist, Chernozhukov & Fernandez-Val, 2006.)

The goal in this article is to suggest and study a method for testing

$$H_0 : \beta_{1\gamma} = \ldots = \beta_{p\gamma} = 0 . \qquad (2)$$

For the related problem of testing

$$H_0 : \beta_j = 0$$

for each j (j=1,...,p), there is a well-known method that appears to perform relatively well in simulations (Koenker, 1994, cf. Koenker & Xiao, 2002, cf. Koenker & Machado, 1999). But when $\gamma$ differs from .5, it seems that there are no results or even suggested methods for testing (2).

For the special case $\gamma$=.5, Birkes and Dodge (1993) suggest testing (2) using a drop in dispersion method. They note that the method requires a relatively large sample size, but they do not specify just how large the sample size must be to achieve reasonably accurate control over the probability of a Type I error. When testing at the .05 level, Bradley (1978) suggests that at a minimum, the actual Type I error probability should be between .025 and .075. When examining the drop in dispersion method (in the simulations described in section 3), it was found that to achieve Bradley's criterion, a sample size of n=100 is required, even under normality. Another concern is that the method assumes a homoscedastic error term. So one goal here is search for a method that gives better results when the sample size is small and another goal is to suggest a method that might be used when the error term is heteroscedastic.

Yet another approach to testing (2) is to use the percentile bootstrap method stemming from results in Liu and Singh (1997). When working with various robust estimators, this approach appears to perform quite well, even with fairly small sample sizes and when there is heteroscedasticity (e.g., Wilcox, 2005). However, this approach was found to be unsatisfactory in the simulations considered here, so it was abandoned.

## Methodology

The Koenker and Bassett (1978) quantile regression method arises as follows.

For some $\gamma$, 0< $\gamma$ <1, let

$$\rho_\gamma(u) = u(\gamma - I_{u<0})$$

where the indicator function $I_{u<0}$=1 if u<0; otherwise $I_{u<0}$=0. Assuming that the $\gamma$ quantile of Y, given X, is given by (1), the Koenker-Bassett quantile regression method estimates the unknown parameters $\beta_{1\gamma},...,\beta_{p\gamma}$ and $\alpha_\gamma$ with the values $b_{1\gamma},...,b_{p\gamma}$ and $a_\gamma$, respectively, that minimize

$$\sum \rho_\gamma(r_i), \qquad (3)$$

where $r_i = Y_i - b_{1\gamma}X_{i1} - .... - b_{p\gamma}X_{ip} - a_\gamma$ are the residuals. Here, the values that minimize (3) were determined with the function rq that is included in the robust library that comes with the software S-PLUS.

The proposed method for dealing with small sample sizes stems in part from the classic generalized $T^2$ statistic used to test the hypothesis that a multivariate normal distribution has a mean vector of zero (e.g., Anderson, 1958, chapter 5). One difficulty here is getting an estimate of the appropriate covariance matrix, and the strategy is to use a bootstrap estimate. (For general results on bootstrap estimates of the standard error, see Buchinsky, 1991; Hahn, 1995.) Results for the special case p=1, reported by Koenker (1994), suggest that this approach will result in an actual Type I error probability that can be substantially less than the nominal level, and this was found to be the case for n<60. However, a simple adjustment is found that corrects this problem in the simulations to be described.

Let $(X^*_{i1},...,X^*_{ip},Y^*)$, $i = 1,...,n$, be a bootstrap sample obtained by randomly sampling, with replacement, n vectors of observations from $X_{i1},\ldots,X_{ip}$, $Y_i$. Given $\gamma$, label the resulting estimate of the slopes $b^*_k$, $k = 1,...,p$. Repeat this process B times yielding $b^*_{1k},...,b^*_{Bk}$. Then from basic principles, an estimate of the variances and covariances associated with $b_{1\gamma},...,b_{p\gamma}$ is

$$S = \frac{1}{B-1}\sum_{c=1}^{B}(b^*_c - \bar{b}^*)^2,$$

where $b^*_c = (b^*_{c1},...b^*_{cp})$, and $\bar{b}^*_k = \sum b^*_{ck}/B$. Then, proceeding in an obvious fashion, the test statistic used here is

$$T^2 = n\bar{b}'S^{-1}\bar{b}.$$

Again from basic principles, a natural strategy is to reject if

$$T^2 \geq \frac{n-1}{n-p}f_{p,n-p},$$

where $f_{p,n-p}$ is the 1-$\alpha$ quantile of an F distribution with p and n-p degrees of freedom. But as previously indicated, preliminary simulations indicated that the actual probability of Type I error is less than the nominal level when the sample size is small. For example, when $\gamma$=.5, p=2, n=20, $\alpha$=.05, and if $X_1$ and $X_2$ have a bivariate normal distribution with correlation $\rho$=0, the actual Type I error probability was estimated to be .026. Increasing p to 6, the estimate is now .001. Very similar results were obtained when $\gamma$=.8. But in all cases considered, with n=60, the actual probability of a Type I error was estimated to be reasonably close to .05.

The results just described suggest the following modification when n<60.

Temporarily assume that the error term is homoscedastic and has a normal distribution. The strategy is to determine an adjusted p-value, $p_a$, so that for n=20, the actual Type I error probability will be approximately $\alpha$ if the null hypothesis is rejected $\hat{p} \leq p_a$ whenever the observed p-value (based on $T^2$) is less than or equal to $p_a$. (In essence, use Gosset's strategy when dealing with the problem of making inferences about means.) For sample sizes between 20 and 60, interpolation is used to

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

| g | h | $\kappa_1$ | $\kappa_2$ |
|---|---|---|---|
| 0 | 0 | 0 | 3 |
| 0 | 0.2 | 0 | 21.46 |
| 0.2 | 0 | 0.61 | 3.68 |
| 0.2 | 0.2 | 2.81 | 155.98 |

determine $p_a$. First consider $\gamma$=.5. For $\alpha$=.1, .05, .025 and .01, simulations indicate that the adjusted p-value is given by $p_a = d_1 n + d_0$, where $d_1$ and $d_0$ are given in Table 1. That is, letting $\hat{p}$ be the p-value based on $T^2$, and assuming that (n-p)$T^2$/(n-1) has an F distribution with p and n-p degrees of freedom, reject if $\hat{p} \le p_a$. Additional simulations indicate that this adjustment continues to perform reasonably well when $\gamma$=.8, provided B=200 is used, as will be seen.

A Simulation Study

Simulations were used to study the small-sample properties of the method just described. The distribution for X was taken to be multivariate normal withcommon correlation $\rho$, and the distribution for Y was taken to be one of four g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If Z has a standard normal distribution, then

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

if g>0

$$Y = Z\exp(hZ^2/2)$$

if g=0 has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal (g=h=0.0), a symmetric 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=0.0). Table 2 shows the skewness ($\kappa_1$) and kurtosis ($\kappa_2$) for each distribution considered. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The two choices for $\rho$ were 0 and .8. It was found that altering $\rho$ had no effect on the simulation results, so for brevity, only results for $\rho$=0 are reported.

To get some indication of the effects of heteroscedasticity, data were also generated according to the model

$$Y = \lambda(X_1)\varepsilon$$

for some specified function $\lambda$, where $\varepsilon$ is independent of $X_1$ and $\varepsilon$ has one of the g-and-h distributions already described. Of course $\lambda(X_1)$=1 corresponds to homoscedasticity. The other two choices were $\lambda(X_1) = |X_1| + 1$ and $\lambda(X_1) = 1/(|X_1| + 1)$. For convenience, these three choices will be called variance patterns VP1, VP2 and VP3, respectively. Note that for all three patterns, the slope remains zero even when $\gamma \ne .5$.

Table 3 shows the estimated probability of a Type I error when testing at the .05 level with n=20, $\gamma$=.5 and .8, and p=2 and 6. For the moment, B=100 is used. It will be seen that generally this suffices, in terms of controlling the probability of a Type I error, but in some cases, B=200 is required. The estimated Type I error probabilities are based on 1,000 replications.

From Robey and Barcikowski (1992), 1,000 replications is sufficient from a power point of view. More specifically, if one tests the hypothesis that the actual Type I error rate is .05,

Table 3: Estimated probability of a Type I error, n=20, $\alpha$=.05, B=100

| | | p=2 | | | p=6 | | |
|---|---|---|---|---|---|---|---|
| g | h | VP1 | VP2 | VP3 | VP1 | VP2 | VP3 |
| | | | | $\gamma$=.5 | | | |
| 0.0 | 0.0 | .043 | .067 | .020 | .044 | .027 | .036 |
| 0.0 | 0.2 | .026 | .032 | .021 | .018 | .023 | .018 |
| 0.2 | 0.0 | .037 | .052 | .023 | .034 | .048 | .028 |
| 0.2 | 0.2 | .030 | .035 | .020 | .017 | .024 | .016 |
| | | | | $\gamma$=.8 | | | |
| 0.0 | 0.0 | .049 | .029 | .076 | .048 | .046 | .032 |
| 0.0 | 0.2 | .044 | .069 | .031 | .034 | .030 | .031 |
| 0.2 | 0.0 | .050 | .089 | .035 | .056 | .052 | .036 |
| 0.2 | 0.2 | .050 | .078 | .030 | .041 | .041 | .036 |

and if one wants power to be .9 when testing at the .05 level and the true $\alpha$ value differs from.05 by .025, then 976 replications are required. As is evident, all indications are that reasonable control over the probability of a Type I error is obtained in nearly all of the situations considered. The main exception is when p=2, $\gamma$=.8 and sampling is from a light-tailed distribution (h=0), in which case, for variance pattern VP2, the estimated probability of a Type I error can exceed .075. The least satisfactory result was obtained when g=.2, in which case the estimate is .089. However, increasing B to 200, the estimate drops to .061. (Leaving B=100 and increasing n to 30 and 40, the estimates were .072 and .06, respectively.) Thus, to be safe, B=200 or larger is recommended.

## Conclusion

The main result is that for the bootstrap method studied here, among all situations considered, the estimated level of the test did not exceed .075 when testing at the .05 level provided $B \geq 200$ is used, even with n=20. With B=100, exceptions occur, as indicated in Table 3, but given the speed of modern computers, using B=200 seems practical. In contrast, the drop-in-dispersion method requires a sample size of at least n=100 to avoid an estimated Type I error probability greater than .075.

It was mentioned that the bootstrap method stemming from Liu and Singh (1997) was unsatisfactory in simulations; the actual probability of a Type I error was well below the nominal level. Perhaps an adjusted p-value, similar to one used here, would correct this problem in a satisfactory manner, but this has not been investigated.

Finally, R and S-Plus software is available from the author for applying the bootstrap method studied here. Ask for the function rqtest.

## References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Angrist, J., Chernozhukov, V, & Fernandez-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, *74*, 539-563.

Birkes, D. & Dodge, Y. (1993). *Alternative methods of regression*. New York: Wiley.

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Buchinsky, M. (1991). *The theory and practice of quantile regression*. PhD Thesis, Dept. of Economics, Harvard University.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, *11*, 105-121.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring data tables, trends, and shapes*., p. 461-515. New York: Wiley.

Koenker, R. (1994). Confidence intervals for regression quantiles. In (P. Mandle M. Huskova, eds.) *Asymptotic Statistics*. Proceedings of the Fifth Prague Symposium, 349-359.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrika*, *46*, 33-50.

Koenker, R. & Machado, A. F. (1999). Goodness of fit related inference for quantile regression. *Journal of the American Statistical Association*, *94*, 1296-1310.

Koenker, R. & Xiao, Z. J. (2002). Inference on the quantile regression process. *Econometrica*, *70*, 1583-1612.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, *92*, 266-277.

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*, 283-288.

Williams, N., Stanchina, J., Bezdjian, S., Skrok, E., Raine A. & Baker, L. (2005). *Porteus' mazes and executive function in children: Standardize Administration and scoring, and relationships to childhood aggression and delinquency*. Unpublished manuscript, Dept. of Psychology, University of Southern California.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Ed. San Diego CA: Academic Press.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 367- 379
Copyright © 2007 JMASM, Inc.
1538 – 9472/05/$95.00

# Time-Series Intervention Analysis Using ITSACORR: Fatal Flaws

Bradley E. Huitema Joseph W. McKean
Western Michigan University

Sean Laraway
San Jose State University

The ITSACORR method (Crosbie, 1993, 1995) is evaluated for the analysis of two-phase interrupted time-series designs. It is shown that each component of the ITSACORR framework (including the structural model, the design matrix, the autocorrelation estimator, the ultimate parameter estimation scheme, and the inferential method) contains fatal flaws.

Key words: Autocorrelation, time-series intervention analysis, time-series regression with auoregressive errors.

## Introduction

Researchers and practitioners working in the behavioral sciences frequently employ interrupted time-series designs to determine the effectiveness of various interventions in both clinical and natural settings. Currently, several methods are available for statistically analyzing data from interrupted time-series designs. Among these methods, autoregressive integrated moving average (ARIMA) intervention models have a long history of endorsement by methodologists (e.g., Glass, Willson, & Gottman, 1975; McCleary & Hay, 1980). Nevertheless, some authors (e.g., Gorman & Allison, 1997) have noted that certain properties of ARIMA models, particularly their analytical complexity and requirement of relatively large sample sizes, make the use of these models troublesome for many behavioral researchers. Concerns regarding these undesirable properties of ARIMA models have prompted the development of several alternatives. These alternatives reportedly (a) reduce the difficulty of analyzing time-series data and (b) enable the analysis of series with relatively few observations, a characteristic of many applications of time-series designs in the

Bradley E. Huitema is Professor of Psychology. His major interests are in the design and analysis of single-case experiments, quasi-experiments, and observational studies. Email: brad.huitema@wmich.edu. Joseph W. McKean is Professor of Statistics. His interests are in robust nonparametric statistical methods. Sean Laraway is Assistant Professor of Psychology. His research interests include behavioral pharmacology and learning and memory.

behavioral sciences. Two commonly cited alternatives to ARIMA intervention models are Gottman's ITSE (Gottman,1981; Rushe & Gottman, 1993) and Crosbie's ITSACORR (Crosbie, 1993, 1995).

Both of these alternatives use the same underlying model and estimate the same intervention parameters. Despite recent corrections, the current version of ITSE does not provide a satisfactory method for analyzing time-series data because it still contains several major defects. These defects are not software bugs; rather, they are problems with the method that are described in a recent critique (Huitema, 2004).

The ITSACORR method builds on the ITSE method; it was designed to analyze short series that likely have autocorrelated errors and that may have trend within one phase or within both phases (Crosbie, 1995). In proposing ITSACORR as a suitable method for analyzing time-series data, Crosbie (1993, 1995) described several supposed advantages of ITSACORR over both ARIMA intervention methods and Gottman's ITSE. First, unlike ARIMA, ITSACORR allegedly yields appropriate results with small sample sizes even in the presence of high levels of autocorrelation (Crosbie, 1995, p. 392). Second, ITSACORR reportedly provides results that agree with those of ARIMA when a large number of observations is available (Crosbie, 1995, pp. 391-392). Third, ITSACORR supposedly has better small-sample inferential properties than does ITSE (Crosbie, 1995).

These claims combined with readily available and uncomplicated software have led to considerable attention for ITSACORR from methodologists and practitioners. Writers in applied fields such as aphasiology, applied behavior analysis, clinical psychology, counseling psychology, and school psychology have strongly encouraged its use. For example, Gottman and Rushe (1993) described ITSACORR as "a new, powerful method for single-case analysis of change over time using the interrupted time-series design . . . this can be done without needing to know sophisticated time-series modeling methods and with very few data before and after the intervention" (p. 909). They further state that ITSACORR ". . . makes time-series methods available to the general clinician for the first time" and that "This approach will have widespread importance in the evaluation of change in patients in clinical trials where it is possible to study people on a case-by-case basis, or in the case work of quantitatively oriented clinical practitioners" (p. 909). This initial endorsement has been followed by additional support (e.g., Gottman, 1995), and ITSACORR has received many positive evaluations published in single-case methodology books (e.g., Franklin, Allison, & Gorman, 1997). Gorman and Allison (1997), for instance, have stated that ITSACORR "combines the best of ARIMA and regression approaches" (p. 94). Similarly, a widely used research methodology textbook (Christensen, 2007) states (p. 345) that Crosbie's method is an effective replacement for the well established methods of Box and Jenkins (1970), Box and Tiao (1965), and Glass, Willson, and Gottman (1975).

In addition to these recommendations from methodologists, ITSACORR has received additional endorsement in expository articles written for practitioners. For example, researchers in the area of aphasiology have stated that "ITSACORR should be the procedure-of-choice, and essentially the standard, for applying hypothesis testing logic to single-subject data" (Robey, Schultz, Crawford, & Sinner, 1999, p. 466). Several other authors (some outside the behavioral sciences) have cited ITSACORR as one of several credible methods for time-series analysis (e.g., Ellis, 1999, p. 573; Hogenraad, McKenzie, & Martindale, 1997, pp. 433-35).

A recent expository article on the design and analysis of time-series studies appeared in The International Journal of Clinical and Experimental Hypnosis; it includes the following endorsement: "ITSACORR is eminently easy to use; it corrects for autocorrelation; it generates statistics that are familiar to reviewers and editors; and it is acceptable for use with as few as 7 to 10 data points per phase" (Borckardt & Nash, 2002, p.127). Following this and other statements, the article presents a half-dozen examples of the use of ITSACORR (pp. 132-142).

It appears that the effect of these books and articles has been widespread acceptance of ITSACORR. One can find many published examples of the application of ITSACORR in journals such as *Aphasiology* (e.g., Robey et al. 1999; Spencer, Doyle, McNeil, Wambaugh, Park, & Carroll, 2000), *British Journal of Clinical Psychology* (e.g., Davidson & Tyrer, 1996), *Journal of Consulting and Clinical Psychology* (e.g., Lucyshyn, Albin, & Nixon, 1997), and *School Psychology Review* (e.g., Stage & Quiroz, 1997). Because ITSACORR is widely recommended and used the descriptive and inferential properties of this method must be understood by methodologists, research workers, and journal editors. The purpose of this article is to explicate these properties.

## Logic of the Two Phase Design

An understanding of the essential descriptive properties associated with the analysis of the interrupted time-series experiment rests on the logic of this design. Consider the simple two-phase (A-B) interrupted time-series design. The data of the first phase can provide a prediction of what would occur during the second phase in the absence of an intervention. The researcher's interest lies in the difference between the predicted (counterfactual) second phase behavior and the behavior that actually occurs during the second phase. There exist two major statistics that characterize this difference. The first is known as level change and the second is known as slope change. Although the interpretation of both of these measures is straightforward, level change is frequently misunderstood and incorrectly computed (Huitema & McKean, 2000a; Huitema, 2004).

## Level Change

One possible measure of level change indicates the amount by which the intervention changes the expected value of the response at the beginning of the intervention phase. If there are n1 observations in the first phase and n2 observations in the second phase, the first observation in the intervention phase occurs at time n1 + 1. The level change can reasonably be defined (under the assumption that an adequate model describes the data for each phase) as the difference between (a) the predicted (counterfactual) value of Y at time n1 +1 based on a model of the first phase data and (b) the expected value of Y at time n1 +1 based on a model of the second-phase data. It is crucial to understand that both of these estimates must be associated with exactly the same time point (viz., n1 +1). Although various time-series intervention models may use different procedures to compute the two level estimates, all acceptable procedures estimate level change at a common time point. It is important to be aware that the concept of level change does not, in general, refer to the difference between the means of the two phases. Level change refers to a shift in elevation that is unexplained by possible within-phase trends.

## Slope Change

Slope change provides the second major way of characterizing the effect of an intervention. Here the term slope has its traditional meaning. It simply refers to the average change in Y given a one-unit change in X, where the X variable is time. If the intervention has an effect, it may produce a change in level, a change in slope, or both. Because a reasonable representation of intervention effects often requires measures of both level change and slope change, an adequate descriptive analysis will usually provide accurate estimates of both of them. Although interventions can also interrupt the structure of time-series data by changing the variance or in other more subtle ways (see, e.g., Stoline, Huitema, & Mitchell, 1980), level change and slope change provide two of the most basic effect measures. The adequacy of ITSACORR with respect to these measures is the focus of this article.

# Methodology

Four linked issues that are relevant in evaluating the adequacy of intervention analyses were studied. First, at the most elementary level, whether ITSACORR produces measures that are consistent with the logic of time-series intervention designs was evaluated. Second, the consistency between the logic of the design and the ITSACORR structural model was examined. Third, the consistency between the ITSACORR

structural model and the ITSACORR design matrix was evaluated. Last, the inferential properties of the tests provided by ITSACORR was evaluated. Details regarding these issues and methods used to study them are described in this section.

Correspondence Between the Logic of the Design and the Parameter Estimates Produced by ITSACORR

The correspondence of the level change and slope change estimates produced by ITSACORR with level change and slope change estimates produced by methods that are consistent with the logic of the interrupted time-series design was evaluated. Three methods that are known to provide parameter estimates consistent with the logic of the interrupted time-series design utilize the same design matrix. This design matrix differs greatly from the matrix used by both ITSE and ITSACORR. Described is the appropriate matrix (denoted as the H-M matrix) in detail elsewhere (e.g., Huitema & McKean, 2000a, 2000b; Huitema, McKean, & McKnight, 1994; McKnight, McKean, & Huitema, 2000). The three methods that use the H-M matrix differ from each other in terms of assumptions and/or method of estimation. The first method (H-M OLS) assumes independent errors and uses ordinary least-squares (OLS) as its estimation procedure. Although some researchers believe that OLS models are never appropriate in the case of time-series designs, this is not true (see Huitema and McKean, 1998). The second and third methods assume first-order autoregressive errors. They differ from each other in that the second method (H-M M-L) uses a maximum-likelihood estimation procedure, whereas the third method (H-M Bootstrap) uses a double bootstrap approach (McKnight et al., 2000).

After results from the first three methods were obtained ITSACORR was applied to the same data and made comparisons among the results of the different methods. All of these comparisons used data from four published studies (see Figure 1). These data are of the type for which ITSACORR was specifically designed. Indeed, all of these data were obtained from expository articles that illustrate and promote the use of ITSACORR (i.e., Borckardt, 2002; Crosbie, 1995; Robey et al., 1999; Spencer et al., 2000).

Correspondence Between the Logic of the Intervention Design and the ITSACORR Structural Model

The evaluation of how well the ITSACORR model corresponds to the logic of the interrupted time-series design involved comparing the level- and slope-change parameters defined in the structural model with the change parameters of interest in the intervention design. This involved answering two questions: (a) Does the ITSACORR model define level change as the difference between the counterfactual level and the observed level? and (b) Does the model define slope change as the difference between the counterfactual slope and the observed slope?

Correspondence Between the Structural Model and the Design Matrix

A coherent methodology will have consistency between the parameters specified in the structural model and the parameters implied by the associated design matrix. This consistency was evaluated by comparing the level change, slope change, and first order autocorrelation parameters specified in the ITSACORR structural model with the corresponding parameters defined by the ITSACORR design matrix.

Evaluation of Inferential Performance

ITSACORR provides inferential tests on the difference between intercepts and slopes.

The inferential aspects of greatest interest in evaluating the performance of hypothesis testing procedures are Type I error and power. A small computer simulation was used to empirically evaluate these properties. The simulation study evaluated these properties under two levels of autocorrelation (.50 and .80) and two intercept change effect sizes (0 and 10 sigma); total sample size (n1 + n2) was set at 20. No slope change was included in any of the simulations. 1,000 simulations were performed under each condition; *a* was set at the nominal value of .05.

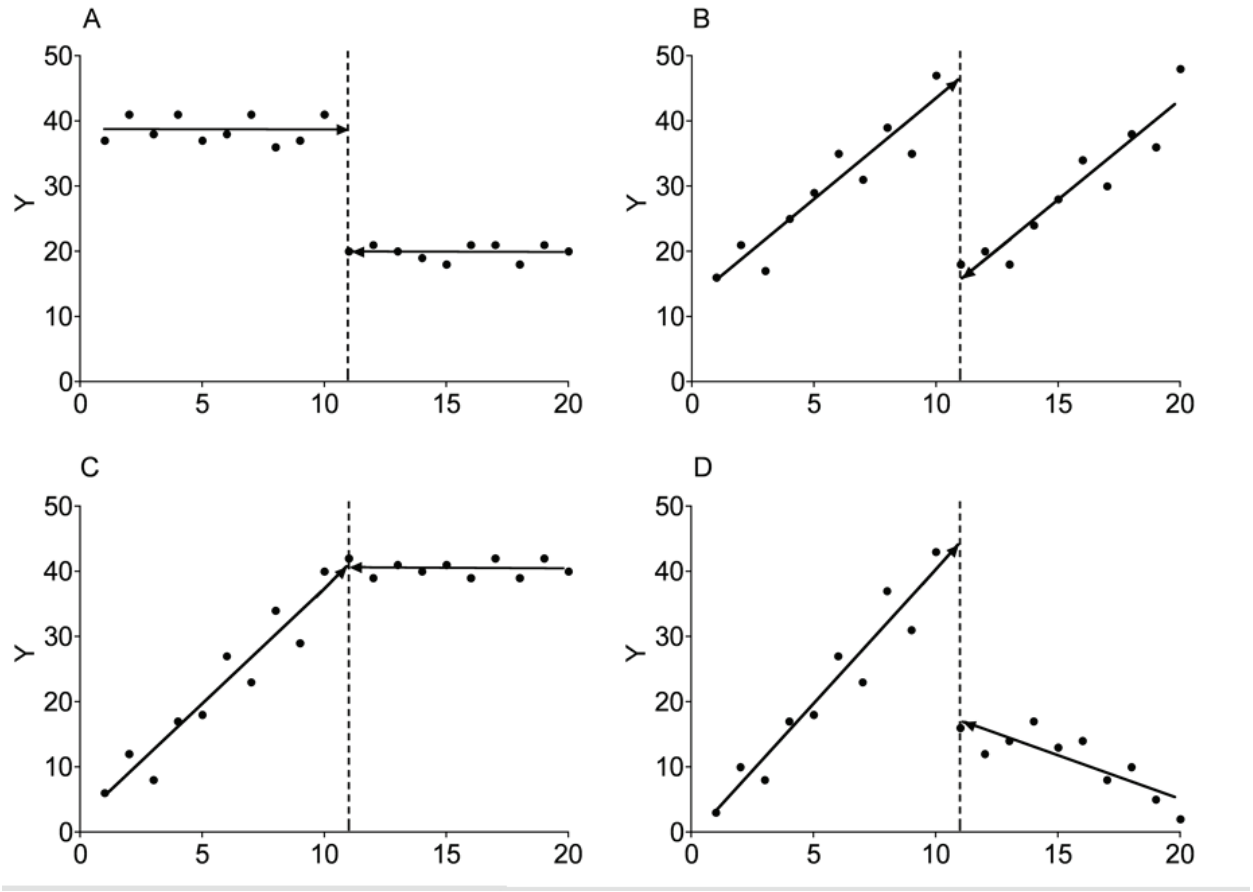



*Figure 1*. Panel A: Perceptual speed data (Holtzman, 1963) that illustrate an apparent change in both level and slope. Panel B: Aphasia data (Robey, et al., 1991) that illustrate an apparent change in both level and slope. Panel C: Weekly diastolic blood pressure readings (Borckardt, 2002) that illustrate little if any change in level and negative change in slope. Panel D: Oral naming accuracy data (Spencer et al., 2000) illustrating a trending series that was not subject to an intervention.

Table 1

Summary of Level Change, Slope Change, and Autocorrelation Estimates Associated with ITSACORR, ITSE, and Three Alternative Methods (H-M OLS, and H-M M-L, and H-M Bootstrap) Applied to Data (Illustrated in Figure 1) from Four Published Sources

| | Method of Analysis | | | | |
|---|---|---|---|---|---|
| | ITSACORR | ITSE | H-M OLS | H-M M-L | H-M Bootstrap |
| Study A | | | | | |
| Level change: | -5.90 | -0.96 | -31.07*** | -30.87*** | -30.61*** |
| Slope change: | -0.87* | -0.99** | -1.01*** | -1.01*** | -1.00*** |
| Autocorrelation: | .68 | .17 | (.15)* | .15 | .22 |
| Study B | | | | | |
| Level change: | 65.91** | 45.13*** | 39.51*** | 40.89*** | 39.71*** |
| Slope change: | 0.74 | 2.98* | 3.65* | 3.73* | 3.48* |
| Autocorrelation: | .54 | .13 | ( -.33) | -.35 | -.18 |
| Study C | | | | | |
| Level change: | -75.28* | -9.14*** | -4.33*** | -2.77 | -2.68 |
| Slope change: | 1.05 | -1.65* | -1.83*** | -1.85*** | -1.96*** |
| Autocorrelation: | -.01 | .56*** | (.51)*** | .61*** | .71* |
| Study D | | | | | |
| Level change: | 55.55*** | 55.55*** | -6.82 | -7.08 | -5.34 |
| Slope change: | -0.24 | -0.25 | -.26 | -.25 | -.26 |
| Autocorrelation: | .12 | -.01 | (-.04) | -.04 | .13 |

*Note*: *$p <$ .05; **$p <$ .01; ***$p <$ .001

## Results

### Inconsistency Between the Logic of the Design and the Estimates Produced by ITSACORR

The intervention effects and autocorrelation estimates associated with ITSACORR, ITSE, and the three methods based on the H-M design matrix appear in Table 1 for the data illustrated in the four panels of Figure 1.

The columns of the table list the methods of analysis and the major rows identify the study; the level change, slope change, and autocorrelation estimates appear in the body of the table.

Study A.

The data illustrated in panel A of Figure 1 are perceptual speed measures obtained from a schizophrenic patient each day before and after the administration of chlorpromazine. These data have appeared in publications by several writers (e.g., Crosbie,1995; Glass, et al., 1975; Holtzman, 1963) to illustrate time-series procedures. Crosbie (1995) used these data to support the claim that, in the case of a large number of observations, ITSACORR, ITSE, and ARIMA methods all reach the same conclusion. An examination of Table 1 reveals that ITSACORR and ITSE provided level decrease estimates of 5.90 and 0.96 points, respectively ($p > .50$ for both methods), whereas each of the three remaining methods estimated the level decrease as about 31 points ($p \leq .001$). An ARIMA analysis of these data by Glass, et al. (1975) (not included in Table 1) estimated a drop in level of approximately 22 points ( $p \leq .001$). Visual inspection of the data suggests a level decrease in the neighborhood of 20 - 30 points. All methods included in Table 1 yielded similar slope change estimates. Because the ARIMA model used by Crosbie (1995) as a basis of comparison with ITSACORR and ITSE does not estimate slopes, one could not compare this ARIMA model with the other analyses in terms of slope change. The autocorrelation estimate produced by ITSACORR was a value of .68 while the other procedures yielded autocorrelation estimates that range from .15 to .22.

Study B.

The data in panel B appeared in an article by Robey et al. (1999) that strongly promoted the use of ITSACORR. After applying ITSACORR to the data these authors stated that "The t test for a change in level is also significant (i.e., $t = 3.341$, $p = .005$); the t test for a change in slope does not achieve statistical significance (i.e., $t = 0.187$, $p = .855$)" (p. 460). Unfortunately, Robey et al. (1999) did not present the descriptive statistics (i.e., intercept and slope estimates) associated with these t and p values. These descriptive statistics are listed in Table 1.

Notice that ITSACORR estimated the level change as approximately 66 points. If one examines panel B of Figure 1 one can see the elevation of the phase 1 line at time point 9 and the elevation of the phase 2 line for the same time point; it is obvious that they differ by approximately 40 points. Indeed, an inspection of the level change statistic for each analysis shown in Table 1 indicates that only the estimate provided by ITSACORR deviates far from 40 points.

The slope-change and autocorrelation estimates provided by ITSACORR also deviate greatly from the results provided by the other methods. In contrast, all of the other methods provide slope-change estimates that are consistent with the visual impression. Table 1 also shows that ITSACORR provides a higher value for the autocorrelation estimate (i.e., .54) than the estimates provided by the other methods (range = -.35 through .13).

Study C.

Borckardt (2002) was written to demonstrate "how clinicians can efficiently conduct scientific analyses of a patient's response to such interventions using time-series designs supported by newly developed analytic procedures." (p. 190). One of the analytic procedures to which he referred was ITSACORR. Weekly diastolic blood pressure data from this study appear in panel C of Figure 1. These data were obtained before and after participants received a multimodal psychotherapy intervention. A visual inspection of the data reveals a minor negative slope during the baseline phase, essentially no level change

after intervention, and a strong negative shift in slope beginning immediately after the intervention. These visual impressions concur with the results of the statistical methods listed in Table 1, with one exception. ITSACORR estimates a huge decrease in level (over 75 points) and a positive shift in slope. Both of these estimates are grossly inconsistent with the visual appearance of the data. Visual inspection suggests that the drop in level can be no more than a few points. Moreover, as the difference between the minimum and maximum values in the entire series is only 32 points, a level change estimate of 75 points can have no real meaning. The easily discerned visible decrease in slope in the second phase suggests that, even in the absence of supporting statistical evidence (e.g., that produced by the other methods described in Table 1), there is strong reason to question the validity of the positive slope-change estimate produced by ITSACORR. Clearly, the level (or intercept) change, slope change, and autocorrelation estimates associated with the ITSACORR method do not describe these data to any reasonable degree.

Study D.

Spencer et al. (2000) applied ITSACORR to a multiple-baseline design that contained three experimental series and one control series. A visual inspection of their complete data (not illustrated here) reveals a major shift to each experimental phase following the intervention and very little change throughout the control series. Although they did not apply ITSACORR to the control series, such an analysis is illuminating. The control data appear in panel D of Figure 1.

If ITSACORR provides reasonable level change and slope change estimates it should confirm the visual impression of little change in the control series other than an upward trend that is quite consistent throughout the duration of the experiment. Although no intervention interrupted this series, a vertical line was inserted to show the time point at which the intervention interrupted one of the experimental series. As seen in Table 1, the level change estimate provided by ITSACORR is almost 56 points ($p < .001$) even though the intervention was not applied to this series. ITSE yielded essentially the same results. In contrast, the other methods estimate a minor decrease in level that fails to reach statistical significance ($p > .05$). All methods essentially agree with respect to the degree of slope change and autocorrelation.

Summary of Observed Differences Between ITSACORR and Other Methods Regarding Parameter Estimates.

A comparison of the ITSACORR level-change estimates with those provided by three acceptable statistical methods (as well as by visual analysis) reveals major inconsistencies for each published study illustrated in Figure 1. In some cases the ITSACORR estimate approximates the estimate provided by ITSE (an unacceptable method), but often these two methods produce very different estimates. A comparison of results from all analyses reveals level-change estimates for ITSACORR that are as much as 50 times as large as the others. In some cases, the ITSACORR estimate is far larger than the difference between the highest and lowest values in the entire series. Although the discrepancies among level change estimates tended to be larger than the discrepancies among slope change and autocorrelation estimates, discrepancies among the latter measures are also pronounced. Because the results of ITSACORR differ so much from those associated with both visual analysis and acceptable statistical methods it is reasonable to ask why. The next two sections provide answers to this question.

Inconsistency Between the Logic of the Design and the Parameters of the Structural Model

This section focuses on the comparison of the intercept parameters specified in the ITSACORR structural model with the level change parameter dictated by the logic of the two phase design. The ITSACORR structural model [identical to the Gottman (1981) ITSE model] comprises two parts, one for the pre intervention data and one for the post intervention data, as shown below.

ITSACORR Model

Pre intervention (1)

$$Y_t = m_1 t + b_1 + \sum_{i=1}^{P} a_i Y_{t-i} + e_t$$

Post intervention (2)

$$Y_t = m_2 t + b_2 + \sum_{i=1}^{P} a_i Y_{t-i} + e_t$$

where, using Gottman's notation, m1 and m2 are the process slopes for phases 1 and 2, respectively, b1 and b2 are the process intercepts for phases 1 and 2, respectively, ai is the ith autoregressive coefficient, P is the autoregressive order of the model, and et is the error. The time indicator t associated with the outcome variable Y takes on values 1, 2, . . ., n1 for observations in the first phase, and values n1 + 1, . . . , n1 + n2 for observations in the second phase (Gottman, 1981, p. 349). The numbering of the time indicator is crucial in understanding the nature of the intercepts defined for this model.

The difference between the two intercept parameters (i.e. b1 and b2) in this model does not measure the change in level at (or near) the appropriate time point n1 + 1. Both b1 and b2 measure elevation at the time point before the first observation in the first phase (i.e., time period zero). The value of b1 results from extrapolating back only one time point, whereas the value of b2 results from extrapolating from time point n1 + 1 all the way back to time point zero. Although both intercepts are associated with the same time point (i.e., zero), the difference between these two measures does not, in general, yield a measure of level change. One can, however, derive the correct level change parameter from the parameters of the ITSACORR model (Huitema & McKean, 2000a, p. 57). The correct expression for the level change parameter is: (b2 - b1) + (n1 + 1)(m2 - m1). It can be seen from this expression that the intercept difference (b2 - b1) is equivalent to the level change parameter only if the two slopes are exactly the same. Because the intercepts in the ITSACORR structural model define elevation at time period zero rather than time period n1 + 1, the model defines change effects that do not coincide with the logic of the two-phase interrupted time-series design.

Inconsistency Between the Structural Model and the Design Matrix

The first stage in the estimation of the parameters of the ITSACORR structural model can be carried out using the full model ITSE-ITSACORR design matrix shown in the Appendix (panel A) . Nevertheless, this matrix is not consistent with the design matrix that conforms to the structural model. The inconsistency can be seen in the numbering of the time periods for the second phase of the design. The second phase numbering follows the sequence t = n1 + 1, . . . , n1 + n2 in the structural model (presented above), whereas the design matrix actually employed in the ITSACORR analysis (see column four in panel A of the Appendix) uses the sequence t = 1, 2, . . . , n2 . This inconsistency means that the ITSACORR method and the resulting parameter estimates deviate from the ITSACORR structural model (which is also inconsistent with the logic of the design) and the intercept parameters it implies. This distinction between the model and the design matrix serves as an important step in conceptually decomposing the problems with the method.

Unacceptable Inferential Performance

It has been shown that ITSACORR provides unacceptable descriptive results. This outcome eliminates most interest in the inferential aspects of the analysis because there is little reason to consider hypothesis tests (or confidence intervals) applied to invalid parameter estimates. Nevertheless, for the sake of completeness, it is shown in this section that the inferential aspects of the analysis remain invalid even if one ignores the unacceptable descriptive properties of the ITSACORR method.

The inferential approach recommended for ITSACORR comprises a two-stage procedure. First, a preliminary omnibus F-test is carried out to test the following compound hypothesis: H0: m1 = m2 and b1 = b2. This hypothesis states that both slopes are identical and both intercepts are identical for the two

phases of the study. The test is based on a comparison of results obtained using the full and reduced model design matrices shown in the Appendix. Rejection of the compound hypothesis is typically interpreted to mean that an intervention effect has occurred in the form of either a slope change or an intercept change (or both). A separate t-test on each sub hypothesis (i.e., H0: m1 = m2 and H0: b1 = b2) is then carried out. Many researchers, however, ignore the preliminary test and attend to only the t's.

At first glance this two stage approach appears consistent with conventional statistical practice outside the time-series context. Upon close inspection, however, it can be seen that the ITSACORR preliminary F-test on the compound hypothesis contains fatal flaws. There has been provided a formal mathematical proof elsewhere (Huitema, McKean, & Laraway, 2007) that illustrates the problem with this test. The essential idea can be conveyed simply. Suppose one has a situation in which there is no level change whatsoever and the slopes are identical (i.e., there is a common slope). As the common slope approaches infinity the difference between ITSACORR intercepts approaches infinity even though the level has not changed. It follows that the difference between intercepts can be infinitely large even though the value of the preliminary F is zero. Because the F-test does not provide information relevant to the evaluation of differences between the intercepts defined for the ITSACORR method, this test has been ignored in the analyses presented in Table 1.

Simulation results regarding the empirical Type I error relevant to the preliminary F-test and the t-tests on change between intercepts and change between slopes are as follows: Type I error for the preliminary omnibus F-test on both intercept and slope change = .25 and .37 when autocorrelation is set at .50 and .80, respectively. The corresponding error rates on the individual test for intercept change equaled .16 and .20, and the corresponding results for the test on slope change equaled .21 and .33. Because the empirical Type I error rates greatly exceed the nominal value the tests do not possess satisfactory inferential properties and the results regarding power are of no interest. Consequently, power results are not provided. Other results, not presented here, show that if realistic levels of slope exist in the first phase, the Type I error rate for the t on intercept change is approximately 1.0.

## Conclusion

The ITSACORR method begins with Gottman's ITSE procedure and adds to it some well-intended modifications. Unfortunately, the descriptive and inferential properties are unacceptable. Each aspect of the whole framework (including the structural model, the design matrix, the autocorrelation estimator, the ultimate parameter estimation scheme, and the inferential method) contains fatal flaws. It can thus be concluded that the ITSACORR method does not provide information that is relevant to the purposes of the interrupted time-series design. Moreover, there is no situation in which one can recommend the use of ITSACORR. This conclusion is clearly at odds with recent recommendations in the literature. Some comments on these published recommendations are in order.

An examination of the foundation supporting the recommendations to use ITSACORR rather than Gottman's ITSE or ARIMA intervention models reveals little more than restatements of claims contained in the original descriptions of the method. Crosbie (1995, p. 391) compared the results produced by ITSACORR with those produced by Gottman's ITSE and an ARIMA moving averages intervention model that Glass et al. (1975) had previously applied to a portion of Holtzman's (1963) perceptual speed data. Crosbie concluded that "all three procedures reach the same conclusion" (p. 392). These methods are not based on the same assumptions regarding the nature of the underlying time-series process and they do not estimate the same parameters. These differences are reflected in the parameters modeled. This is why there are no slopes in the cited ARIMA analysis. Therefore, the claim that ITSACORR, ITSE, and ARIMA procedures "reach the same conclusion" (Crosbie, p. 392) is without foundation. Unfortunately there are several textbooks (e.g., Franklin, Allison, &

Gorman, 1997 and Christensen, 2007) and many recent journal articles that perpetuate this mistaken notion.

Another misunderstanding regarding ITSACORR relative other procedures have recently appeared. Jenson, Clark, Kircher, and Kristjansson (2007) have stated that "ITSACORR yields conservative estimates of intervention effects" (p. 488). Examples presented have been based on published data where this is far from true. Studies C and D in the present article yield ITSACORR estimates of intervention effects that are approximately 10 to 25 times the size of the correct estimates.

Because it has been shown that both the descriptive and inferential properties of ITSACORR are unacceptable it is recommend that this method not be used. More adequate methods include certain ARIMA and regression-based approaches cited in this article; it is recommended that they be given serious consideration when choosing an analysis for interrupted time-series designs.

## References


Borckardt, J. J. (2002). Case study examining the efficacy of a multi-modal psychotherapeutic intervention for hypertension. International *Journal of Clinical and Experimental Hypnosis*, *50*, 189-201.

Borckardt, J. J., & Nash, M. R. (2002). How practitioners (and others) can make scientifically viable contributions to clinical-outcome research using the single-case time-series design. *International Journal of Clinical and Experimental Hypnosis*, *50*, 114-148.

Box, G. E. P., & Jenkins, G. M. (1970). *Time-series analysis: Forecasting, and control*. San Francisco: Holden-Day.

Box, G. E. P., & Tiao, G. L. (1965). A change in level of a non-stationary time series. *Biometrics*, *52*, 181-192.

Christensen, L. B. (2007). *Experimental Methodology* (10th ed.). Boston: Allyn and Bacon.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *J*ournal of Consulting and Clinical Psychology, 61, 966-974.

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; How it can be improved. In J. M. Gottman (Ed.), *The Analysis of Change* (p. 361-395). Mahwah, NJ: Lawrence Erlbaum.

Davidson, K. M., & Tyrer, P. (1996). Cognitive therapy for antisocial and borderline personality disorders: Single case study series. *British Journal of Clinical Psychology*, *35*, 413-429.

Ellis, M.V. (1999). Repeated measures designs. *The Counseling Psychologist*, *27*, 552-578.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975*). Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and Analysis of Single-case Research* (p. 159-214). Mahwah, NJ: Lawrence Erlbaum.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. New York: Cambridge University Press.

Gottman, J. M. (Ed.). (1995). *The analysis of change*. Mahwah, NJ: Lawrence Erlbaum.

Gottman, J. M., & Rushe, R. H. (1993). The analysis of change: Issues, fallacies, and new ideas. *Journal of Consulting and Clinical Psychology*, *61*, 907-910.

Hogenraad, R., McKenzie, D. P., & Martindale, C. (1997). The enemy within: Autocorrelation bias in content analysis of narratives. *Computers and the Humanities*, *30*, 433-439.

Holtzman, W. (1963). Statistical models for the study of change in the single case. In C. W. Harris (Ed.) *Problems in Measuring Change* (p. 99-211). Madison: University of Wisconsin Press.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *3*, 27-46.

Huitema, B. E. & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*, 104-116.

Huitema, B. E., & McKean, J. W. (2000a). Design specification issues in time-series intervention models. *Psychological Measurement*, *60*, 38-58.

Huitema, B. E. & McKean, J. W. (2000b). A simple and powerful test for autocorrelated errors in OLS intervention. *Psychological Reports*, *87*, 3-20.

Huitema, B. E. & McKean, J. W. (2007). *Time-Series Intervention Analysis using ITSACORR: Fatal Flaws Expanded version*. Unpublished Manuscript.

Huitema, B. E., McKean, J. W., & McKnight, S. (1994, August). *Small-sample time-series intervention analysis: Problems and solutions*. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, *44*, 483-493.

Lucyshyn, J. M., & Albin, R. W. (1997). Embedding comprehensive behavioral support in family ecology: An experimental, single-case analysis. *Journal of Consulting and Clinical Psychology*, *65*, 241-251.

McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences*. Newbury Park, CA: Sage.

McKnight, S., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, *5*, 87-101.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single-subject clinical-outcome research: designs, data, effect sizes, and analyses. *Aphasiology*, *13*, 445-473.

Rushe, R. H., & Gottman, J. M. (1993). Essentials in the design and analysis of time-series experiments. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical Issues* (p. 493-528). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Spencer, K. A., Doyle, P. J., McNeil, M. R., Wambaugh, J. L., Park, G., & Carroll, B. (2000). Examining the facilitative effects of rhyme in a patient with output lexicon damage. *Aphasiology*, *14*, 567-584.

Stage, S. A., & Quiroz, D. R. (1997). A meta-analysis of interventions to decrease disruptive classroom behavior in education settings. *School Psychology Review*, *26*, 333-368.

Stoline, M. R., Huitema, B. E., & Mitchell, B. (1980). Intervention time-series model with different pre- and post-intervention first-order autoregressive parameters. *Psychological Bulletin*, *88*, 46-53.

Appendix

(A)
ITSE - ITSACORR Full Model Design Matrix (X) and Y Vector

$$\mathbf{X} = \left[\begin{array}{cccc|c} 1 & 1 & 0 & 0 & Y_1 \\ 1 & 2 & 0 & 0 & Y_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & n_1 - 1 & 0 & 0 & Y_{n_1-1} \\ \hline 0 & 0 & 1 & 1 & Y_{n_1} \\ 0 & 0 & 1 & 2 & Y_{n_1+1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 1 & n_2 & Y_{N-1} \end{array}\right] \qquad \mathbf{Y} = \left[\begin{array}{c} Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ Y_{n_1+2} \\ . \\ . \\ . \\ Y_N \end{array}\right]$$

(B)
ITSE - ITSACORR Reduced Model Design Matrix (XR) and Y Vector

$$\mathbf{X}_R = \begin{bmatrix} 1 & 2 & Y_1 \\ 1 & 3 & Y_2 \\ 1 & 4 & Y_3 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & N & Y_{N-1} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} Y_2 \\ Y_3 \\ Y_4 \\ . \\ . \\ . \\ Y_N \end{bmatrix}$$

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 380-398
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

*Regular Articles*

# A Comparison of Procedures for the Analysis of Multivariate Repeated Measurements

Lisa M. Lix
University of Manitoba

Anita M. Lloyd
University of Alberta

Three procedures for analyzing within-subjects effects in multivariate repeated measures designs are compared when group covariances are heterogeneous: the multiple regression model (MRM) with a structured covariance, Johansen's (1980) procedure, and the multivariate Brown and Forsythe (1974) procedure. A preliminary likelihood ratio test of a Kronecker product covariance structure is sensitive to sample size and derivational assumption violations. Error rates of the procedures are generally well-controlled except when the distribution is skewed. The MRM procedure displayed few power advantages over the other procedures.

Key words: doubly multivariate data; robustness; Kronecker product; assumption violations; general linear model; Kenward-Roger approximation.

## Introduction

Multivariate repeated measures data arise when measurements are obtained from study participants on $P$ dependent variables at each of $T$ occasions. The choice of a procedure for testing multivariate within-subjects main and interaction effects depends, in part, on the assumptions made about $\text{cov}(\mathbf{Y}_{ij}) = \mathbf{\Omega}_{ij}$, where $\mathbf{Y}_{ij} = [Y_{ij11}\ Y_{ij12}\ \ldots\ Y_{ij1P}\ \ldots\ Y_{ijTP}]^{\mathrm{T}}$, the vector of measurements for the $i$th subject ($i = 1, \ldots, n_j$) in the $j$th group ($j = 1, \ldots, J$), and $^{\mathrm{T}}$ is the transpose operator.

Two procedures for testing within-subjects effects in multivariate repeated measures data are the doubly multivariate model (DMM) and multivariate mixed model (MMM) procedures (Boik, 1988, 1991; Crawford & Johnson, 1994; Naik & Rao, 2001; Thomas, 1983), which are extensions of multivariate analysis of variance (MANOVA) and analysis of variance (ANOVA) for repeated measurements, respectively, to the case of two or more dependent variables. Both procedures define the multivariate mean response as a function of the measurement occasions and the between-subjects (i.e., grouping) factor levels. The DMM makes no assumptions about the structure of $\mathbf{\Omega}_{ij} = \mathbf{\Omega}$, where $\mathbf{\Omega}$ is the pooled covariance, other than it is positive definite. The MMM assumes a multivariate spherical (M-spherical) structure for $\mathbf{\Omega}$, in which pairs of repeated measurements exhibit a common variance across the dependent variables. The MMM is more powerful than the DMM for testing multivariate within-subjects effects if the assumption of M-sphericity is satisfied and the data follow a multivariate normal distribution (Boik, 1988; 1991). However if M-sphericity is not a tenable assumption, Type I error rates of the MMM tests may be substantially inflated; the magnitude of the deviation from the nominal level of significance, $\alpha$, will increase as the degree of departure from an M-spherical structure increases (Boik, 1988). Accordingly, Boik recommended the DMM over the MMM

Lisa Lix is an Associate Professor, Department of Community Health Sciences, and Director, Biostatistical Consulting Unit, University of Manitoba. Email her at: lisa_lix@cpe.umanitoba.ca. Anita Lloyd is a biostatistician.

provided that total sample size is sufficiently large.

The multiple regression model (MRM) with a structured covariance is one alternative to the ANOVA and MANOVA procedures for the analysis of repeated measures data. Jennrich and Schluchter (1986) and Zimmerman and Nunez-Anton (2001) (see also Fitzmaurice, Laird, & Ware, 2004; Littell, Pendergast, & Natarajan, 2000) have described this procedure for the case of a single dependent variable. The MRM procedure allows the researcher to specify a parametric form for both the mean and covariance of the repeated measurements. When a parsimonious structure is specified for the covariance, the MRM procedure should result in a more powerful test of within-subjects effects than the MANOVA procedure because there are fewer parameters to estimate and greater denominator degrees of freedom. However if the covariance structure is incorrectly specified, tests of within-subjects effects may be biased (Guerin & Stroup, 2000).

In multivariate repeated measures data, a parsimonious structure for $\mathbf{\Omega}$ in the MRM procedure is a Kronecker product structure (Galecki, 1994), that is, $\mathbf{\Omega} = \mathbf{\Sigma}_T \otimes \mathbf{\Sigma}_P$ (Chaganty & Naik, 2002; Mitchell, Genton, & Gumpertz, 2006; Naik & Rao, 2001), where $\mathbf{\Sigma}_T$ is the covariance of the repeated measurements, $\mathbf{\Sigma}_P$ is the covariance of the dependent variables and $\otimes$ is the Kronecker product operator. This structure is also referred to as a separable covariance structure (Mitchell et al., 2006). A likelihood ratio test (LRT) of a Kronecker product structure has been proposed for choosing between the MRM and DMM procedures (Naik & Rao, 2001; Roy & Khattree, 2005; Timm, 2002). If the MRM procedure is adopted, selection of the best-fitting model from the set of candidate models with different covariance structures is accomplished either by assessing the statistical significance of a LRT for two nested models, or by comparing the values of a penalized log likelihood-based information criterion, such as the Akaike criterion (Akaike, 1974), for these candidate models (Fitzmaurice et al., 2004; Littell et al., 2000).

There has been only limited investigation of the MRM with a structured covariance when $P \geq 2$ (e.g., Chinchilli & Carter, 1984; Reinsel, 1982), and not for the case when group covariances are heterogeneous. Previous research on methods for the analysis of multivariate repeated measures data when covariances are heterogeneous has focused on the properties of DMM tests of multivariate within-subjects main and interaction effects and robust alternatives to DMM tests. Robust alternatives include Johansen's (1980) approximate degrees of freedom (ADF) multivariate test and a multivariate extension of the Brown and Forsythe (1974) ADF test (Keselman & Lix, 1997; Lix, Algina, & Keselman, 2003; Vallejo, Fidalgo, & Fernandez, 2001). These ADF tests have been implemented with least-squares estimators when the data follow a multivariate normal distribution, as well as with trimmed estimators (i.e., trimmed means and Winsorized covariances) for the case when the data follow a multivariate heavy-tailed or skewed distribution. While the Johansen and Brown and Forsythe ADF tests are insensitive to covariance heterogeneity, they assume an unstructured form for $\mathbf{\Omega}$ and should, in theory, be less powerful than the MRM, provided that all procedures can control the rate of Type I errors to $\alpha$. However, Johansen's procedure is also known to produce inflated error rates when sample size is small (Keselman & Lix, 1997). Thus, at present it is not clear which procedure(s) should be recommended for analyzing multivariate repeated measurements data when covariances are heterogeneous.

The objectives of this article are to: (a) examine the Type I error performance of a LRT of a Kronecker product structure for $\mathbf{\Omega}$, and (b) compare the Type I error and power of the MRM, Johansen (1980), and multivariate Brown and Forsythe (1974) procedures for testing multivariate within-subjects main and interaction effects when covariances are heterogeneous. As part of the second objective, several information criteria are investigated for selecting the best-fitting model from amongst candidate models with different covariance structures for the MRM in the presence of covariance heterogeneity.

Description of Procedures

Notation

The procedures are described for a multivariate design with $T$ repeated measurements, $P$ dependent variables, and $J$ levels of a between-subjects factor. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y} = \left[\mathbf{Y}_{11}^{\mathrm{T}} \ldots \mathbf{Y}_{n_J J}^{\mathrm{T}}\right]^{\mathrm{T}}$ is the $N$ x $TP$ matrix of responses with $\mathbf{Y}_{ij}$ as defined previously, $\mathbf{X}$ is an $N$ x $q$ design matrix, $\boldsymbol{\beta}$ is the $q$ x $TP$ matrix of fixed effect parameters to be estimated, and $\boldsymbol{\varepsilon}$ is the $N$ x $TP$ matrix of residual errors. The rows of $\boldsymbol{\varepsilon}$ are assumed to be independent and to follow a normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Omega}_j$, the covariance for the $j$th group.

Likelihood Ratio Test of a Kronecker Product Structure for $\boldsymbol{\Omega}$

There are $T(T + 1)/2 + P(P + 1)/2$ parameters to estimate when $\boldsymbol{\Omega}$ has a Kronecker product structure, compared to $TP(TP + 1)/2$ parameters to estimate when $\boldsymbol{\Omega}$ is unstructured. For example, with $T = 4$ and $P = 2$, there are a total of 10 + 3 = 13 parameters to be estimated in the former case, compared to 8(9)/2 = 36 parameters to be estimated in the latter case.

Tests of different forms of a Kronecker product structure have been described in the literature (e.g., Boik, 1991; Naik & Rao, 2001). Mitchell et al. (2006) derived a LRT for a general Kronecker product structure, which makes no assumptions about the form of either $\boldsymbol{\Sigma}_T$ or $\boldsymbol{\Sigma}_P$. To test the null hypothesis $H_{01} : \boldsymbol{\Omega} = \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_P$ against the alternative $H_{A1} : \Omega \neq \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_P$ the test statistic is

$$\lambda = \frac{\left|\hat{\boldsymbol{\Sigma}}_T\right|^{NP/2} \left|\hat{\boldsymbol{\Sigma}}_P\right|^{NT/2}}{\left|\mathbf{E}\right|^{N/2}} \tag{2}$$

where

$$\mathbf{E} = \frac{1}{N}\mathbf{Y}^{\mathrm{T}}\left[\mathbf{I}_N - \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{X}\right]\mathbf{Y} \tag{3}$$

and $\mathbf{I}_N$ is an identity matrix of dimension $N$. The statistic $-2\ln\lambda$ asymptotically follows a $\chi^2_f$ distribution, where $f = PT(PT + 1)/2 - P(P + 1)/2 - T(T + 1)/2$, under the assumptions of a multivariate normal distribution of responses and covariance homogeneity. Maximum likelihood (ML) estimates of $\boldsymbol{\Sigma}_P$ and $\boldsymbol{\Sigma}_T$ can be obtained via algorithms proposed by Boik (1991), Dutilleul (1999), or Mardia & Goodall (1993).

The Multiple Regression Model, Johansen (1980), and Multivariate Brown and Forsythe (1974) Procedures

For multivariate repeated measures data, the MRM procedure with a structured covariance is defined as

$$\mathrm{vec}(\mathbf{Y}^{\mathrm{T}}) = (\mathbf{X} \otimes \mathbf{I}_{TP})\mathrm{vec}(\boldsymbol{\beta}^{\mathrm{T}}) + \mathrm{vec}(\boldsymbol{\varepsilon}^{\mathrm{T}}) \tag{4}$$

where $\mathrm{vec}(\cdot)$ is the vec operator, and

$$\begin{aligned}\mathrm{cov}(\mathrm{vec}(\mathbf{Y}^{\mathrm{T}})) &= \mathbf{I}_N \otimes \boldsymbol{\Omega} \\ &= \mathbf{V},\end{aligned} \tag{5}$$

when homogeneity of group covariances is assumed. If $\mathbf{V}$ is known, the least-squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}&\mathrm{vec}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}) = \\ &\left[(\mathbf{X} \otimes \mathbf{I}_{TP})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{X} \otimes \mathbf{I}_{TP})\right]^{-1} (\mathbf{X} \otimes \mathbf{I}_{TP})^{\mathrm{T}} \mathbf{V}^{-1} \mathrm{vec}(\mathbf{Y}^{\mathrm{T}})\end{aligned} \tag{6}$$

When the data follow a multivariate normal distribution, $\hat{\boldsymbol{\beta}}$ also follows a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}) = \left[(\mathbf{X} \otimes \mathbf{I}_{TP})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{X} \otimes \mathbf{I}_{TP})\right]^{-1}. \tag{7}$$

When $\mathbf{V}$ is unknown, covariance parameters (i.e., $\hat{\mathbf{V}}$) are estimated using ML or restricted maximum likelihood (REML). Then $\hat{\mathbf{V}}$ is substituted for $\mathbf{V}$ in equations 6 and 7.

When this substitution is made, $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$ under asymptotic theory (Fitzmaurice et al., 2004). Kackar and Harville (1981) have also shown that, provided the population distribution is symmetric, $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimate of $\boldsymbol{\beta}$. However, for small sample sizes, the accuracy of the approximation may be poor, particularly when $T$ and/or $P$ are large. When covariances are not assumed homogeneous across groups, separate parameters are estimated for each level of the between-subjects factor, such that $\hat{\boldsymbol{\Omega}}_j$ denotes the estimated covariance for the $j$th group.

Hypotheses about multivariate within-subjects main and interaction effects are of the form $H_{02}$: $\mathbf{L}\text{vec}(\boldsymbol{\beta}^{\mathrm{T}}) = \mathbf{0}$ where $\mathbf{L}$ of dimension $r$ x $qTP$ contains weights that define one or more linear contrasts among the elements of $\boldsymbol{\beta}$. To test hypotheses about the individual $\beta_m$ ($m$ = 1 ,…, $qTP$), a number of different test statistics may be adopted (Fouladi & Shieh, 2004), including an approximate $t$ statistic. If the rank of $\mathbf{L}$ is greater than one, an $F$ statistic to test the null hypothesis is,

$$F = \frac{1}{\text{rank}(\mathbf{L})}\left(\mathbf{L}\text{vec}(\hat{\boldsymbol{\beta}}^{\mathrm{T}})\right)^{\mathrm{T}}\left[\mathbf{L}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})\mathbf{L}^{\mathrm{T}}\right]^{-1}\left(\mathbf{L}\text{vec}(\hat{\boldsymbol{\beta}})\right) \quad (8)$$

where $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\beta}})$ estimates $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}})$. This statistic is compared to the critical value, $F[(1 - \alpha); v_1, v_2]$, where $v_1$ = rank($\mathbf{L}$) and $v_2$ is approximated from the data (Kenward & Roger, 1997; Satterthwaite; 1946). Guerin and Stroup (2000) recommend adopting the Kenward-Roger approximation because the former can result in inflated error rates for small sample sizes.

As noted previously, for the MRM procedure applied to $P \geq 2$ dependent variables, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_T \otimes \boldsymbol{\Sigma}_P$ defines a parsimonious structure for $\boldsymbol{\Omega}$. The matrix $\boldsymbol{\Sigma}_P$ is typically assumed to have an unstructured form with $P(P + 1)/2$ unique elements $\sigma_{ll'}$ ($l$, $l'$ = 1 ,…, $P$), while the parameters in $\boldsymbol{\Sigma}_T$ are assumed to be a function of the measurement occasions $u$ and $v$ ($u$, $v$ = 1 ,…, $T$) and the levels of one or more between-subjects factor(s) when covariance homogeneity is not assumed (Zimmerman & Nunez-Anton, 2001). Examples of possible structures for $\boldsymbol{\Sigma}_T$ have been enumerated in several sources, including Fitzmaurice et al. (2004), Littell et al. (2000), and Littell, Stroup, and Freund (2002). For example, the compound symmetric (CS) structure has the following variance and correlation specification: $\sigma_{kk} = \sigma^2$ and $\rho_{kk'} = \rho$ ($k$ = 1 ,…, $T$ ; $k \neq k'$). This parsimonious structure assumes constant variances and correlations across measurement occasions. Multivariate compound symmetry is a more restrictive assumption than that of M-sphericity (Crawford & Johnson, 1994). The variance and correlation specifications for the first order autoregressive (AR-1) structure is $\sigma_{kk} = \sigma^2$, for $k$ = 1 ,…, $T$ and $\rho_{kk'} = \rho^{k'-k}$ for $k' > k$ and $k'$ = 2, …, $T$. The unstructured (UN) covariance has $T(T + 1)/2$ unique elements denoted $\sigma_{kk'}$ ($k$, $k'$ = 1 ,…, $T$).

Several information criteria for assessing model fit have been proposed, including the Akaike (AIC; Akaike, 1974), Bayesian-Schwarz (BIC: Schwarz, 1978), finite population-corrected AIC (CAIC; Bozdogan, 1987), and Hannan & Quinn (HQIC; Hannan & Quin, 1979) criteria. These are respectively defined as

$$\begin{aligned} \text{AIC} &= -2l + 2d, \\ \text{BIC} &= -2l + d\log(N^*), \\ \text{CAIC} &= -2l + d\log(N^* + 1), \\ \text{HQIC} &= -2l + 2d\log\log(N^*), \end{aligned} \quad (9)$$

where $l$ is the logarithm of the ML or REML function for the specified model, $d$ is the number of covariance parameters to be estimated, $N^* = N$ for ML estimation and $N^* = N - q$ for REML estimation. Amongst candidate models with different covariance structures, the best-fitting model is the one with the smallest value for the selected information criterion. The criteria will not always select the same model. For example, the BIC penalizes the model more severely for the number of parameters than does the AIC, and therefore tends to choose less complex models than the AIC.

Computational formulae for the Johansen (1980) and multivariate Brown and Forsythe (1974) ADF procedures have been enumerated in a number of sources, and

therefore have not been repeated in this manuscript (e.g., Keselman & Lix, 1997; Lix et al., 2003). Vallejo, Fidalgo, and Fernandez (2001) extended the Brown and Forsythe (1974) procedure to the case of multivariate repeated measurements, but a recent modification proposed by Vallejo and Ato (2006), to address the conservative Type I error properties of this procedure for testing within-subjects effects, was adopted in the current study.

## Methodology

### Methods for Investigating the Properties of a Likelihood Ratio Test of a Kronecker Product Covariance Structure

Monte Carlo techniques were used to investigate the Type I error properties of the LRT for testing the null hypothesis that $\mathbf{\Omega}$ has a Kronecker product structure. The data were generated for a multivariate design containing a single between-subjects factor with two levels and a single within-subjects factor. The parameters manipulated in the study were: (a) total sample size, (b) number of repeated measurements, (c) number of dependent variables, (d) degree of covariance heterogeneity, and (e) degree of departure from a multivariate normal distribution. The value of the LRT statistic does not depend on the form of either $\mathbf{\Sigma}_T$ or $\mathbf{\Sigma}_P$ (Mitchell et al., 2003), so the data were generated from a population in which both $\mathbf{\Sigma}_T$ or $\mathbf{\Sigma}_P$ had CS covariance structures.

Dutilleul's (1999) algorithm (see also Dutilleul & Pinel-Alloul, 1996) was used to obtain ML estimates of $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_P$. This algorithm finds solutions to the following system of equations,

$$\hat{\mathbf{\Sigma}}_P = \frac{1}{TN}\sum_{i=1}^{n_j}\sum_{j=1}^{J}(\mathbf{W}_{ij}-\bar{\mathbf{W}})\hat{\mathbf{\Sigma}}_T^{-1}(\mathbf{W}_{ij}-\bar{\mathbf{W}})^{\mathrm{T}} \quad (10)$$

and

$$\hat{\mathbf{\Sigma}}_T = \frac{1}{PN}\sum_{i=1}^{n_j}\sum_{j=1}^{J}(\mathbf{W}_{ij}-\bar{\mathbf{W}})\hat{\mathbf{\Sigma}}_P^{-1}(\mathbf{W}_{ij}-\bar{\mathbf{W}})^{\mathrm{T}} \quad (11)$$

where $\mathrm{W}_{ij}$ is the $P$ x $T$ matrix obtained by re-shaping $\mathrm{Y}_{ij}$, and $\overline{\mathbf{W}}$ is the matrix of means obtained by averaging across all such observation matrices.

Three levels of total sample size were investigated: $N$ = 40, 60, and 100. The number of repeated measurements was set at $T$ = 4 and 6, while the number of dependent variables was set at $P$ = 2, 3, 4, and 6. These conditions reflect the range of simulation parameters that have been investigated in previous research on methods for the analysis of multivariate repeated measures data (Boik, 1991; Lix et al., 2003; Vallejo et al., 2001).

The tests were investigated for homogeneous covariances (i.e., $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$), as well as for two cases of covariance heterogeneity: $\Omega_1 = 5\Omega_2$, and $\Omega_1 = 9\Omega_2$.

Multivariate data were generated from both normal and non-normal distributions. Pseudorandom observation vectors $\mathrm{Y}_{ij}$ from a multivariate normal distribution with mean vector $\mathbf{\beta}_j$ and covariance matrix $\mathbf{\Omega}_j$ were obtained by the following method. A column vector of standard normal deviates (i.e., $\mathrm{d}_{ij}$) was transformed to a vector of multivariate observations via $\mathbf{Y}_{ij} = \mathbf{\beta}_j + \mathbf{R}\mathbf{d}_{ij}$ where R is an upper triangular matrix of dimension $TP$ with the property $\mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{\Omega}_j$.

Two multivariate non-normal distributions were investigated (Lix, Keselman, & Hinds, 2005). The first was a symmetric distribution with a mild degree of heavy-tailedness and skewness ($\gamma_1$) and kurtosis ($\gamma_2$) values of 0 and 1.7 respectively, (the normal distribution has $\gamma_1$ = 0 and $\gamma_2$ = 0) while the second distribution had $\gamma_1$ = 2.0 and $\gamma_2$ = 6.0, which are equivalent to the shape parameters of an exponential distribution. A vector of constants $\mathbf{w} = [a\ b\ c\ d]^{\mathrm{T}}$ was obtained using Fleishman's (1978) method, to provide the desired degree of skewness and kurtosis for each of these distributions. An intermediate covariance matrix (i.e., $\zeta$) was then computed so that $\mathbf{Y}_{ij}$ would have the desired final covariance structure. Elements of this intermediate matrix were computed using Vale and Maurelli's (1983) method. The vector of univariate deviates was transformed to a vector of multivariate

normal deviates via, $\mathbf{Z}(\zeta)_{ij} = \boldsymbol{\beta}_j + \mathbf{R}_\zeta \mathbf{d}_{ij}$, where $\mathbf{Z}(\zeta)_{ij}$ is the vector of transformed variates, and $\mathbf{R}_\zeta$ is an upper triangular matrix such that $\mathbf{R}_\zeta^{\mathrm{T}} \mathbf{R}_\zeta = \boldsymbol{\zeta}$. Next, each element of $\mathbf{Y}_{ij}$ was obtained by computing the zero through third powers of the corresponding elements of $\mathbf{Z}(\zeta)_{ij}$, so that $\mathbf{Z}(\zeta)_{ijm} = [1\ \mathbf{Z}(\zeta)_{ijm}\ \mathbf{Z}(\zeta)_{ijm}^2\ \mathbf{Z}(\zeta)_{ijm}^3]$ ($m$ = 1 ,…, $TP$) represents the vector of powers. From this, $Y_{ijkl} = \mathbf{Z}(\zeta)_{im}\mathbf{w}$ ($k$ = 1 ,…, $T$; $l$ = 1 ,…, $P$).

Five thousand replications of each combination of conditions were performed using $\alpha$ = .05 as the criterion for assessing statistical significance. The simulation program was written in SAS/IML (SAS Institute Inc., 2004a). Descriptive techniques were used to summarize the Type I error rates.

Methods for Investigating the Properties of the Multiple Regression Model, Johansen, and Multivariate Brown and Forsythe Procedures

Monte Carlo techniques were used to evaluate the Type I error and power of the MRM, Johansen (1980), and multivariate Brown and Forsythe (1974) procedures for testing multivariate within-subjects main and interaction effects as well as to investigate the properties of the information criteria for assessing model fit. The data were generated for a design with a single between-subjects factor with two levels, a single within-subjects factor with four levels, and three dependent variables. The simulation parameters were: (a) total sample size, (b) degree of group size imbalance, (c) degree of covariance heterogeneity, (d) pairing of group sizes and covariances, (e) degree of departure from a multivariate normal distribution, (f) structure of $\boldsymbol{\Omega}$, and (g) configuration of the population means.

The analyses were conducted for total sample size conditions of 60 and 100. Group sizes were both equal and unequal. Table 1 lists the group sizes that were investigated for each value of $N$. The tests were investigated for homogeneous covariances (i.e., $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$), as well as for two cases of covariance heterogeneity: $\boldsymbol{\Omega}_1 = 5\boldsymbol{\Omega}_2$, and $\boldsymbol{\Omega}_1 = 9\boldsymbol{\Omega}_2$.

Positive and negative pairings of group sizes and group covariances were investigated. A positive pairing refers to the case in which the largest $n_j$ is associated with the covariance matrix containing the largest element values; a negative pairing refers to the case in which the largest $n_j$ is associated with the covariance matrix with the smallest element values.

Multivariate normal and non-normal data were generated using the method described in the previous section. Two multivariate non-normal distributions were investigated. The first was a symmetric heavy-tailed distribution with shape parameters equivalent to those of a double exponential distribution (i.e., $\gamma_1$ = 0; $\gamma_2$ = 3.0.) The second was a skewed distribution that represented an extreme degree of departure from multivariate normality, with shape parameters equivalent to those of a multivariate lognormal distribution (i.e., $\gamma_1$ = 6.2; $\gamma_2$ = 110.9).

In this phase of the study, $\boldsymbol{\Sigma}_T$ had either a CS or AR-1 structure. Both structures had $\sigma^2$ = 1 and $\rho$ = 0.5. $\boldsymbol{\Sigma}_P$ had a CS structure with $\sigma^2$ = 1 and $\rho$ = 0.4.

All procedures were investigated when the configuration of population means was null and non-null. For the non-null case, the following configuration of means was investigated: $\boldsymbol{\beta}_1$ = [.25 0 0 -.25 0 0 0 0 .25 0 0 -.25]$^{\mathrm{T}}$ and $\boldsymbol{\beta}_2$ = [.25 0 0 .25 0 0 0 0 .25 0 0 .25]$^{\mathrm{T}}$.

Multivariate datasets were generated using a program written in SAS/IML (SAS Institute Inc., 2004a). A SAS/IML program was also written to analyze each dataset with the Johansen (1980) and multivariate Brown and Forsythe (1974) procedures. A PROC MIXED (SAS Institute Inc., 2004b) macro was written to analyze each dataset using the MRM procedure, and output the $F$ statistics, $p$-values, and degrees of freedom for tests of the within-subjects main and interaction effects, as well as the numeric values for each of the four investigated information criterion. Only one thousand replications were performed for each combination of conditions because of the lengthy execution time required for PROC MIXED. The syntax to implement the MRM procedure is reported in Appendix A; it is the same as that reported by Timm (2002) and Thiebaut, Jacquim-Gadda, Chene, Leport and Commenges (2002). All parameters were estimated using REML. In the PROC MIXED macro, each dataset was analyzed using

Table 1. Group Sizes Investigated in the Simulation Study

| $N$ | $n_1, n_2$ | $\Delta n_j$ |
|---|---|---|
| 60 | 30, 30 | 0.0 |
| | 24, 36 | 0.2 |
| | 20, 40 | 0.3 |
| 100 | 50, 50 | 0.0 |
| | 40, 60 | 0.2 |
| | 35, 65 | 0.3 |

Note. $N$ = total sample size. $\Delta n_j$ is the coefficient of variation for group sizes (see Lix & Keselman, 1997).

Three different models; each model had the same fixed effects, but a different Kronecker product covariance structure in the REPEATED statement (see Appendix A). The best-fitting model, among the three models that were fit to the data, was the model that resulted in the lowest numeric value of an information criterion. The percentage of times that the best-fitting model had the same covariance structure as the population covariance structure was recorded for each criterion; this is denoted as the percentage of correct model selection.

Type I error and power rates were calculated for the MRM, Johansen (1980), and multivariate Brown and Forsythe (1974) procedures. The percent bias in Type I error rates, $B = 100(\hat{\alpha} - \alpha)/\alpha$, where $\hat{\alpha}$ is the empirical error rate for a test, was also calculated. Type I error rates, percentages of bias, and power rates were summarized descriptively. For each of the three procedures, regression analyses were used to model the effect of the simulation parameters on the Type I error rates. For the MRM, the regression model had a random simulation effect, because there were repeated measurements on each model covariance structure. For each procedure, separate models were defined for equal and unequal group size cases for within-subjects main and interaction effects, respectively. All models included main effects as well as two-way interactions among the simulation parameters.

## Results

### Likelihood Ratio Test of a Kronecker Product Covariance Structure

Figure 1 contains the empirical Type I error rates for the LRT for each of the investigated values of $P$ when $T = 4$ and $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$. The sensitivity of the LRT to total sample size is apparent. When the data were multivariate normal and $P = 2$, the empirical error rate was 0.11 for $N = 40$, but quickly converged to 0.05 when $N = 100$. As the dimension of $\mathbf{\Omega}$ increased from $TP = 8$ to 36, error rates also increased across the range of values of $N$. For example, with $P = 3$ and $N = 100$, the error rate was 0.09. As the dimension of the data increased, Type I error rates of the LRT rapidly approached the upper bound of 1.00.

Error rates were also highly sensitive to the presence of multivariate non-normality. For example, when the data were sampled from a heavy-tailed distribution when $P = 2$, the error rate was 0.18 for $N = 40$ and 0.12 for $N = 100$. Error rates were also more inflated for the skewed distribution than the heavy-tailed distribution. For the latter distribution, error rates attained or approached the upper bound of 1.00 for many of the investigated conditions.

When the number of repeated measurements increased to six and $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, the same pattern of results was observed, although error rates were even more inflated than for $T = 4$. For example, for $N = 100$, the empirical error

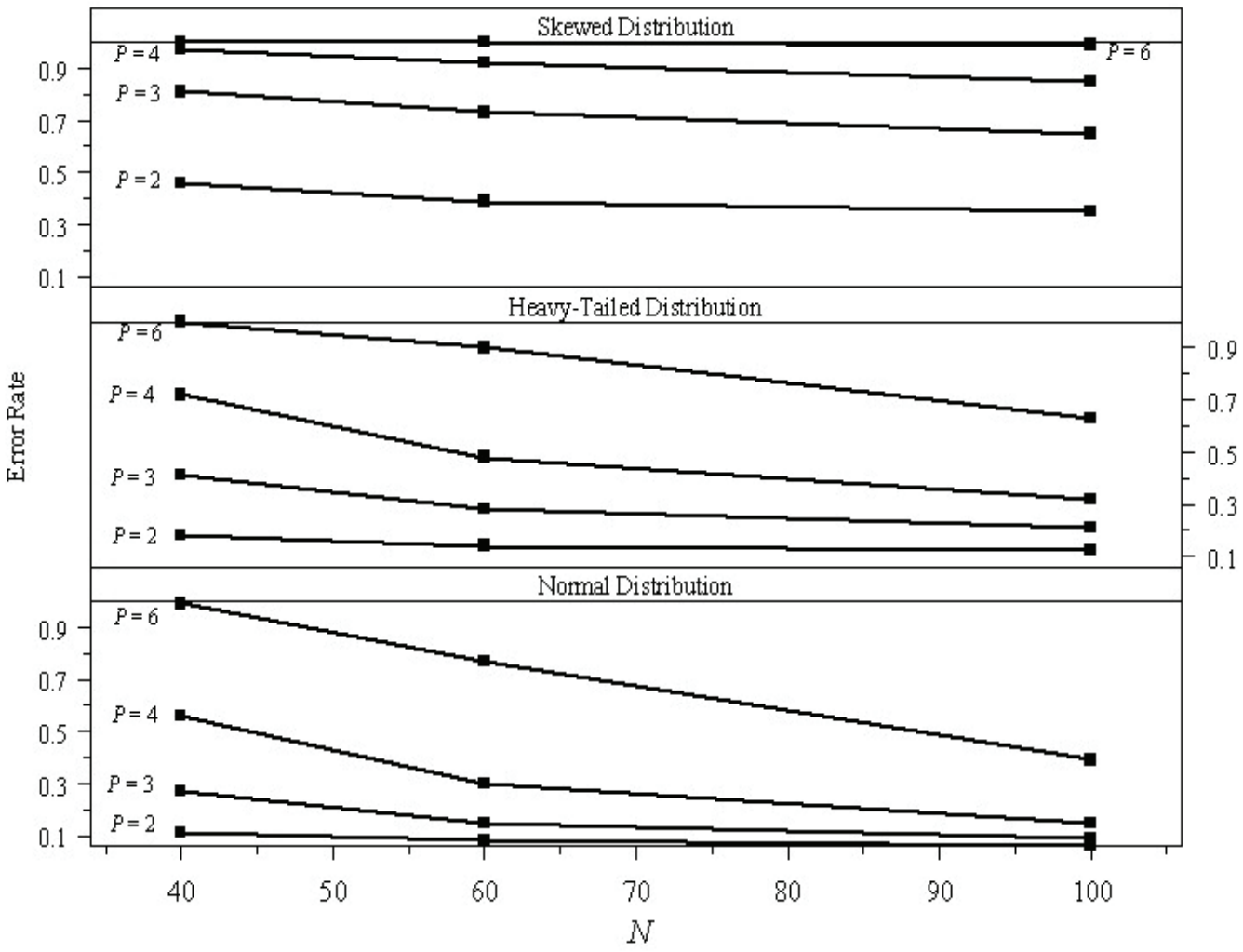



Figure 1. Type I Error Rates for the Likelihood Ratio Test (LRT) of a Kronecker Product Structure, $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, $T = 4$

*Note*. Heavy-tailed distribution has $\gamma_1 = 0$ and $\gamma_2 = 1.7$. Skewed distribution has $\gamma_1 = 2.0$ and $\gamma_2 = 6.0$.

rate was 0.07 for $P = 2$ but increased to 0.89 for $P = 6$ when the data were sampled from a multivariate normal distribution. When the data were from multivariate non-normal distributions, error rates always exceeded $\alpha$ and frequently attained the upper bound of 1.00.

When covariances were heterogeneous, the error rates of the LRT were always inflated regardless of the total sample size, dimension of $\mathbf{\Omega}$, or degree of departure from a multivariate normal distribution. For example, when the data were multivariate normal with $T = 4$ and $P = 2$, the error rate was 0.52 for $N = 40$ and 0.37 for $N = 100$.

Information Criteria for Assessing Model Fit

The results for the four information criteria are reported in Table 2; they have been averaged over the conditions of total sample size, degree of covariance heterogeneity, and degree of group size imbalance because a similar

pattern of results was observed for these conditions.

When the data followed a multivariate normal distribution and $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, the differences among the four criteria were small and on average, the correct (i.e., population) covariance structure was selected in as many as 96.4% of the models. When covariances were heterogeneous and the distribution was multivariate normal, the percentages of correct model selection were lower when $\mathbf{\Sigma}_T$ had a CS structure (approximately 65% for all criteria) than when $\mathbf{\Sigma}_T$ had an AR-1 structure (approximately 90% for all criteria).

When $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$ and the data were sampled from non-normal distributions, the average percentages of correct model selection were lower than when the data were sampled from a multivariate normal distribution. For the heavy-tailed distribution when $\mathbf{\Sigma}_T$ had a CS structure, correct model selection was observed, on average, for 64.8% of the models for the AIC and 75.2% of the models for the BIC. When $\mathbf{\Sigma}_T$ had an AR-1 structure, the average percentages for the AIC and BIC were 77.4% and 90.3%, respectively, for this distribution. For the skewed distribution, the percentages were substantially lower; for example, when $\mathbf{\Sigma}_T$ had a CS structure, correct model selection was observed for only 22.2% of the models for the AIC. Moreover, the AIC and HQIC were more sensitive to multivariate non-normality than the BIC and CAIC. The latter two procedures always resulted in higher average percentages of correct models than the former two procedures when the data were obtained from heavy-tailed or skewed distributions.

When covariances were heterogeneous and the data were obtained from multivariate non-normal distributions, a similar pattern of results was observed. All four information criteria produced similar average percentages of correct model selection regardless of whether group sizes were equal or unequal. The values obtained when $\mathbf{\Sigma}_T$ had an AR-1 structure were higher than those obtained when $\mathbf{\Sigma}_T$ had a CS structure. The AIC and HQIC were more sensitive to multivariate non-normality than the BIC and CAIC.

Tests of Multivariate Within-Subjects Main and Interaction Effects

Type I error rates

Percentages of bias in Type I error rates are reported in Tables 3 and 4; the results are averaged across the two values of total sample size because a similar pattern of results was observed. For both the main and interaction effect tests when group sizes were equal (Table 3), the average bias was small for Johansen's (1980) procedure as well as for the multivariate Brown and Forsythe (1974) procedure when the data were from symmetric distributions. For Johansen's test, average bias ranged from 10.0% to 19.5% for the within-subjects main effect and from -1.5% to 22.0% for the within-subjects interaction effect. For the Brown and Forsythe (1974) test, average bias ranged from -2.0% to 13.0% for the within-subjects main effect and from -11.0% to 7.5% for the within-subjects interaction. Overall, the Type I error rates of the MRM procedure were more biased than error rates of either ADF procedure when the distribution was symmetric. Average bias ranged from -28.0% to 29.0% for the within-subjects main effect and from -15.0% to 23.0% for the within-subjects interaction effect. However, the results for symmetric distributions reveal that the magnitude of bias varied across the three model covariance structures. As expected, when $\mathbf{\Sigma}_T$ had a CS structure, there was generally less bias in the error rates when a model with a CS or UN covariance structure for the repeated measurements was adopted than when a model with an AR-1 structure was adopted. Similarly, when $\mathbf{\Sigma}_T$ had an AR-1 structure, there was less bias when a model with either an AR-1 or UN structure was selected than when a model with a CS structure was selected. However, for the former case, the AR-1 structure tended to result in negatively biased error rates for the within-subjects main effect and positively biased results for the within-subjects interaction effect, while in the latter case the CS structure results in

Table 2. Average Percentages of Correct Model Selection for Four Information Criteria

| | | AIC | HQIC | BIC | CAIC |
|---|---|---|---|---|---|
| | | | $\boldsymbol{\Sigma}_T$ = CS | | |
| Nor | $=n_j/=\boldsymbol{\Omega}_j$ | 86.4 | 87.5 | 87.5 | 87.5 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 63.0 | 63.2 | 63.3 | 63.3 |
| | + pair | 65.9 | 66.3 | 66.3 | 66.3 |
| | - pair | 64.9 | 65.3 | 65.3 | 65.3 |
| HT | $=n_j/=\boldsymbol{\Omega}_j$ | 64.8 | 73.7 | 75.2 | 75.3 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 59.2 | 63.8 | 64.5 | 64.5 |
| | + pair | 58.9 | 64.5 | 65.2 | 65.2 |
| | - pair | 58.9 | 64.5 | 65.3 | 65.3 |
| SK | $=n_j/=\boldsymbol{\Omega}_j$ | 22.2 | 24.3 | 30.6 | 38.4 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 20.6 | 22.4 | 29.2 | 36.5 |
| | + pair | 21.6 | 23.4 | 30.4 | 37.9 |
| | - pair | 21.5 | 23.6 | 31.0 | 38.8 |
| | | | $\boldsymbol{\Sigma}_T$ = AR-1 | | |
| Nor | $=n_j/=\boldsymbol{\Omega}_j$ | 95.6 | 96.4 | 96.4 | 96.4 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 89.9 | 90.4 | 90.4 | 90.4 |
| | + pair | 92.1 | 92.8 | 92.8 | 92.8 |
| | - pair | 90.5 | 91.1 | 91.1 | 91.1 |
| HT | $=n_j/=\boldsymbol{\Omega}_j$ | 77.4 | 88.5 | 90.3 | 90.4 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 75.5 | 83.5 | 84.6 | 84.6 |
| | + pair | 75.4 | 85.8 | 87.3 | 87.4 |
| | - pair | 75.6 | 83.0 | 84.1 | 84.2 |
| SK | $=n_j/=\boldsymbol{\Omega}_j$ | 20.4 | 23.1 | 30.5 | 38.4 |
| | $=n_j/\neq\boldsymbol{\Omega}_j$ | 19.8 | 22.0 | 30.0 | 37.8 |
| | + pair | 20.0 | 22.9 | 31.9 | 40.1 |
| | - pair | 20.8 | 23.7 | 31.9 | 40.7 |

*Note*. CS = compound symmetric; AR-1 = first-order autoregressive. Nor = multivariate normal distribution with $\gamma_1 = 0$ and $\gamma_2 = 0$; HT = multivariate heavy-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 3$; SK= multivariate skewed distribution with $\gamma_1 = 6.2$ and $\gamma_2 = 110.9$. $=n_j/=\boldsymbol{\Omega}_j$ = equal group sizes and equal group covariances; $=n_j/\neq\boldsymbol{\Omega}_j$ = equal group sizes and unequal covariances; + pair = positive pairing of group sizes and covariances; - pair = negative pairing of group sizes and covariances.

positively biased error rates for both main and interaction tests. For the multivariate skewed distribution, Type I error rates were almost always negatively biased for all procedures when group sizes were equal. This finding was observed for both main and interaction effects. For example, for the multivariate within-subjects interaction, average bias for Johansen's (1980) procedure ranged from -23.0% to -48.0% and for the multivariate Brown and Forsythe (1974) procedure average bias ranged from -31.0% to -51.0%. For the MRM procedure, average bias was similar, and ranged from -13.5% to -45.0% across the three models for the interaction effect.

Percentages of bias for unequal group sizes are reported in Table 4; separate summaries are given for positive and negative pairings of group sizes and covariances and the results are averaged across conditions of total sample size, degree of covariance heterogeneity, and degree of group size imbalance because

Table 4. Average Percentages of Bias in Type I Error Rates for Multivariate Within-Subjects Effects when Group Sizes are Unequal

| | | $\boldsymbol{\Sigma}_T$ = CS | | | | | $\boldsymbol{\Sigma}_T$ = AR-1 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MRM CS | MRM AR-1 | MRM UN | J | BF | MRM CS | MRM AR-1 | MRM UN | J | BF |
| | | | | | | Main Effect | | | | | |
| Nor | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | 7.0 | -4.0 | 4.0 | 18.0 | 13.0 | 24.0 | 7.0 | 16.0 | 6.0 | 4.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | 9.7 | -23.2 | 6.6 | 11.5 | 4.0 | 29.0 | 4.0 | 17.0 | 19.5 | 7.5 |
| HT | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | -5.0 | -28.0 | -22.0 | 11.0 | 7.0 | 27.0 | 10.0 | -1.0 | 10.0 | 9.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | -11.9 | -23.4 | -11.4 | 10.0 | -1.5 | 19.0 | -4.0 | 8.0 | 11.0 | -2.0 |
| SK | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | -17.0 | -47.0 | -24.0 | -39.0 | -39.0 | -16.0 | -29.0 | -32.0 | -44.0 | -46.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | -36.0 | -54.0 | -40.5 | -22.0 | -29.0 | 4.5 | -12.0 | -31.5 | -26.5 | -35.5 |
| | | | | | | Interaction Effect | | | | | |
| Nor | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | 10.0 | 13.0 | -11.0 | 5.0 | 1.0 | 13.0 | -4.0 | -14.0 | 10.0 | 6.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | -3.6 | 9.6 | -3.1 | 9.0 | -1.5 | 21.5 | 2.0 | 10.0 | 22.0 | 7.5 |
| HT | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | -15.0 | 12.0 | -10.0 | 8.0 | 4.0 | 8.0 | -11.0 | -4.0 | 3.0 | 0.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | -7.6 | -7.7 | -9.1 | -1.5 | -11.0 | 23.0 | -0.5 | 8.5 | 8.5 | -2.5 |
| SK | $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ | -23.0 | -14.0 | -36.0 | -46.0 | -50.0 | -19.0 | -44.0 | -45.0 | -48.0 | -51.0 |
| | $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$ | -29.5 | -31.5 | -39.0 | -24.5 | -33.0 | -13.5 | -20.5 | -32.0 | -23.0 | -31.0 |

*Note*. CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured. MRM = multiple regression model; J = Johansen's (1980) procedure; BF = multivariate Brown and Forsythe (1974) procedure; Nor = multivariate normal distribution with $\gamma_1 = 0$ and $\gamma_2 = 0$; HT = multivariate heavy-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 3$; SK = multivariate skewed distribution with $\gamma_1 = 6.2$ and $\gamma_2 = 110.9$. + pair = positive pairing of group sizes and covariances; - pair = negative pairing of group sizes and covariances.

similar patterns of results were observed. For Johansen's (1980) procedure, bias was almost always positive for the multivariate normal and heavy-tailed distributions for both within-subjects main and interaction effects; it was highest for negative pairings, where average bias ranged from 31.8% to 51.5%. For the multivariate Brown and Forsythe (1974) procedure, bias was small for the normal and heavy-tailed distributions and ranged from -9.0% to 8.5% across the main and interaction effects. For the MRM procedure, average bias ranged from -29.0% to 23.5% for the main effect and from -14.4% to 28.3% for the interaction effect for symmetric distributions across the three model covariance structures. The same pattern of results was observed for unequal group sizes as was observed for equal group sizes. When $\boldsymbol{\Sigma}_T$ = AR-1 and a model with a CS structure was adopted, positive bias was observed for both the multivariate normal and heavy-tailed distributions. However, when $\boldsymbol{\Sigma}_T$ = AR-1 and a model with either an AR-1 or UN covariance structure was adopted, error rates were less biased and ranged from -6.0% to 10.8% for the main effect and from 1.5% to 14.8% for the interaction effect. When $\boldsymbol{\Sigma}_T$ = CS and a model with an AR-1 structure was adopted, bias ranged from -29.0% to -19.9% for the main effect and from -5.9% to 7.1% for the interaction effect.

When the distribution was multivariate skewed and group sizes were unequal, error rates were almost always negatively biased for the three multivariate procedures. For example, for the MRM procedure, bias ranged from -50.8% to -5.6% for the main effect and from

-86.5% to -19.0% for the interaction effect. The range of values for average bias was similar for equal and unequal group sizes.

As noted previously, for both balanced and unbalanced designs, it was generally the case that less bias was observed in Type I error rates when the model covariance structure corresponded to the population covariance structure. For MRM tests of the multivariate within-subjects main effect, the average bias across equal group sizes for $\mathbf{\Sigma}_T$ = CS when the data were multivariate normal was 8.8% for the model with the CS covariance structure, -16.8% for the AR-1 model, and 5.8% for the UN structure. When $\mathbf{\Sigma}_T$ = AR-1 under a normal multivariate distribution, the average bias across equal group size conditions was 5.0% for the model with the AR-1 covariance structure, 27.3% for the model with the CS structure, and 16.7% for the model with the UN covariance structure. The average percentages of bias were similar for multivariate within-subjects main and interaction effects. As sample size increased from $N$ = 60 to 100 when group sizes were equal, bias tended to decrease for both the normal and skewed distributions under either population covariance structure while bias tended to increase when data were from a heavy-tailed distribution. When group sizes were unequal and sample size increased from $N$ = 60 to 100, bias tended to increase when data was from a normal distribution while bias tended to decrease when data was from a skewed distribution. For data from a heavy-tailed distribution, a trend in average bias values was not evident.

When group sizes were equal, the regression model for Johansen's (1980) procedure accounted for 87.7% of the variation in Type I error rates for the within-subjects main effect and 90.5% of the variation in error rates for the within-subjects interaction effect. For the multivariate Brown and Forsythe (1974) procedure, the model accounted for 87.5% of the variation in Type I error rates for the within-subjects main effect and 86.2% of the variation in Type I error rates for the interaction effect. For both procedures, the majority of this variation was attributed to the main effect of population distribution (i.e., between 85.5% and 99.0% of the total explained variation). For Johansen's procedure for the test of the within-subjects interaction, total sample size and the two-way interaction of population distribution and degree of covariance heterogeneity accounted for 6.1% and 5.1% of the variation in error rates, respectively. Other main and two-way interaction effects in the models accounted for a small percentage of the explained variation.

For the MRM when group sizes were equal, the regression model that contained main effects and two-way interactions accounted for 92.8% of the variation in Type I error rates for the within-subjects main effect test and 91.5% of the variation in error rates for the within-subjects interaction effect test. For the main effect test, population distribution, population covariance structure for $\mathbf{\Sigma}_T$, and model covariance structure respectively accounted for 39.1%, 20.1%, and 12.1% of the explained variation. None of the other model effects individually accounted for more than 5% of the variation. For the within-subjects interaction effect test, the type of population distribution, two-way interaction between total sample size and degree of covariance heterogeneity, two-way interaction between population covariance structure and model covariance structure, and two-way interaction of total sample size and type of population distribution accounted for 47.5%, 6.6% and 5.7%, and 5.4% of the explained variance, respectively. None of the other model effects accounted for more than 5% of the variation.

When group sizes were unequal, the regression analyses for Johansen's (1980) procedure revealed that the model containing main effects and two-way interactions accounted for 90.2% of the variation in Type I error rates for the within-subjects main effect and 89.3% of the variation for the within-subjects interaction. For both tests, the majority of the explained variation was due to the main effects of population distribution, total sample size, and pairing of group sizes and covariances, and to the two-way interaction of total sample size and pairing of group sizes and covariances for the within-subjects interaction effect tests. For the multivariate Brown and Forsythe (1974) procedure, the regression model accounted for 80.3% and 84.3% of the variation in Type I error rates for the multivariate within-subjects main

and interaction effects, respectively. Almost all of this explained variation (i.e., > 90%) was due to the main effect of population distribution, although the two-way interaction of degree of covariance heterogeneity and degree of group size imbalance also accounted for slightly more than 5% of the explained variation.

For the MRM procedure, the regression model accounted for 91.5% of the variation in Type I error rates for the test of the within-subjects main effect and 93.9% of the variation in error rates for the within-subjects interaction effect. The model effects that accounted for the most explained variation for the within-subjects main effect were population type (34.1%), population covariance structure (21.2%), and model covariance structure (10.8%). For the within-subjects interaction, variables that accounted for most of the explained variance were population type (55.2%) and population covariance structure (6.8%).

## Power

Percentages of power are reported in Table 5. They are summarized separately for equal and unequal group sizes for multivariate normal and heavy-tailed distributions. Power results are not reported for the skewed distribution because the three procedures could not control the Type I error rate for this condition. Only the results for N = 60 are reported because the pattern of results was similar for N = 100, and because power approached its upper bound for several of the simulation conditions for this latter value.

The Johansen (1980) and multivariate Brown and Forsythe (1974) procedures produced similar percentages of power for all conditions, except when group sizes and covariances were negatively paired. In this case, the Johansen (1980) procedure was more powerful than the multivariate Brown and Forsythe (1974) procedure. This is likely a result of the slightly liberal Type I error rates that were observed for the former procedure for negative pairing conditions. Average power was 56.9% and 56.2% for Johansen's procedure for the within-subjects main and interaction effects, respectively, and the corresponding values for the multivariate Brown and Forsythe (1974) procedure were 53.3% and 51.9%.

In general, the MRM procedure resulted in lower power than either the Johansen (1980) or multivariate Brown and Forsythe (1974) procedures for the within-subjects main effect; the average differences ranged between 5% and 20% for most of the investigated conditions although the difference was occasionally greater than this. For the interaction effect, the MRM procedure was often more powerful than either of the ADF tests, although the differences were never more than 10%. However, Johansen's procedure for testing the interaction effect was, on average, more powerful than the MRM procedure when group sizes and covariances were negatively paired.

When $\mathbf{\Sigma}_T$ = CS, the MRM procedure with either a CS or UN model covariance structure resulted in substantially higher power than the MRM with an AR-1 model structure. For example, when the data were normally distributed and group sizes were equal, the average power was 41.1% and 40.2%, respectively for the CS and UN model covariance structures, and only 26.5% for the AR-1 structure. This large difference in power was observed for both equal and unequal group sizes regardless of the shape of the population distribution. In contrast, when $\mathbf{\Sigma}_T$ = AR-1, the MRM procedure resulted in similar percentages of power for the CS, UN, and AR-1 structures. For example, when the data were from a heavy-tailed distribution and group sizes and covariances were negatively paired, the average power was 33.1%, 32.7%, and 33.2% for the multivariate interaction effect under CS, AR-1 and UN structures respectively.

## Conclusion

Multivariate repeated measurements arise in the social, behavioral, and health sciences when researchers collect data on multiple psychological or physiological characteristics of study participants over time or across multiple experimental conditions. Global tests of hypotheses for multivariate within-subjects main or interaction effects take account of the correlation that exists among the repeated measurements and dependent variables. These tests may be conducted within the context of the

Table 5. Average Percentages of Power for Multivariate Within-Subjects Effects when Group Sizes are Equal and Unequal, $N = 60$

| | | $\boldsymbol{\Sigma}_T$ = CS | | | | | $\boldsymbol{\Sigma}_T$ = AR-1 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MRM CS | MRM AR-1 | MRM UN | J | BF | MRM CS | MRM AR-1 | MRM UN | J | BF |
| Multivariate Main Effect | | | | | | | | | | | |
| Nor | $= n_j$ | 41.1 | 26.5 | 40.2 | 50.2 | 47.3 | 30.9 | 30.7 | 30.1 | 38.2 | 35.5 |
| | + pair | 47.3 | 31.0 | 46.1 | 57.5 | 55.8 | 35.5 | 34.2 | 33.5 | 43.9 | 42.2 |
| | - pair | 31.3 | 19.5 | 30.0 | 41.2 | 32.6 | 24.1 | 22.5 | 22.0 | 32.2 | 24.3 |
| HT | $= n_j$ | 39.1 | 25.5 | 38.6 | 52.2 | 49.4 | 31.5 | 30.2 | 29.9 | 40.1 | 37.2 |
| | + pair | 47.1 | 32.3 | 46.6 | 59.3 | 57.9 | 36.8 | 35.0 | 35.8 | 44.7 | 42.9 |
| | - pair | 30.8 | 19.7 | 30.6 | 42.5 | 34.5 | 24.9 | 23.1 | 23.3 | 33.2 | 25.8 |
| Multivariate Interaction Effect | | | | | | | | | | | |
| Nor | $= n_j$ | 55.5 | 42.2 | 53.6 | 50.7 | 48.3 | 42.5 | 43.9 | 41.5 | 37.23 | 34.8 |
| | + pair | 61.8 | 48.7 | 60.2 | 58.5 | 57.2 | 50.4 | 51.2 | 48.0 | 43.8 | 42.5 |
| | - pair | 39.4 | 29.3 | 38.0 | 42.3 | 31.1 | 31.7 | 32.0 | 29.6 | 31.5 | 22.7 |
| HT | $= n_j$ | 54.0 | 42.2 | 54.8 | 53.0 | 50.4 | 42.5 | 42.4 | 42.2 | 39.8 | 36.9 |
| | + pair | 61.5 | 47.9 | 60.2 | 59.8 | 58.6 | 50.2 | 51.7 | 49.6 | 45.0 | 43.9 |
| | - pair | 40.7 | 31.5 | 38.8 | 43.9 | 33.8 | 33.1 | 32.7 | 33.2 | 33.5 | 24.6 |

*Note*. CS = compound symmetric; AR-1 = first-order autoregressive; UN = unstructured. MRM = multiple regression model; J = Johansen's(1980) procedure; BF = multivariate Brown and Forsythe (1974) procedure; Nor = multivariate normal distribution with $\gamma_1 = 0$ and $\gamma_2 = 0$; HT = multivariate heavy-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 3$. $=n_j =$ equal group sizes; + pair = positive pairing of group sizes and covariances; - pair= negative pairing of group sizes and covariances.

general linear model using one of several procedures, including multivariate extensions of ANOVA and MANOVA. The choice of a procedure depends, in part, on the assumptions the researcher is willing to make about the covariance structure of the data.

A likelihood ratio test of a Kronecker product covariance structure, which might be used as a preliminary test to choose between the multiple regression model procedure that assumes a structured covariance and a procedure that makes no assumptions about the structure of the covariance matrix, requires a large sample size, relative to the dimension of the data, to control the rate of Type I errors to the nominal level of significance when the data are sampled from a multivariate normal distribution and covariances are homogeneous. When covariances are heterogeneous or the data are sampled from multivariate non-normal distributions, this test can result in severely inflated Type I error rates. Thus, the likelihood ratio test is not useful as a preliminary test because it will almost always reject the null hypothesis of a Kronecker product structure in favor of the alternative hypothesis of an unstructured covariance.

Consistent with the results of previous research, the Johansen (1980) and multivariate Brown and Forsythe (1974) procedures provided good control of the Type I error rate across the majority of the investigated conditions when the data were sampled from multivariate normal and heavy-tailed distributions. When group sizes and covariance matrices were negatively paired, Johansen's (1980) test produced slightly inflated Type I error rates, although the magnitude of this positive bias decreased as total sample size increased. Both procedures produced conservative Type I error rates when the data

were sampled from a multivariate skewed distribution.

The multiple regression model procedure also provided good control of Type I error rates across the majority of the investigated conditions when the data were sampled from multivariate normal and heavy-tailed distributions. Like the other two procedures, it resulted in conservative error rates when the data were sampled from a multivariate skewed distribution. As expected, Type I error rates deviated less from the nominal level of significance when the selected model covariance structure corresponded to the population covariance structure, or when an unstructured covariance was selected.

For tests of the within-subjects main effect, the Johansen (1980) and multivariate Brown and Forsythe (1974) procedures were more powerful than the multiple regression model procedure regardless of which model covariance structure was selected for the latter. The differences in power were moderate to large. For the multiple regression model, power was higher when the selected model covariance structure was the same as the population covariance structure, than when an unstructured covariance model was selected, but the differences were small (i.e., less than five percentage points). For tests of the within-subjects interaction effect, the multiple regression model procedure was often more powerful than the other two procedures, but the differences were modest. Moreover, the multiple regression model procedure could also be less powerful than the Johansen or the multivariate Brown and Forsythe procedures if the covariance structure was misspecified.

For both within-subjects main and interaction effects, the magnitude of power differences among the multiple regression model procedures when the three model covariance structures for the repeated measurements were compared indicated that the power advantages gained by correctly specifying the covariance structure varies as a function of the form of the population covariance. When the population covariance structure of the repeated measurements was compound symmetric, the model with a compound symmetric structure was much more powerful than the model with an autoregressive structure. However, when the population covariance structure of the repeated measurements was autoregressive, there was only a small difference in power between the models with compound symmetric and autoregressive covariance structures.

Comparison of four penalized log-likelihood information criteria for assessing model fit revealed that all of the criterion performed better when group covariances were homogeneous than when they were heterogeneous, and when the data were sampled from symmetric distributions than when they were sampled from skewed distributions. The BIC and CAIC more often selected the model with the correct covariance structure than the AIC and HQIC.

Given these results, there appear to be limited benefits associated with adopting a multiple regression model procedure for testing multivariate within-subjects main and interaction effects in multivariate repeated measurement designs when covariances are heterogeneous and sample size is small or moderate. The Johansen (1980) and multivariate Brown and Forsythe (1974) approximate degrees of freedom procedures controlled the Type I error rates and were often more powerful than the multiple regression model procedures. Moreover, previous research has demonstrated that when the distribution is non-normal, robust versions of both procedures can control the rate of Type I errors to the nominal level of significance and can result in increased power to detect within-subjects effects.

While the results of this study suggest that either of the Johansen (1980) or multivariate Brown and Forsythe (1974) procedures could be recommended for analyzing within-subjects effects, researchers should be cautious in generalizing these results to all data-analytic conditions encountered in the analysis of multivariate repeated measures data. First, the properties of the three test procedures were only examined when the covariance of the repeated measurements and dependent variables had a Kronecker product structure. There have been no studies of the degree to which data encountered in the social, behavioral, and health sciences conform to a Kronecker product structure, nor of the magnitude of positive or negative bias in

error rates of the multiple regression model procedure when the data do not conform to a Kronecker product structure. Second, the test procedures were compared only for datasets with no missing observations. Unbiased estimates of regression parameters can be obtained for the multiple regression model procedure provided the observations are either missing completely at random or missing at random (Little & Rubin, 1987). The other two procedures investigated in this research do not accommodate study participants with missing observations; rather, incomplete cases are removed from the analysis, which can result in reduced power to detect within-subjects effects. Finally, power analyses were conducted only for a single effect size and a single configuration of the population means.

A number of opportunities for further research arise from this study. A likelihood ratio test of a Kronecker product structure that is less sensitive to sample size and is robust to multivariate non-normality and/or covariance heterogeneity requires investigation. Boik (1991) proposed an approximate likelihood ratio test of multivariate sphericity for small or moderate sample sizes under the assumption of a multivariate normal distribution. The approximation is based on the work of Box (1949), who proposed finding the moments of the likelihood statistics to derive the approximation. Mitchell et al. (2006) proposed a bootstrap likelihood ratio test that is less sensitive to sample size, but did not investigate its properties in the presence of covariance heterogeneity or multivariate non-normality. Zhu, Ng, and Jing (2002) compared likelihood ratio tests based on bootstrap and permutation re-sampling methods to test equality of covariances in the presence of multivariate non-normality. They found that the permutation test performed better than the bootstrap test.

Graphic techniques and statistical tests to assess model fit and select among candidate model covariance structures for the multiple regression model also need to be investigated and described for the case of multivariate repeated measures data. Littell et al. (2000) and Zimmerman and Nunez-Anton (2001) provide a thorough discussion of graphic and descriptive techniques for the case of a single dependent variable, but they have not been extended to multivariate data. Techniques for assessing model fit in the presence of multivariate non-normality include bias-corrected versions of the AIC and empirical cross-validation techniques, have been proposed (e.g., Yanagihara, 2006), and could be investigated in the context of multivariate repeated measurements.

Parametric and non-parametric procedures for the analysis of multivariate repeated measurements with structured covariances which are robust to the presence of non-normal distributions and covariance heterogeneity require development and evaluation (Wang & Zhu, 2006; Reilly, 2005). Furthermore, comparisons among these procedures under the types of data-analytic conditions that may be encountered in practice are necessary to develop recommendations on choosing a statistical procedure.

Finally, models with structured covariances that bridge the gap between the restrictive compound symmetric Kronecker product structure and the less efficient unstructured Kronecker product structure require further development for the multivariate case. One flexible covariance structure described by Zimmerman and Nunez-Anton (2001) for models with a single dependent variable is the antedependence structure. It allows for a pattern of monotonically decreasing correlation among the repeated measurements, which is common in repeated measurements, as well as for non-constant variances of the repeated measurements. The authors describe software to implement a multiple regression model with the antedependence covariance structure for the case of a single dependent variable, and this could be investigated for possible extension to the multivariate case.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC -19, 716 -723.

Boik, R. J. (1988). The mixed model for multivariate repeated measures: Validity conditions and an approximate test. *Psychometrika*, *53*, 469-486.

Boik, R. J. (1991). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communications in Statistics –Theory and Methods*, *20*, 1233-1255.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*, 317-346.

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

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

Chaganty, N. R., & Naik, D. N. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, *103*, 421-436.

Chinchilli, V. M., & Carter, W. H., Jr. (1984). A likelihood ratio test for a patterned covariance matrix in a multivariate growth curve model. *Biometrics*, *40*, 151-156.

Crawford, A. M. K., & Johnson, W. D. (1994). Statistical considerations for multivariate repeated measures with patterned covariance. *Proceedings of the 1993 Biopharmaceutical Section of the American Statistical Association,* 61-66.

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, *64*, 105-123.

Dutilleul, P., & Pinel-Alloul, B. (1996). A doubly multivariate model for statistical analysis of spatio-temporal environmental data. *Environmetrics*, *7*, 551-565.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Fouladi, R. T., & Shieh, Y-Y. (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics - Simulation and Computation*, *33*, 807-824.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics –Theory and Methods*, *23*, 3105-3119.

Guerin, L., & Stroup, W. W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture*. Manhattan, KS: Kansas State University.

Hannan, E. J., & Quinn, B.G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society* – Series B, *41*, 190 -195.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*, 805-820.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*, 85-92.

Kackar, R. N., & Harville, D. A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics – Theory and Methods*, *10*, 1249-1261.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983-997.

Keselman, H. J., & Lix, L. M. (1997). Analysing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, *50*, 319-339.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Tutorial in biostatistics: Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, *19*, 1793-1819.

Littell, R. C., Stroup, W. W., & Freund, R. J. (2002). *SAS for linear models, Fourth edition*. Cary, NC: SAS Institute, Inc.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research*, *38*, 403-431.

Lix, L. M., Keselman, H. J., & Hinds, A. (2005). Robust tests for the multivariate Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine*, *77*, 129-139.

Mardia, K. V., & Goodall, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In G. P. Patil, & C. R. Rao (Eds.), *Multivariate Environmental Statistics* (pp. 347-386). Amsterdam: Elsevier Science.

Mitchell, M. W., Genton, M. G., & Gumpertz, M. L. (2006). A likelihood ratio test for separability of covariances. *Journal of Multivariate Analysis, 97*, 1025-1043.

Naik, D. N., & Rao, S. S. (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *Journal of Applied Statistics*, *28*, 91-105.

Reilly, C. (2005). A nonparametric approach to the analysis of longitudinal data via a set of level crossing problems with application to the analysis of microarray time course experiments. *Biostatistics*, *6*, 271-278.

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, *77*, 190-195.

Roy, A. & Khattree, R. (2005). On implementations of a test for Kronecker product covariance structure for the multivariate repeated measures data. *Statistical Methodology, 2*, 297-306.

SAS Institute, Inc. (2004a). *SAS/IML user's guide, version 9.1*. Cary, NC: Author.

SAS Institute, Inc. (2004b). *SAS/STAT user's guide, version 9.1*. Cary, NC: Author.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

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

Thiebaut, R., Jacquim-Gadda, H., Chene, G., Leport, C., Commenges, D. (2002). Bivariate linear mixed models using SAS Proc MIXED. *Computer Methods and Programs in Biomedicine, 69,* 249-256.

Thomas, D. R. (1983). Univariate repeated measures techniques applied to multivariate data. *Psychometrika*, *48*, 451-464.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

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

Vallejo, G., & Ato, M. (2006). Modified Brown-Forsythe test for testing interaction effects in split-plot designs. *Multivariate Behavioral Research, 41*, 549-578.

Vallejo, G., Fidalgo, A., & Fernandez, P. (2001). Effects of covariance heterogeneity on three procedures for analyzing multivariate repeated measures designs. *Multivariate Behavioral Research*, *36*, 1-27.

Wang, Y-G., Zhu, M. (2006). Rank-based regression for analysis of repeated measures. *Biometrika*, *93*, 459-464.

Yanagihara, H. (2006). Corrected version of AIC for selecting multivariate normal linear regression models in a general nonnormal case. *Journal of Multivariate Analysis, 97*, 1070-1089.

Zhu, L.-X., Ng, K. W., & Jing, P. (2002). Resampling methods for homogeneity tests of covariance matrices. *Statistica Sinica*, *12*, 769-783.

Zimmerman, D. L., & Nunez-Anton, V. (2001). Parametric modeling of growth curve data: An overview. *Test*, *10*, 1-73.

## Appendix A. PROC MIXED Code to Implement the MRM

This is the syntax that was used to implement the MRM when the data are assumed to have a Kronecker product covariance structure, with an unstructured model for $\mathbf{\Sigma}_T$ as well as an unstructured model for $\mathbf{\Sigma}_P$.

```
proc sort data=datafin;
by id;
run;

proc mixed data=datafin method=reml ic;
class group id dv rm;
model val=group dv rm dv*rm group*dv*rm /noint
ddfm=kenwardroger;
repeated dv rm / type=un@un subject=id(group) group=group;
run;
```

Where:
id = subject identification variable
group = variable to identify levels of between-subjects grouping factor
dv = variable to identify "levels" of the dependent variable factor
rm = variable to identify levels of within-subjects factor

The data are assumed to be arranged in a "long" structure, with one value of the dependent variable *val* per line. Accordingly, each row of *datafin* contains a single observation and the corresponding values of the variables id, group, dv, and rm.

A model with a compound symmetric form for $\mathbf{\Sigma}_T$ is obtained by specifying *type = cs@un* in the *repeated* statement. A model with a first-order autoregressive form for $\mathbf{\Sigma}_T$ is obtained by specifying *type = ar(1)@un* in the *repeated* statement. The only available model for $\mathbf{\Sigma}_P$ is unstructured (UN).

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 399- 412
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Multiple Comparison Of Medians Using Permutation Tests

Scott J. Richter
University of North Carolina at Greensboro

Melinda H. McCann
Oklahoma State University

A robust method is proposed for simultaneous pairwise comparison using permutation tests and median differences. The new procedure provides strong control of familywise error rate and has better power properties than the median procedure of Nemenyi/Levy. It can be more powerful than the Tukey-Kramer procedure using mean differences, especially for nonnormal distributions and unequal sample sizes.

Key words: Simultaneous inference, pairwise comparisons, median difference, permutation test.

## Introduction

The technique of using permutation methods for multiple comparisons has received relatively little attention in the literature. Nemenyi (1963) and later Levy (1979) proposed a procedure using medians, with the maximum of the differences of pairwise Mood statistics used to construct the reference distribution. Miller (1966, 1981), and more recently Higgins (2004), proposed a permutation version of the Tukey-Kramer method (Tukey, 1949; Kramer, 1956), where the range of the sample means is calculated for each permutation of observations among the $k$ groups to obtain the reference distribution. The mean difference for each pair of means is then compared to this reference distribution to determine statistically significant differences. However, when distributions are skewed or there are outliers in the data, it may be desirable to make comparisons of medians rather than means. Thus, a logical extension of Miller's procedure is to replace means by medians. Consider the following example.

Scott J. Richter is Associate Professor and Director of the Statistical Consulting Center. His research interests are in robust methods and applied statistics. Email him at sjricht2@uncg.edu. Melinda McCann is Associate Professor of Statistics. Her research interests involve multiple comparisons procedures and their applications.

## Example

Manly (1997) reported the data in Table 1 based on articles by Powell & Russell (1984, 1985) and Linton et al (1989). The data represent dry biomass (in mg) of ants for 24 eastern horned lizards, taken in three months in 1980.

It is desired to determine which, if any, of the months have different consumptions. The relation between the means and medians for each month suggests that the distributions of biomass are skewed, and that the means may not be representative of monthly consumption. Thus, comparisons based on medians may be more appropriate.

Both the median procedure of Nemenyi and Levy and Miller's procedure permute freely across all groups (unrestricted randomization). However, this unrestricted randomization scheme has been criticized. Petrondas and Gabriel (1983) contend that Miller's approach does not control the familywise error rate (FWE): the probability of making at least one false declaration of inequality, since the test for any subset hypothesis that a pair of means is equal should be based on permuting observations only among the groups whose distributions are assumed equal under the null hypothesis. The FWE actually is controlled under the overall null hypothesis that all $k$ distributions have the same location—that is, in the weak sense (Hochberg & Tamhane, 1987), but not necessarily under a subset pairwise null hypothesis that requires only the two distributions being considered to have equal

Table 1. Dry biomass of ants for 24 eastern horned lizards, taken in three months in 1980.

| Month | Dry biomass (mg) | Median | Mean |
|---|---|---|---|
| June | 13, 242, 105 | 105.0 | 120.0 |
| July | 8, 59, 20, 2, 245 | 20.0 | 66.8 |
| August | 515, 488, 88, 233, 50, 600, 82, 40, 52, 1889 | 160.5 | 403.7 |

location, that is, in the strong sense (Hochberg & Tamhane, 1987). Accordingly, both Petrondas and Gabriel (1983) and Hochberg and Tamhane (1987) suggest performing each pairwise test separately using a Bonferroni adjustment. Similarly, Hochberg and Tamhane (1987) and Ryan and Ryan (1980) note that the median procedure of Nemenyi/Levy is not based on a joint testing family, and thus does not control the FWE. Hochberg and Tamhane (1987) instead suggest permuting separately within each pair (restricted randomization) and utilizing the maximum of pairwise Mood statistics to derive the reference distribution.

A new testing procedure is proposed based on the procedure of Nemenyi/Levy, using median difference statistics instead of differences between Mood statistics, and Type I error and power properties are compared to the new procedure to those of the Nemenyi/Levy procedure, pairwise tests using a Bonferroni adjustment, and also to the Tukey-Kramer procedure based on mean differences, which assumes normally distributed populations.

## Methodology

Throughout, consider a one-way layout with $k$ groups, where $F_i$ is the common continuous distribution function for the $i^{\text{th}}$ group, $n_i$ is the sample size of the $i^{\text{th}}$ group, and $N = n_1 + n_2 + \cdots + n_k$. Further, let $\mu_i$ be the location parameter associated with the $i^{\text{th}}$ distribution and $\hat{\mu}_i$ be the sample median for the $i^{\text{th}}$ group. Distributions are assumed identical for all treatments except for possible location differences.

Permutation-based Multiple Comparison Procedures:

Miller (1966, 1981) proposed a permutation analog to the Tukey-Kramer procedure for multiple pairwise comparison of several means. The reference distribution for Miller's method was based on the statistic, $\max_{1\le i<j\le k}\left|\overline{Y}_i - \overline{Y}_j\right|$, where $\overline{Y}_i$ and $\overline{Y}_j$ are the respective sample means of groups $i$ and $j$. The reference distribution consists of the values of this statistic for all $\dfrac{N!}{n_1!n_2!\cdots n_k!}$ possible permutations of the observed data. Each pairwise absolute difference is compared to this distribution to determine statistical significance. Bonferroni-adjusted pairwise tests suggested by Hochberg and Tamhane (1987) and Petrondas and Gabriel (1983) will also be considered.

Nemenyi (1963) and later Levy (1979) also proposed an analog to the Tukey-Kramer procedure, but based on Mood's (1950) median test, as follows. First, calculate the grand median for the pooled sample of $N = n_1 + n_2 + \cdots + n_k$ observations. Then determine $M_i$, the number of observations in the $i^{\text{th}}$ sample that exceed the grand median. The test statistic for comparing any pair is $\left|\dfrac{M_i}{n_i} - \dfrac{M_j}{n_j}\right|$. The reference distribution is based on the distribution of $\max_{1\le i<j\le k}\left|\dfrac{M_i}{n_i} - \dfrac{M_j}{n_j}\right|$, the maximum value of the test statistic over all pairs, which is calculated for a large set of random reassignments of observations to groups. As with Miller's method, an observation may be

reassigned to any of the $k$ groups to form a new permutation. Hochberg and Tamhane (1987) suggest computing a separate grand median for each pair and calculating the test statistic above. The maximum over all pairs is then found for a large set of random reassignments, where reassignments are restricted to within each pair, and these values form the reference distribution.

A New Method Using Median Differences:

In situations involving skewed distributions or outliers it may be more appropriate to consider medians instead of means. Thus, we propose multiple comparison procedures based on median differences. The method of Nemenyi/Levy, based on Mood statistics, does utilize medians, but does not incorporate the magnitude of the difference between medians. It is believed that there may be situations when incorporating this information could lead to a more sensitive procedure.

Analogous to the mean-based procedure of Miller, the reference distribution for our new procedure is based on the distribution of $\max_{1\le i<j\le k}\left|\hat{\mu}_i-\hat{\mu}_j\right|$, the maximum of all pairwise *median* differences, calculated for a large set of random reassignments of observations to groups. Each pairwise absolute median difference is compared to this reference distribution to determine statistical significance. Both methods of permuting discussed in Section 2.1, namely restricted and unrestricted, are investigated.

Restricted Randomization Guarantees FWE Control:

The strongest argument against unrestricted permuting is that it does not necessarily provide strong control of the FWE. Restricted permuting, however, does provide strong control.

Consider $k$ independent samples from distributions that differ by at most a location parameter. That is, for $i, j = 1, 2, ..., k$ with $i < j$, $F_i(x) = F_j\left(x - \Delta_{ij}\right)$. (Throughout Section 2.3 let $i, j = 1, 2, ..., k$ with $i < j$.) The null hypothesis then involves $\binom{k}{2}$ pairwise hypotheses of the form $H_{0ij}: \Delta_{ij} = 0$. Now consider the permutation distribution of median differences from samples $i$ and $j$, and let $D_{ij}(\alpha)$ be the $1-\alpha$ percentile of this permutation distribution. Similarly, define $D_{\max}(\alpha)$ to be the $1-\alpha$ percentile of the permutation distribution for the maximum median difference among all $\binom{k}{2}$ pairs.

First consider the case under the complete null hypothesis where all $\Delta_{ij} = 0$. Let the calculated median difference from samples $i$ and $j$ be denoted by $\tilde{D}_{ij}$. Under the complete null hypothesis the probability that a calculated median difference from a particular pair of samples in a given permutation is the maximum difference is $\binom{k}{2}^{-1}$. Thus, each pair of samples will contribute $\alpha\binom{k}{2}^{-1}$ of the values from the pairwise difference permutation distribution to the maximum difference permutation distribution.

Consequently, the probability that any observed difference from a particular pair exceeds $D_{\max}(\alpha)$, the comparisonwise error rate, is $\alpha\binom{k}{2}^{-1}$. Alternatively, the familywise error rate is given by

$$P(\text{declare at least one pair different in location} \mid \text{all pairs have equal location})$$
$$= \sum_{\left[i,j=1,\ldots,\binom{k}{2},\, i<j\right]} P\left(\tilde{D}_{ij} \ge D_{\max}(\alpha)\right) = \binom{k}{2}\left(\alpha / \binom{k}{2}\right)$$
$$= \alpha.$$

This shows that using the permutation distribution of the maximum difference controls

the FWE in the weak sense (Hochberg & Tamhane, 1987).

Now consider the case where only $t < \binom{k}{2}$ of the pairwise null hypotheses are indeed true. For any permutation, a difference from one of these $t$ pairs with a true pairwise null hypothesis is less likely to be the maximum difference than differences from the $\binom{k}{2} - t$ pairs where $\Delta_{ij} \neq 0$. Consequently, the comparisonwise error rate is $P\left(\tilde{D}_{ij} \geq D_{\max}(\alpha)\right) \leq \alpha \binom{k}{2}^{-1}$. Thus, the familywise error rate, the probability of rejecting at least one of the $t$ true null hypotheses, is

$$P(\text{reject at least one true null hypothesis} \mid t \text{ true null hypotheses}) \leq t\left(\alpha / \binom{k}{2}\right) < \alpha.$$

Thus, the FWE is controlled at level $\alpha$ for any combination of $t$ true and $\binom{k}{2} - t$ false hypotheses, and the FWE is controlled in the *strong* sense (Hochberg & Tamhane, 1987).

Alternatively, the FWE may be controlled by performing separate two-sample permutation tests and utilizing $\alpha \binom{k}{2}^{-1}$, a Bonferroni adjustment, as the significance level for each individual comparison. Based on their performance in the normal theory setting, it is expected that a Tukey-type permutation procedure will generally be less conservative than a procedure utilizing pairwise permutation tests with a Bonferroni adjustment.

Simulation Study

A simulation was conducted to evaluate five permutation procedures:

1. A modification of Miller's (1966, 1981) procedure, using medians instead of means and unrestricted randomization (MEDUR);
2. A modification of (1) using restricted randomization (MEDR);
3. Separate Bonferroni-adjusted pairwise permutation tests for median differences (MEDBON);
4. The procedure of Nemenyi (1963)/Levy (1979) based on differences between Mood statistics and unrestricted randomization (MOODUR);
5. A modification of (4), using restricted randomization (MOODR).

The following model was assumed to generate the data:

$$y_{ij} = \mu_i + e_{ij},$$

where $y_{ij} =$ the $j^{\text{th}}$ observation for the $i^{\text{th}}$ treatment $\mu_i =$ the location parameter for the $i^{\text{th}}$ treatment $e_{ij} =$ the random error associated with the $j^{\text{th}}$ observation for the $i^{\text{th}}$ treatment. The $e_{ij}$ are assumed independent and identically distributed.

Several different error distributions were examined:

- Normal ( $\mu = 0, \sigma^2 = 1$);
- Uniform [-3,3];
- Exponential ( $\lambda = 3$ );
- Double exponential (Exp( $\lambda = 3$ ) – Exp( $\lambda = 3$ ));
- Location-contaminated normal ($N$ (0,1) with 10% contamination from $N$ (9,1)).

These choices encompass two symmetric, nonnormal distributions: the uniform (lighter-tailed than normal) and the double exponential (heavier-tailed than normal); and two skewed distributions: the exponential and contaminated normal. Models contained either three or five groups, and both equal and unequal sample sizes were examined. In most cases the total number of permutations possible is prohibitive, and thus a random sample of permutations was used to estimate the $p$-value for any given test. Keller-McNulty and Higgins (1987) examined the issue

of randomly sampling the permutations, and concluded that little is to be gained by taking more than 1600 randomly sampled permutations. Thus, each permutation test was based on a reference distribution estimated via a slightly conservative 2000 randomly sampled permutations, and the estimated proportions of rejections were based on 2000 randomly generated samples. The simulations were implemented using Resampling Stats version 5.0 (Resampling Stats Inc., 2000).

The familywise error rate (FWE) and any-pair power (Shaffer, 1995), the probability of detecting at least one true difference, are reported in the Tables 2-12. For the Tukey-type procedures based on medians, in cases where either all groups have identical locations or all groups had different locations, these were estimated by comparing the maximum pairwise difference from among the samples to the respective reference distribution, and counting the number of random samples where this maximum was in the top 5% of the reference distribution. In cases where some pairs had identical locations while others pairs differed in location, the FWE was estimated as the proportion of permutations where at least one of the true null hypotheses was rejected (strong FWE).

## Results

### Comparison of Median-based Procedures

#### Type I Error

All median-based procedures controlled the FWE in the strong sense (See Tables 2-4). In fact, in the cases where some pairs had equal locations and some did not, the probability of at least one false rejection was usually lower than the case where all locations were equal. As Petrondas and Gabriel (1983) admitted, their counterexample was very small, and, "for realistic, larger examples the corresponding tests (using unrestricted permuting) may be both valid and useful." It is also worth noting, however, that even though the unrestricted permuting method did not exhibit inflated FWE rates for either the median difference statistic or the Mood statistic, in cases where there was a difference between unrestricted and restricted FWE rates, the unrestricted FWE was almost always higher. This was true especially with unequal sample sizes, where error rates more than twice as large for unrestricted permuting were not uncommon. As we shall see in the next section, however, higher FWE rates did not typically lead to more powerful tests. In light of this evidence and the earlier cited criticisms of unrestricted randomization, as well as the fact that power is generally at least as good under restricted randomization, only procedures using restricted randomization will be considered in the remainder of the discussion.

#### Power

Consider first the case of equal sample sizes. With small group sample size ($n = 5$) and small location differences ( $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$ or $\Delta_1 = \Delta_2 = 2,\ \Delta_3 = \Delta_4 = \Delta_5 = 0$), MEDR always had the highest power among the median procedures (See Tables 5 and 7). When there were larger location differences ($\Delta_1 = \Delta_2 = 2, \Delta_3 = 5$ or $\Delta_1 = \Delta_2 = 2,\ \Delta_3 = 3,\ \Delta_4 = \Delta_5 = 0$),MOODR often had highest power for normal and contaminated normal data (e.g., see Table 6). On the other hand, MEDBON had no power with $n = 5$ (See Tables 5-7). With group sample size $n = 10$ (e.g., see Table 8), MEDR was often most powerful for heavier-tailed distributions (exponential, double exponential), especially with larger location differences and more groups (e.g., 3 groups, $n = 10$, $\Delta_1 = \Delta_2 = 2, \Delta_3 = 5$; 5 groups, $n = 10,\ \Delta_1 = \Delta_2 = 2,\ \Delta_3 = \Delta_4 = \Delta_5 = 0$) while MOODR was most powerful for the latter five group scenarios for contaminated normal data. MEDBON often had higher power than MOODR, but always trailed MEDR. For $n = 20$, MEDBON was most powerful for uniform and exponential data, and all three median-based procedures had similar power for the other distributions (See Table 9). MEDR performed most consistently across different scenarios, was never much less powerful than any other procedure for nonnormal data, and was often substantially more powerful. For example, in Table 11, MEDR had power almost 200 times the power of MOODR (0.591 versus 0.003), while the largest power advantage for

Table 2. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = \Delta_3 = 0$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.053 | 0.046 | 0.047 | 0.037 | 0.027 |
| MEDUR | 0.035 | 0.041 | 0.054 | 0.040 | 0.019 |
| MOODR | 0.013 | 0.018 | 0.017 | 0.019 | 0.007 |
| MOODUR | 0.009 | 0.013 | 0.011 | 0.013 | 0.003 |
| TUKEY | 0.053 | 0.059 | 0.060 | 0.044 | 0.026 |

Table 3. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, five groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.000 | 0.009 | 0.009 | 0.014 | 0.000 |
| MEDUR | 0.000 | 0.023 | 0.017 | 0.021 | 0.001 |
| MOODR | 0.001 | 0.008 | 0.005 | 0.003 | 0.001 |
| MOODUR | 0.001 | 0.008 | 0.005 | 0.003 | 0.001 |
| TUKEY | 0.024 | 0.025 | 0.025 | 0.023 | 0.025 |

Table 4. FWE – Proportion of times at least one true null hypothesis was rejected at $\alpha = 0.05$, five groups, $n_1 = 3, n_2 = 4, n_3 = 5, n_4 = 6, n_5 = 7$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.001 | 0.005 | 0.008 | 0.006 | 0.003 |
| MEDUR | 0.003 | 0.013 | 0.025 | 0.014 | 0.026 |
| MOODR | 0.001 | 0.005 | 0.007 | 0.001 | 0.002 |
| MOODUR | 0.001 | 0.005 | 0.007 | 0.001 | 0.002 |
| TUKEY | 0.000 | 0.000 | 0.000 | 0.001 | 0.001 |

Table 5. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.579 | 0.269 | 0.098 | 0.151 | 0.336 |
| MEDUR | 0.487 | 0.256 | 0.095 | 0.113 | 0.297 |
| MEDBON | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| MOODR | 0.238 | 0.064 | 0.049 | 0.080 | 0.133 |
| MOODUR | 0.131 | 0.045 | 0.039 | 0.055 | 0.070 |
| TUKEY | 0.818 | 0.342 | 0.125 | 0.186 | 0.478 |

Table 6. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 5$, locations $\Delta_1 = 0, \Delta_2 = 2, \Delta_3 = 5$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | D-Exp | Exponential | Cont-Normal |
| Procedure | | | | | |
| MEDR | 0.786 | 0.707 | 0.262 | 0.410 | 0.455 |
| MEDUR | 0.976 | 0.716 | 0.220 | 0.422 | 0.581 |
| MEDBON | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| MOODR | 0.888 | 0.469 | 0.156 | 0.302 | 0.537 |
| MOODUR | 0.820 | 0.377 | 0.127 | 0.248 | 0.499 |
| TUKEY | 1.000 | 0.979 | 0.350 | 0.620 | 0.590 |

Table 7. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, five groups, $n_i = 5$, locations $\Delta_1 = \Delta_2 = 2; \Delta_3 = \Delta_4 = \Delta_5 = 0$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.637 | 0.369 | 0.059 | 0.137 | 0.396 |
| MEDUR | 0.400 | 0.293 | 0.078 | 0.104 | 0.245 |
| MEDBON | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| MOODR | 0.477 | 0.112 | 0.096 | 0.135 | 0.303 |
| MOODUR | 0.477 | 0.112 | 0.096 | 0.135 | 0.303 |
| TUKEY | 0.886 | 0.422 | 0.000 | 0.186 | 0.540 |

Table 8. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 10$, locations $\Delta_1 = 0, \Delta_2 = 2, \Delta_3 = 5$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 1.000 | 0.996 | 0.661 | 0.949 | 0.923 |
| MEDUR | 1.000 | 0.990 | 0.635 | 0.904 | 0.911 |
| MEDBON | 1.000 | 1.000 | 0.574 | 0.947 | 0.854 |
| MOODR | 0.888 | 0.469 | 0.156 | 0.302 | 0.537 |
| MOODUR | 0.820 | 0.377 | 0.127 | 0.248 | 0.499 |
| TUKEY | 1.000 | 1.000 | 0.627 | 0.890 | 0.940 |

Table 9. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_i = 20$, locations $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double-Exp. | Exponential | Cont-Normal |
| Procedure | | | | | |
| MEDR | 1.000 | 0.664 | 0.374 | 0.664 | 0.991 |
| MEDUR | 1.000 | 0.676 | 0.361 | 0.676 | 0.979 |
| MEDBON | 1.000 | 0.776 | 0.342 | 0.776 | 0.983 |
| MOODR | 0.998 | 0.569 | 0.384 | 0.648 | 0.996 |
| MOODUR | 0.997 | 0.529 | 0.352 | 0.614 | 0.992 |
| TUKEY | 1.000 | 0.550 | 0.278 | 0.550 | 0.436 |

Table 10. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_1 = 4, n_2 = 5, n_3 = 6$, locations $\Delta_1 = \Delta_3 = 0, \Delta_2 = 2$.

| | Distribution | | | | |
|---|---|---|---|---|---|
| | Normal | Uniform | Double–Exp. | Exponential | Cont.-Normal |
| Procedure | | | | | |
| MEDR | 0.607 | 0.260 | 0.090 | 0.129 | 0.287 |
| MEDUR | 0.558 | 0.262 | 0.093 | 0.121 | 0.264 |
| MEDBON | 0.332 | 0.108 | 0.047 | 0.100 | 0.203 |
| MOODR | 0.147 | 0.041 | 0.060 | 0.070 | 0.125 |
| MOODUR | 0.147 | 0.041 | 0.060 | 0.070 | 0.125 |
| TUKEY | 0.220 | 0.035 | 0.005 | 0.012 | 0.051 |

Table 11. Power – Proportion of times at least one pairwise difference detected at $\alpha = 0.05$, three groups, $n_1 = 4, n_2 = 5, n_3 = 6$, normally distributed data.

| | Location pattern | | |
|---|---|---|---|
| | $\Delta_1 = 2, \Delta_2 = \Delta_3 = 0$ | $\Delta_1 = \Delta_3 = 0, \Delta_2 = 2$ | $\Delta_1 = \Delta_2 = 0, \Delta_3 = 2$ |
| Procedure | | | |
| MEDR | 0.591 | 0.607 | 0.711 |
| MEDUR | 0.656 | 0.558 | 0.478 |
| MEDBON | 0.302 | 0.332 | 0.458 |
| MOODR | 0.003 | 0.147 | 0.654 |
| MOODUR | 0.003 | 0.147 | 0.654 |
| TUKEY | 0.219 | 0.220 | 0.228 |

Table 12. Power – Proportion of times at least one difference detected at $\alpha = 0.05$, five groups, $n_1 = 3, n_2 = 4, n_3 = 5, n_4 = 6, n_5 = 7$, normally distributed data.

| | Location pattern | | | |
|---|---|---|---|---|
| | $\Delta_1 = \Delta_2 = 2;$ $\Delta_3 = \Delta_4 = \Delta_5 = 0$ | $\Delta_1 = 0; \Delta_2 = \Delta_3 = 2;$ $\Delta_4 = \Delta_5 = 0$ | $\Delta_1 = \Delta_2 = 0;$ $\Delta_3 = \Delta_4 = 2; \Delta_5 = 0$ | $\Delta_1 = \Delta_2 = \Delta_3 = 0;$ $\Delta_4 = \Delta_5 = 2$ |
| Procedure | | | | |
| MEDR | 0.546 | 0.451 | 0.556 | 0.702 |
| MEDUR | 0.516 | 0.372 | 0.322 | 0.298 |
| MEDBON | 0.003 | 0.000 | 0.041 | 0.002 |
| MOODR | 0.001 | 0.001 | 0.416 | 0.832 |
| MOODUR | 0.001 | 0.001 | 0.430 | 0.831 |
| TUKEY | 0.000 | 0.032 | 0.025 | 0.024 |

MOODR was less than 1.2 times that of MEDR, 0.537 versus 0.455 See Table 6). Table 8 shows, however, that when the sample size increased from $n = 5$ to $n = 10$, MOODR no longer had a power advantage over MEDR (in fact had substantially less power) for the same location pattern as in Table 6.

When sample sizes were unequal and group locations were different, the power of all tests depended on the pattern of location parameters. MOODR was by far the most affected by the pattern of differences, with virtually no power in the most extreme case (smallest samples with nonzero location parameters and largest with zero location parameters), while sometimes having the highest power with the situation reversed. In contrast, MEDR maintained respectable power for all location patterns (See Tables 11 and 12). MEDBON displayed low power when sample sizes were small, especially with five groups (10 comparisons). Power was higher with larger sample sizes, but still generally trailed the other two procedures. Many other scenarios were examined. These results are available at www.uncg.edu/~sjricht2/Research.html.

Table 13. *P*-values for pairwise comparisons.

| Comparison | Median difference | Procedure | | | | |
|---|---|---|---|---|---|---|
| | | MEDR | MOODR | MEDUR | MOODUR | TUKEY |
| 1vs2 | 85.0 | 0.950 | 1.000 | 0.794 | 0.974 | 0.985 |
| 1vs3 | 55.5 | 0.996 | 0.566 | 0.834 | 0.534 | 0.605 |
| 2vs3 | 140.5 | 0.691 | 0.295 | 0.645 | 0.345 | 0.372 |

Table 14. Average times to complete an interview for four interviewers.

| Interviewer | Average time (min.) | Median | Mean |
|---|---|---|---|
| 1 | 10.0, 25.0, 40.1, 29.2, 4.1 | 25.0 | 21.6 |
| 2 | 15.0, 5.2, 55.3, 15.1, 23.2 | 15.1 | 22.8 |
| 3 | 19.1, 25.4, 8.3 | 19.1 | 17.6 |
| 4 | 5.1, 9.2, 14.1 | 9.2 | 9.5 |

Table 15. *P*-values for pairwise comparisons.

| Comparison | Median difference | Procedure | | | | |
|---|---|---|---|---|---|---|
| | | MEDR | MOODR | MEDUR | MOODUR | TUKEY |
| 1vs2 | 9.9 | 0.851 | 1.000 | 0.920 | 1.000 | 0.999 |
| 1vs3 | 5.9 | 1.000 | 1.000 | 0.978 | 0.915 | 0.980 |
| 1vs4 | 15.8 | 0.211 | 0.450 | 0.525 | 0.362 | 0.666 |
| 2vs3 | 4.0 | 1.000 | 1.000 | 1.000 | 0.915 | 0.961 |
| 2vs4 | 5.9 | 1.000 | 0.450 | 0.978 | 0.362 | 0.607 |
| 3vs4 | 9.9 | 0.851 | 0.824 | 0.920 | 0.915 | 0.900 |

## Power Advantages of Median-based Procedures

The power of the median-based procedures was compared to that of the Tukey-Kramer procedure using means. For normally distributed data and equal sample sizes, TUKEY always had higher power than the median-based procedures (See Tables 4-6). However, with unequal sample sizes, the median based procedures often had higher power even for normally distributed data (See Tables 10, 11 and 12). This may not be surprising, since the Tukey-Kramer procedure has been shown to be conservative for unequal sample sizes (Hayter, 1984). For nonnormally distributed data, the median-based procedures often had higher power, especially with larger sample sizes.

## Conclusion

The maximum median difference test (MEDR) is recommended as a robust pairwise comparison procedure when strong control of FWE is desired. The maximum Mood difference test (MOODR) is not recommended, due to poor power properties, especially for unequal sample sizes. Likewise, the procedure of using separate median difference tests with a Bonferroni adjustment (MEDBON) generally had less power and no power in some cases with small sample sizes. Tukey's HSD (TUKEY) is preferred when groups have small and equal samples sizes ($n = 5$), even for nonnormal data, and also with normal data, regardless of the sample size. In all other cases, the maximum median difference test (MEDR) is preferred. With nonnormal data and large ($n \geq 20$) equal

sample sizes, and in all cases with unequal sample sizes, MEDR had higher power than TUKEY. MEDR never performed poorly with regard to power, and was often much more powerful than the other median-based procedures considered.

Example 1

The first example is based on the data in the Introduction (See Table 1.) Table 13 gives *p*-values for the three pairwise comparisons, for the MEDR, MEDUR, MOODR, MOODUR and TUKEY procedures. Notice that the Mood tests yield the most evidence for a difference between months two and three. This is an example of a scenario studied in the simulations, namely small samples with differences between all pairs, with larger differences associated with the larger samples, a case where the Mood tests often had the highest power.

Example 2:

Consider data reported by Gibbons (1985, p. 202) in Table 14. The data represent average times spent to complete an interview for four interviewers.

It is desired to test if there is evidence that certain interviewers tend to have longer interview times. Table 15 gives *p*-values for the six pairwise comparisons. Here MEDR provides the strongest evidence of location difference between the pair with the largest observed difference, interviewers 1 and 4. Resampling Stats code for calculating the permutation *p*-values in this example is provided in the Appendix.

## References


Gibbons, J. D. (1985). *Nonparametric Methods for Quantitative Analysis, 2nd edition.* Columbus, OH: American Sciences Press, Inc.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics*, *12*, 61–75.

Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA: Brooks/Cole.

Hochberg Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures.* New York: John Wiley and Sons.

Keller-McNulty, S. & Higgins, J. J. (1987). Effect of tail weight and outliers on power and type-I error of robust permutation tests for location. *Communications in Statistics: Simulation and Computation*, *16*(1):17-36.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, *12*, 309–310.

Levy (1979). Pairwise Comparisons associated with the k independent sample median test. *The American Statistician*, *33*(3), 138-139.

Linton, L. R., Edgington, E. S. & Davies, R. W. (1989). A view of niche overlap amenable to statistical analysis. *Canadian Journal of Zoology*, *67*, 55-60.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* (2nd ed.). London: Chapman & Hall.

Miller, R. G. (1966). *Simultaneous statistical inference*. New York: McGraw-Hill.

Miller, R. G. (1981). *Simultaneous statistical inference.* (2nd ed.). New York: Springer-Verlag.

Mood, A. M. (1950). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Nemenyi, P. (1963). *Distribution-free multiple comparisons*. Unpublished doctoral dissertation, Princeton University, Princeton, NJ.

Petrondas, D. A. & Gabriel, K. R. (1983). Multiple comparisons by rerandomization tests. *Journal of the American Statistical Association*, *78*, 949-957.

Powell, G. L. & Russell, A. P. (1984). The diet of the eastern short-horned lizard (*Phrynosoma douglassi breviroste*) in Alberta and its relationship to sexual size dimorphism. *Canadian Journal of Zoology*, *62*, 428-440.

Powell, G. L. & Russell, A. P. (1985). Growth and sexual size dimorphism in Alberta populations of the eastern short-horned lizard, *Phrynosoma douglassi breviroste*. *Canadian Journal of Zoology*, *63*, 139-154.

Resampling Stats (2000). *Resampling Stats Inc*., Arlington, Virginia.

Ryan, T. A. & Ryan, T. A., Jr. (1980). *K* Independent sample median test. A letter to the Editor, *The American Statistician*, *34*, 123.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561-584.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*, 99–114.

## Appendix

Below is Resampling Stats® code to calculate the permutation *p*-values in Example 2. The program can be modified to handle different numbers of groups.

```
'set maximum vector size
        maxsize default 500000
        seed 1234

'create data vectors
        data (10 25 40.1 29.2 4.1) d1
        data (15 5.2 55.3 15.1 23.2) d2
        data (19.1 25.4 8.3) d3
        data (5.1 9.2 14.1) d4

'combine data vectors for unrestricted randomization
        concat d1 d2 d3 d4 dat

'create pairwise data vectors for restricted randomization
        concat d1 d2 dat12
        concat d1 d3 dat13
        concat d1 d4 dat14
        concat d2 d3 dat23
        concat d2 d4 dat24
        concat d3 d4 dat34

'obtain permutation distribution
        let nrand=2000
        repeat nrand

'unrestricted randomization
        shuffle dat sdat
        take sdat 1,5 sdat1
        take sdat 6,10 sdat2
        take sdat 11,13 sdat3
        take sdat 14,16 sdat4

'restricted randomization
        shuffle dat12 sdat12
        take sdat12 1,5 sdat121
        take sdat12 6,10 sdat122
        shuffle dat13 sdat13
        take sdat13 1,5 sdat131
        take sdat13 6,8 sdat133
        shuffle dat14 sdat14
        take sdat14 1,5 sdat141
        take sdat14 6,8 sdat144
        shuffle dat23 sdat23
        take sdat23 1,5 sdat232
        take sdat23 6,8 sdat233
        shuffle dat24 sdat24
        take sdat24 1,5 sdat242
        take sdat24 6,8 sdat244
        shuffle dat34 sdat34
        take sdat34 1,3 sdat343
        take sdat34 4,6 sdat344

'compute medians of shuffled data
        median sdat1 med1
        median sdat2 med2
        median sdat3 med3
        median sdat4 med4
        median sdat121 med121
        median sdat122 med122
        median sdat131 med131
        median sdat133 med133
        median sdat141 med141
        median sdat144 med144
        median sdat232 med232
        median sdat233 med233
        median sdat242 med242
        median sdat244 med244
        median sdat343 med343
        median sdat344 med344

'compute median differences of shuffled data, unrestricted randomization
        subtract med1 med2 med12
        subtract med1 med3 med13
        subtract med1 med4 med14
        subtract med2 med3 med23
        subtract med2 med4 med24
        subtract med3 med4 med34

'create one vector, take absolute values
        concat med12 med13 med14 med23 med24 med34
```

```
        medvec abs medvec medvec

'compute median differences of shuffled data,
restricted randomization
        subtract med121 med122 med12r
        subtract med131 med133 med13r
        subtract med141 med144 med14r
        subtract med232 med233 med23r
        subtract med242 med244 med24r
        subtract med343 med344 med34r

'create one vector, take absolute value
        concat med12r med13r med23r medvecr
        abs medvecr medvecr

'compute maximum absolute difference
        max medvec qmedsim
        max medvecr qmedsimr

'compute Mood statistics, unrestricted
randomization
        median sdat grndmed
        count sdat1 >= grndmed m1
        count sdat2 >= grndmed m2
        count sdat3 >= grndmed m3
        count sdat4 >= grndmed m4
        median sdat12 gm12
        count sdat1 >= gm12 m121
        count sdat2 >= gm12 m122
        median sdat13 gm13
        count sdat1 >= gm13 m131
        count sdat3 >= gm13 m133
        median sdat14 gm14
        count sdat1 >= gm14 m141
        count sdat4 >= gm14 m144
        median sdat23 gm23
        count sdat2 >= gm23 m232
        count sdat3 >= gm23 m233
        median sdat24 gm24
        count sdat2 >= gm24 m242
        count sdat4 >= gm24 m244
        median sdat34 gm34
        count sdat3 >= gm34 m343
        count sdat4 >= gm34 m344
        subtract m1 m2 m12
        subtract m1 m3 m13
        subtract m1 m4 m14
        subtract m2 m3 m23
        subtract m2 m4 m24
        subtract m3 m4 m34
```

```
'Mood statistics are m12-m34

'create one vector, take absolute values
        concat m12 m13 m14 m23 m24 m34
mood
        abs mood mood

'compute maximum absolute difference
        max mood maxmood

'Compute Mood statistics, restricted
randomization
        subtract m121 m122 m12r
        subtract m131 m133 m13r
        subtract m141 m144 m14r
        subtract m232 m233 m23r
        subtract m242 m244 m24r
        subtract m343 m344 m34r

'Mood statistics are m12r-m34r

'create one vector, take absolute values
        concat m12r m13r m14r m23r m24r
m34r
        moodr abs moodr moodr

'compute maximum absolute difference
        max moodr maxmoodr

'save statistic values for reference distributions
        score qmedsim qmddist
        score qmedsimr qmddistr
        score maxmood qmood
        score maxmoodr qmoodr
    end

'compute medians and differences of observed
data
        median d1 obsmed1
        median d2 obsmed2
        median d3 obsmed3
        median d4 obsmed4

        subtract obsmed1 obsmed2 mddiff12
        abs mddiff12 mddiff12
        subtract obsmed1 obsmed3 mddiff13
        abs mddiff13 mddiff13
        subtract obsmed1 obsmed4 mddiff14
        abs mddiff14 mddiff14
        subtract obsmed2 obsmed3 mddiff23
        abs mddiff23 mddiff23
```

```
        subtract obsmed2 obsmed4 mddiff24
        abs mddiff24 mddiff24
        subtract obsmed3 obsmed4 mddiff34
        abs mddiff34 mddiff34

'compute Mood statistic for observed data
        median dat grndmed
        count d1 >= grndmed obsm1
        count d2 >= grndmed obsm2
        count d3 >= grndmed obsm3
        count d4 >= grndmed obsm4
        subtract obsm1 obsm2 obsm12
        abs obsm12 obsm12
        subtract obsm1 obsm3 obsm13
        abs obsm13 obsm13
        subtract obsm1 obsm4 obsm14
        abs obsm14 obsm14
        subtract obsm2 obsm3 obsm23
        abs obsm23 obsm23
        subtract obsm2 obsm4 obsm24
        abs obsm24 obsm24
        subtract obsm3 obsm4 obsm34
        abs obsm34 obsm34

'compute p-values
**************************************
'MEDUR
        count qmddist >=  mddiff12 mdsg12q
        divide mdsg12q nrand medur12
        count qmddist >=  mddiff13 mdsg13q
        divide mdsg13q nrand medur13
        count qmddist >=  mddiff14 mdsg14q
        divide mdsg14q nrand medur14
        count qmddist >=  mddiff23 mdsg23q
        divide mdsg23q nrand medur23
        count qmddist >=  mddiff24 mdsg24q
        divide mdsg24q nrand medur24
        count qmddist >=  mddiff34 mdsg34q
        divide mdsg34q nrand medur34
'MEDR
        count qmddistr >= mddiff12 mdsg12qr
        divide mdsg12qr nrand medr12
        count qmddistr >= mddiff13 mdsg13qr
        divide mdsg13qr nrand medr13
        count qmddistr >= mddiff14 mdsg14qr
        divide mdsg14qr nrand medr14
        count qmddistr >= mddiff23 mdsg23qr
        divide mdsg23qr nrand medr23
        count qmddistr >= mddiff24 mdsg24qr
        divide mdsg24qr nrand medr24
        count qmddistr >= mddiff34 mdsg34qr
        divide mdsg34qr nrand medr34
'MOODUR
        count qmood >= obsm12 mood12q
        divide mood12q nrand moodur12
        count qmood >= obsm13 mood13q
        divide mood13q nrand moodur13
        count qmood >= obsm14 mood14q
        divide mood14q nrand moodur14
        count qmood >= obsm23 mood23q
        divide mood23q nrand moodur23
        count qmood >= obsm24 mood24q
        divide mood24q nrand moodur24
        count qmood >= obsm34 mood34q
        divide mood34q nrand moodur34
'MOODR
        count qmoodr >= obsm12 mood12qr
        divide mood12qr nrand moodr12
        count qmoodr >= obsm13 mood13qr
        divide mood13qr nrand moodr13
        count qmoodr >= obsm14 mood14qr
        divide mood14qr nrand moodr14
        count qmoodr >= obsm23 mood23qr
        divide mood23qr nrand moodr23
        count qmoodr >= obsm24 mood24qr
        divide mood24qr nrand moodr24
        count qmoodr >= obsm34 mood34qr
        divide mood34qr nrand moodr34

**************************************
'print output here
print medur12  medur13  medur14  medur23
medur24  medur34
print medr12  medr13  medr14  medr23
medr24  medr34
print moodur12 moodur13 moodur14 moodur23
moodur24 moodur34
print moodr12  moodr13  moodr14  moodr23
moodr24  moodr34
```

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 413- 420
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# The Non-Parametric Difference Score: A Workable Solution for Analyzing Two-Wave Change When The Measures Themselves Change Across Waves

Jennifer E. V. Lloyd Bruno D. Zumbo
University of British Columbia

The non-parametric difference score is introduced. It is a workable solution to the problem of analyzing change over two waves (i.e., a pretest-posttest design) when the measures themselves vary over time. An example highlighting the solution's implementation is provided, as is a discussion of the solution's assumptions, strengths, and limitations.

Key words: Non-parametric, difference score, two-wave, change, quantitative analysis.

## Introduction

Individual change is the subject of significant attention in education, health, and the social sciences. The analysis of such change is aimed at quantifying the amount by which individuals grow, mature, improve, and progress over time. By measuring and tracking changes, it is possible to reveal the temporal nature of development (Singer & Willett, 2003).

This temporal nature of development may be studied over varied spans of time: hours, days, weeks, months, and even years. Waves are the measurement occasions or periods of data collection that are plan-fully interspersed throughout these spans of time. Two-wave designs, often known as pretest-posttest designs, are the specific focus of this article. Such designs allow for relatively straightforward appraisal of a treatment effect by detecting differences in a given outcome across two waves – typically before the treatment and after it. Such differences normally represent the comparison of test-takers' scores at the second wave of data collection to their respective baseline or initial measure scores (Zumbo, 1999). Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) explore the problem of analyzing change and growth when the measures themselves change across multiple (i.e., three or more) waves.

Jennifer E.V. Lloyd is a Research Scientist and Post-Doctoral Fellow for the Human Early Learning Partnership at the University of British Columbia (UBC). Email her at jennifer.lloyd@ubc.ca. Bruno D. Zumbo is Professor, Measurement, Evaluation, and Research Methodology (MERM) Program, and Department of Statistics and the Institute of Applied Mathematics at the University of British Columbia. Email him at bruno.zumbo@ubc.ca. The authors thank the Social Sciences and Humanities Research Council of Canada for funding this research project, and the British Columbia Ministry of Education for approving the use of their data. Thanks are also extended to Dr. Anita Hubley, Dr. Kimberly Schonert-Reichl, and Alex Mann for their comments on an earlier draft of this paper.

## Repeated Measures Analyses: Three Research Scenarios

Several familiar parametric methodologies, called repeated measures analyses, centre upon quantifying change over time. As described by Lloyd (2006) and Lloyd, Zumbo, and Siegel (2006), these methodologies are generally used in three research scenarios:

### Scenario 1: Exact same measure across both waves

In this scenario, one's construct of choice makes possible the use and re-use of the

exact same measure across both waves, regardless of the ever-emergent age, cognitive development, and personal and scholarly experiences of one's test-takers. The measures' content, item wording, response categories, and response formats do not change whatsoever across waves.

Scenario 2: Linkable time-variable measures

Time-variable measures are those whose content, wording, response categories, and/or response formats vary across waves in repeated measures designs. In this scenario, although the time-variable measures are not completely identical across waves, there is at least one anchor item shared by each of the measures, on whose linked (or equated) scores traditional analyses can be performed (Kolen & Brennan, 2004).

Scenario 3: Non-linkable time-variable measures

This scenario involves using measures whose content, item wording, response categories, and/or response formats vary completely across waves. Imagine, for example, a reading achievement test administered at Grade 5 and then Grade 6: The measure administered at Grade 5 cannot be same as that used in Grade 6. If they were the same, the reliability and validity of the test scores would likely be compromised, rendering the study ineffectual (Singer & Willett, 2003). This scenario may also be encountered when one's sample size is small or when one cannot compare the sample's scores to those of a norming group. In such cases, even if the measures share common items, it is not always advisable to link or equate the measures' scores.

Objective

Repeated measures analyses are often characterized by one set of individuals being measured more than once on the same or commensurable dependent variable. Many researchers understand the phrase "same or commensurable dependent variable" to mean that the exact same measure must be used across all waves study.

As Scenario 1 (exact same measure across both waves) illustrates, some constructs can in fact be measured using the exact same measure over time. As Scenario 2 (linkable time-variable measures) and particularly Scenario 3 (non-linkable time-variable measures) describe, however, there are often situations in which one's construct of choice makes the use and re-use of the exact same measure across waves difficult – and even impossible. Seeing as traditional linking/equating techniques are not possible when the measures cannot be made to be identical (Kolen & Brennan, 2004), what is a researcher to do, then, if the use of time-variable measures is necessary?

Therefore, this article focuses on the analysis of two-wave change with linkable – and particularly non-linkable – time-variable measures. Many of the current strategies used to handle time-variable measures (such as vertical scaling and item response theory techniques; see Kolen & Brennan, 2004) are often only useful to large testing organizations that have access to very large numbers of test-takers and expansive item pools, or in those situations in which the time-variable measures share some number of common items. Therefore, the objective of this article is to introduce a workable solution to the problem of analyzing change with time-variable measures administered over two waves – a solution that can be implemented easily in everyday research settings.

The Non-Parametric Difference Score (NPAR-DIFF)

The NPAR-DIFF involves rank transforming or ordering individuals' original test scores within wave, and then using the change (difference) score computed from the respective ranks as the dependent variable in subsequent parametric independent sample *t*-tests. It is this use of ranks, instead of original scores, that makes the NPAR-DIFF a non-parametric solution.

Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) refer to the general approach of converting original scores into ranks pre-analysis as the Conover solution, in recognition of the influential work of W. J. Conover (e.g., Conover, 1999; Conover & Iman, 1981), whose research not only inspired the NPAR-DIFF, but also provides evidence for the solution's viability.

A rank represents the position of a test-taker on a variable relative to the positions held by all other test-takers on that same variable. Ranking or rank transforming refers to the process of transforming a test-taker's original score to rank relative to other test-takers – suggesting a one-to-one function $f$ from the sample values [e.g.,$\{X_1, X_2, ..., X_N\}$] to the first $N$ positive integers [e.g., $\{1, 2,\ldots, N\}$], (Zimmerman & Zumbo, 1993).

For example, if Test-taker $X$ earned a score of 20 on a given variable, Test-taker $Y$ earned a score of 21, and Test-taker $Z$ earned a score of 22, then the test-takers' respective ranks would be 1, 2, and 3 (where a rank of 1 is given to the test-taker with the lowest score). One may also assign ranks such that the test-taker with the highest score receives a rank of 1; however, it is often easier to think of test-takers receiving the highest score as also receiving the highest rank value.

The NPAR-DIFF's Assumptions

As with all methodological tools, the NPAR-DIFF comes with its own set of assumptions. First, the scales for the measures' original scores must be at least ordinal in nature. Second, the ranks must show heterogeneous change, meaning that all test-takers do not change the same amount across waves (Zumbo, 1999). Imagine that Test-Taker $X$ earns a rank score = 1 across both waves and Test-Taker $Y$ earns a rank score = 2 across both waves. For both test-takers, the change scores computed from the rank equal zero, suggesting homogeneous change – which, for reasons outlined by Zumbo (1999), cannot be used in change analyses. It should be noted that an inability to handle homogeneous change is not a problem endemic to the NPAR-DIFF; homogeneous change also renders ineffectual the calculation of simple difference scores.

Finally, the NPAR-DIFF requires that a commensurable (or comparable or similar) construct is measured across all waves of the study. Commensurability is generally thought to mean that the same primary dimension or latent variable is driving the test-takers' responses at each wave. A latent variable is an unobserved variable that accounts for the correlation among one's observed or manifest variables. In ideal circumstances, measures are designed such that the latent variable that drives test-takers' responses represents the construct of interest.

Example

Suppose a researcher is interested in exploring whether there are gender differences in test-takers' rank-based numeracy assessment difference scores (scores that represent the comparison of test-takers' scores at the second wave of data collection to their respective baseline or initial measure scores). Note that the research question changes slightly when one applies the NPAR-DIFF: No longer are the inferences made from the original scores; rather they are made from the ranks.

To illustrate the implementation of the NPAR-DIFF, Foundation Skills Assessment (FSA) numeracy subtest data from the British Columbia Ministry of Education were obtained. The FSA, an annual assessment administered by the Ministry, is designed to measure the reading comprehension, writing, and numeracy skills of 4th- and 7th-grade students throughout British Columbia. The FSA is administered in public and funded independent schools across the province in late April/early May of each year. Approximately 40,000 students per grade level write the FSA each year.

Obtained was the entire population of standardized numeracy subtest scores of 41,675 test-takers who wrote the FSA in both 1999/2000 (Wave 1, Grade 4) and 2002/2003 (Wave 2, Grade 7). Test-takers who were missing a wave of FSA data were excluded from analyses. Of this population of test-takers, a random 10% convenience sample of 4097 test-takers ($n_{\text{female}} = 2055$; $n_{\text{male}} = 2042$) was retained for analyses. Each test-taker's record included an arbitrary case number, and a gender flag. The Ministry has standardized test-takers' FSA scores such that each wave's score distribution has $M = 0$ and $SD = 1$.

Willett, Singer, and Martin (1998) state that standardized test scores should never be used in the place of raw scores in individual growth modeling analyses (readers are referred to their article for the specific reasons why). In this case, however, ranks are being used in the

Table 1

*Descriptive Statistics for Each of the Two Waves of FSA Original Scores (N = 4097)*

| *Gender* | *Original Variable Name* | *Min* | *Max* | *M* | *SD* | *Skew* | *Kurtosis* |
|---|---|---|---|---|---|---|---|
| Female | *grade4original* | -4.83 | 4.66 | -.26 | 1.10 | -.88 | 4.30 |
| (*n* = 2055) | *grade7original* | -2.08 | 2.85 | .06 | .90 | .31 | -.30 |
| Male | *grade4original* | -4.83 | 5.33 | -.16 | 1.11 | -.78 | 4.21 |
| (*n* = 2042) | *grade7original* | -2.58 | 2.85 | .11 | .92 | .31 | -.25 |

place of the original test scores. Thus, it is unimportant whether or not the original test scores come in the form of standardized scores. Furthermore, the Ministry of Education does not supply researchers with raw FSA scores – only standardized scores.

As Table 1 illustrates, the descriptive statistics for each wave of FSA original scores vary across gender and wave. When performing the NPAR-DIFF, data must be entered into the data matrix (spreadsheet) in person-level format, in which one row represents one individual, with time-related variables represented along the horizontal of the spreadsheet (as in Table 2). The key to implementing the NPAR-DIFF is that one first rank transforms the data within wave, with the mean rank being assigned to ties. Table 2 illustrates that Test-Taker *X*, for example, earns a Rank = 2 for Wave 1 (Grade 4) because his original Wave 1 score (0.20) is between those of Test-Taker *Y* (-.15, Rank = 1) and Test-Taker *Z* (1.45, Rank = 3). Recall from an earlier section that a Rank = 1 is assigned to the test-taker with the lowest within-wave score.

Two-Wave Designs: Two Common Change Scores

As described earlier, two-wave designs are characterized by some comparison of an individual's score at the second wave of data scores involved in two-wave designs are:

(a) the simple difference score and
(b) the residualized change score (Zumbo, 1999).

Simple difference score

The most common of all change indices is the simple difference score, which is calculated by simply subtracting a test-taker's score at Wave 1 from his or her score at Wave 2. A positive simple difference score typically indicates an increase over time, whereas a negative score indicates a decrease over time.

Residualized change score

As Zumbo (1999) describes more fully, it has been argued that simple difference scores are unfair because of their base-dependence (i.e., scores at Wave 2 are correlated negatively with scores at Wave 1). As such, the residualized change score was developed as an alternative to the simple difference score. Although there are different ways to create such scores, the most common residualized change score is estimated from the regression analysis of the Wave 2 score on the Wave 1 score. In other words, the estimated Wave 2 score is subtracted from the actual Wave 2 score (whether it be an original or rank).

Table 2. An example person-level data matrix showing two waves of hypothetical original FSA scores and their corresponding within-wave rank scores.

| | Example Original Variables | | Corresponding Rank Variables | |
|---|---|---|---|---|
| | *grade4original* | *grade7original* | *grade4rank* | *grade7rank* |
| Test-Taker *X* | 0.20 | 0.45 | 2 | 1.5 |
| Test-Taker *Y* | -.15 | 1.35 | 1 | 3 |
| Test-Taker *Z* | 1.45 | 0.45 | 3 | 1.5 |

The intrinsic fairness, usefulness, reliability, and validity of the two-wave research design have been debated for decades (Zumbo, 1999). In their seminal article, Cronbach and Furby (1970) disparage the use of two-wave designs, arguing that change scores are rarely useful, no matter how they are adjusted or refined (Cronbach & Furby, 1970). Their disdain of two-wave designs was so strong that they stated that researchers who ask questions using simple difference scores are better advised to frame their questions in other ways (Cronbach & Furby, 1970). As Zumbo (1999) notes, it is somewhat puzzling that there exists the notion that one should avoid two-wave designs at "all costs", given that variations of the difference score lie at the heart of various widely-used and commonly-accepted statistical tests, such as the paired samples t-test.

Determining the Appropriate Change Score to Serve as the Dependent Variable

In order to determine which specific change score should serve as the dependent variable in this particular FSA example, it is necessary to follow the guidelines of Zumbo (1999), who writes that "one should utilize the simple difference score instead of the residualized difference if and only if $\rho(X_1, X2) > \sigma X_1 / \sigma X_2$" (p. 293) – that is, if the correlation between the Wave 1 and 2 scores is greater than the ratio of the respective standard deviations. It is important to stress that, when implementing the NPAR-DIFF solution for two-wave data, one's decision about using the simple difference and residualized change score must be based on test-takers' ranks– not their original scores.

The computed across-gender correlation between the Grade 4 and 7 ranks [$\rho(X_1,X_2)$] was computed as 0.66, compared to 0.99 (1182.84/1182.84) for the ratio of the two standard deviations of rank scores [$\sigma X_1 / \sigma X_2$]. Because the correlation value is less than the ratio value, the rank-based residualized change score is used in the place of the rank-based simple difference score as the dependent variable in the subsequent parametric analysis (Zumbo, 1999).

Explanation of the Statistical Output

A regular independent samples *t*-test was then performed on test-takers' rank-based residualized change scores, with gender identified as the predictor variable. It should be reiterated that the unique aspect of the analysis is that test-takers' rank-based change scores are used in the place of the change scores computed from test-takers' original scores. Original scores are, in a sense, only collected as a means of computing test-takers' ranks. The research question, results, and inferences made from the results must reflect the fact that the scores have

been transformed and, hence, the focus is no longer on the original scores.

The independent sample's *t*-test output revealed that the mean rank-based residualized change score for males was -7.64 (*SD* = 882.31) as opposed to 7.59 for females (*SD* = 876.71), meaning that the average Wave 2 rank less the rank at Wave 2 predicted from the Wave 1 rank score is higher for females than for males. This finding suggests that the female test-taker gained 7.5 points in relative standing across the two waves, whereas the average male test-taker's relative standing decreased approximately 7.6 points.

Despite the mean differences in residualized change scores for males and females, the independent samples *t*-test results showed that there is no statistically significant gender difference in the residualized change scores, $t(4095) = -.555$, $p = .579$ (assuming equal variances; two-tailed). Thus, the male test-takers' mean rank-based residualized change score did not differ significantly from that of the female test-takers – suggesting that neither gender's relative standing over time differ significantly from the other.

Even though there was no statistically-significant gender difference found, an effect size was still computed, for reasons outlined by Zumbo and Hubley (1998). A Cohen's *d* effect size was calculated by subtracting the mean residualized change score of one group (females) from that of the other group (males) and dividing that difference by the pooled rank-based standard deviation. The resultant effect size was computed as 0.02, which represents a small effect size (Cohen, 1988).

Strengths of the NPAR-DIFF

The non-parametric difference score, a solution for the problem of analyzing change and growth with time-variable measures collected over two waves is an effective tool for researchers in everyday research settings for the following reasons:

Ease of use

One strength of the NPAR-DIFF is that it is easy to implement. As Conover and Iman (1981) observe, it is often more convenient to use ranks in a parametric statistical program than it is to write a program for a non-parametric analysis. Furthermore, all of the steps required for the implementation of the NPAR-DIFF (i.e., rank transforming data within waves, conducting independent samples t-tests, etc.) can be easily performed using commonly-used statistical software packages.

Marries non-parametric and parametric methods:

Second, by rank transforming the data pre-analysis, parametric and non-parametric statistical methods are combined, providing "a vehicle for presenting both the parametric and nonparametric methods in a unified manner" (Conover & Iman, 1981, p. 128).

Makes use of the ordinal nature of data

Third, the NPAR-DIFF makes use of the ordinal nature of continuous-scored data: A test-taker with a low original score relative to other test-takers in his wave will also yield a low relative rank. Similarly, a test-taker with a high test-score will also yield a high rank. As a result, within-wave order among the test-takers is preserved.

Requires no common/linkable items

Unlike many of the traditional test linking methods and strategies, the NPAR-DIFF can be implemented not only in situations in which one's study involves time-variable measures that can be linked (Scenario 2), but also situations in which the time-variable measures share no linkable items whatsoever (Scenario 3). Hence, unlike vertical scaling, equating, and their linking counterparts, the NPAR-DIFF provides a means by which researchers can study change – whether or not the measures contain linkable items.

Requires no norming group

Due to time and financial constraints, it is not always possible to compare the scores of one's sample to those of an external norming sample. As such, an additional strength of the NPAR-DIFF is that it can be conducted using simply the scores of the sample of test-takers, thereby eliminating the need for a group to which to compare the sample's scores.

Limitations of the NPAR-DIFF

As with any methodological tool, the NPAR-DIFF has various limitations. Within-wave ranks are bounded. Rank transforming refers to the process of converting a test-taker's original score to rank relative to other test-takers. The values assigned by the function to each sample value in its domain are the number of sample values having lesser or equal magnitude. Consequently, the ranks are bounded from above by *N*. As a result, "any outliers among the original sample values are not represented by deviant values in the rank" (Zimmerman & Zumbo, 1993, p. 487).

Suppose on a standardized test of intelligence, Test-Taker *W* earns a score 100, Test-Taker *X* earns a score of 101, Test-Taker *Y* earns a score of 102, and Test-Taker Z earns a score of 167. Test-Taker *Z*'s score, relative to the other test-takers, is exceptional. Despite the exceptional performance on the measure, the test score is masked by the application of ranks: Test-taker $W = 1$, Test-taker $X = 2$, Test-taker $Y = 3$, and Test-taker $Z = 4$.

As a result, one limitation of the NPAR-DIFF is that there may be problems associated with the inherent restriction of range it places on data. Differences between any two ranks range between 1 and $N - 1$, whereas the differences between original sample values range between 0 and infinity (Zimmerman & Zumbo, 1993).

Difficulties associated with handling missing data

Recall that only those test-takers for whom data were available at both waves were retained in the analyses. As most educational, health, and social science researchers will agree, no discussion about change and growth is complete without a complementary discussion about one unavoidable problem: missing data. In longitudinal designs, particularly those that span months or years, it is extremely common to face problems associated with participant dropout, attrition, and as well as participants who join, or return to the study, in later waves.

One possible strategy for circumventing, or at least mitigating the effect of, missing data is to impute the missing original scores prior to rank-transforming the data within-wave pre-analysis. Schumacker and Lomax (2004) discuss various missing data imputation methods.

Makes use of the ordinal nature of data:

Recall that the fact that the NPAR-DIFF makes use of the ordinal nature of continuous-scored data was previously identified as one of the solution's strengths. As Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) observe, precisely what the NPAR-DIFF wins by, it also loses by: Because of the rank transformation of the original scores, differences between raw scores are not necessarily preserved by the corresponding ranks. For example, a difference between the raw scores corresponding to the 15th and the 16th ranks is not necessarily the same as the difference between the raw scores corresponding to the 61st and 62nd ranks in a collection of 500 test scores (Zimmerman & Zumbo, 2005, p. 618).

## Conclusion

Investigating the problem of analyzing change and growth with time-variable measures is important for two reasons. First, as Willett et al. (1998) and von Davier, Holland, and Thayer (2004) describe, the rules about which tests are permissible for repeated measures designs are precise and strict. Given these conditions, it is necessary to investigate how repeated measures analyses can be made possible – psychometrically and practically – when the measures themselves change across waves.

Second, given the sizeable growth in longitudinal large-scale testing in recent years, it is necessary to find a viable and coherent solution to the problem so that researchers can make the most accurate inferences possible about their test scores.

Recognizing the importance of this problem, this article introduced a workable solution for handling the analysis of change over two waves, when the measures used at each wave are not the same. Although useful in many research settings, the non-parametric difference score (NPAR-DIFF) is by no means a universal panacea and should, therefore, be used judiciously and in accordance with the aforementioned assumptions. Given that the problem of time-variable measures has, to date,

gone relatively unaddressed in the change/growth and test linking literatures, it is imperative that future research explores this profoundly important, problem to a much fuller degree.

## References


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

Conover, W. J. (1999). *Practical nonparametric statistics (3rd ed.)*. New York: John Wiley & Sons.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" - Or should we? *Psychological Bulletin, 74,* 68-80.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.)*. New York: Springer-Verlag.

Lloyd, J. E. V. (2006). *On modeling change and growth when the measures themselves change across waves: Methodological and measurement issues and a novel non-parametric solution.* Unpublished doctoral dissertation, University of British Columbia.

Lloyd, J. E. V., Zumbo, B. D., & Siegel, L. S. (2006). *The non-parametric HLM: A workable solution for analyzing change and growth when the measures themselves change across waves.* Manuscript submitted for publication.

Schumacker, R. E, & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling, 2nd edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating.* New York: Springer.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395-426.

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Volume 1: Methodological issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum.

Zimmerman, D. W., & Zumbo, B. D. (2005). Can percentiles replace raw scores in statistical analysis of test data? *Educational and Psychological Measurement, 65,* 616-638.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In Bruce Thompson (Ed.). *Advances in Social Science Methodology, Volume 5,* (pp. 269-304). Greenwich, CT: JAI Press.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *Journal of the Royal Statistical Society, Series D (The Statistician), 47,* 385-388.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 421- 442
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Probability Coverage and Interval Length for Welch's and Yuen's Techniques: Shift in Location, Change in Scale, and (Un)Equal Sizes

S. Jonathan Mends-cole
Walden University

Coverage for Welch's technique was less than the confidence-level when size was inversely proportional to variance and skewness was extreme. Under negative kurtosis, coverage for Yuen's technique was attenuated. Under skewness and heteroscedasticity, coverage for Yuen's technique was more accurate than Welch's technique.

Key words: Yuen's procedure, Welch's procedure, confidence interval, interval length, probability coverage, Monte Carlo simulation

## Introduction

When assessing how well the sample effect $(\overline{X}_1 - \overline{X}_2)$ estimates the population effect $(\mu_1 - \mu_2)$, a confidence interval is the appropriate statistical technique. The interval-length conveys the magnitude of the standard error of the effect. When comparing intervals for measuring an effect, wider interval-lengths imply greater standard errors. The confidence-level expresses the long-run probability that the limits include the population parameter.

The use of confidence intervals has been strongly suggested in some disciplines (Cohen, 1994; Wilkinson & Task Force on Statistical Inference, 1999). Some spurious reasons include (a) they provide statistical inference without specifying an a priori threshold and (b) it is presumed that confidence intervals provide a degree of certainty about the population parameter that hypothesis tests do not. However, Sawilowsky (2003) was opposed to (a) as being contrary to the principles of the scientific method, and noted that the Type I and Type II probabilities of hypothesis tests are the same as for confidence intervals.

Type I and Type II errors do apply to confidence intervals as follows.

1. Is zero truly within the interval yet the interval does not enclose zero (Type I error)?
2. Is zero not truly within the interval yet the interval does enclose zero (Type II error)?

Monte Carlo simulations have been used to assess the extent to which the Type I and Type II error rates deviate from the α and β levels. Magnitudes of interval-length and probability-coverage $(1-\hat{\alpha})$ serve as criteria concerning the appropriateness of confidence intervals. The traditional test for bi-group comparisons is the independent samples t-test. The calculation of the confidence interval for the mean difference is outlined as follows. Where $n_i$ is the sample size for group i, $\overline{X}_i$ is the mean for group i, and $X_{ji}$ is the jth observation for group i, the standard error of the effect is given as follows:

$$SE_{\bar{x}1-\bar{x}2} = \sqrt{S^2(n_1^{-1} + n_2^{-1})} \qquad (1)$$

$$S^2 = \left[\sum\left(X_{j1} - \overline{X}_1\right)^2 + \sum\left(X_{j2} - \overline{X}_2\right)^2\right] \Big/ (n_1 + n_2 - 2) \qquad (2)$$

S. Jonathan Mends-cole is an Instructor at Walden University, sjmendscole@aim.com. Mail: P.O. Box 07285, Detroit, MI 48207.

Where $t_{1-\alpha/2}$ refers to the critical value of the test distribution with $n_1 + n_2 - 2$ degrees of freedom, the confidence interval is:

$$\overline{X}_1 - \overline{X}_2 \mp t_{1-\alpha/2} SE_{\bar{x}1-\bar{x}2} \qquad (3)$$

Along with the assumptions that observations were randomly sampled from defined populations and that the samples were independent, some assumptions of parametric tests are homoscedasticity and normality (Wilcox, 1996). When heteroscedasticity and skewness are present in data, the error rates for a technique are inaccurate.

Violations of Parametric Test Assumptions

Skewness

Samples from skewed populations occur with some frequency as observed by Blair (1981) and Micceri (1989). Specifically, Micceri (1989) surveyed 440 published data sets. The p-value of the Kolmogorov-Smirnov test showed the distributions of each data set to be significantly different from a normal distribution ($p < .01$). Monte Carlo Type I error results (Sawilowsky & Blair, 1992) suggested that the probability-coverage would be greater than $1-\alpha$ for skewed distributions, i.e., for skewness ranging from 1.25 to 1.75. Setting the alpha level at 0.05, if Type I error rate is less than 0.05, probability-coverage is greater than 0.95. Sawilowsky and Blair observed that the independent samples t-test was robust: (a) if the test was two tailed rather than one tailed, (b) if sample sizes were about equal, and (c) if sample sizes were 25 or more.

Heteroscedasticity

Usually, when group means differ, group variances also differ (Sawilowsky & Blair, 1992, p. 358; Wilcox, 1996, p. 149). Why is heteroscedasticity likely to occur? Edwards (1972) attributed it to the absence of random assignment. If the variable for the treatment group exhibited greater variation before the application of the treatment after applying the treatment the difference is likely to remain unchanged. Another possibility is the multiplicative effect of the treatment. That is, if prior to the application of the treatment, $\sigma_2 / \sigma_1 < 2.0$, but after applying the treatment $\sigma_2 / \sigma_1 \geq 2.0$, the treatment may have acted multiplicatively to increase the variance.

Skewness & Heteroscedasticity

Heteroscedasticity has different effects on probability-coverage (Algina, Oshima, & Lin, 1994; Penfield, 1994). (a) If sizes are equal, the effect on probability-coverage is negligible, i.e., $0.925 \leq 1-\hat{\alpha} \leq 0.975$. (b) Small group sizes, e.g., $(n_1, n_2) = (5,15)$, skewness, and proportional heteroscedasticity augment probability-coverage (Penfield, 1994). (c) Small sizes, extreme skewness, and disproportional heteroscedasticity attenuate probability-coverage. If the confidence level was set at 0.95, the t-test displayed coverage-probabilities of 0.90 or less (Algina, Oshima, & Lin, 1994; Penfield, 1994). Although increasing sample sizes decreases the magnitude of separation between the Type I error rate and alpha level, Bradley (1978) observed that group sample sizes as large as 1,024 were needed for the independent samples t-test to maintain a 0.01 Type I error rate, if the application of the treatment increases the variance, heteroscedasticity increases interval-length. The larger group variance increases the standard error, thereby increasing the interval-length.

## Use of Transformations

Using transformations to remedy the error rate problems of skewness and heteroscedasticity is problematic. The interpretation of statistical significance for the transformed scale no longer holds for the untransformed scale (Games, 1983). Yet, the untransformed scale was selected based upon an underlying rationale for doing the study.

Welch's and Yuen's Techniques

Both Welch's and Yuen's techniques have been recommended for amending the Type I and Type II error rate problems resulting from heteroscedasticity and skewness (Wilcox, 1996). The confidence interval for Welch's technique uses a separate variance estimate of the standard

error. Where $s_i^2$ is the variance for group *i*, and $s_{\bar{x}i}^2 = s_i^2 / n_i$, the standard error is estimated as

$$SE_{\bar{x}1-\bar{x}2} = \sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2} \qquad (4)$$

The degrees of freedom are calculated as

$$df_{welch} = (s_{\bar{x}1}^2 + s_{\bar{x}2}^2)^2 \Big/ \left( (s_{\bar{x}1}^2)^2/(n_1 - 1) + (s_{\bar{x}2}^2)^2/(n_2 - 1) \right) \qquad (5)$$

Yuen's technique assesses the difference between the trimmed means. The technique is outlined as follows. Trimming a group sample involves omitting a fixed proportion of the largest scores and an equivalent number of the smallest scores from the sample. Winsorization involves replacing a fixed proportion of the largest scores with the maximum score for the trimmed version of the same sample, and replacing an equivalent number of the smallest scores with the minimum score for the trimmed version of the same sample. Wilcox (2003) suggested that 20% trimming is "a good choice for general use" (p. 251). (a) Where tau ($\tau_i$) is the integer portion of $0.20(n_i)$, the trimmed sample size is $h_i = n_i - 2(\tau_i)$. The trimmed mean $(\overline{X}_{ti})$ is the mean of observations for the trimmed sample size. (b) The Winsorized mean $(\overline{X}_{wi})$ is the mean of observations for the Winsorized sample. The Winsorized sum of squared deviations is estimated as

$$SSD_{wi} = (\tau+1)[(X_{(\tau+1)} - \overline{X}_w)^2 + (X_{(n-\tau)} - \overline{X}_w)^2] + \sum_{i=\tau+2}^{n-\tau-1} (X_{(i)} - \overline{X}_w)^2 \qquad (6)$$

Note that the subscripts in parentheses, e.g., $(\tau+1)$, $(n-\tau)$, and $(i)$ represent the ascending order of the *X* values. (c) Where the Winsorized variance is estimated as $S_{wi}^2 = SSD_{wi}/(h_i - 1)$ and the standard error of the trimmed mean is $S_{\bar{x}ti}^2 = S_{wi}^2 / h_i$, the standard error of the effect is estimated by:

$$SE_{\bar{x}t1-\bar{x}t2} = \sqrt{S_{\bar{x}t1}^2 + S_{\bar{x}t2}^2} \qquad (7)$$

The degrees of freedom is calculated as follows.

$$df_{yuen} = (S_{\bar{x}t1}^2 + S_{\bar{x}t2}^2) \Big/ (S_{\bar{x}t1}^2/(h_1 - 1) + S_{\bar{x}t2}^2/(h_2 - 1)) \qquad (8)$$

The confidence interval of trimmed means for bi-groups (Wilcox, 1996) is:

$$(\overline{X}_{t1} - \overline{X}_{t2}) \mp t_{1-\alpha/2} SE_{\bar{x}t1-\bar{x}t2} \qquad (9)$$

Welch's and Yuen's techniques exhibited appropriate coverage for extreme skewness, and homoscedasticity, i.e., $0.925 \le 1-\hat{\alpha} \le 0.975$ (Algina et al., 1994; Wilcox, 1994). Under conditions of skewness and disproportional heteroscedasticity, Welch's coverage was less than 0.925 (Luh & Guo, 2000). Yuen's coverage was less than the confidence-level but to a lesser extent than Welch's technique was, i.e., $1-\hat{\alpha} = 0.92$ versus 0.85. The probabilities of coverage were outlined in the table below.

Objections to the studies of Table 1 are related to the random samples assessed and the outcome measures used. The first objection is that the techniques were recommended based on random numbers generated using mathematical functions. The skewness and kurtosis properties of the random numbers may not generalize to the samples observed in applied situations in education and psychology. To the extent that Monte Carlo samples represent applied situations, the results are generalizable to similar situations (Sawilowsky & Fahoome, 2003).

The second objection with the manner in which the preceding studies were conducted is that the techniques were recommended based on Type I and Type II error rates alone. The Type I and the Type II error rates indirectly relate to confidence intervals; whereas, the probability-coverage and interval-length serve as outcome measures for confidence intervals. Though interval-length serves as an outcome measure for confidence intervals, journals in education and in psychology did not provide the interval-length for assessing Welch's and Yuen's techniques.

Table 1. Probability of Coverage of Yuen's & Welch's Techniques Reported in the Literature.

| Test | Citation | n1 | n2 | σ1 | σ2 | Skew. | Kurt. | PC |
|---|---|---|---|---|---|---|---|---|
| Welch's | Yuen (1974) | 10 | 10 | 1.00 | 0.71 | 0.00 | -1.20 | 0.95 |
| | | 20 | 10 | 1.00 | 0.71 | 0.00 | -1.20 | 0.95 |
| | | 20 | 20 | 1.41 | 0.71 | 0.00 | -1.20 | 0.95 |
| | | 10 | 20 | 4.00 | 1.00 | 0.00 | 0.00 | 0.95 |
| | | 10 | 10 | 2.00 | 1.00 | 0.00 | 0.00 | 0.95 |
| | Algina (1994) | 33 | 67 | 3.00 | 1.00 | 6.10 | np | 0.88 |
| | et al. | 33 | 67 | 2.00 | 1.00 | 6.10 | np | 0.90 |
| | | 33 | 67 | 1.00 | 1.00 | 6.10 | np | 0.94 |
| | Penfield (1994) | 10 | 20 | 1.00 | 1.00 | 0.00 | np | 0.96 |
| | | 10 | 20 | 1.00 | 1.00 | 1.50 | np | 0.96 |
| | | 20 | 20 | 1.00 | 2.00 | 0.00 | np | 0.95 |
| | | 20 | 20 | 1.00 | 2.00 | 1.50 | np | 0.95 |
| | | 10 | 20 | 1.00 | 2.00 | 1.50 | np | 0.95 |
| Welch's | | 10 | 20 | 1.00 | 2.00 | 1.00 | np | 0.95 |
| | | 10 | 20 | 2.00 | 1.00 | 0.00 | np | 0.95 |
| | Penfield (1994) | 10 | 20 | 2.00 | 1.00 | 1.50 | np | 0.96 |
| | Luh & Guo (2000) | 12 | 24 | 1.00 | 4.00 | 6.20 | 111.00 | 0.91 |
| | | 12 | 24 | 4.00 | 1.00 | 6.20 | 111.00 | 0.85 |
| | Guo & Luh (2000) | 18 | 12 | 1.00 | 6.00 | 1.75 | 5.90 | 0.92 |
| | | 18 | 12 | 1.00 | 6.00 | 6.20 | 111.00 | 0.85 |
| Yuen's | Luh & Guo (2000) | 12 | 24 | 1.00 | 4.00 | 6.20 | 111.00 | 0.95 |
| | | 12 | 24 | 4.00 | 1.00 | 6.20 | 111.00 | 0.92 |
| | Wilcox (1994) | 12 | 12 | 1.00 | 1.00 | 2.00 | 6.00 | 0.95 |
| | | 40 | 12 | 1.00 | 1.00 | 2.00 | 6.00 | 0.95 |
| | | 80 | 20 | 1.00 | 1.00 | 2.00 | 6.00 | 0.94 |
| | | 12 | 12 | 1.00 | 1.00 | 3.90 | 42.20 | 0.95 |
| | | 40 | 12 | 1.00 | 1.00 | 3.90 | 42.20 | 0.95 |
| | | 80 | 20 | 1.00 | 1.00 | 3.90 | 42.20 | 0.95 |

Purpose

The purpose of the study was to assess the probability-coverage and the interval-length for Welch's and Yuen's techniques. The techniques were assessed (a) using empirical data sets that were not normally distributed (i.e., Sawilowsky & Blair, 1992), (b) under conditions of heteroscedasticity, and (c) for unequal group sample sizes.

## Methodology

Micceri (1986) identified eight distributions prevalent in educational and psychological research. Table 2 provides the means, standard deviations and third and fourth moment estimates of skewness and kurtosis of the eight distributions. The kurtosis was adjusted so that the value for a normal distribution would be 0.00. Estimates of interval-length and probability-coverage were obtained by sampling from the seven distributions. Random samples were obtained independently and with replacement using the International Mathematical and Statistical Libraries (1998): RNUND and RNSET subroutines. One million repetitions were performed.

The procedure involved obtaining random samples from the empirical distributions, standardizing the scores, modeling the effect and modeling heterogeneity, trimming and Winsorizing the dataset, computing the interval, summing values of interval length and probability-coverage, and averaging values of interval length and the values of probability-coverage.

Sample size ratios of 1:1, 3:1, and 1:3 were selected. The respective sample sizes were ($n_1$, $n_2$) = (13, 13), (13, 39), (39, 13), and (39, 39). Variance ratios of 1:1, 1:2 and 1:4 allowed for a comparison of the probability-coverage and interval-lengths for each technique under homoscedasticity and heteroscedasticity. Coverage-probabilities and interval-length were examined at the 0.01, and 0.05 alpha levels.

Where $\mu'$ is the mean for the transformed score, $\sigma'$ is the standard deviation for the transformed score, and $Z$ is a standard score, the transformed score was obtained as follows.

$$X' = \mu' + \sigma' Z \quad (10)$$

The mean of the second group was set to one. The levels of skewness, size, variance, and effect under study represent a subset of conditions in applied situations.

The ratio of the average length for Student's technique divided by the average length for the comparison technique, i.e., Welch's or Yuen's technique, was calculated to compare interval lengths.

## Results

Probability-coverage

The results showed inflated probability-coverage for Yuen's techniques was observed with extreme skewness. Probability-coverage was greater than the confidence-level when skewness was above 1.25, sample sizes were equal and less than 25 or sample sizes were unequal. The results were observed under homoscedasticity. In addition, probability-coverage was greater than the confidence-level when skewness was above 1.25 and heteroscedasticity was proportional to size or sample sizes were equal, less than 25, and heteroscedastic. The probability-coverage exceeded the upper bound of the Bradley-criterion, i.e., $(1-\hat{\alpha}) > (1-0.5\alpha)$. The results were not observed where $\sigma_2 / \sigma_1 = 4$. Results were presented in Table 3 through Table 9.

Welch's technique:

Attenuated coverage-probabilities were observed for both extreme skewness (i.e., absolute skewness greater than 1.25) and heteroscedasticity ($\sigma_2 / \sigma_1 = 4$). That is, coverage-probabilities were less than 0.925 ($\alpha = 0.05$) or 0.985 ($\alpha = 0.01$). The results occurred where sample sizes were inversely proportional to variances; alternatively, the results occurred where group sample sizes were less than 25.

Table 2. Descriptive Information Pertaining to Eight Real World Distributions.

| Distribution | M | SD | Skew. | Kurt. |
|---|---|---|---|---|
| Mass at Zero | 12.92 | 4.42 | -0.03 | 0.31 |
| Extreme Asymmetry-Psychometric | 13.67 | 5.75 | 1.64 | 1.52 |
| Extreme Asymmetry-Achievement | 24.5 | 5.79 | -1.33 | 1.11 |
| Extreme Bimodality | 2.97 | 1.69 | -0.08 | -1.70 |
| Multimodal & Lumpy | 21.15 | 11.9 | 0.19 | -1.20 |
| Digit Preference | 536.95 | 37.64 | -0.07 | -0.24 |
| Smooth Symmetric | 13.19 | 4.91 | 0.01 | -0.34 |

*Note*. Adapted from "A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality", by S. S. Sawilowsky and R. C. Blair, 1992, Psychological Bulletin, 2, p. 353. Copyright 1992 by the American Psychological Association

Table 3. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from an Extreme Asymmetry – Achievement Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.952 | 0.956a | 0.964a | 0.992a | 0.994a | 0.995a |
| 13 | 39 | 1 | 0.952 | 0.937c | 0.948 | 0.991 | 0.980d | 0.988c |
| 39 | 13 | 1 | 0.952 | 0.937c | 0.948 | 0.991 | 0.980d | 0.988c |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.955 | 0.990 | 0.991 | 0.993a |
| 13 | 13 | 2 | 0.933c | 0.935c | 0.947 | 0.978d | 0.979d | 0.988c |
| 13 | 39 | 2 | 0.986b | 0.952 | 0.958a | 0.997b | 0.992a | 0.994a |
| 39 | 13 | 2 | 0.838d | 0.922d | 0.932c | 0.934d | 0.965d | 0.974d |
| 39 | 39 | 2 | 0.945 | 0.946 | 0.948 | 0.986c | 0.986c | 0.988c |
| 13 | 13 | 4 | 0.914d | 0.921d | 0.931c | 0.961d | 0.965d | 0.974d |
| 13 | 39 | 4 | 0.992b | 0.945 | 0.948 | 0.998b | 0.986c | 0.988c |
| 39 | 13 | 4 | 0.753d | 0.918d | 0.929c | 0.863d | 0.962d | 0.972d |
| 39 | 39 | 4 | 0.938c | 0.941c | 0.943c | 0.981d | 0.982d | 0.983d |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 4. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from an Extreme Bimodal Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.962a | 0.962a | 0.961a | 0.994a | 0.994a | 0.993a |
| 13 | 39 | 1 | 0.959a | 0.958a | 0.952 | 0.994a | 0.993a | 0.988c |
| 39 | 13 | 1 | 0.959a | 0.958a | 0.952 | 0.994a | 0.993a | 0.988c |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.949 | 0.990 | 0.990 | 0.989 |
| 13 | 13 | 2 | 0.958a | 0.961a | 0.953 | 0.993a | 0.994a | 0.989 |
| 13 | 39 | 2 | 0.993b | 0.955 | 0.953 | 0.999b | 0.991 | 0.990 |
| 39 | 13 | 2 | 0.857d | 0.960a | 0.948 | 0.949d | 0.994a | 0.984d |
| 39 | 39 | 2 | 0.948 | 0.950 | 0.947 | 0.989 | 0.989 | 0.987c |
| 13 | 13 | 4 | 0.953 | 0.961a | 0.949 | 0.991 | 0.994a | 0.984d |
| 13 | 39 | 4 | 0.997b | 0.951 | 0.948 | 1.000b | 0.990 | 0.987c |
| 39 | 13 | 4 | 0.781d | 0.961a | 0.949 | 0.893d | 0.994a | 0.982d |
| 39 | 39 | 4 | 0.946 | 0.949 | 0.946 | 0.988c | 0.989 | 0.985c |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 5. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from a Digit Preference Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.950 | 0.951 | 0.951 | 0.990 | 0.991 | 0.990 |
| 13 | 39 | 1 | 0.950 | 0.949 | 0.947 | 0.990 | 0.989 | 0.988c |
| 39 | 13 | 1 | 0.950 | 0.949 | 0.947 | 0.990 | 0.989 | 0.988c |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.949 | 0.990 | 0.990 | 0.990 |
| 13 | 13 | 2 | 0.946 | 0.950 | 0.948 | 0.988c | 0.990 | 0.989 |
| 13 | 39 | 2 | 0.991b | 0.950 | 0.949 | 0.999b | 0.990 | 0.989 |
| 39 | 13 | 2 | 0.846d | 0.949 | 0.945 | 0.940d | 0.989 | 0.988c |
| 39 | 39 | 2 | 0.948 | 0.950 | 0.948 | 0.989 | 0.990 | 0.989 |
| 13 | 13 | 4 | 0.941c | 0.949 | 0.945 | 0.985c | 0.989 | 0.988c |
| 13 | 39 | 4 | 0.998b | 0.950 | 0.949 | 1.000b | 0.990 | 0.989 |
| 39 | 13 | 4 | 0.765d | 0.949 | 0.946 | 0.881d | 0.989 | 0.988c |
| 39 | 39 | 4 | 0.947 | 0.950 | 0.948 | 0.989 | 0.990 | 0.989 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 6. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from a Mass at Zero Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.950 | 0.951 | 0.951 | 0.990 | 0.991 | 0.990 |
| 13 | 39 | 1 | 0.950 | 0.950 | 0.947 | 0.990 | 0.990 | 0.988c |
| 39 | 13 | 1 | 0.950 | 0.950 | 0.947 | 0.990 | 0.990 | 0.988c |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.950 | 0.990 | 0.990 | 0.990 |
| 13 | 13 | 2 | 0.947 | 0.951 | 0.948 | 0.989 | 0.990 | 0.989 |
| 13 | 39 | 2 | 0.991b | 0.950 | 0.950 | 0.999b | 0.990 | 0.990 |
| 39 | 13 | 2 | 0.847d | 0.950 | 0.945 | 0.941d | 0.990 | 0.987c |
| 39 | 39 | 2 | 0.949 | 0.950 | 0.948 | 0.990 | 0.990 | 0.989 |
| 13 | 13 | 4 | 0.942c | 0.950 | 0.945 | 0.986c | 0.990 | 0.987c |
| 13 | 39 | 4 | 0.998b | 0.950 | 0.948 | 1.000b | 0.990 | 0.989 |
| 39 | 13 | 4 | 0.765d | 0.950 | 0.945 | 0.882d | 0.990 | 0.988c |
| 39 | 39 | 4 | 0.947 | 0.950 | 0.948 | 0.989 | 0.990 | 0.989 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 7. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from a Smooth Symmetric Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.950 | 0.950 | 0.950 | 0.990 | 0.990 | 0.990 |
| 13 | 39 | 1 | 0.950 | 0.948 | 0.946 | 0.990 | 0.989 | 0.988c |
| 39 | 13 | 1 | 0.950 | 0.949 | 0.947 | 0.990 | 0.989 | 0.988c |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.950 | 0.990 | 0.990 | 0.990 |
| 13 | 13 | 2 | 0.945 | 0.949 | 0.947 | 0.988c | 0.989 | 0.988c |
| 13 | 39 | 2 | 0.991b | 0.950 | 0.950 | 0.999b | 0.990 | 0.990 |
| 39 | 13 | 2 | 0.846d | 0.949 | 0.945 | 0.939d | 0.989 | 0.987c |
| 39 | 39 | 2 | 0.949 | 0.950 | 0.949 | 0.989 | 0.990 | 0.989 |
| 13 | 13 | 4 | 0.941c | 0.949 | 0.946 | 0.985c | 0.989 | 0.988c |
| 13 | 39 | 4 | 0.998b | 0.950 | 0.949 | 1.000b | 0.990 | 0.989 |
| 39 | 13 | 4 | 0.765d | 0.949 | 0.946 | 0.881d | 0.989 | 0.988c |
| 39 | 39 | 4 | 0.947 | 0.950 | 0.948 | 0.988c | 0.990 | 0.989 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 8. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from a Multimodal Lumpy Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.949 | 0.949 | 0.949 | 0.989 | 0.989 | 0.989 |
| 13 | 39 | 1 | 0.950 | 0.947 | 0.939c | 0.990 | 0.987c | 0.981d |
| 39 | 13 | 1 | 0.950 | 0.947 | 0.939c | 0.990 | 0.987c | 0.981d |
| 39 | 39 | 1 | 0.950 | 0.950 | 0.950 | 0.990 | 0.990 | 0.989 |
| 13 | 13 | 2 | 0.944c | 0.947 | 0.937c | 0.986c | 0.987c | 0.980d |
| 13 | 39 | 2 | 0.991b | 0.950 | 0.948 | 0.999b | 0.989 | 0.989 |
| 39 | 13 | 2 | 0.845d | 0.947 | 0.930c | 0.937d | 0.986c | 0.972d |
| 39 | 39 | 2 | 0.948 | 0.949 | 0.947 | 0.989 | 0.989 | 0.987c |
| 13 | 13 | 4 | 0.938c | 0.946 | 0.929c | 0.982d | 0.986c | 0.971d |
| 13 | 39 | 4 | 0.997b | 0.950 | 0.948 | 1.000b | 0.989 | 0.987c |
| 39 | 13 | 4 | 0.767d | 0.946 | 0.929c | 0.880d | 0.986c | 0.971d |
| 39 | 39 | 4 | 0.947 | 0.949 | 0.946 | 0.988c | 0.989 | 0.986c |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 9. Coverage-probabilities for Each Technique by Sizes, Standard Deviations, and Alpha Levels when Sampling from a Extreme Asymmetry - Psychometric Distribution.

| | | | $\alpha = 0.05$ | | | $\alpha = 0.01$ | | |
|---|---|---|---|---|---|---|---|---|
| n1 | n2 | σ2/σ1 | Student | Welch | Yuen | Student | Welch | Yuen |
| 13 | 13 | 1 | 0.969a | 0.973a | 0.989b | 0.996b | 0.998b | 0.999b |
| 13 | 39 | 1 | 0.961a | 0.949 | 0.979b | 0.991 | 0.985c | 0.998b |
| 39 | 13 | 1 | 0.960a | 0.948 | 0.979b | 0.991 | 0.985c | 0.998b |
| 39 | 39 | 1 | 0.952 | 0.953 | 0.969a | 0.992a | 0.992a | 0.997b |
| 13 | 13 | 2 | 0.943c | 0.946 | 0.982b | 0.983d | 0.984d | 0.999b |
| 13 | 39 | 2 | 0.983b | 0.960a | 0.976b | 0.996b | 0.995a | 0.998b |
| 39 | 13 | 2 | 0.861d | 0.928c | 0.958a | 0.950d | 0.965d | 0.994a |
| 39 | 39 | 2 | 0.944c | 0.945 | 0.947 | 0.985c | 0.985c | 0.989 |
| 13 | 13 | 4 | 0.921d | 0.925c | 0.952 | 0.961d | 0.963d | 0.995a |
| 13 | 39 | 4 | 0.989b | 0.944c | 0.944c | 0.997b | 0.984d | 0.987c |
| 39 | 13 | 4 | 0.779d | 0.923d | 0.940c | 0.880d | 0.960d | 0.987c |
| 39 | 39 | 4 | 0.936c | 0.938c | 0.927c | 0.977d | 0.979d | 0.968d |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 10. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from an Extreme Asymmetry – Achievement Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.993a | 0.733a | 0.989a | 0.712a |
| 13 | 39 | 1 | 0.975c | 0.727 | 0.954d | 0.696c |
| 39 | 13 | 1 | 0.975c | 0.727 | 0.954d | 0.696c |
| 39 | 39 | 1 | 0.999 | 0.763 | 0.999 | 0.757a |
| 13 | 13 | 2 | 0.979c | 0.724 | 0.966d | 0.695c |
| 13 | 39 | 2 | 1.359 | 1.023a | 1.353a | 1.005a |
| 39 | 13 | 2 | 0.694d | 0.514c | 0.667d | 0.480d |
| 39 | 39 | 2 | 0.994 | 0.760 | 0.991c | 0.751c |
| 13 | 13 | 4 | 0.960d | 0.709c | 0.935d | 0.670d |
| 13 | 39 | 4 | 1.610 | 1.224 | 1.608c | 1.211c |
| 39 | 13 | 4 | 0.573d | 0.424c | 0.547d | 0.390d |
| 39 | 39 | 4 | 0.987c | 0.755c | 0.980d | 0.742d |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 11. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from an Extreme Bimodal Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.999a | 0.886a | 0.999a | 0.870a |
| 13 | 39 | 1 | 0.964a | 0.840 | 0.943a | 0.809c |
| 39 | 13 | 1 | 0.964a | 0.840 | 0.944a | 0.809c |
| 39 | 39 | 1 | 1.000 | 0.827 | 1.000 | 0.822 |
| 13 | 13 | 2 | 0.981a | 0.864 | 0.969a | 0.833 |
| 13 | 39 | 2 | 1.360 | 1.162 | 1.356 | 1.148 |
| 39 | 13 | 2 | 0.682a | 0.594 | 0.655a | 0.556d |
| 39 | 39 | 2 | 0.994 | 0.820 | 0.991 | 0.811c |
| 13 | 13 | 4 | 0.959a | 0.837 | 0.934a | 0.790d |
| 13 | 39 | 4 | 1.613 | 1.343 | 1.612 | 1.330c |
| 39 | 13 | 4 | 0.568a | 0.493 | 0.542a | 0.455d |
| 39 | 39 | 4 | 0.987 | 0.811 | 0.980 | 0.797c |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 12. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from a Digit Preference Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.996 | 0.768 | 0.994 | 0.750 |
| 13 | 39 | 1 | 0.971 | 0.749 | 0.950 | 0.718c |
| 39 | 13 | 1 | 0.971 | 0.749 | 0.950 | 0.718c |
| 39 | 39 | 1 | 1.000 | 0.775 | 0.999 | 0.770 |
| 13 | 13 | 2 | 0.980 | 0.753 | 0.968 | 0.725 |
| 13 | 39 | 2 | 1.361 | 1.054 | 1.356 | 1.038 |
| 39 | 13 | 2 | 0.688 | 0.527 | 0.662 | 0.492c |
| 39 | 39 | 2 | 0.994 | 0.769 | 0.991 | 0.761 |
| 13 | 13 | 4 | 0.959 | 0.733 | 0.935 | 0.692c |
| 13 | 39 | 4 | 1.612 | 1.247 | 1.610 | 1.234 |
| 39 | 13 | 4 | 0.570 | 0.435 | 0.544 | 0.400c |
| 39 | 39 | 4 | 0.987 | 0.763 | 0.980 | 0.749 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 13. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from a Mass at Zero Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.995 | 0.812 | 0.993 | 0.793 |
| 13 | 39 | 1 | 0.972 | 0.793 | 0.951 | 0.761c |
| 39 | 13 | 1 | 0.972 | 0.794 | 0.951 | 0.761c |
| 39 | 39 | 1 | 1.000 | 0.824 | 0.999 | 0.819 |
| 13 | 13 | 2 | 0.980 | 0.796 | 0.967 | 0.767 |
| 13 | 39 | 2 | 1.360 | 1.119 | 1.355 | 1.102 |
| 39 | 13 | 2 | 0.690 | 0.558 | 0.663 | 0.521c |
| 39 | 39 | 2 | 0.994 | 0.819 | 0.991 | 0.809 |
| 13 | 13 | 4 | 0.959 | 0.775 | 0.935 | 0.731c |
| 13 | 39 | 4 | 1.611 | 1.325 | 1.610 | 1.312 |
| 39 | 13 | 4 | 0.571 | 0.460 | 0.545 | 0.423c |
| 39 | 39 | 4 | 0.987 | 0.811 | 0.980 | 0.797 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 14. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from a Smooth Symmetric Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.996 | 0.735 | 0.994 | 0.718 |
| 13 | 39 | 1 | 0.971 | 0.716 | 0.950 | 0.687c |
| 39 | 13 | 1 | 0.971 | 0.716 | 0.950 | 0.687c |
| 39 | 39 | 1 | 1.000 | 0.742 | 0.999 | 0.738 |
| 13 | 13 | 2 | 0.980 | 0.721 | 0.968 | 0.694c |
| 13 | 39 | 2 | 1.361 | 1.009 | 1.356 | 0.994 |
| 39 | 13 | 2 | 0.688 | 0.504 | 0.661 | 0.470c |
| 39 | 39 | 2 | 0.994 | 0.737 | 0.991 | 0.729 |
| 13 | 13 | 4 | 0.959 | 0.702 | 0.935 | 0.662c |
| 13 | 39 | 4 | 1.612 | 1.195 | 1.611 | 1.183 |
| 39 | 13 | 4 | 0.570 | 0.416 | 0.544 | 0.383c |
| 39 | 39 | 4 | 0.987 | 0.731 | 0.980 | 0.718 |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 15. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from a Multimodal Lumpy Distribution.

| . | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.998 | 0.808 | 0.997 | 0.790 |
| 13 | 39 | 1 | 0.968 | 0.777c | 0.947c | 0.747d |
| 39 | 13 | 1 | 0.968 | 0.777c | 0.947c | 0.747d |
| 39 | 39 | 1 | 1.000 | 0.784 | 1.000 | 0.779 |
| 13 | 13 | 2 | 0.980 | 0.790c | 0.969c | 0.762d |
| 13 | 39 | 2 | 1.362 | 1.085 | 1.357 | 1.071 |
| 39 | 13 | 2 | 0.685 | 0.548c | 0.658c | 0.512d |
| 39 | 39 | 2 | 0.994 | 0.778 | 0.991 | 0.769c |
| 13 | 13 | 4 | 0.959 | 0.768c | 0.935c | 0.725d |
| 13 | 39 | 4 | 1.613 | 1.268 | 1.612 | 1.256c |
| 39 | 13 | 4 | 0.569 | 0.454c | 0.543c | 0.418d |
| 39 | 39 | 4 | 0.987 | 0.770 | 0.980 | 0.756c |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Table 16. Ratio of the Average Lengths for Student's Technique to that for Welch's and Yuen's Techniques when Sampling from an Extreme Asymmetry – Psychometric Distribution.

| | | | $\alpha = 0.05$ | | $\alpha = 0.01$ | |
|---|---|---|---|---|---|---|
| $n_1$ | $n_2$ | $\sigma_2/\sigma_1$ | Welch | Yuen | Welch | Yuen |
| 13 | 13 | 1 | 0.992a | 0.698b | 0.988b | 0.673b |
| 13 | 39 | 1 | 0.967 | 0.702b | 0.946c | 0.669b |
| 39 | 13 | 1 | 0.967 | 0.702b | 0.945c | 0.668b |
| 39 | 39 | 1 | 0.999 | 0.772a | 0.999a | 0.765b |
| 13 | 13 | 2 | 0.978 | 0.696b | 0.965d | 0.666b |
| 13 | 39 | 2 | 1.349a | 0.998b | 1.343a | 0.974b |
| 39 | 13 | 2 | 0.691c | 0.501a | 0.664d | 0.467a |
| 39 | 39 | 2 | 0.994 | 0.774 | 0.990c | 0.765 |
| 13 | 13 | 4 | 0.960c | 0.692 | 0.935d | 0.654a |
| 13 | 39 | 4 | 1.605c | 1.230c | 1.604d | 1.213c |
| 39 | 13 | 4 | 0.573d | 0.416c | 0.547d | 0.383c |
| 39 | 39 | 4 | 0.988c | 0.777c | 0.980d | 0.763d |

a. $1-\hat{\alpha} > 0.955, \alpha = 0.05$ or $1-\hat{\alpha} > 0.991, \alpha = 0.01$

b. $1-\hat{\alpha} > 0.975, \alpha = 0.05$ or $1-\hat{\alpha} > 0.995, \alpha = 0.01$

c. $1-\hat{\alpha} < 0.945, \alpha = 0.05$ or $1-\hat{\alpha} < 0.989, \alpha = 0.01$

d. $1-\hat{\alpha} < 0.925, \alpha = 0.05$ or $1-\hat{\alpha} < 0.985, \alpha = 0.01$

Yuen's technique

For Yuen's techniques, attenuated coverage-probabilities were observed for extreme negative kurtosis. For the population defined by the trimmed mean, kurtosis values less than –1.25 were observed with coverage-probabilities less than 0.985. Given extreme bimodality, coverage-probabilities that were within the range of 0.925-0.975 ($\alpha = 0.05$) were below 0.985 ($\alpha = 0.01$). The kurtosis of the extreme bimodal distribution after trimming was –1.454. The results occurred under both homoscedastic and heteroscedastic conditions. Where size was inversely paired with variance, under a multimodal lumpy distribution, coverage-probabilities were within the range 0.925-0.975 at the 0.05 alpha level. At the 0.01 alpha level, coverage-probabilities were less than 0.985. The kurtosis of the multimodal lumpy distribution after trimming was -1.269.

Interval Length

The results for interval length showed where $S^2_{\min}$ was divided by $n_{\min}$, the interval-lengths for Welch's and Yuen's techniques were less than the interval-lengths for Student's technique. If $S^2_{\max}$ was divided by $n_{\min}$, the reverse was true. Results were presented in Table 10 through Table 16. Second, interval-lengths for Yuen's technique were wider than the interval-lengths for Welch's technique. The interval-length ratios for Yuen's technique were smaller than the ratios of Welch's technique. Larger interval-lengths were observed for the heteroscedastic than for the homoscedastic condition.

## Conclusion

Similar to findings by Sawilowsky and Blair (1992, p. 359) showing that skewness attenuated the Type I error rates for the t-test, the results of the present study showed that if skewness was above 1.25, e.g., skewness of the extreme asymmetric - psychometric distribution was 1.417 after trimming, coverage-probabilities were augmented (i.e., $(1-\hat{\alpha}) > (1-0.5\alpha)$).

Similar to findings by Luh and Guo (2000) and Algina et al. (1994) showing that when size was inversely proportional to heteroscedasticity and skewness was greater or equal to 2.00, Welch's technique displayed coverage-probabilities less than the confidence-level when size was inversely proportional to heteroscedasticity and skewness was –1.33 or 1.64.

Finally, the augmentation or attenuation of probability-coverage for both techniques occurred more at 0.01 than at 0.05 alpha levels; this finding was consistent with results from Bradley (1978, p. 147) showing that larger sample sizes were required for the t-test to exhibit robustness at the 0.01 level than at the 0.05 level.

## References

References marked with an asterisk indicate studies included in Table 1.

*Algina, J., Oshima, T. C., & Lin, W-Y. (1994). Type I error rates for Welch's tests and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics. 19*, 275-291.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." *Review of Educational Research, 51*, 499-507.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology, 31*, 144-152.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

Edwards, A. L. (1972). *Experimental design in psychological research*. New York: Holt, Rinehart, and Winston, Inc.

Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin, 93,* 382-387.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

*Guo, J-H. & Luh, W-M. (2000). An invertable tranformation two-sample trimmed t-

statistic under heterogeneity and nonnormality. *Statistics & Probability Letters, 49*, 1-7.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences.* Boston, MA: Hougton Mifflin Company.

International Mathematical and Statistical Libraries (1998). IMSL(R) F90 MP library 3.0. [Computer software]. Incline Village, NV: Lahey Computer Systems, Inc.

Lahey Computer Systems. (1995-2000). Essential Lahey FORTRAN 90 [Computer software]. Incline Village, NV: Lahey Computer Systems, Inc.

*Luh, W-M. & Guo, J-H. (2000). Johnson's transformation two-sample trimmed *t* and its bootstrap method for heterogeneity and nonnormality. *Journal of Applied Statistics, 27*, 965-973.

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *American Statistician, 44,* 322-326.

Micceri, T. (1986). A futile search for that statistical chimera of normality. Paper presented at the 31st Annual Convention of the Florida Educational Research Association, Tampa.

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

* Penfield, D. A. (1994). Choosing a two-sample location test. *Journal of Experimental Education, 62*, 343-360.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352-360.

Sawilowsky, S. S. & Fahoome, G. (2003). *Statistics through Monte Carlo experimentation with FORTRAN*. Oak Park, MI: JMASM, Inc.

Sawilowsky, S. S. (2003). A different future for the social and behavioral sciences. *Journal of Modern Applied Statistical Methods, 2*, 128-132.

*Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal. 3*, 259-273.

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, CA: Academic Press.

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

*Yuen, K. K. (1974)/ The two-sampled trimmed *t* for unequal population variances. *Biometrika,* 61, 165-170.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 443- 455
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# The Effect Of Different Degrees Of Freedom Of The Chi-square Distribution On The Statistical Power Of The t, Permutation t, And Wilcoxon Tests

Michèle Weber
San Jose, California

The Chi-square distribution is used quite often in Monte Carlo studies to examine statistical power of competing statistics. The power spectrum of the t-test, Wilcoxon test, and permutation t test are compared under various degrees of freedom for this distribution. The two t tests have similar power, which is generally less than the Wilcoxon.

Key words: t-test, Wilcoxon test, permutation test, Chi Square, Monte Carlo simulation

## Introduction

Weber (2006) found that the power of the Wilcoxon test was only somewhat higher than that of the t and permutation t-tests in the Chi-square distribution with 6 degree of freedom. It was expected that the Wilcoxon test is much more powerful than the t-test under non-normality (Blair & Higgins, 1985; Hodges & Lehmann, 1956; Sawilowsky, 1990; Sawilowsky & Fahoome, 2003).

## Purpose

This study investigates the effect on statistical power of the t, permutation t and Wilcoxon tests when data are sampled from the Chi squared distribution.

## Methodology

A Monte Carlo simulation was used to study the properties of the two independent samples Student's t test, the permutation t test, and the Wilcoxon Rank Sum test with regard to their statistical power.

Michèle Weber, PhD, LMSW, LMFT, is an independent researcher, who resides in San Jose, California. Her areas of interest are in Monte Carlo methods and hypothesis tests. Email her at M.Fatal-Weber@worldnet.att.net

Nominal alpha was selected at $\alpha = 0.05$. The samples were composed of random numbers obtained from Rangen 2.0, which is a collection of subroutines to generate pseudo-random numbers (Sawilowsky & Fahoome, 2003). Small sample sizes used for this study were $n_1 = n_2 = 10$ and $n_1 = 5$ and $n_2 = 15$. The larger sample sizes were $n_1 = n_2 = 20$ and $n_1 = 10$ and $n_2 = 30$. One thousand five hundred repetitions were done to obtain the power, with shifts in means of $\mu = .2\sigma, .5\sigma, .8\sigma$ and $1.2\sigma$.

The procedures were performed on the Chi-square distribution with degree of freedom 1, 2, 3, 4, 5, 6, 8, 10, 20 and 40.

## Results

### Comparison of Power between t and Permutation t tests

All results are presented in the Appendix. For the even sample sizes, $n_1 = n_2 = 10$ and $n_1 = n_2 = 20$, the t and permutation t-tests reflected the same power regardless of the degree of freedom of the Chi-square distribution. Indeed, as shown in the different graphs, the power starts somewhat low at .10 for small sample sizes and .15 for $n_1 = n_2 = 20$ when the shift $\mu$ is $.2\sigma$. As the shifts in means increase, the power increases to an average of .97 for $n_1 = n_2 = 10$ and .93 for larger sample sizes when $\mu = 1.2\sigma$.

For uneven sample sizes, the t-test is less powerful than that of the permutation t-test. As the shift in means increases, the difference between the power of these two tests decreases.

For $n_1 = 5$ and $n_2 = 15$, the difference is .10 at $\mu = .2\sigma$ and becomes .01 when $\mu = 1.2\sigma$. The disparity is less noticeable for the larger uneven sample sizes, from .06 to .01 with the increase in the shift.

Comparison of Wilcoxon Test with t and Permutation t-tests

For the Chi-square distribution with degree of freedom 1, the power of the Wilcoxon test starts off somewhat higher than the t and permutation t-tests at .25 for $n_1 = n_2 = 10$. For $n_1 = n_2 = 20$, the power of the Wilcoxon test is .45 and .4 for the larger uneven sample size when $\mu = .2\sigma$. However, as the shift in means increases, the difference in power between the Wilcoxon and the t and permutation t-tests becomes smaller, until it reaches a plateau.

As the degree of freedom becomes larger, the power difference between the Wilcoxon and the t and permutation t-tests decreases. When df = 8, the power of the three tests is so similar that the graphs do not reflect a difference between them, especially for the other sample sizes. When the degree of freedom is 10, the Wilcoxon test becomes less powerful than the two t-tests. For $n_1 = n_2 = 20$, the power of the Wilcoxon test is below the power of the two tests at df = 40.

## Discussion

The Chi-square distribution with 1 degree of freedom is extremely asymmetric. Thus, it was expected and found that the Wilcoxon test is, indeed, much more powerful than the t and permutation t-test, as supported by Sawilowsky and Blair (1992).

As the degree of freedom increases, the Chi-square becomes less asymmetric and light tailed. Thus, the power properties of the Student's t and the permutation t-tests were expected to be rehabilitated as the distribution become more normal like, which was confirmed by the results. Indeed, with degrees of freedom 4, 5, 6 and 8, the increased symmetry and decreased tail weights made the permutation t-test and the Student's t-test more competitive, although still decreasingly less powerful than the Wilcoxon Rank-Sum test.

For df = 20 and 40, the t and permutation t-tests only had a modest power increase over the Wilcoxon test as predicted by the Asymptotic Relative Efficiencies (ARE).

As stated previously, the uneven sample sizes ($n_1 = 5$ and $n_2 = 15$, $n_1 = 10$ and $n_2 = 30$) offered a different outcome in the smaller degree of freedom of the Chi-square distribution. The difference in power between the Student's t-test and its permutation counterpart can be explained by the fact that the t-test performs worse than the permutation t-test under non-normality and unevenness of the sample size (Sawilowksy & Fahoome, 2003). As for the superiority of the Wilcoxon test, the results are similar to the Sawilowsky and Blair (1992)'s study.

The power of the Wilcoxon test is superior compared with both the t and permutation t-tests for the Chi-square distribution, though decreasing progressively as the degree of freedom increases from 1 to approximately 10. Then, the t and permutation t-tests regain their superiority as the distribution with higher degree of freedom becomes more symmetric with lighter tails simulating the normal distribution. Therefore in situations where the data suggest a Chi-square distribution with lower (less than 10) degree of freedom, the Wilcoxon test is preferable to its competitors when the nature of the treatment changes the mean of two independent samples.

## 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, 119-128.

Bukszar, J. & van den Oord, E. J. (2006). Accurate and efficient power calculations for 2 × m tables in unmatched case-control designs. *Statistics in Medicine*, 25 (15), 2632.

Gimenez, P.; Bolfarine, H. & Colosimo, E. A. (2000). Hypotheses testing for error-in-variables models. *Annals of The Institute of Statistical Mathematics* (Tokyo), 52 (4), 698-711.

Hodges, J. & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*, 27, 324-335.

Shinozabi, O. Y. (2006). Estimation of error variance in ANOVA model and order restricted scale parameters. *Annals of The Institute of Statistical Mathematics* (Tokyo), 34 (4), 739.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, 60 (1), 91-126.

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111, 353-360.

Sawilowsky, S. S. & Fahoome, G. F. (2003). *Statistics Through Monte Carlo Simulation With Fortran.* Oak Park, MI: JMASM, Inc.

Weber, M. (2006). *Robustness and Power of the t, Permutation t and Wilcoxon Tests* (Doctoral dissertation, Wayne State University, 2006).

APPENDIX

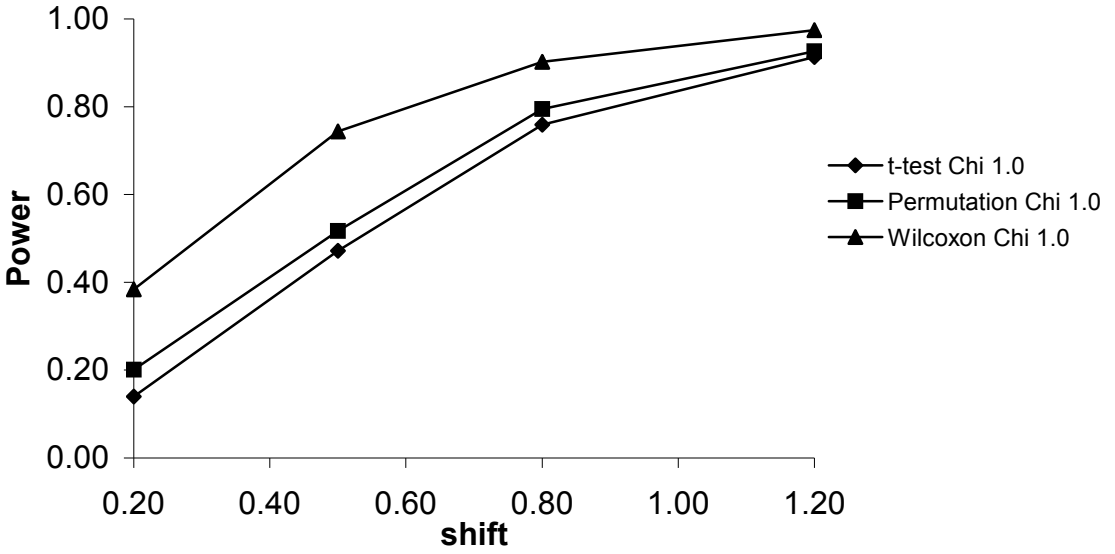



Figure 21. Shift vs. Power in the Chi-Square Distribution (df = 1) for Sample Size $n_1 = 10$, $n_2 = 30$

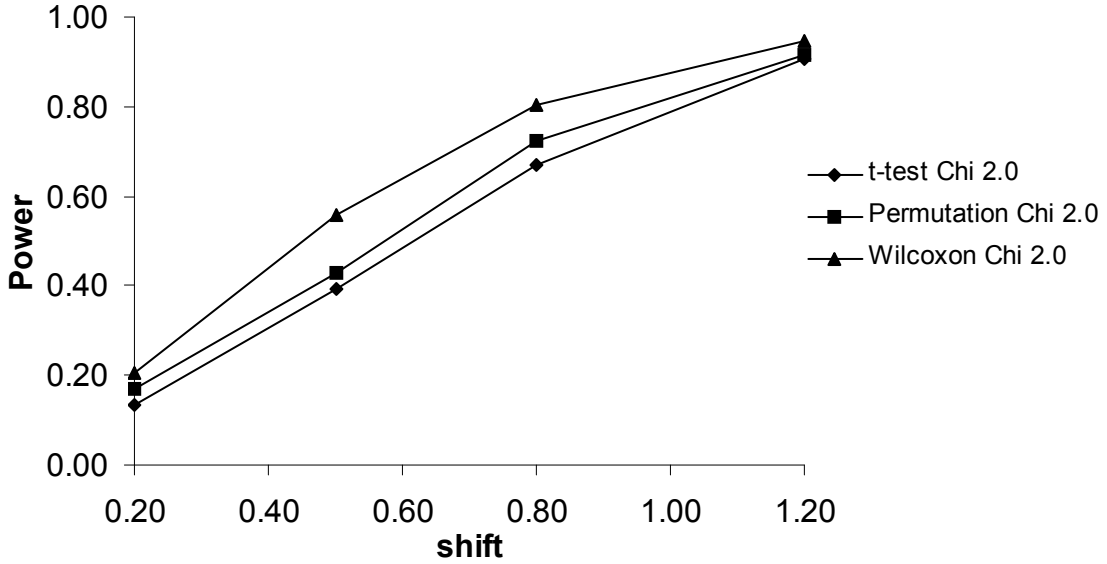



Figure 22. Shift vs. Power in the Chi-Square Distribution (df = 2) for Sample Size $n_1 = 10$, $n_2 = 30$

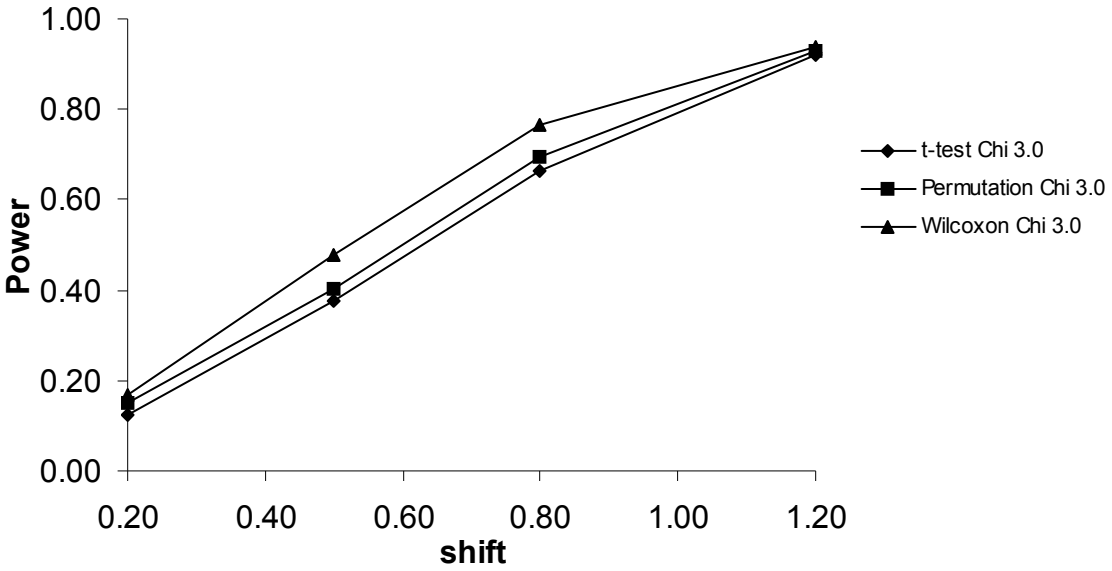



Figure 23. Shift vs. Power in the Chi-Square Distribution (df = 3) for Sample Size

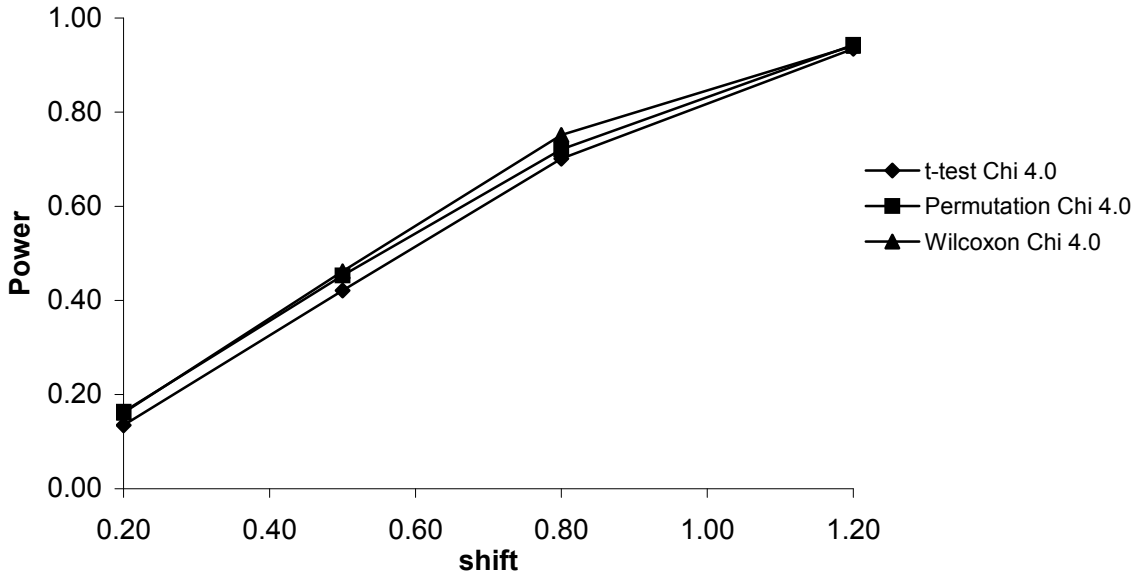



Figure 24. Shift vs. Power in the Chi-Square Distribution (df = 4) for Sample Size $n_1 = 10$, $n_2 = 30$

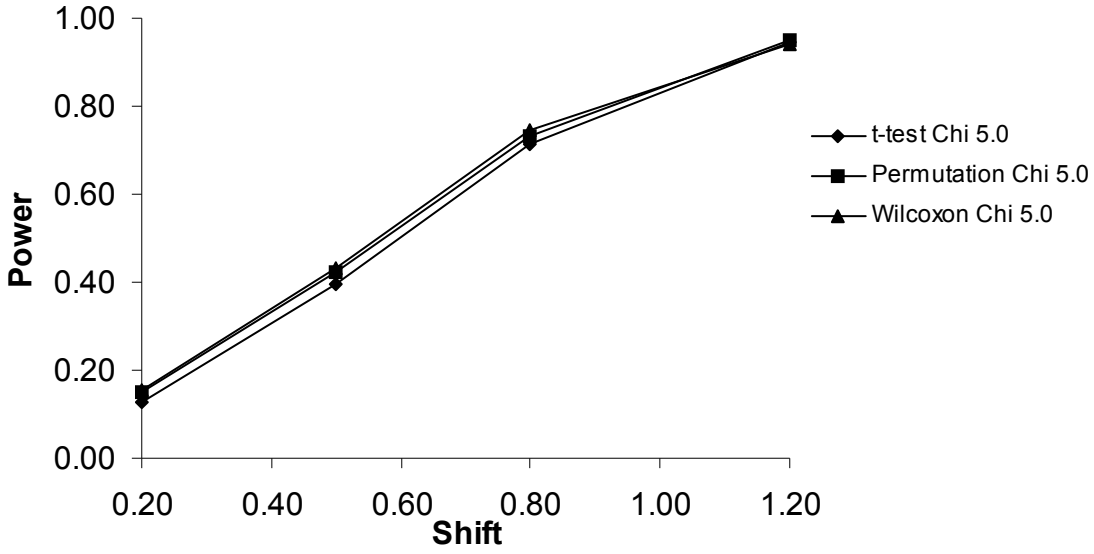



Figure 25: Shift vs. Power in the Chi-Square Distribution (df = 5) for Sample Size $n_1 = 10$, $n_2 = 30$

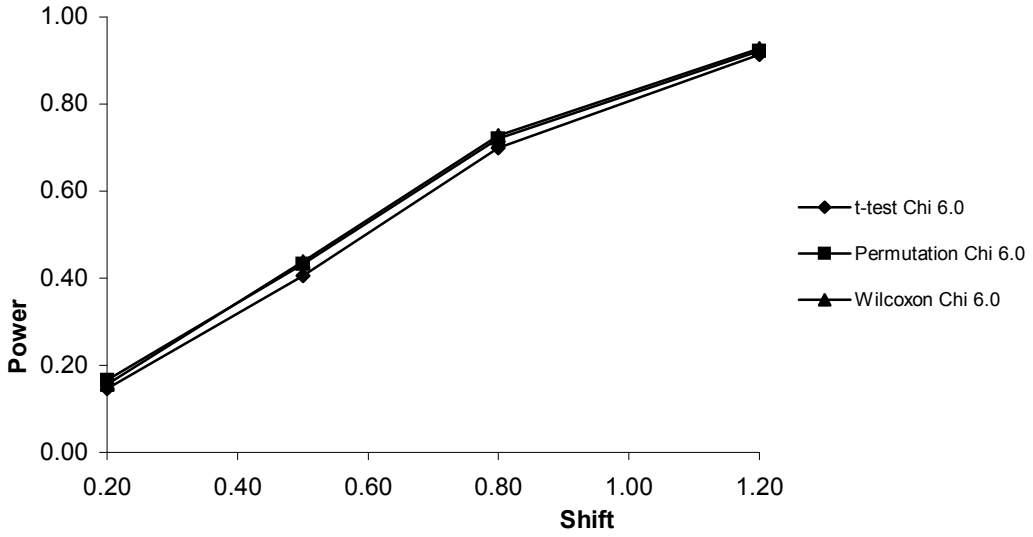



Figure 26: Shift vs. Power in the Chi-Square Distribution (df = 6) for Sample Size $n_1 = 10$, $n_2 = 30$

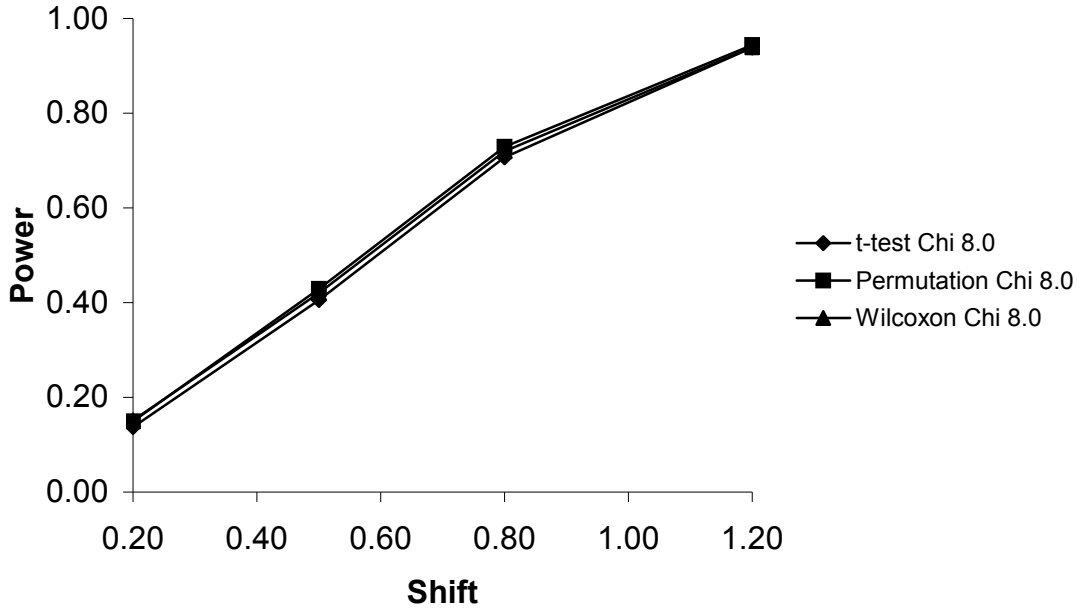



Figure 27: Shift vs. Power in the Chi-Square Distribution (df = 8) for Sample Size $n_1 = 10$, $n_2 = 30$

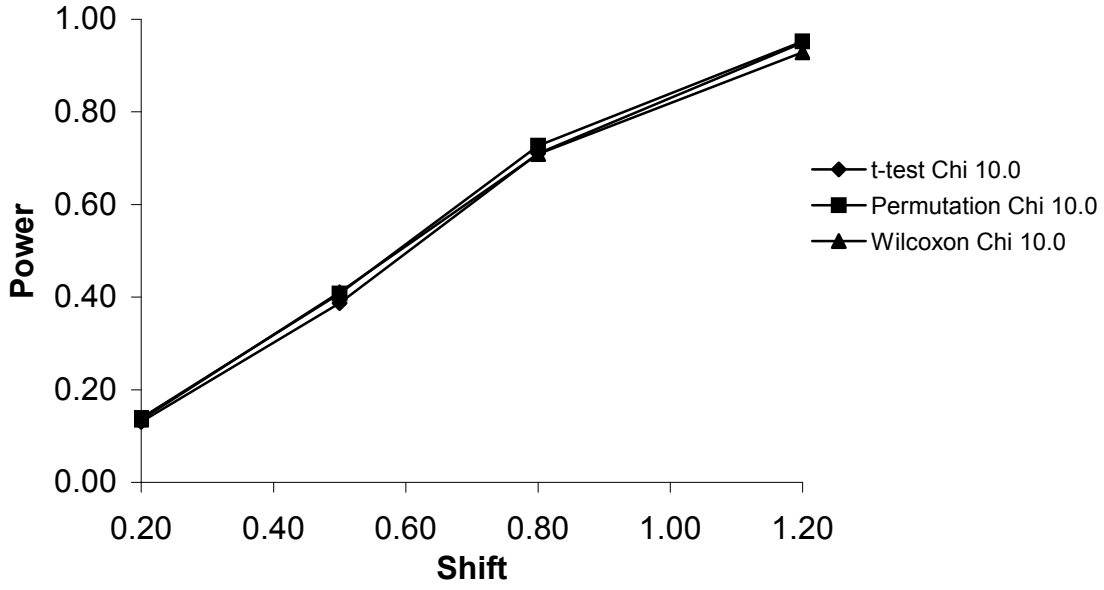



Figure 28: Shift vs. Power in the Chi-Square Distribution (df = 10) for Sample Size $n_1 = 10$, $n_2 = 30$

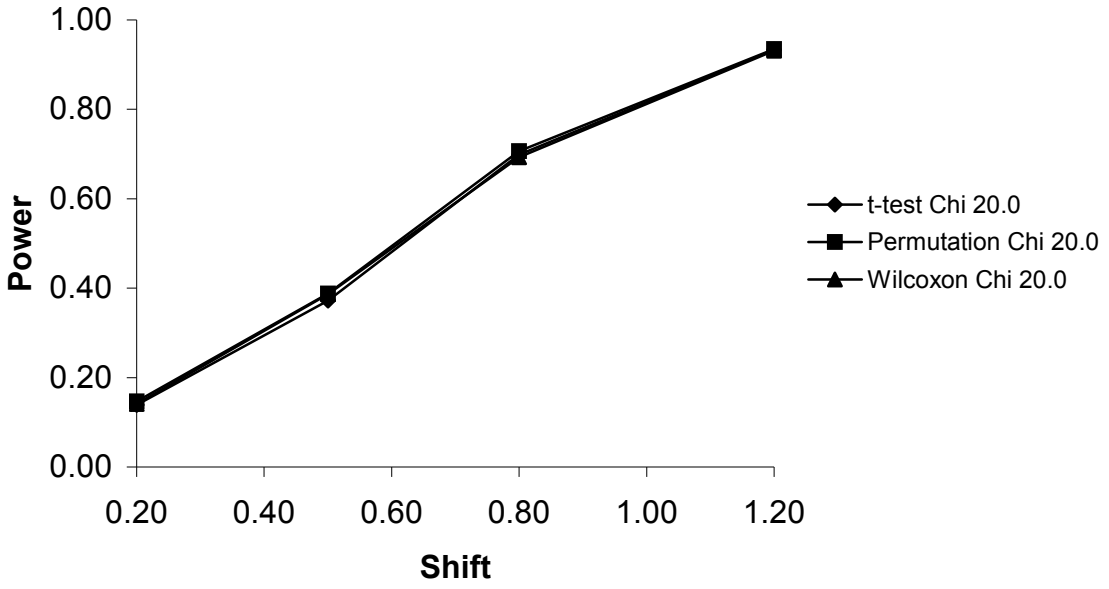



Figure 29: Shift vs. Power in the Chi-Square Distribution (df = 20) for Sample Size $n_1 = 10$, $n_2 = 30$

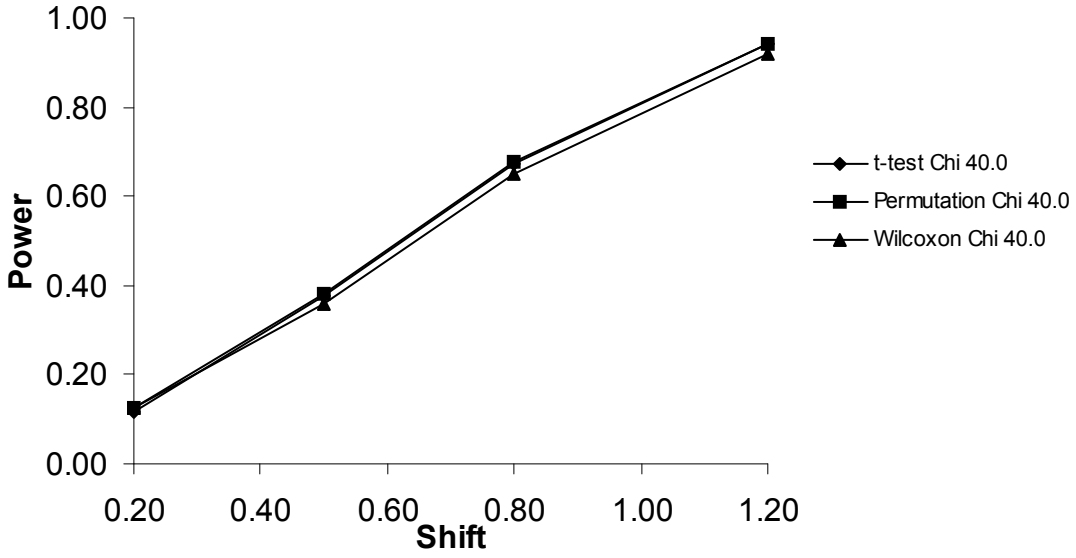



Figure 30: Shift vs. Power in the Chi-Square Distribution (df = 40) for Sample Size $n_1 = 10$, $n_2 = 30$

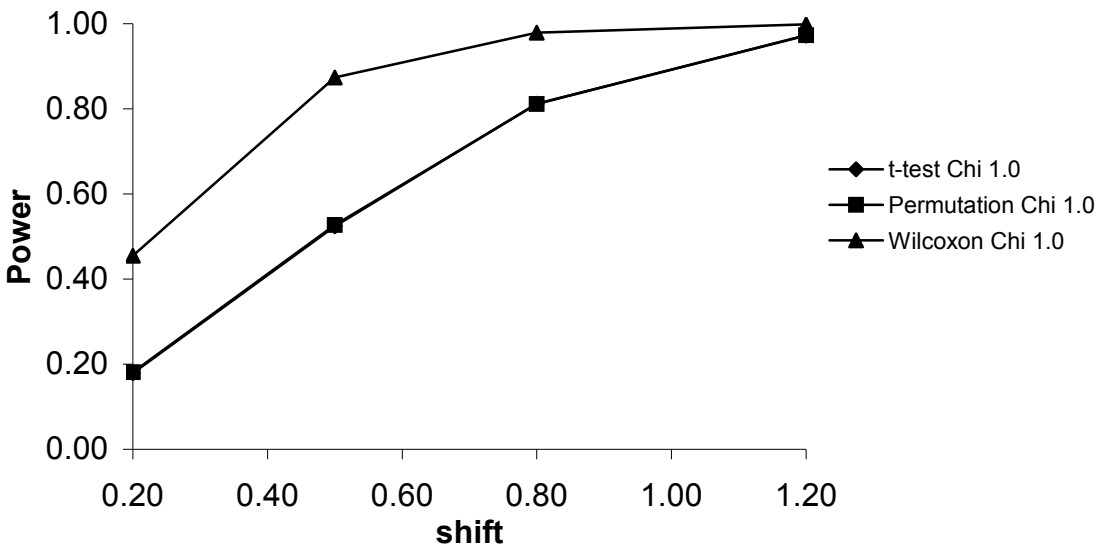



Figure 31: Shift vs. Power in the Chi-Square Distribution (df = 1) for Sample Size $n_1 = n_2 = 20$

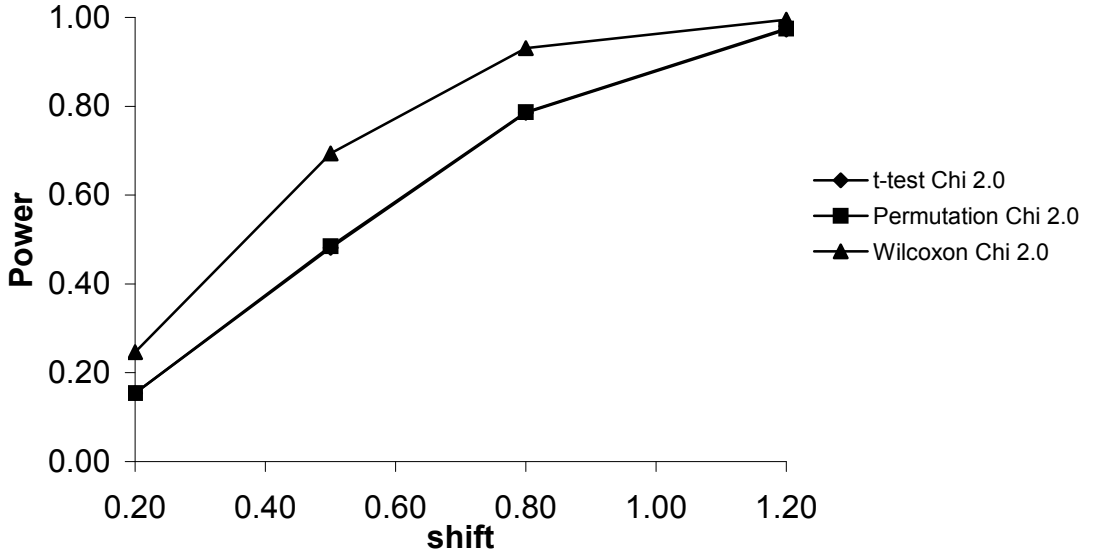



Figure 32: Shift vs. Power in the Chi-Square Distribution (df = 2) for Sample Size $n_1 = n_2 = 20$

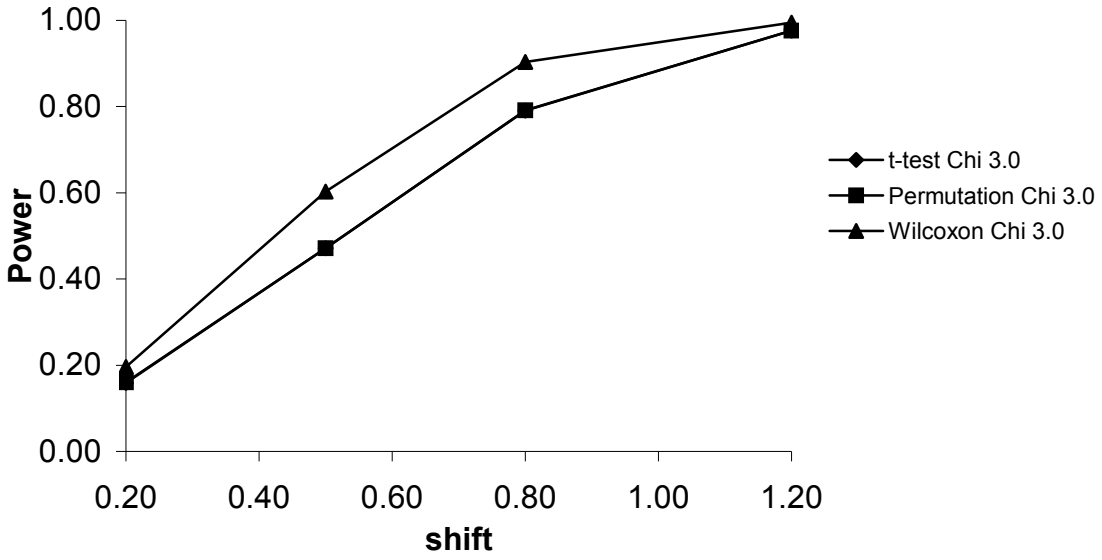



Figure 33: Shift vs. Power in the Chi-Square Distribution (df = 3) for Sample Size $n_1 = n_2 = 20$

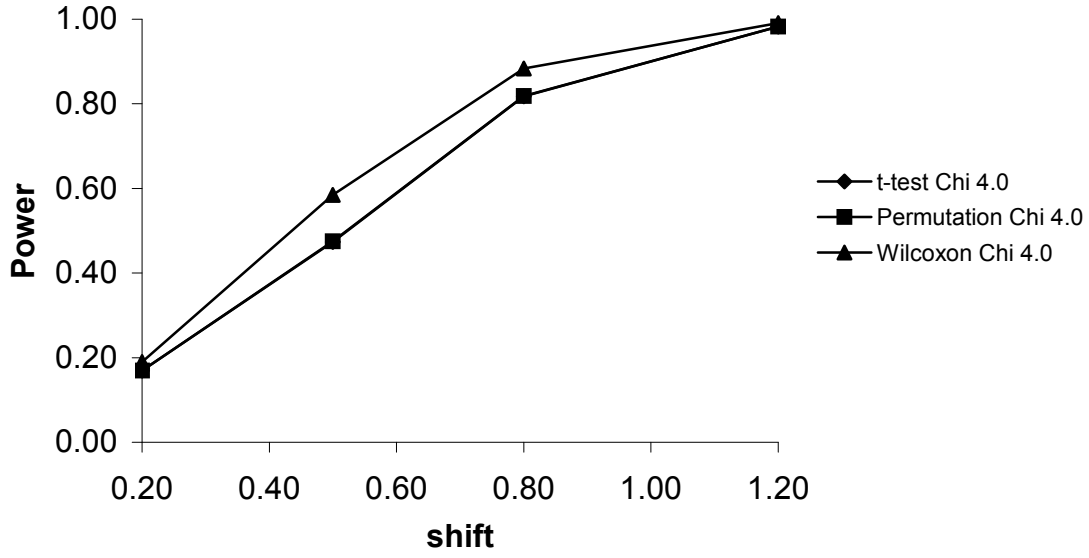



Figure 34: Shift vs. Power in the Chi-Square Distribution (df = 4) for Sample Size $n_1 = n_2 = 20$

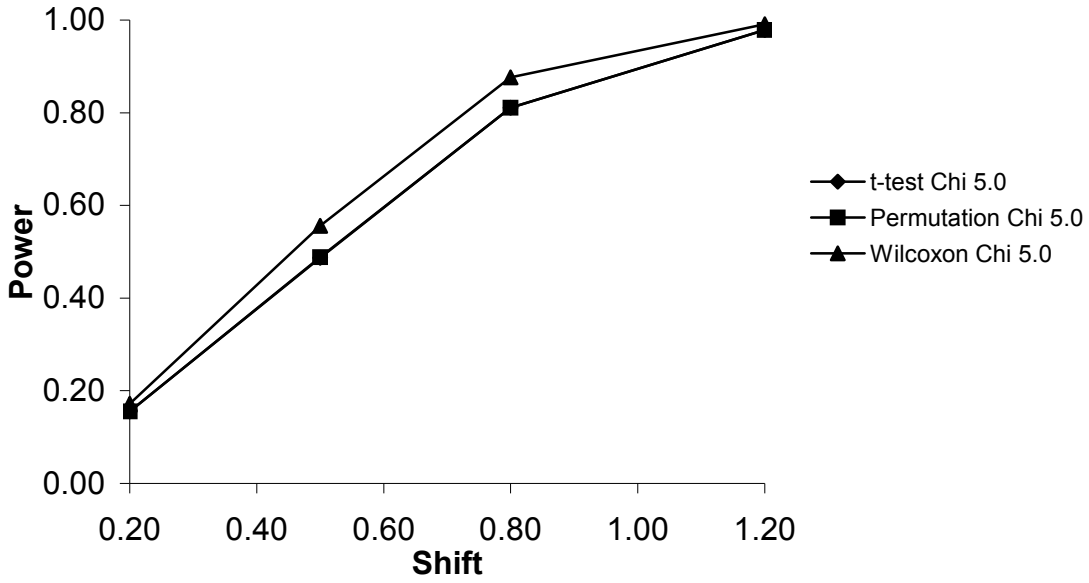



Figure 35: Shift vs. Power in the Chi-Square Distribution (df = 5) for Sample Size $n_1 = n_2 = 20$

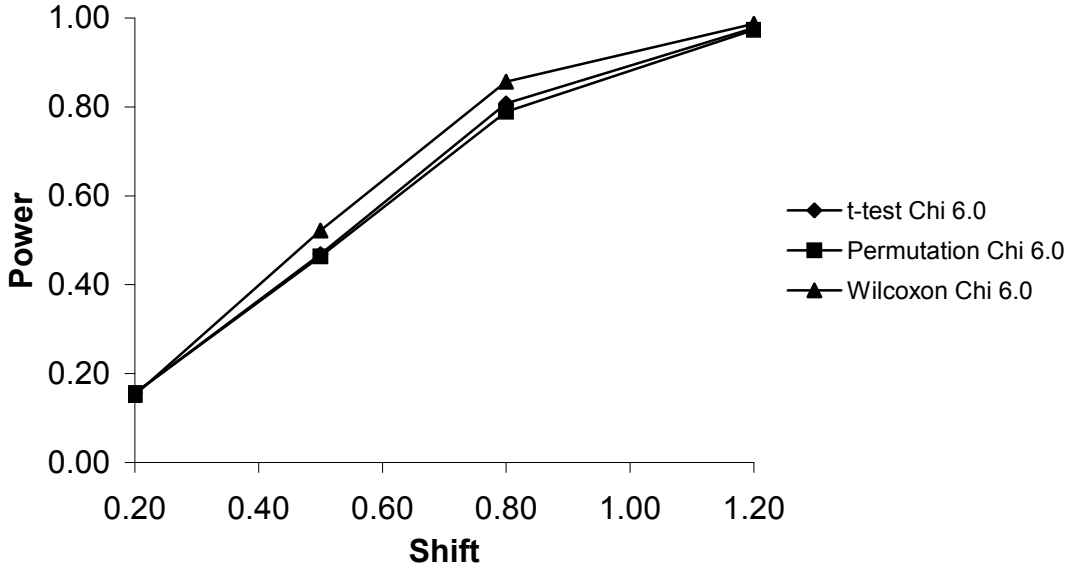



Figure 36: Shift vs. Power in the Chi-Square Distribution (df = 6) for Sample Size $n_1 = n_2 = 20$

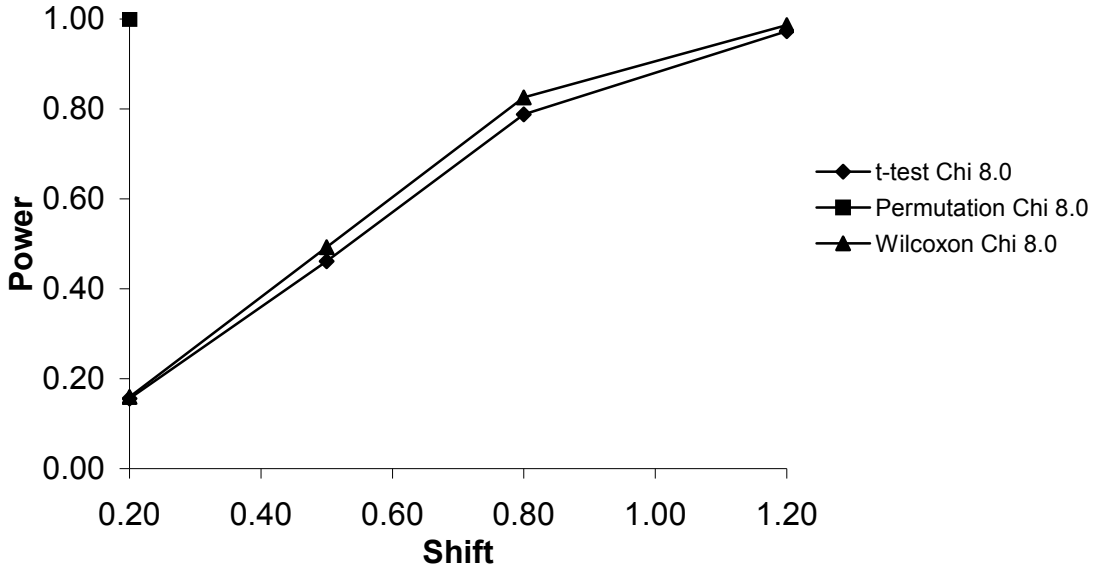



Figure 37: Shift vs. Power in the Chi-Square Distribution (df = 8) for Sample Size $n_1 = n_2 = 20$

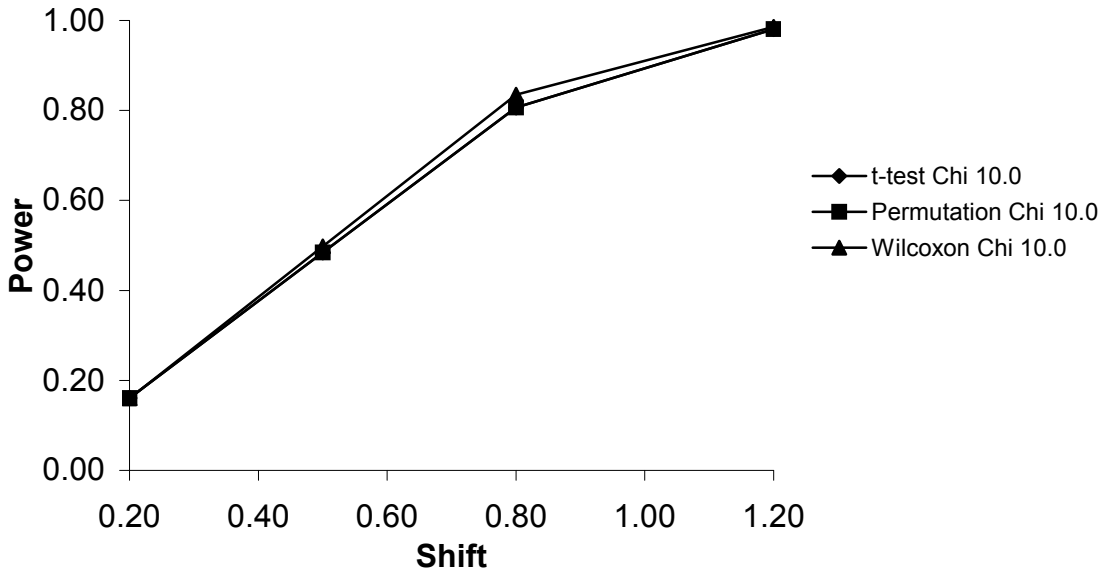



Figure 38: Shift vs. Power in the Chi-Square Distribution (df = 10) for Sample Size $n_1 = n_2 = 20$

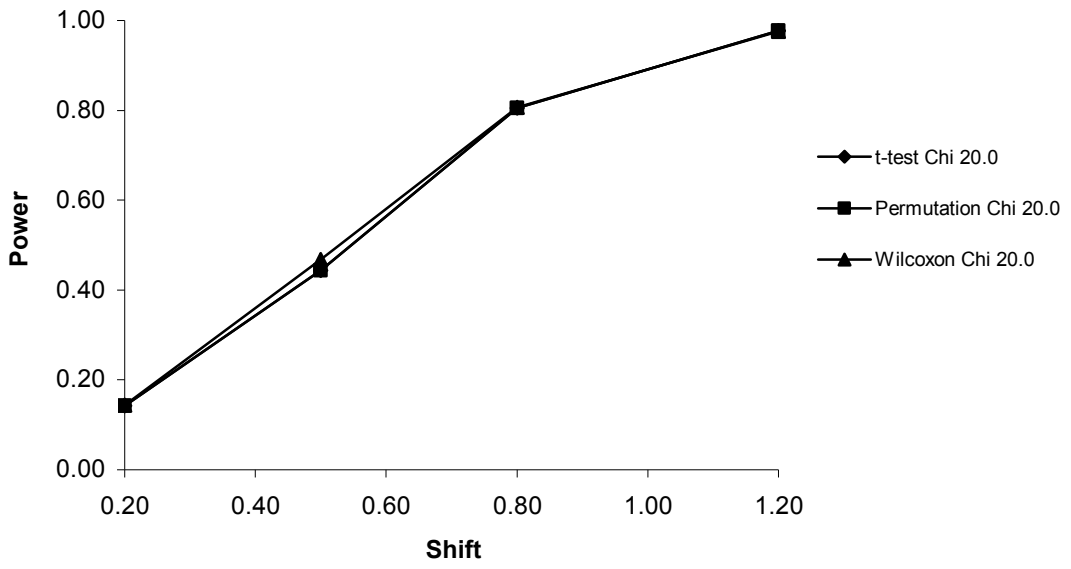



Figure 39: Shift vs. Power in the Chi-Square Distribution (df = 20) for Sample Size $n_1 = n_2 = 20$

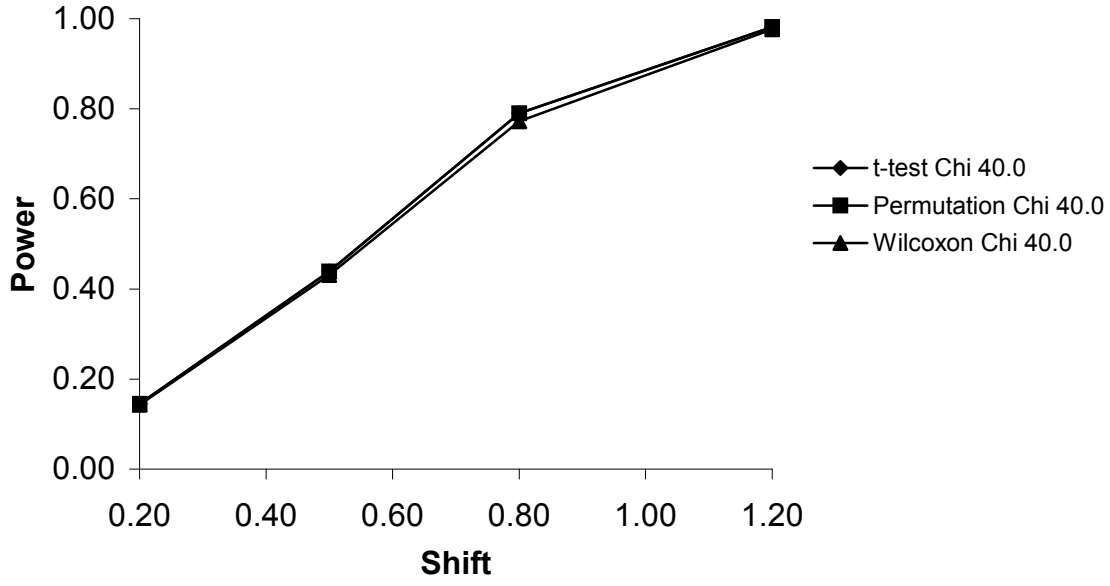



Figure 40: Shift vs. Power in the Chi-Square Distribution (df = 40) for Sample Size $n_1 = n_2 = 20$

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 456-468
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Tests for 2 × 2 Tables in Clinical Trials

Vic Hasselblad Yuliya Lokhnygina
Duke University

Five standard tests are compared: chi-squared, Fisher's exact, Yates' correction, Fisher's exact mid-p, and Barnard's. Yates' is always inferior to Fisher's exact. Fisher's exact is so conservative that one should look for alternatives. For certain sample sizes, Fisher's mid-p or Barnard's test maintain the nominal alpha and have superior power.

Key words: Power, sample size, dichotomous endpoint, alpha level

## Introduction

The literature on tests for 2×2 tables is extremely vast and controversial. However, the issues can be focused somewhat when considering the use of these tests for clinical trials. In this situation, the trials have two arms and the sample size of each arm is fixed. Tests are almost always made at the 0.05 nominal alpha level. There is no requirement that the tests be computationally simple, but only that they are available in standard commercial statistical software. The following two examples illustrate many of the issues of interest.

Cotter et al (2000) conducted a small, randomized pilot study (15 patients per treatment arm) comparing $N^{\omega}$-nitro-L-arginine methyl ester (L-NAME) to placebo in patients with cardiogenic shock. Mortality results are given in Table 1.

Cotter et al. reported a p-value of 0.028 (no test specified), a value which is consistent with the standard chi-square test. However, if Fisher's exact test had been used in the standard manner, the p-value would have been 0.0656. If Fisher's mid-p or Barnard's test had been used, then the p-value would have been 0.0374 or 0.0352, respectively. The results of this trial, along with other preliminary data, were suggestive of an effect, and so a second study, SHOCK II (Dzavik et al., submitted), was conducted. Ironically, the SHOCK II Trial showed no evidence of a treatment effect, but there were significant differences between the SHOCK II Trial and the Cotter Trial.

The second example is taken from the A to Z Trial (Blazing et al., 2004). This trial compared enoxaparin with un-fractionated heparin for the treatment of 3905 patients with acute coronary syndrome (ACS). Based on other studies, there was a concern that enoxaparin might lead to an increase in the number of bleeding events. Given in Table 2 are the counts of patients with TIMI major bleeding events by treatment arm.

Note that the bleeding rates are quite low in both arms (less than one percent). The Statistical Analysis Plan specified that "Statistical comparison will be conducted using Fisher's exact test …"

Vic Hasselblad earned a Ph. D. in Biostatistics in 1967 from the University of California at Los Angeles. An expert in the area of synthesis of evidence (meta-analysis), he coauthored a book and more than 20 peer-reviewed articles. E-mail: victor.hasselblad@duke.edu. Yuliya Lokhnygina is Assistant Professor of Biostatistics at Duke University. Her research interests include statistical methods in clinical trials, survival analysis, causal inference in observational studies and dynamic (adaptive) treatment strategies. E-mail: yuliya.lokhnygina@duke.edu

Table 1. Deaths by Treatment Arm in the L-NAME Trial

| | L-NAME | No L-NAME | Total |
|---|---|---|---|
| Died | 4 | 10 | 14 |
| Survived | 11 | 5 | 16 |
| Total | 15 | 15 | 30 |

Table 2. Bleeding Events by Treatment Arm for the A to Z Trial

| | Enoxaparin | Un-fractionated Heparin | Total |
|---|---|---|---|
| Bleed | 18 | 8 | 26 |
| No bleed | 1922 | 1957 | 3879 |
| Total | 1940 | 1965 | 3905 |

In this case, Fisher's exact test gives a p-value between 0.0285 and 0.0501. The problem with Fisher's exact test is that it is .0393 or 0.0352, respectively. It is clear that summarizing the results of the above table as non-significant would not accurately describe the information.

These two examples point out some of the difficulties in choosing a statistical test in the simplest of trials, namely the two-arm dichotomous trials. There are several possible tests that can be used and they have different implications for both the nominal alpha level as well as the power. We will restrict our consideration to those tests available in commercial software packages such as SAS® (SAS Institute, 1999) or StatXact (StatXAct with Cytel Studio, 2005).

## Methodology

Assume a study where the number of positives and negatives are measured for a control group and a treated group, and that the results are summarized in a standard 2 x 2 contingency table where *A, B, C,* and *D* are the observed counts. Let $T = A + B + C + D$. The rate in the treated group, $p_1$, is estimated by $A / N_1$ and the rate in the control group, $p_2$, is estimated by $C / N_2$. The null hypothesis is that $p_1 = p_2$ and the usual alternative hypothesis is

| | Treated | Control | Total |
|---|---|---|---|
| Positive | $A$ | $C$ | $S_1$ |
| Negative | $B$ | $D$ | $S_2$ |
| Total | $N_1$ | $N_2$ | $T$ |

that $p_1 \neq p_2$. The object is to find a test statistic which is a function of $A$, $B$, $C$, and $D$ such that the value of the test is very different when $p_1 = p_2$ as compared with when $p_1 \neq p_2$. There are several test statistics which could be used to test the null hypothesis, and the properties of five such tests will be investigated: 1) the uncorrected chi-squared test, 2) Fisher's exact test, 3) Yates' correction to the chi-squared test, 4) Fisher's exact mid-p test, and 5) Barnard's test.

Uncorrected chi-squared test

The standard uncorrected chi-squared statistic (Pearson, 1900) is:

$$CS = \frac{T(AD - BC)^2}{N_1 N_2 S_1 S_2}. \qquad (1)$$

For an intended α-level of 0.05, the test rejects the null hypothesis whenever $CS > 3.8415$ and accepts otherwise. The power of the test is the probability that $CS > 3.8415$ given particular values of $p_1$, $p_2$, $N_1$, and $N_2$.

Fisher's exact test

In 1925, Fisher (1925) gave an exact test which requires a bit more effort to compute. The test is based on the hyper geometric distribution. Assume that the four marginal totals, $N_1$, $N_2$, $S_1$, and $S_2$, are fixed. Under the null hypothesis, the probability that $A = i$ for $i = 0, 1, \ldots, \min(N_1, S_1)$ is:

$$Prob(A|N_1, N_2, S_1, S_2) = \binom{S_1}{A}\binom{S_2}{N_1 - A} \Big/ \binom{T}{N_1}. \qquad (2)$$

The (two sided) probability of an observed or more extreme than observed result is given by

$$\sum_{Prob(i|N_1,N_2,S_1,S_2) < Prob(A|N_1,N_2,S_1,S_2)} Prob(i|N_1, N_2, S_1, S_2) + Prob(A|N_1, N_2, S_1, S_2) \qquad (3)$$

For example, the values for Cotter et al (2000) are 0.0092 + 0.0564 = 0.0656. The two values, 0.0092 and 0.0656, are the only two reasonable values for the size of Fisher's exact test in this particular case (see Kendall and Stuart, Vol. 2, pp. 553, 1961). A non-randomized test cannot be constructed at any arbitrary level. But by convention, the largest value, 0.0656, is often taken as the p-value from the test. This value is often described as conservative, but it is only conservative if the object is to reject the null hypothesis. Thus, the null hypothesis would not be rejected at the 0.05 level using this test in this particular manner. The test could be made exact by choosing a random number between the values of 0.0092 and 0.0656 as the p-value. However, using randomization as part of the hypothesis testing procedure has never been accepted in clinical literature. This example demonstrates that using a conservative test is not necessarily a conservative strategy when the endpoint in question is a safety endpoint.

Yates' corrected chi-square test

The third test is Yates' (1934) correction to the Pearson chi-squared statistic:

$$CSC = \frac{T(|AD - BC| - T/2)^2}{N_1 N_2 S_1 S_2}. \qquad (4)$$

This correction is designed to make the chi-squared statistic give a p-value which is often very close to the p-values calculated from Fisher's exact test.

Fisher's mid-p test

The fourth test is a modification of Fisher's exact test, known as Fisher's mid-p value, as defined by Lancaster (1961). The calculations are made exactly as those done for Fisher's exact test, except that the probability of a result more extreme is averaged with the probability of a result as extreme or more so. In the Cotter et al. (2000) example, this would be (0.0092 + 0.0656)/2 = 0.0374. StatXact (2005) and LogXact (LogXact with Cytel Studio, 2005) report mid-p values as part of their output.

Barnard's test

Barnard (1947) proposed an unconditional exact test based on a minimax elimination of the nuisance parameter. The reference set was defined to be the set of all 2 x 2 tables with fixed row margins and all possible column margins. Because the reference set for Barnard's test does not fix the column margins, the distribution of the test statistic is less discrete than would be obtained by permuting the conditional reference set in which both margins are fixed. However, Barnard was not satisfied with his test, and disavowed it two years later (Barnard, 1949). There is an interesting discussion by Barnard of the reasons for his disavowal in Yates (1984, with discussion). Barnard invoked Fisher's principle of ancillarity (see Fisher, 1973, Chapter IV), whereby inference should be based on hypothetical repetitions of the original experiment, fixing those aspects of the experiment that are unrelated to the hypothesis under test. Little (1989) gives a clear discussion of this topic. In two more recent publications, Barnard (1989, 1990) provided additional arguments against the test. However, Little (1989) showed that the row totals are not ancillary statistics.

If the true value of p was known under the null hypothesis ($p_1 = p_2 = p$), then the probability of any possible outcome could be calculated, e.g. the probability of $x_1$ events in the first arm (of size $N_1$), and $x_2$ events in the second arm (of size $N_2$):

$$\Pr(x_1, x_2) = \binom{N_1}{x_1} p^{x_1} (1-p)^{N_1 - x_1} \binom{N_2}{x_2} p^{x_2} (1-p)^{N_2 - x_2} \qquad (5)$$

Next, order the outcomes. One possible ordering would be to use the D statistic:

$$D = \frac{\frac{x_2}{N_2} - \frac{x_1}{N_1}}{\sqrt{\left(\frac{x_1 + x_2}{N_1 + N_2}\right)\left(\frac{N_1 + N_2 - x_1 - x_2}{N_1 + N_2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \qquad (6)$$

Using this ordering, the probabilities can be found of all tables at least as extreme, or more so, than the observed table for a given $p$. The sum of all these probabilities is the p-value associated with the specified $p$. Calculate this p-value for all possible specified $p$'s and take their maximum. This is Barnard's p-value. A plot of the extreme values as a function of $p$ for the Cotter et al (2000) example is in Figure 1.

Note that the statistic reaches a maximum of 0.0352, and this is Barnard's p-value for the Cotter et al study (2000). Barnard's test is actually guaranteed to be conservative for certain specific sample sizes. The reason that the test is not always conservative is that it uses a normal approximation to order the outcomes.

Power Formulas

The formula for the probability of rejection for any test of equality of proportions is given by:

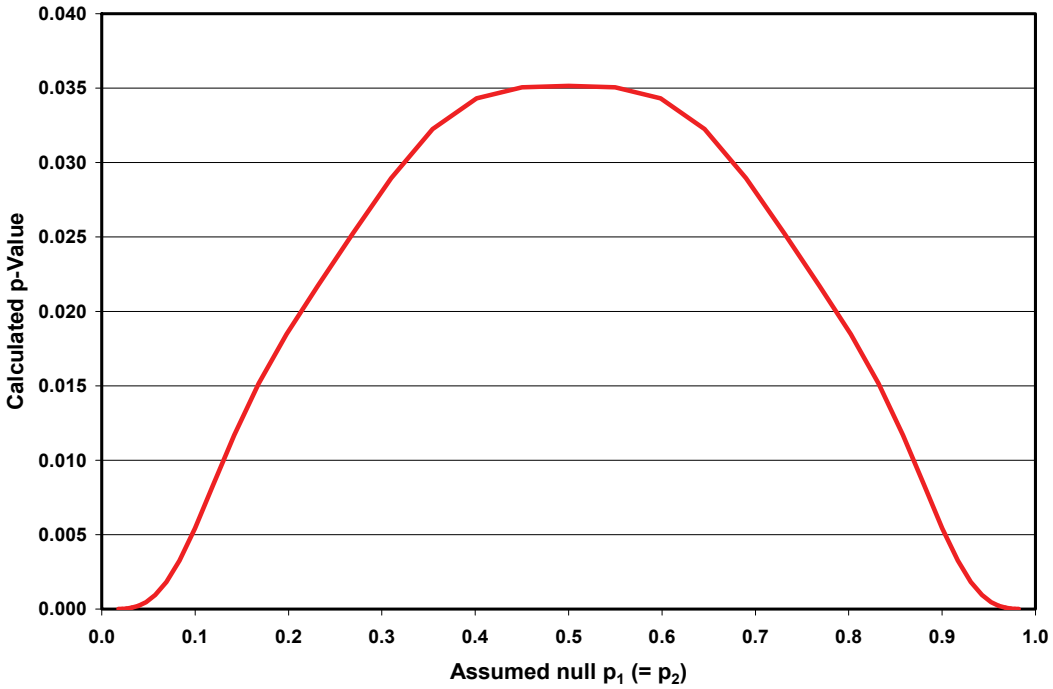



Figure 1. Calculation of Barnard's Statistic for Specified Null Probabilities ($p_1$)

$$\Pr[reject] = \sum_{i=0}^{N_1}\sum_{j=0}^{N_2} \delta_{ij} \binom{N_1}{i}\binom{N_2}{j} p_1^i (1-p_1)^{N_1-i} p_2^j (1-p_2)^{N_2-j} \quad (7)$$

where $N_1$ and $N_2$ are the sample sizes of the two arms respectively, where $p_1$ and $p_2$ are the true event rates in each arm, and where $\delta_{ij}$ is one if the test statistic based on $i$, $N_1$, $j$, $N_2$ is statistically significant, and zero otherwise.

This formula can be used to determine either the nominal alpha level for a given test (by assuming that $p_1$ equals $p_2$) or to determine the power (by not assuming equality). The formula is an exact one – no simulations are necessary. All results presented in the next section are exact calculations.

## Results

The actual alpha-levels are calculated for all five tests assuming that the intended alpha-level was 0.05 and $N_1 = N_2 = 25$, $N_1 = N_2 = 50$, $N_1 = N_2 = 100$, and $N_1 = 25$, $N_2 = 50$. The calculations were made for the entire range of $p_1$ *(with* $p_2 = p_1$*)* and these are shown in Figures 2, 3, 4 and 5.

Note that the actual alpha-levels for the standard chi-square, Fisher's mid-p, and Barnard's tests are reasonably close to the intended alpha-level for $0.2 < p_1 < 0.8$. The maximum actual alpha-level for any test never exceeds .065 for any $p_1$. Note also that Fisher's exact test has very low alpha-levels. The maximum alpha-level for Fisher's exact test for $N_1 = N_2 = 25$ is 0.0328. Yates' correction to the standard chi-square test yields alpha levels as low as or lower than Fisher's exact test. Fisher's mid-p test falls below the nominal alpha level of 0.05 everywhere, but is uniformly larger than either Fisher's exact or Yates' correction. Barnard's test is as large, or larger, than Fisher's mid-p, but it does exceed 0.05 for event rates between 0.107 and 0.172 and between 0.828 and 0.893.

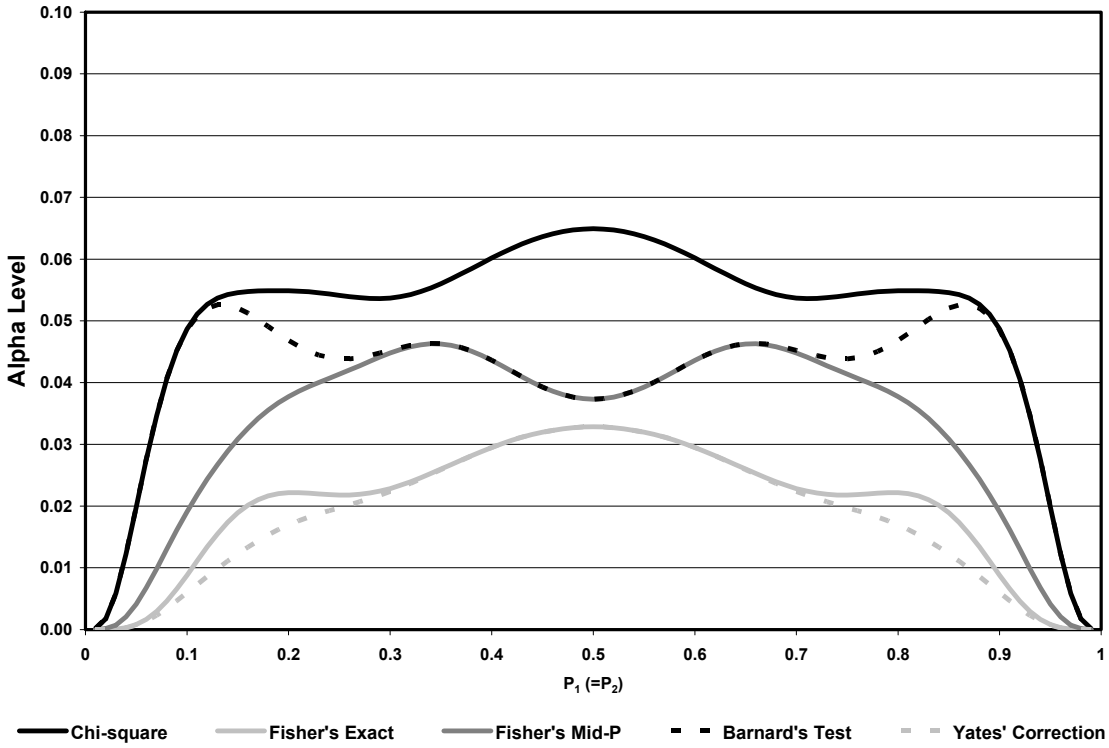



Figure 2. Alpha Levels for Two Arm Dichotomous Tests for $N_1 = N_2 = 25$

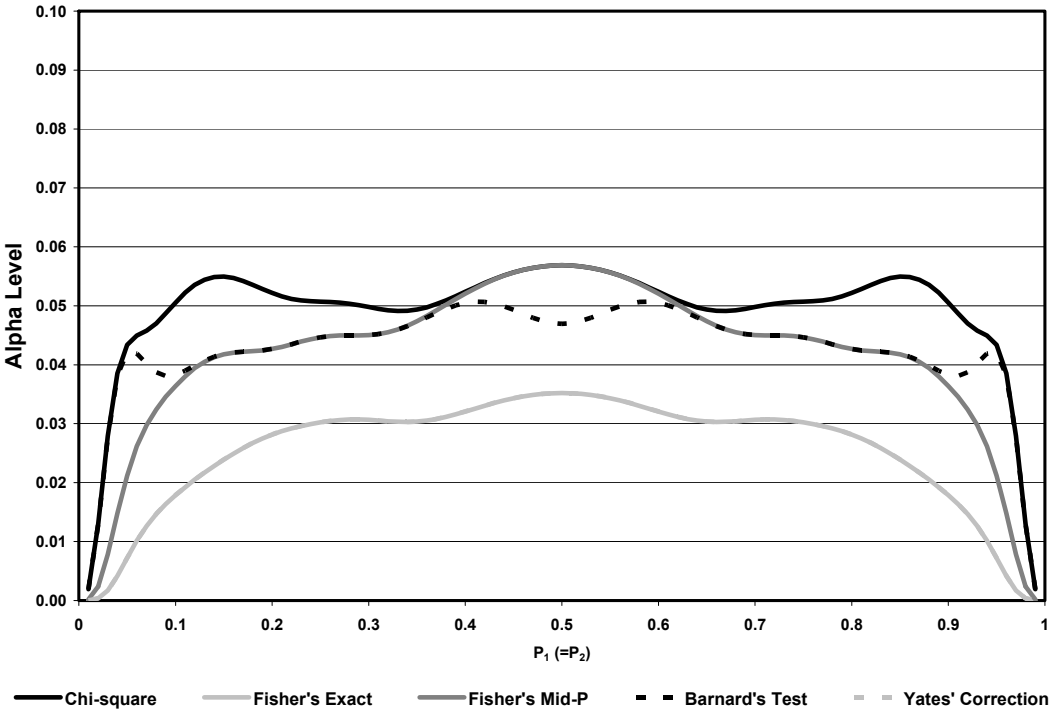



Figure 3. Alpha Levels for Two Arm Dichotomous Tests for $N_1 = N_2 = 50$

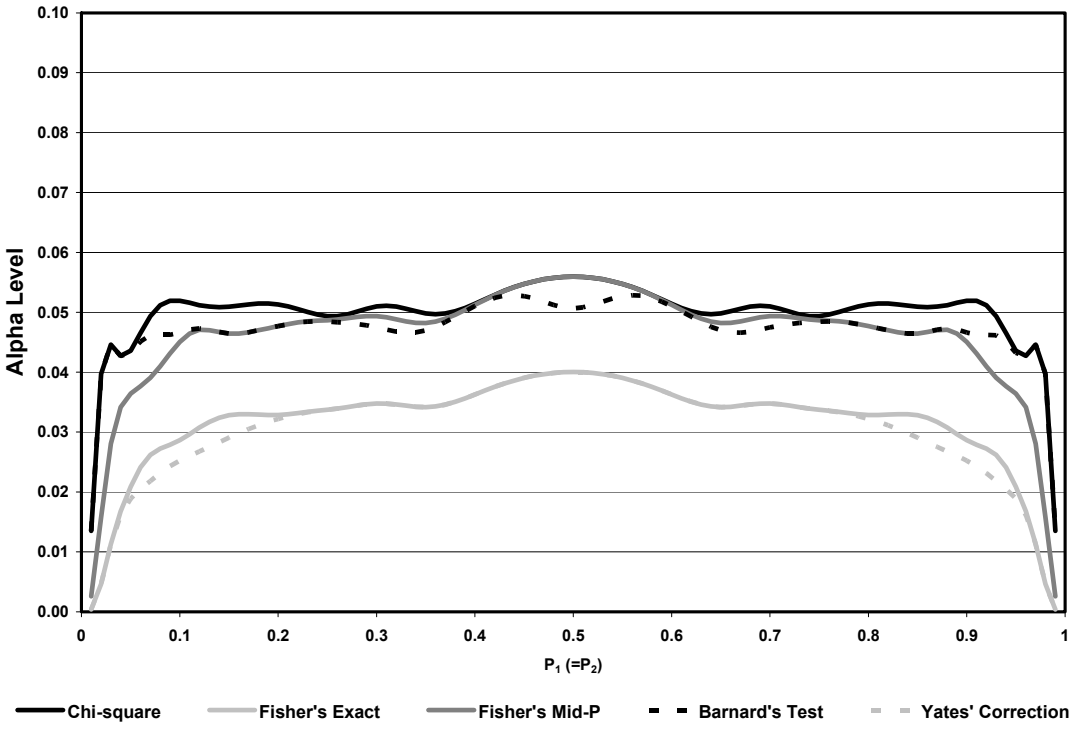



Figure 4. Alpha Levels for Two Arm Dichotomous Tests for $N_1 = 100$, $N_2 = 100$

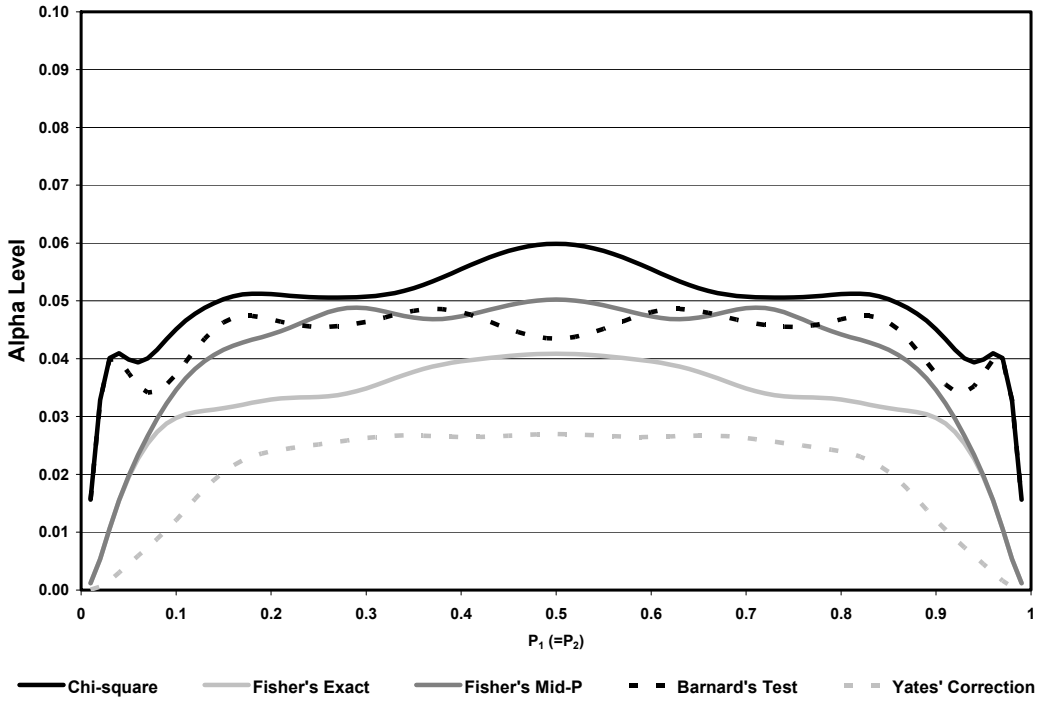



Figure 5. Alpha Levels for Two Arm Dichotomous Tests for $N_1 = 25$, $N_2 = 50$

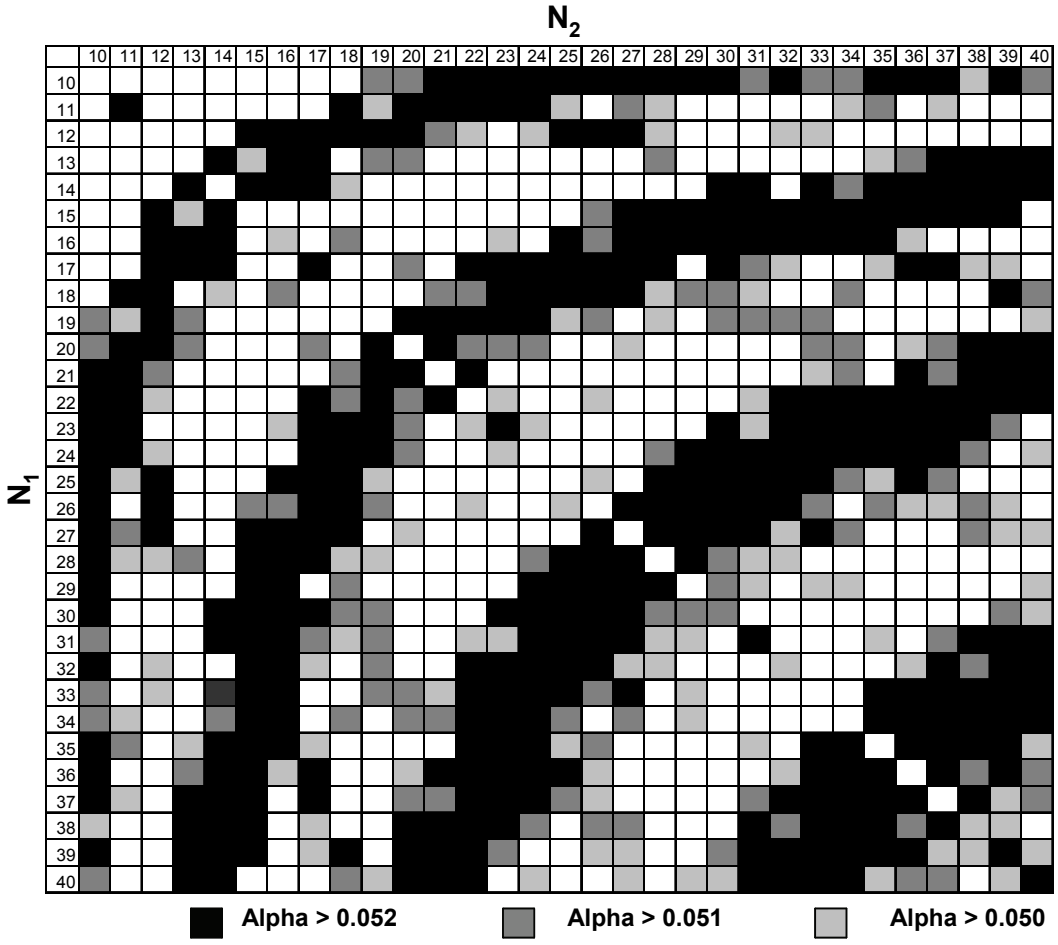



Figure 6. Diagram of Maximum Alpha Levels for Fisher's Mid-P Test with Nominal Alpha Level of 0.05.

For sample sizes of 50 per arm, the actual alpha-levels for the standard chi-square, Fisher's mid-p, and Barnard's tests approach the nominal alpha-level for $0.2 < p_1 < 0.8$. The maximum actual alpha-level for any test never exceeds .057 for any $p_1$. Fisher's exact test still has very low alpha-levels, falling below 0.035 everywhere. Fisher's mid-p test remains below the nominal alpha level of 0.05 for event rates below 0.3, but does reach a maximum of 0.057. Barnard's test never exceeds 0.0507, and is generally closer to 0.05 than any of the other tests.

For sample sizes of 100 per arm, the actual alpha-levels for the standard chi-square, Fisher's mid-p, and Barnard's tests approach the nominal alpha-level for $0.1 < p_1 < 0.9$. The maximum actual alpha-level for any test never exceeds .056 for any $p_1$. Fisher's exact test is increased, but still falls below 0.040 everywhere. Fisher's mid-p test falls below the nominal alpha level of 0.05 for event rates below 0.3, but does reach a maximum of 0.056. Barnard's test never exceeds 0.053, and is generally closer to 0.05 than any of the other tests.

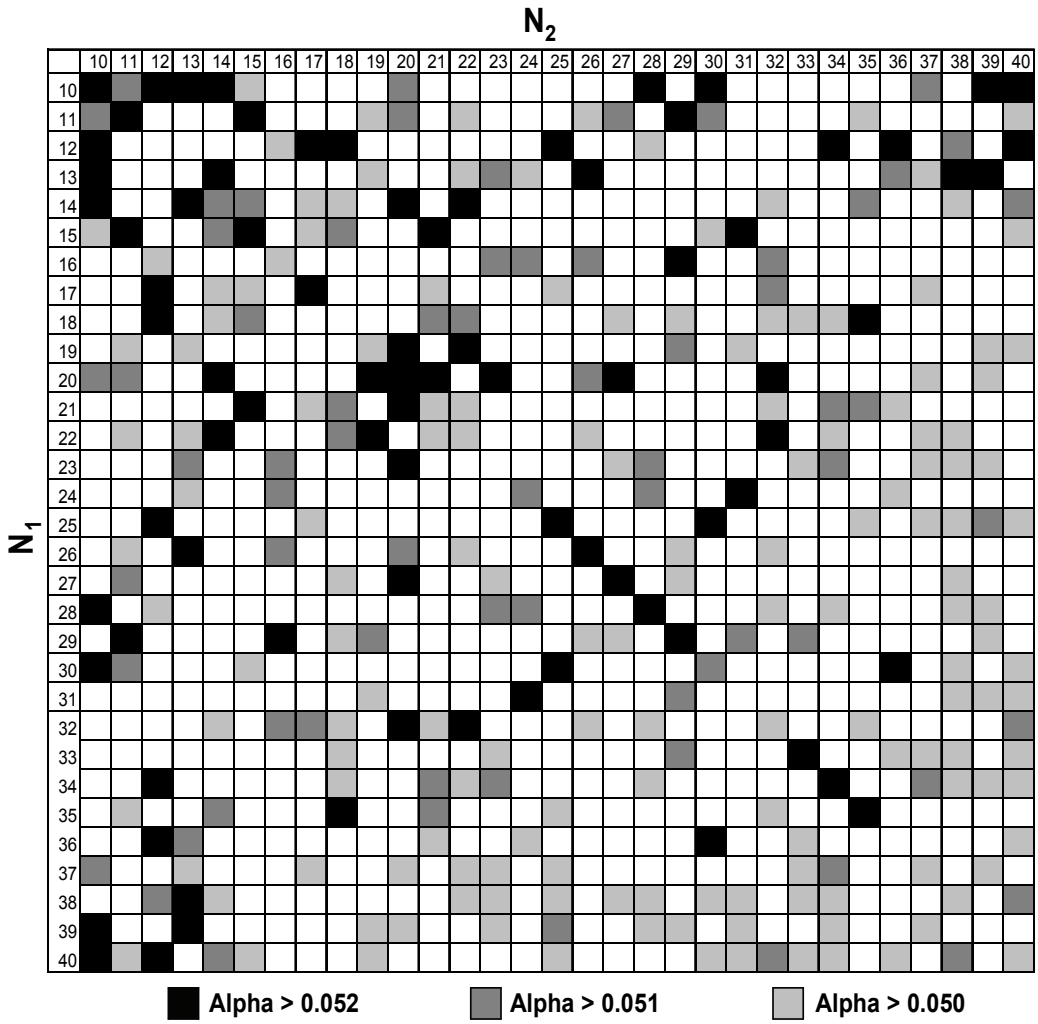



Figure 7. Diagram of Maximum Alpha Levels for Barnard's Test with Nominal Alpha Level of 0.05.

Fisher's exact test had alpha levels a bit closer to that of the other tests, but Yates' correction had very low alpha levels, achieving a maximum of 0.0270.

For unequal samples of 25 and 50 per arm, the results were somewhat similar to the previous results. Barnard's test had a maximum alpha level of 0.0484 and Fisher's mid-p test had a maximum alpha level of 0.0503. However, the chi-square test had a maximum of 0.0599.

The results from Figures 2 – 5 are consistent with the results presented by Hasselblad and Allen (2003). Their results suggested that an expected number of events of approximately 40 is required to insure that the actual alpha level for the chi-square test is between 0.049 and 0.051 when the intended alpha level is 0.05.

Fisher's mid-p and Barnard's tests are examined in greater detail. Specifically, the interest is to determine if those tests were conservative for all values of $p_1$ *(with* $p_2 = p_1$*)* for specific values of $N_1$ and $N_2$. The results for Fisher's mid-p for $N_1 = 10, \ldots, 40$ and $N_2 = 10, \ldots, 40$ are in Figure 6. Those squares which are white correspond to an actual alpha level less

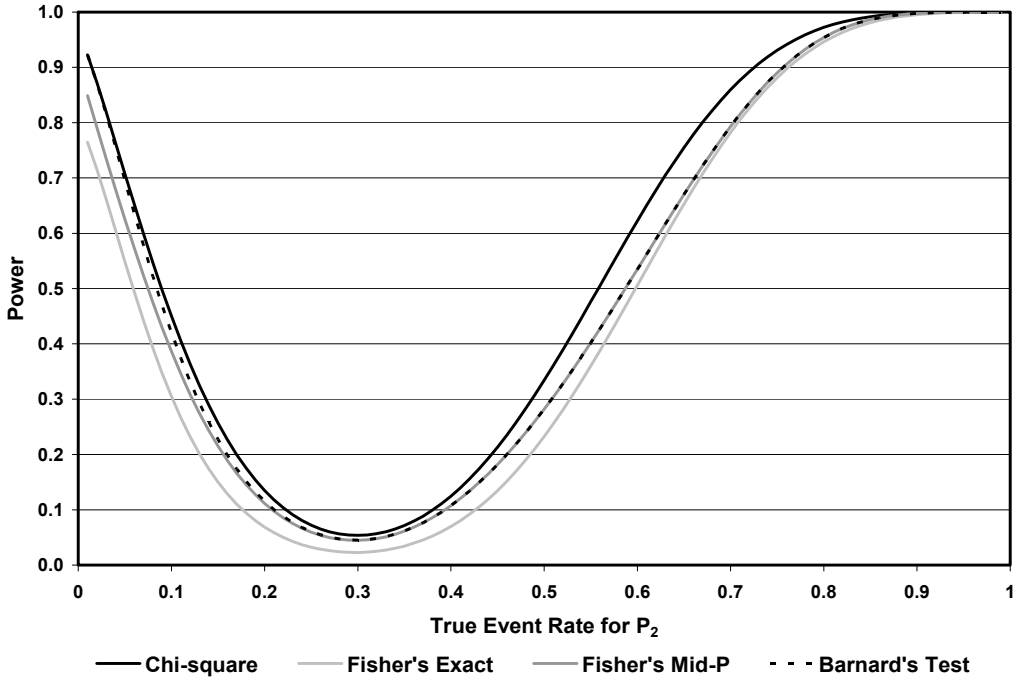



Figure 8. Power of Various Tests for Sample Sizes of 25 Per Arm

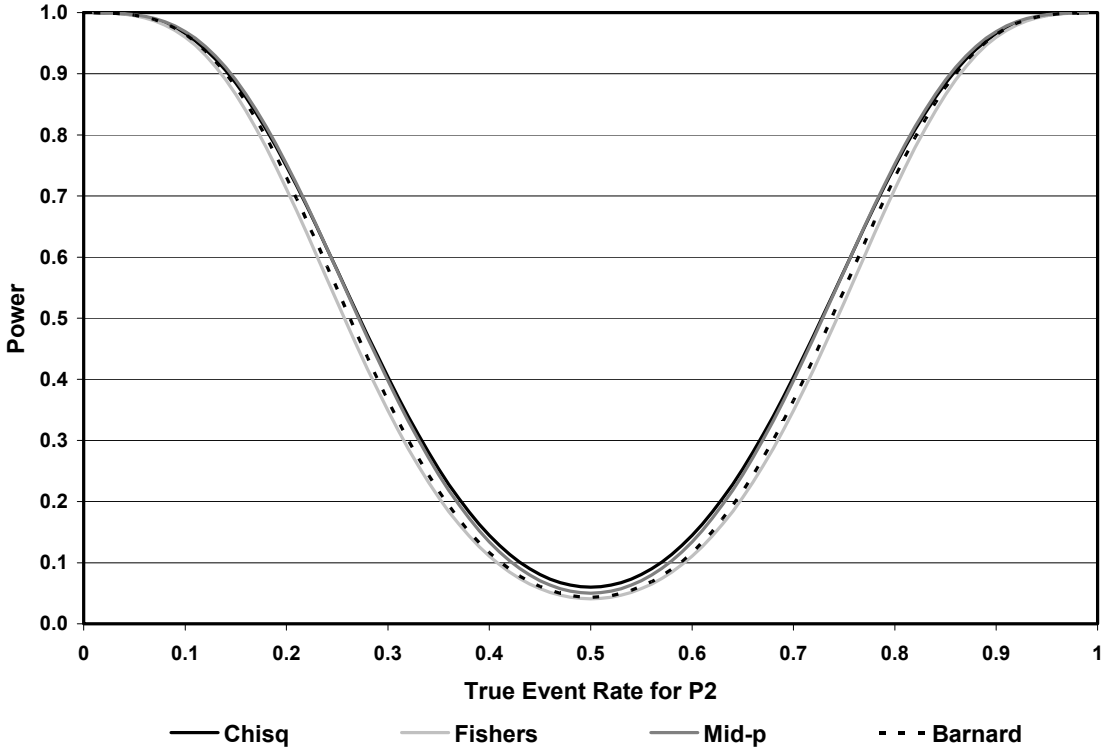



Figure 9. Power of Various Tests for Sample Sizes of 25 in the Control Arm
And 50 in the Treated Arm

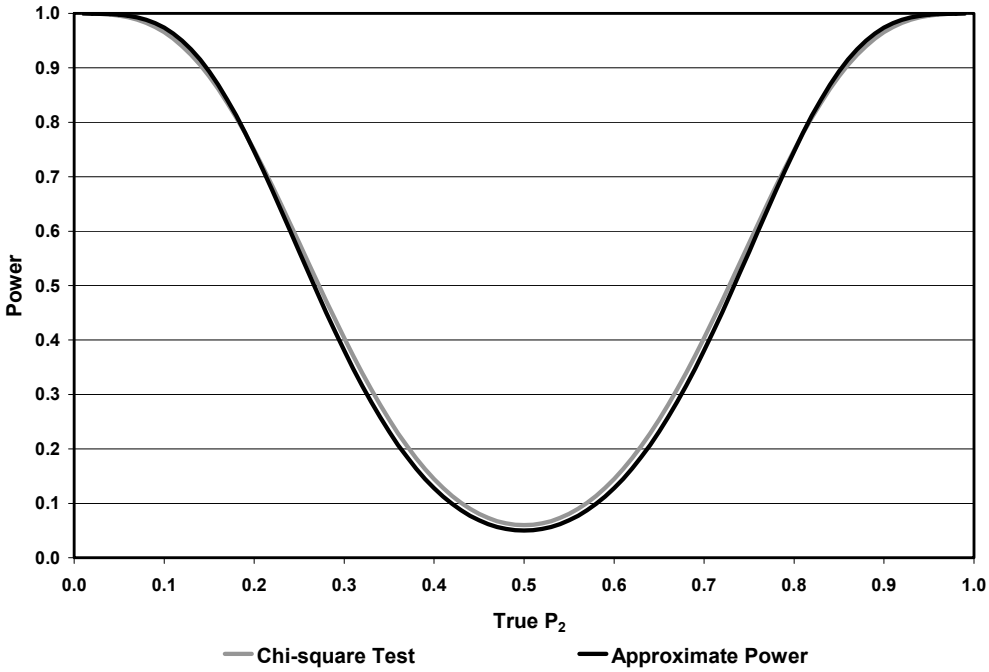



Figure 10. Estimated Power Using Approximation versus Actual Power of the Chi-square Test for Sample Sizes of 25 in the Control Arm and 50 in the Treated Arm

than 0.05 for all p.

For example, if one sample size is 15 and the other is 15, 16,…, or 25, then Fisher's mid-p test is conservative. On the other hand, if both sample sizes are 26, then the test may not be conservative, depending on the true null rate. However, for most null rates, the test will still be conservative. Figure 6 only shows the worst possible case. Of the 496 different sample size combinations shown in Figure 6, 40.9 percent had a nominal alpha level less than 0.05.

The results for Barnard's test for $N_1$ = 10, … ,40 and $N_2$ = 10, … ,40 are in Figure 7. For example, if one sample size is 25 and the other is 18, 19,…, or 24, then Barnard's test is conservative. On the other hand, if both sample sizes are 25, then the test may not be conservative, depending on the true null rate. Of the 496 different sample size combinations shown in Figure 7, 66.5 percent had a nominal alpha level less than 0.05.

The power for four of the tests described previously was calculated for $N_1 = N_2 = 25$ and $p_1 = 0.3$ (Yates' test was dropped to make the graph more readable). The results are in Figure 8.

Note that the power curves behave as expected, that is, they reach a minimum at $p_1 = p_2 = 0.3$ and then increase rapidly as $p_2$ moves away from $p_1$. The shapes of the power curves are all quite similar. The differences at $p_1 = p_2 = 0.3$ are exactly the differences in the alpha-levels of the tests. The power curves show one other key point – the tests do not cross each other. That is, if a test has a lower nominal alpha level, then it will have lower power for the alternatives.

The power for four of the tests was also calculated for $N_1 = 25$, $N_2 = 50$ and $p_1 = 0.5$. The results are in Figure 9.

Figure 9 shows the same general patterns as did Figure 8.

There are approximate formulas for power that are reasonably accurate. One formula given by Fleiss (1981, p. 27) is

$$\beta = \Phi\left(\left(c_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}-(p_2-p_1)\right)\Big/\sqrt{\frac{p_1(1-p_1)}{N_1}+\frac{p_2(1-p_2)}{N_2}}\right) \quad (8)$$

where $\bar{p} = (p_1 + p_2)/2$ and $\Phi$ is the cumulative normal distribution function. This approximate function is shown in Figure 10, where it is drawn as a function of $p_2$. The exact and approximate formulas are reasonably similar, and they get closer as the sample size increases. There are several other formulas that have various correction formulas in order to make the approximation better. There is, however, a limit to the accuracy of these approximations because they are not based on the test statistic itself.

## Conclusion

There are some conclusions which can be made as a result of the calculations presented:

- Even though Fleiss (1981, p. 27) states that "[Yates'] correction should always be used", the test is always inferior to (its nominal alpha level is less than or equal to) Fisher's exact test, and for that reason it should not be used.

- Fisher's exact test is so conservative that one should always look for an alternative even if one requires that the alpha level of the test not exceed the nominal level (by even the smallest amount). For certain sample sizes, either Fisher's mid-p or Barnard's test will satisfy the requirement, and those tests have much superior power. For example, knowing that the test is conservative when both arms have 15 observations, the data of Cotter et al. (2000) could have been analyzed using Fisher's mid-p test.

- For tests of safety, being conservative is not desirable. Because event rates are often very low for safety issues, Fisher's mid-p test is a very appealing alternative. For example, the maximal nominal alpha level for this test for the A to Z bleeding data is 0.05007 (assuming that the true event rates are less than 20 percent).

- The chi-square test works adequately for very large sample sizes, but the standard rule of an expected minimum value of 5 (which is commonly used) is not acceptable. Even if the expected number of counts exceeds 40 per cell, the alpha level (for a nominal alpha level of 0.05) is approximately bounded by 0.049 and 0.051. Barnard's test is certainly an attractive alternative in the moderate sample size situation when the event rates are not especially small.

As mentioned previously, only tests available in widely used commercial software packages were considered. Such restrictions leave out some recently developed unconditional tests for which no commercially developed and tested software is available. An example is a test based on the confidence interval p-value developed by Berger and Boos (1994, 1996). This test can be seen as a modification of Barnard's test. Although Barnard's p-value is obtained by maximizing the p-value for given nuisance parameter $p$ over the unit interval, the p-value of the test by Berger and Boos is obtained as a sum of the supremum of p-values over the 100(1-β)% confidence interval for $p$ calculated from the data and β. This test can be more powerful then Barnard's and requires less computational effort.

## Acknowledgements

The authors wish to thank Paula Smith for manuscript editing and formatting assistance.

## References

Barnard, G.A. (1947). Significance tests for 2 x 2 tables. *Biometrika*, *34(1-2)*, 123-138.

Barnard, G.A. (1949). Statistical inference. *Journal of the Royal Statistical Society*, *Series B*, 11, 115-139.

Barnard, G.A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine 8*, 1469-1477.

Barnard, G.A. (1990). Must clinical trials be large? The interpretation of p-values and the combination of test results. *Statistics in Medicine*, *9*, 601-614.

Berger, R.L. & Boos, D.D. (1994). P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, *89*, 1012-1016.

Berger, R.L. (1996). More powerful tests from confidence interval p-values. *The American Statistician*, *50*, 314-318.

Blazing, M.A., de Lemos, J.A., White, H.D., Fox, K.A.A., Verheugt, F.W.A., Ardissino, D., DiBattiste, P.M., Palmisano, J., Bilheimer, D.W., Snapinn, S.M., Ramsey, K.E., Gardner, L.H., Hasselblad, V., Pfeffer, M.A., Lewis, E.F., Braunwald, E., & Califf, R.M., for the A to Z Investigators (2004). Safety and efficacy of enoxaparin vs unfractionated heparin in patients with non–ST-segment elevation acute coronary syndromes who receive tirofiban and aspirin: a randomized controlled trial. *Journal of the American Medical Association, 292*, 55-64.

Cotter, G., Kaluski, E., Blatt, A., Milovanov, O., Moshkovitz, Y., & Zaidenstein, R. (2000). L-NMMA (a nitric oxide synthase inhibitor) is effective in the treatment of cardiogenic shock. *Circulation*, *101(12)*, 1358-1361.

Cytel Inc. (2005). *StatXact 7 with cytel studio statistical software for exact nonparametric inference*. Cambridge, MA.

Cytel Inc. (2005). *LogXact 7 with Cytel Studio. Discrete Regression Software Featuring Exact Methods*. Cambridge, MA.

Fisher, R.A. (1973). *Statistical Methods and Scientific Inference, Third Edition*. London: Collier Macmillan Publishers.

Fisher, R.A. (1925). *Statistical methods for research writers*. Edinburgh: Oliver and Boyd.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons.

Hasselblad, V., & Allen, A.S. (2003). Power calculations for large multi-arm placebo-controlled studies of dichotomous outcomes. *Statistics in Medicine*, *22*, 1943-1954.

Kendall, M.G., & Stuart, A. (1961). *The advanced theory of statistics vol. 2*. London: Charles Griffin & Company Limited.

Lancaster, H.O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, *56*, 223-234.

Little, R.J.A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician*, *43*, 283-288.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50(5),* 157-172.

SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8*. Cary, NC.

Yates, F. (1934). Contingency tables involving small numbers and the $_x2$ test. *Journal of the Royal Statistical Society*, *(Supp 1)*, 217-235.

Yates, F. (1984). Tests of significance for 2 x 2 contingency tables (with discussion). *Journal of the Royal Statistical Society, Series A*, *147*, 426-463.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 469- 475
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Sensitivity Curves for Asymmetric Trimming Hinge Estimators

D. B. Stark
The University of Akron

J. F. Reed III
Lehigh Valley Hospital and Health Services

Robust estimators have been developed and tested for symmetric distributions via simulation studies. The primary objective was to show that they are more efficient than the sample mean when used in conjunction with asymmetric distributions. Little attention has been given to how they perform on data that are from asymmetric distributions, or from distributions that have inherent anomalies (messy data). Thus, the behavior of hinge estimators using sensitivity curve are examined.

Key words: Robust estimators, adaptive estimators, ancillary statistics, selector statistics.

## Introduction

In spite of the considerable bad press that the sample mean (the least square estimator of μ) has received, the standard normal theory statistic performs well when real data are nearly normal. However, robust estimators of location have been and continue to be developed and tested for symmetric long and short-tailed distributions by means of extensive simulation studies. The ancestry of these estimators may be traced to the work of Hogg (1967) and the Princeton Robust Study (Andrews, et al., 1972). The objective of these and other studies was to demonstrate that adaptive location estimators would be more efficient than the sample mean. In general, an adaptive procedure may be characterized as an application approach to data analysis rather than a theoretical application.

The objective is to supplement previous simulation studies of robust estimators, hinge estimators, by examining the behavior of these adaptive location estimators using sensitivity curves first developed by the Princeton Robust Study

David B. Stark is Associate Professor of Statistics. His research interests are experimental design, linear models, mathematical statistics, regression, robust statistics, and topology. Email: dstark@uakron.edu James Reed III is Interim Chief of Health Studies and Director of Research at Lehigh Valley Hospital & Health Network. His interests include applied statistical analyses, medical education, and statistical methods in simulation studies. Email him at: James_F.Reed@lvh.com

## Selector Statistics and Adaptive Location Estimators

Characteristics such as skewness, tail length, and peakedness describe the distribution characteristics. In defining tail length and skewness, the notation here is from Hogg (1967, 1982). Define $L_\alpha$ = mean of the smallest α n observations and $U_\alpha$ = mean of the largest α n observations. (For instance, if α = 0.05, then $L_{(0.05)}$ is the mean of the smallest [0.05n] observations). Let B = mean of the next largest [0.15n] observations, C = mean of the next largest [0.30n] observations, D = mean of the next largest [0.30n] observations and E = mean of the next largest [0.15n] observations.

Hogg (1967) defined two measures of tail length, Q and $Q_1$. These two statistics as selector statistics are used to classify symmetric distributions as light-tailed (uniform [0,1]), medium-tailed (normal (0,1)), or heavy-tailed (double exponential). Both Q, $Q = ( U_{(0.05)} - L_{(0.05)} ) / ( U_{(0.50)} - L_{(0.50)})$, and $Q_1$, $Q_1 = U_{(0.20)} - L_{(0.20)}) / (U_{(0.50)} - L_{(0.50)})$, are location-free and are then uncorrelated with location statistics like the trimmed means. Values of $Q < 2.0$ imply a light-tailed (uniform) distribution, $2.0 \le Q \le 2$ 2.6 implies a medium tailed-distribution (normal), $2.6 < Q \le 3.2$ implies a heavy tailed distribution (double exponential), and a $Q > 3.2$ implies a Cauchy like distribution (very heavy-tailed distribution). When using $Q_1$ a suggested

classification scheme is: $Q_1 < 1.81$ (light tailed, $1.81 \le Q_1 \le 1.87$ (medium-tailed), and $Q_1 > 1.87$ (heavy-tailed). Hogg (1982) also defined a third measure of tail length $H_3$ ,where $H_3 = (U_{(0.05)} - L_{(0.05)}) / (E - B)$. $H_3 < 1.26$ is associated with a uniform distribution, $1.26 \le H_3 \le 1.76$ is generally associated with a normal distribution, and $H_3 > 1.76$ is associated with a double exponential distribution.

Hertsgaard (1979) used $Q_2$, $Q2 = (U_{(0.05)} - T_{25}) / (T_{25} - L_{(0.50)})$, to classify distributions as left skewed ($Q_2 < 0.7$, symmetric ($0.7 \le Q_2 < 1.4$) and right skewed ($Q_2 \ge 1.4$). $H_1$, $H_1 = (U_{(0.05)} - D) / (C - L_{(0.05)})$, was proposed by Hogg (1982) and was also found to be useful in classifying skewness of a wide variety of distributions. And, Reed and Stark (1996) proposed two quick-and-dirty skewness measures $SK_2$ , $SK_2 = ( X_{(1)} - XMD) / (XMD - X_{(n)})$, and $SK_5$ , $SK_5 = (X_{(1)} - XM ) / (XM - X_{(n)})$. The form of $SK_2$ and $SK_5$ are identical to $Q_2$ and $H_1$. The advantage of using either the median XMD ($Q_2$) or mean XM ($H_1$) lies in the familiarity of these common location estimators. Note: XMD is the median, XM is the arithmetic mean, $T_{25}$ is the [0.25n] trimmed mean ($T_\alpha$), $X_{(1)}$ and $X_{(n)}$ are the first and last order statistics.

Reed and Stark (1996) proposed a set of asymmetric linear estimators or hinge estimators, defined using the following scheme. Set a total trimming proportion to be trimmed from the sample, α. Determine a proportion to be trimmed from the lower end of the sample ($\alpha_l$) using the following: $\alpha_l = \alpha [ UW_x / (UW_x + LW_x)]$, and the upper trimming proportion, $\alpha_u = \alpha - \alpha_l$., where $UW_x$ and $LW_x$ be the numerator and denominator portions of the selector statistic X and define eight adaptive location estimators as:

| Estimator | α | $\alpha_l$ |
|---|---|---|
| HQ | 0.10 | |
| $\alpha_l = \alpha [UW_Q / (UW_Q + LW_Q)]$ | | |
| $HQ_1$ | 0.10 | |
| $\alpha_l = \alpha [UW_{Q1} / (UW_{Q1} + LW_{Q1})]$ | | |
| $HH_3$ | 0.10 | |
| $\alpha_l = \alpha [UW_{H3} / (UW_{H3} + LW_{H3})]$ | | |
| $HQ_2$ | 0.25 | |
| $\alpha_l = \alpha [UW_{Q2} / (UW_{Q2} + LW_{Q2})]$ | | |
| $HH_1$ | 0.10 | |
| $\alpha_l = \alpha [UW_{H1} / (UW_{H1} + LW_{H1})]$ | | |
| $HSK_2$ | 0.10 | |
| $\alpha_l = \alpha [UW_{SK2} / (UW_{SK2} + LW_{SK2})]$ | | |
| $HSK_5$ | 0.25 | |
| $\alpha_l = \alpha [UW_{SK5} / (UW_{SK5} + LW_{SK5})]$ | | |

In the Princeton Robust Study (Andrews, et al, 1972), sensitivity curves were introduced to provide a basis for comparing estimators. The notion behind a sensitivity curve is to show how the value of a particular estimator is affected by an outlier. The method of construction is fairly straight forward. Start with a symmetric sample that is centered about a given value. In this article, the sample consisted of forty nine points (beginning at -4.8 and ending at 4.8) symmetrical about zero. Then add another point to the sample to see how the value of the estimator is affected. The added point ranged from -9.0 to 9.0. The horizontal axis represents the value of the added point while the vertical axis represents the value of the estimator at that value of the added point.

## Results

The sensitivity curve for the sample mean is shown in Figure 1. Note that the curve is a straight line, suggesting that the value of the mean changes linearly with the value of the added point. As the value of the added point increases away from zero, the value of the mean does also. The larger the added value is, the larger the change in the mean value. There is no bound to the influence of the added point.

### The Median

The sensitivity curve is given in Figure 2. Note here, that the change in the value of the median is bounded. If the added point is one of the two middle values then it has a direct influence. However, if it is not one of those two values its influence is bounded regardless of the size of the value.

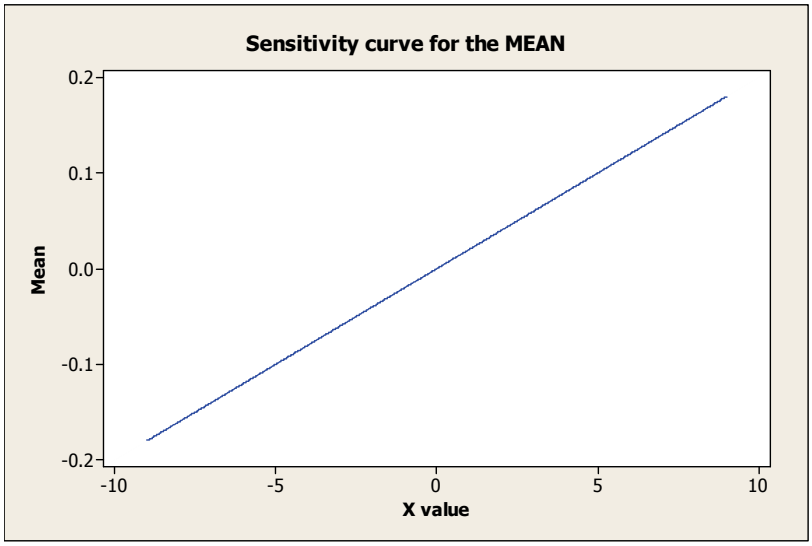



Figure 1.

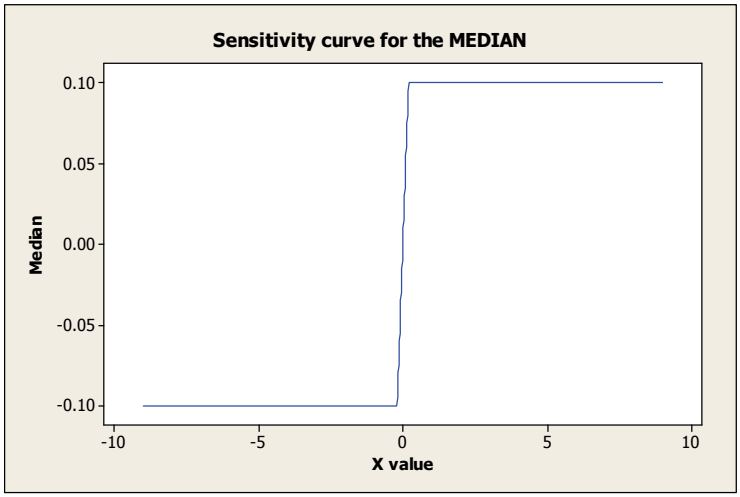



Figure 2.

Next, consider a 15% trimmed mean. The sensitivity curve is shown in Figure 3. As might be expected, the added point has a wider range direct influence. However, once outside of that range of values, the influence of the added point is bounded.

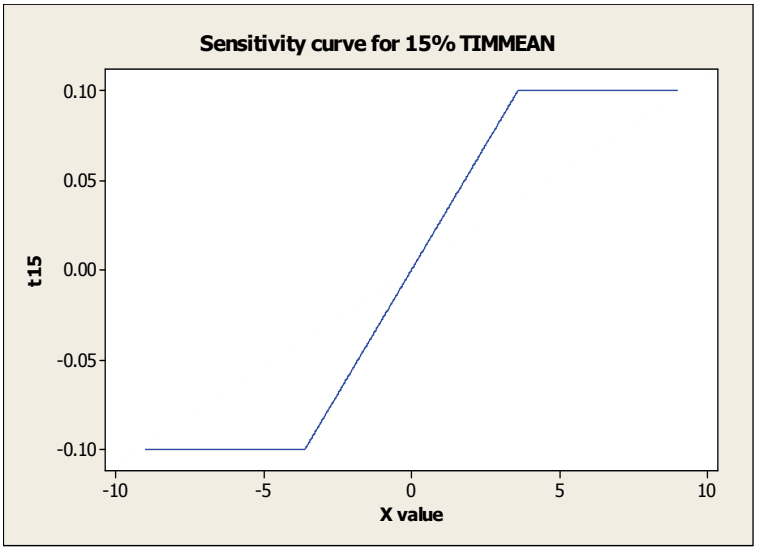



Figure 3.

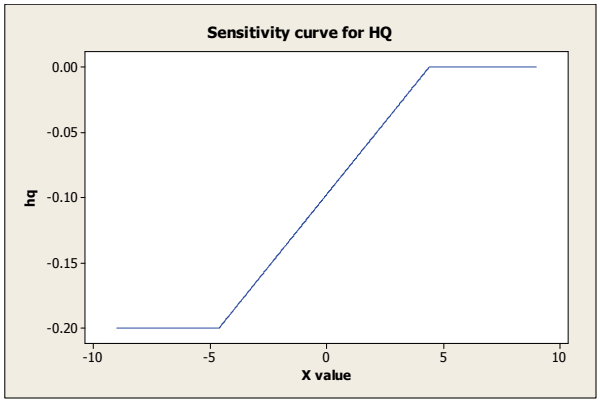



Figure 4.

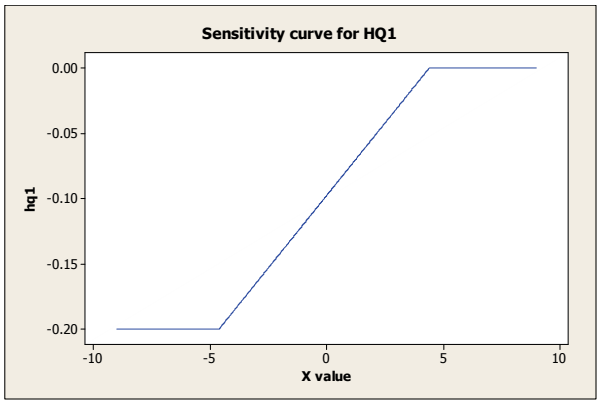



Figure 5.

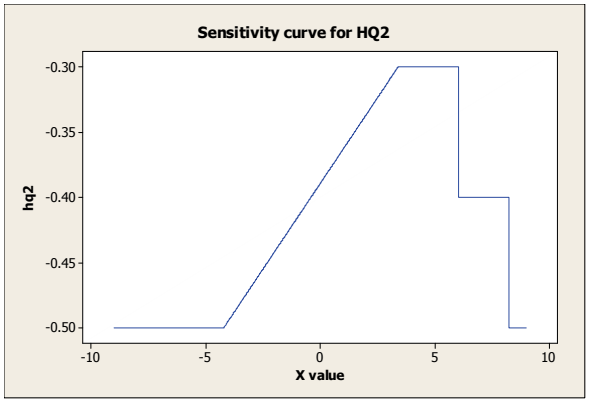



Figure 6.

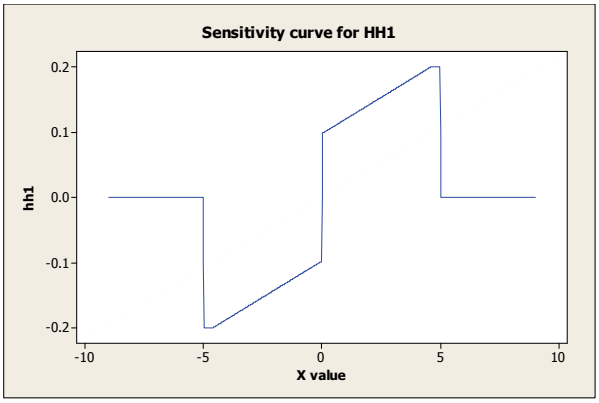



Figure 7.

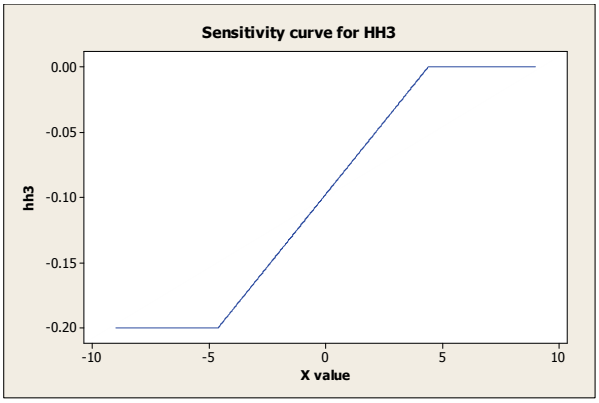



Figure 8.

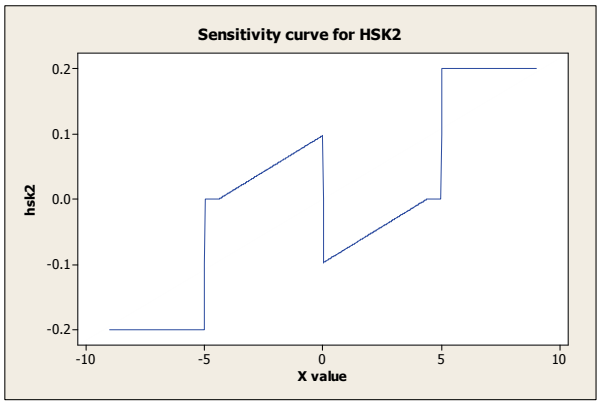



Figure 9.

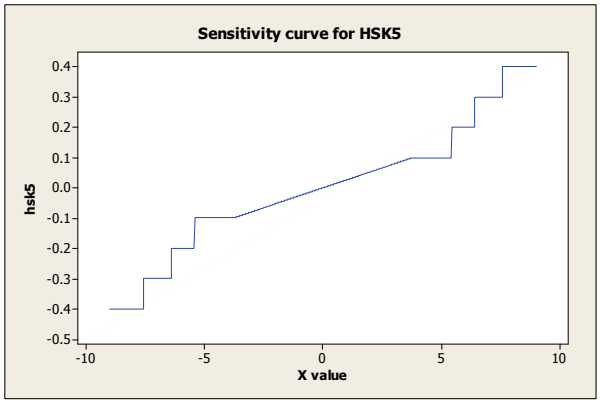



Figure 10.

The sensitivity curves for the seven estimators defined above are given in Figure 4-10. Note the sensitivity curves for HQ, HQ1, HH3 suggest that the adaptive trimming causes the value of the estimator to decrease only. These estimators are not reacting symmetrically to the sample. The estimator HQ2 is just kind of weird. However, HH1, HH3, HSK2 and HSK5 are at least symmetric in their reaction to the value of the added point. The estimator HH1 has a somewhat unique property that when the value of the added point gets a bit outside the symmetric part of the sample, its influence is zero. For data with contaminated with large outliers, this might be a very attractive trait. The estimator HH3 appears to act like a trimmed mean.

## Conclusion

Real-world data sets may be described as messy with everything but a normal distribution presenting to the data analyst. From a methodology point rather than a theoretical basis, reasonable alternatives should be available. In the asymmetric data distributions faced on a daily basis, estimators that adapt themselves to the data may be formulated and used. Adaptively trimmed means can correct for uncontrollable data anomalies.

## References

Andrews D. F., Bickel P. H, Hampel F. R., Huber P. H., Rogers W. H., and Tukey J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.

Hertsgaard D. M. (1979). Distribution of asymmetric trimmed means. *Comm. Stat-Simul.Comp, 8*, 359-367.

Hogg R. V. (1967). Some observations on robust estimation. *Journal of the American Statistical Association, 62*, 1179-1182.

Hogg R. V. (1982). On adaptive statistical inferences. *Comm. Stat. - Theory Method, 11*, 2531-2542.

Reed J. F. and Stark D. B. (1996). Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine*, *49*, 11-17.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 476- 486
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# The Effect Of GARCH (1,1) On The Granger Causality Test In Stable VAR Models

Panagiotis Mantalos
Lund University

Ghazi Shukur
Jönköping and Växjö Universities

Pär Sjölander
Jönköping University

Using Monte Carlo methods, the properties of Granger causality test in stable VAR models are studied under the presence of different magnitudes of GARCH effects in the error terms. Analysis reveals that substantial GARCH effects influence the size properties of the Granger causality test, especially in small samples. The power functions of the test are usually slightly lower when GARCH effects are imposed among the residuals compared with the case of white noise residuals.

Key words: Causality test, GARCH, size and power.

## Introduction

One of the most important issues in the subject of time series econometrics is the ability to statistically perform causality test. By causality it is meant causality in the Granger (1969) sense. That is, one would like to know if one variable precedes the other variable or if they are contemporaneous. The Granger approach to the question whether a variable say $y_1$ causes another variable say $y_2$ is to see how much of the current value of the second variables can be explained by past values of the first variable. $Y_2$ is said to be Granger-caused by $y_1$ if $y_1$ helps in the prediction of $y_2$, or equivalently, if the coefficients of the lagged $y_1$ are statistically significant in a regression of $y_2$ on $y_1$. Empirically, one way to test for causality in Granger sense is by means of vector autoregressive (VAR) model.

Panagiotis Mantalos is Associate Professor of Statistics in the Dept. of Economics & Statistics. Email: panagiotis.mantalos@stat.lu.se. Ghazi Shukur is Professor of Statistics in the Dept. of Economics and Statistics. Email: ghazi.shukur@jibs.hj.se, Par Sjolander is Professor of Economics in the Dept. of Statistics. Email: par.sjolander@jibs.hg.se.

The main purpose of this article is to investigate the properties of the Granger causality test in stationary and stable VAR models under conditions when there exists some kind of volatility among the error terms, more specifically, Generalised Autoregressive Conditional Heteroscedasticity (GARCH) effects. It is well known that the analysis of causality is very sensitive to model specification and is almost only valid under conditions when the error terms are fairly close to white noise. At the same time it is also known that a considerable proportion of the time series variables follow some type of GARCH process. Hence, it is important to investigate the properties of this commonly used causality test under the presence of generalized conditional heteroscedasticity.

## The Model and the Monte Carlo Experiment

Consider the data-generating process (DGP) consists of a two dimensional time series generated by a stabile VAR(p) process:

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + \varepsilon_t \qquad (1)$$

where $\varepsilon_t = \left(\varepsilon_{1t}, \ldots, \varepsilon_{kt}\right)'$ is a zero mean independent white noise process with nonsingular covariance matrix $\Sigma_\varepsilon$ and, for j = 1, ... , k, $\mathrm{E}\left|\varepsilon_{jt}\right|^{2+\tau} < \infty$ for some $\tau > 0$. The order p of the process is assumed to be known. Let

$\alpha_p = \text{vec}[\text{A}_1, \cdots, \text{A}_\text{p}]$ be the vector of the true parameters, where vec[.] denotes the vectorization operator that stacks the columns of the argument matrix. Now, suppose that one is interested in testing q independent linear restrictions:

$$H_o: \text{R}\alpha_p = \text{s}$$
vs. (2)
$$H_1: \text{R}\alpha_p \neq \text{s}$$

where q and s are fixed (q x 1) vectors and R is a fixed [q x $k^2(p)$ ] matrix with rank q.

The process $\{y_t\}$ is generated by the VAR(p) process in (1), with the $\hat{A}_i$ (i = 1, … p) the Ordinary Least Squares (OLS) estimators and $\hat{\alpha}_p^{p-1}$ the $[k^2(p-1)]$dimensional vector, consisting of the $k^2(p-1)$ elements of $\hat{\alpha} = vec[\hat{A}_1, \cdots, \hat{A}_p]$, that are obtained by deleting the matrix $\hat{A}_i$ i ∈ {1, … p}. Then:

$$T^{1/2}(\hat{\alpha}_p - \alpha_p) \Rightarrow N(0, \Sigma_p) \quad (3)$$

where $\Rightarrow$ denotes weak convergence in distribution and the [$k^2(p)$ x $k^2(p)$ ] covariance matrix $\Sigma_p$ is non-singular. The $\alpha_p$ is the $[k^2(p)]$dimensional vector of the true parameters. Moreover given a consistent estimator $\hat{\Sigma}_p$, then the Wald test of the null hypothesis in (2):

$$\lambda_w = T(R\hat{\alpha}_p - s)'(R\Sigma_p R')^{-1}(R\hat{\alpha}_p - s) \quad (4)$$

has an asymptotic $\chi^2(q)$-distribution under the null hypothesis. And with $y_t$ portioned in (m) and (k-m) dimensional sub vectors $y_t^1$ and $y_t^2$, and $A_i$ matrices portioned conformably, then $y_t^2$ does not Granger-cause the $y_t^1$ if the following hypothesis is true:

$$H_0 = \text{A}_{12,\text{i}} = 0$$

for

$$\text{i} = 1, \cdots, p\text{-}1. \quad (5)$$

The error components $(\varepsilon_{1t}, \varepsilon_{2t})'$ in (1) and (2) are generated by GARCH(1,1) models, i.e.,

$$\begin{gathered} \varepsilon_{it} = h_{it}\upsilon_{it} \quad i = 1,2 \\ \upsilon_{it} \text{ i.i.d., } \text{E}(\upsilon_{it}) = 0, \text{E}(\upsilon_{\text{it}}^2) = 1 \\ h_{it}^2 = \gamma_i + \phi_i h_{it-1}^2 + \varphi_i \varepsilon_{it-1}^2 \end{gathered} \quad (6)$$

and $\text{Cov}(\varepsilon_{1t}\ \varepsilon_{2t}) = 0$. The condition for finite variance is $\phi_i + \varphi_i < 1$ and the condition for finite fourth moment is $3\phi_{i1}^2 + 2\phi_i\varphi_i + \varphi_i^2 < 1$. Furthermore, if $\gamma_i > 0$ and $\phi_i + \varphi_i < 1$, then the unconditional variance of the $\varepsilon_i$ exist and equals $\sigma_{\varepsilon i}^2 = (\gamma_i / 1 - \phi_i - \varphi_i)$. Note that when $\phi = \varphi = 0$, the $\varepsilon_{it}$ is reduced to iid white noises.

To illustrate and study the possible effects of a GARCH(1,1) process on the Granger-causality test in a stable VAR(1) system Monte Carlo methods. The estimated size is calculated by simply observing how many times the null is rejected in repeated samples under conditions where the null is true. To judge the reasonability of the results use an approximated 95% confidence interval for the actual size ($\pi$):

$$\hat{\pi} \pm 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N}} \quad (7)$$

where $\hat{\pi}$ is the estimated size and $N$ is the number of replications.

The Monte Carlo experiment has been performed by generating data according to the model defined by (1) and (2),

$$y_t = \begin{bmatrix} 0.02 \\ 0.03 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.3 \\ T^{-1/2}\lambda & 0.5 \end{bmatrix} y_{t-1} + \varepsilon_t \quad (8)$$

If $\lambda = 0$, $y_{1t}$ is Granger-non-causal for $y_{2t}$ and if $\lambda \neq 0$, $y_{1t}$ causes $y_{2t}$. Therefore, the $\lambda = 0$ is used to study the size of the test.

Three GARCH versions are simulated with a) high persistence, HP, (0.01, 0.09, 0.9), b) medium persistence, MP, (0.05, 0.05, 0.9) and c) low persistence, LP, (0.20,0.05,0.75). The processes includes a constant term and fit a VAR(1) : $y_t = v + A_1 y_{t-1} + \varepsilon_t$.

This means that order p of the process is assumed to be known and since this assumption might be too optimistic, however, also fit a VAR (2) : $y_t = v + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$.

For each model perform 10 000 replications and use three different nominal sizes, namely 1%, 5% and 10%. However, different authors have put forward reasons for using both larger and smaller significance levels. Maddala (1992) suggests using significance levels of as high as 25% in diagnostic testing, while MacKinnon (1992) suggest going in the other direction to avoid mass significance. To reduce this problem, in this study, also use graphical methods that may provide more information about the size and the power of the test. Simple graphical methods are used, developed and illustrated by Davidson and MacKinnon (1998), which are based on the empirical distribution function (EDF) of the P-values and are easy to interpret. The P value plot is used to study the size and the Size-Power curves to study the power of the test.

Furthermore, to judge the reasonability of the results use a 95% confidence interval for the actual size ($\pi$) as: $\pi_0 \pm 2\sqrt{\frac{\pi_0(1-\pi_0)}{N}}$, where *N* is the number of replications. Results that lie between these bounds will be considered satisfactory.

Several factors are expected to affect the size and power properties of causality tests. Samples typical for small, medium, large and very large sizes have been investigated. For each time series 20 pre-sample values are generated with zero initial conditions, and with net sample sizes of T = 50, 100, 200, 500, 1000. Table 1 shows the different parameters of our Monte Carlo design. The number of replications per model is 10 000 for the size, and 1000 for the power of the test. The calculations were performed using GAUSS 6.0.

Results of the Size of the Test

Presented in this section are the most important results of our Monte Carlo experiment concerning the size of the test. Regarding the P value plots, under the condition when the distribution used to compute the $p_s$ is correct, each of the $p_s$ should be distributed as uniform (0,1) and therefore the resulting graph should be close to the $45^o$ line as in Figure 1a below.

Size of the test for the VAR (1), given that the true model is a VAR (1)

In this sub-section the results are presented when the estimated and the true model is a VAR (1). As can be seen from the results, in Table 1a in the Appendix, the calculated sizes of the test over estimate the nominal sizes in all situations more or less regardless whether there exist low, medium or high GARCH effects. This is the case when a small sample of 50 observations are studied. This is also confirmed when the P-value plots are observed in Figure 1a, in the Appendix, in which one only presents the size when white noise and high GARCH effects are imposed. Here one can see that in both cases the test over rejects the size, but that the calculated sizes still lay near to the 95% confidence interval for nominal size with a slightly higher over rejection when the high GARCH magnitudes are present.

When the sample size increases to 100 observations, as is illustrated in Table 2a and Figure 2a, the properties of the test become better but there still some over rejection present. When enlarging the sample size to 200 observations the test performs well in all cases except for the case with high GARCH effect. In this case the test slightly over rejects the nominal size, as can be seen in Table 3a. Figure 3a shows that the over rejection become more severe for larger nominal sizes.

Table 1 Monte Carlo Parameters of the GARCH Effects

| | $\Lambda$ | $\gamma$ | $\phi$ | $\varphi$ |
|---|---|---|---|---|
| High Persistence | 0 | 0.01 | 0.09 | 0.90 |
| Medium | 0 | 0.05 | 0.05 | 0.90 |
| Low | 0 | 0.20 | 0.05 | 0.75 |
| High Persistence | 2 | 0.01 | 0.09 | 0.90 |
| Medium | 2 | 0.05 | 0.05 | 0.90 |
| Low | 2 | 0.20 | 0.05 | 0.75 |

The same is also true when the sample size is equal to 500 observations, as is illustrated in Table and Figure 4a in the Appendix.

In a very large sample, i.e. 1000 observations in Table and Figure 5a, the test performs satisfactorily in almost all situations, but with one exception in the case when a high GARCH effect is present

Size of the test for the extra lag; VAR (2), given that the real model is a VAR (1)

Here the results are presented when the estimated model contains an extra lag, i.e. a VAR(2), while the true model is a VAR(1). In this case to investigate the effect of possible over parameterization of the true model is what is desired. Table and Figure 1b in the Appendix, the sizes of the test, as in the previous sub-section, over estimate the nominal sizes in all situations almost regardless whether there exist low, medium or high GARCH effects. In Figure 1b, the clear over rejection is illustrated for both white noise and high GARCH effects.

However, as the results confirm in Table 2b and Figure 2b, the over rejection become less severe when the number of observations increases to 100 observations. The results are almost similar when increase the sample size is increased to 200 observations, see Table 3b and Figure 3b in the Appendix.

When the sample size increases to 500 observations, as in Table 4b and Figure 4b, the test performs well in almost all situations except for in the case of high GARCH effects. Finally, in Table 5b and Figure 5b, the results show that the test performs satisfactorily but still with a slight over rejection in the case of high GARCH effects.

In general, the results from these two sub-sections are generally similar. Moreover, one could not find the over rejection to be that severe even in the case of the existence of high GARCH effects in comparison with that of the white noise. The test is consistent and converges slowly to its nominal size as the sample size increases.

Analysis of the Power of the Test

In this section the results of the Monte Carlo experiment regarding the power of the Granger-causality test are discussed. The power of the test was analyzed using sample sizes of 50, 100, 200, 500 and 1000 observations. The power functions have been calculating for the test in the case of white nose and under different GARCH effects. The power functions have shown to be fairly similar in the cases of the white noise, low persistence and medium persistence GARCH. Based on this and since one could not find any noticeable differences in the performances of the test between these combinations regarding the size properties, only show and compare the power functions of the white nose and the high GARCH.

The power functions are estimated by calculating the rejection frequencies in 1000 replications using values of the $\lambda$ coefficients in equation (8) equal to 2. The estimated power functions of the test have been compared only graphically. One may follow the same procedure as for the size investigation to evaluate the EDF's denoted $\hat{F}^{\oplus}\left(x_j\right)$, by using the same sequence of random numbers as in the case of the size of the test. For plotting the estimated power functions against the nominal size, there are the Size-Power Curves. Presented is the power of the test in cases when the model is

exactly identified, i.e. the true and estimated models are VAR (1) and in the case when the model is over parameterized, i.e. the estimated model is VAR (2) while the true is VAR (1).

The power of the Granger causality test, as expected, depends on how well the model is specified. This can be seen when comparing the power functions in the upper and lower parts of Figures 6-10 in the Appendix. This is the effect of over parameterization.

Moreover, from the figures it can be seen that the power functions satisfy the expected properties of increasing with the sample size. Lower powers are observed when the samples are small and higher when the samples are large. A closer examination of the figures shows, that most frequently, the power functions are slightly lower in the case of the GARCH residuals (the dashed lines) than the white noise.

## Conclusion

The results regarding the size of the tests have been presented both in form of tables and P-value plots. Our analysis revealed that the Granger-causality test slightly over rejects the nominal sizes in small samples and under the existence of high GARCH effects. This over rejection becomes even lower when the sample size increases and when the GARCH effects are not high. These results are similar in both of the exactly parameterized VAR (1) model and the over parameterized VAR (2) model. Moreover, the test is consistent in the sense that the size of the test converges slowly to its nominal size as the sample size increases.

The power functions have been presented only graphically. As expected, the analysis of the power indicates that these power functions increase with an increasing sample size. Furthermore, most of the times these power functions are slightly lower in the case of the GARCH residuals than under white noise. The power of the test, as expected, becomes lower when including an extra lag in the VAR model, i.e. in the case of VAR(2).

## References

Davidson, R. and J. G. MacKinnon (1998). Graphical methods for investigating the size and power of test statistics. *The Manchester School*, 66, 1-26.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica,* 55, 251-276.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 24-36.

MacKinnon, J. G. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature*, 30, 102-146.

Maddala, G. S. (1992). *Introduction to Econometrics.* Second edition, New York, Wiley.

## APPENDIX

Table 1a. Size of the test for 50 observations

| Nominal | White Noise | | GARCH(1,1) | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0160 | 0.0151 | 0.0156 | 0.0152 |
| 0.05 | 0.0642 | 0.0643 | 0.0658 | 0.0668 |
| 0.10 | 0.1169 | 0.1222 | 0.1231 | 0.1225 |

Table 2a. Size of the test for 100 observations

| Nominal | White Noise | | GARCH(1,1) | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0133 | 0.0126 | 0.0126 | 0.0141 |
| 0.05 | 0.0584 | 0.0579 | 0.0578 | 0.0593 |
| 0.10 | 0.1093 | 0.1069 | 0.1051 | 0.1087 |

Table 3a. Size of the test for 200 observations

| Nominal | White Noise | | GARCH(1,1) | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0112 | 0.0119 | 0.0119 | 0.0146 |
| 0.05 | 0.0546 | 0.0528 | 0.0527 | 0.0584 |
| 0.10 | 0.1056 | 0.1036 | 0.1054 | 0.1109 |

Table 4a. Size of the test for 500 observations

| Nominal | White Noise | | GARCH(1,1) | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0095 | 0.0107 | 0.0108 | 0.0141 |
| 0.05 | 0.0558 | 0.0544 | 0.0535 | 0.0639 |
| 0.10 | 0.1068 | 0.1031 | 0.1038 | 0.1121 |

Table 5a. Size of the test for 1000 observations

| Nominal | White Noise | | GARCH(1,1) | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0096 | 0.0083 | 0.0084 | 0.0150 |
| 0.05 | 0.0476 | 0.0479 | 0.0496 | 0.0628 |
| 0.10 | 0.0979 | 0.1034 | 0.0997 | 0.1183 |

P-value plots HP (GARCH)

Figure 1a. 50 observations

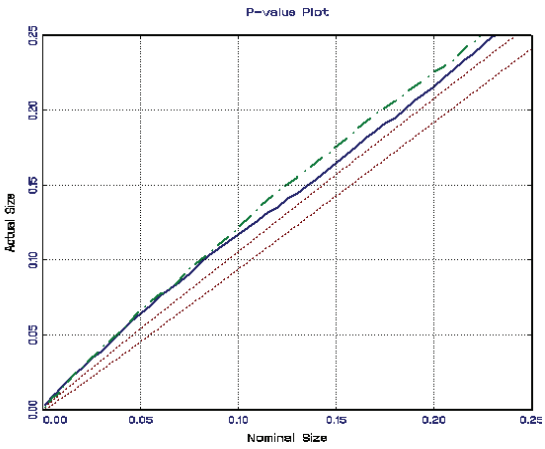



Figure 2a. 100 observations

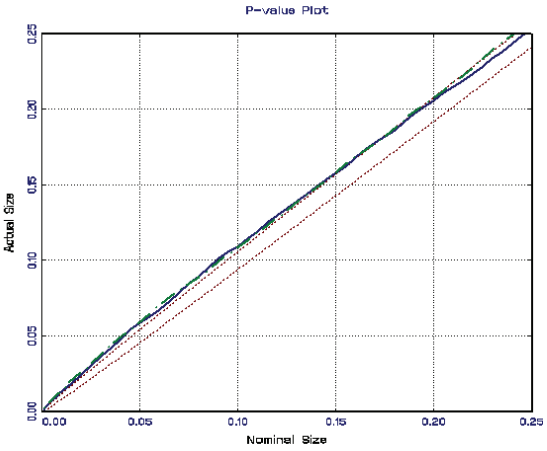



Figure 3a. 200 observations

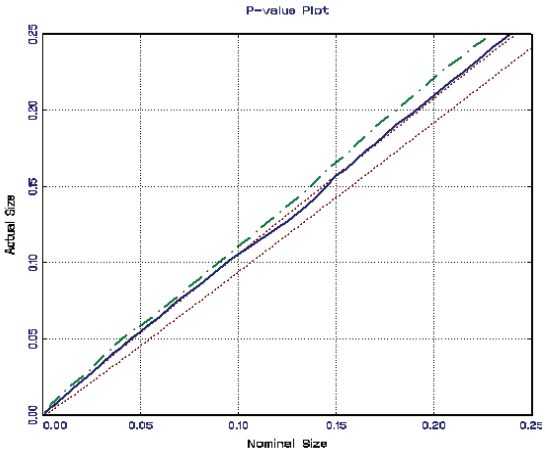



Figure 4a. 500 observations

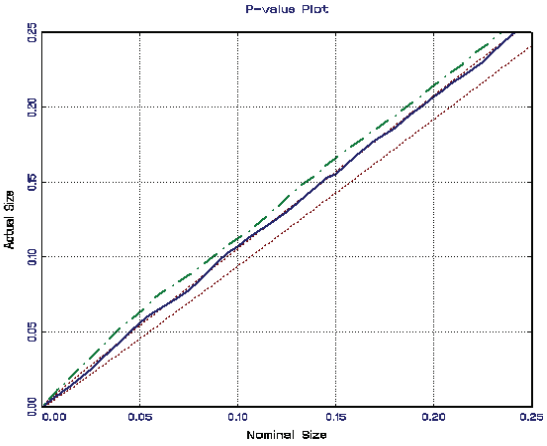
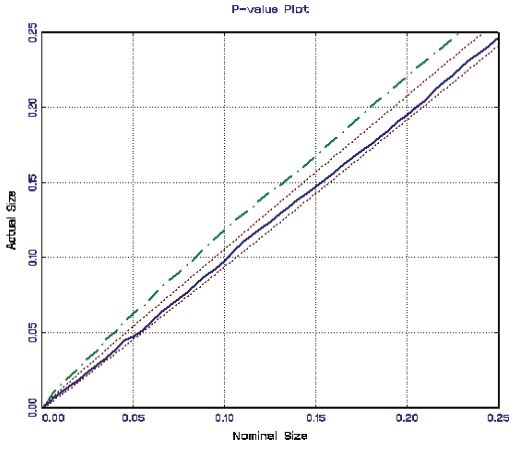



Figure 5a. 1000 observations




Solid lines = White noise. Dot dash line = GARCH. Dot lines = 95% confidence interval for nominal size.

Table 1b. Size of the test for 50 observations

| Nominal | White Noise | GARCH(1,1) | | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0169 | 0.0151 | 0.0179 | 0.0185 |
| 0.05 | 0.0633 | 0.0643 | 0.0648 | 0.0671 |
| 0.10 | 0.1192 | 0.1222 | 0.1221 | 0.1248 |

Table 2b. Size of the test for 100 observations

| Nominal | White Noise | GARCH(1,1) | | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0125 | 0.0115 | 0.0109 | 0.0119 |
| 0.05 | 0.0542 | 0.0566 | 0.0566 | 0.0593 |
| 0.10 | 0.1067 | 0.10890.1089 | 0.1095 | 0.1126 |

Table 3b. Size of the test for 200 observations

| Nominal | White Noise | GARCH(1,1) | | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0138 | 0.0122 | 0.0129 | 0.0143 |
| 0.05 | 0.0582 | 0.0542 | 0.0542 | 0.0611 |
| 0.10 | 0.1098 | 0.1053 | 0.1062 | 0.1118 |

Table 4b. Size of the test for 500 observations

| Nominal | White Noise | GARCH(1,1) | | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0111 | 0.0111 | 0.0111 | 0.0125 |
| 0.05 | 0.0524 | 0.0532 | 0.0529 | 0.0568 |
| 0.10 | 0.1006 | 0.1017 | 0.1026 | 0.1127 |

Table 5b. Size of the test for 1000 observations

| Nominal | White Noise | GARCH(1,1) | | |
|---|---|---|---|---|
| | | LP | MP | HP |
| 0.01 | 0.0092 | 0.0095 | 0.0091 | 0.0130 |
| 0.05 | 0.0449 | 0.0487 | 0.0484 | 0.0581 |
| 0.10 | 0.0950 | 0.0969 | 0.0943 | 0.1084 |

P-value plots HP (GARCH)

Figure 1b. 50 observations

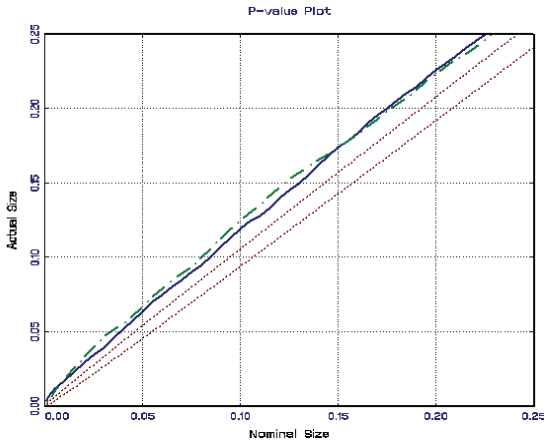



Figure 2b. 100 observations

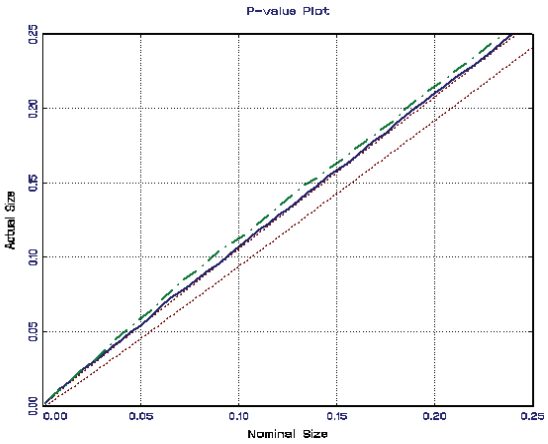



Figure 3b. 200 observations

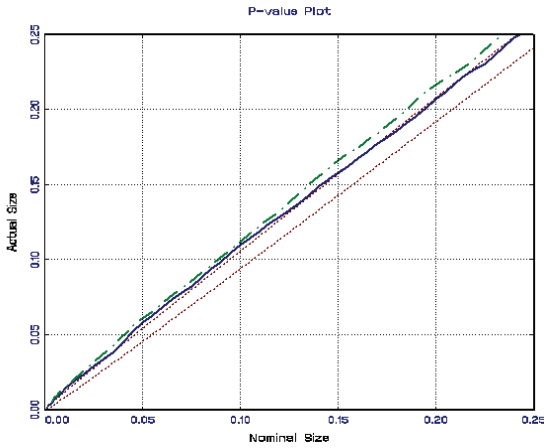



Figure 4b. 500 observations

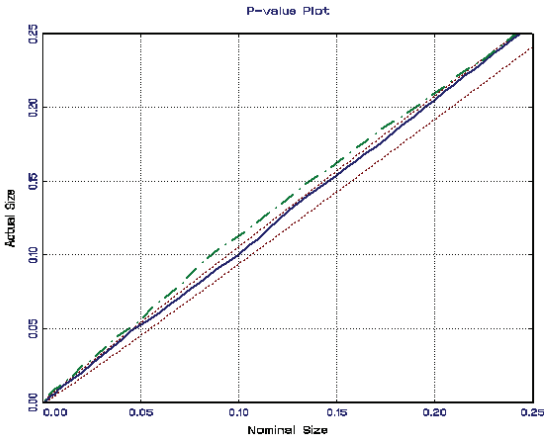
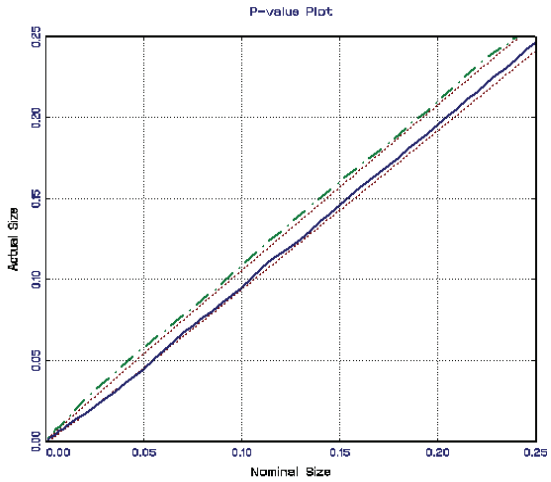
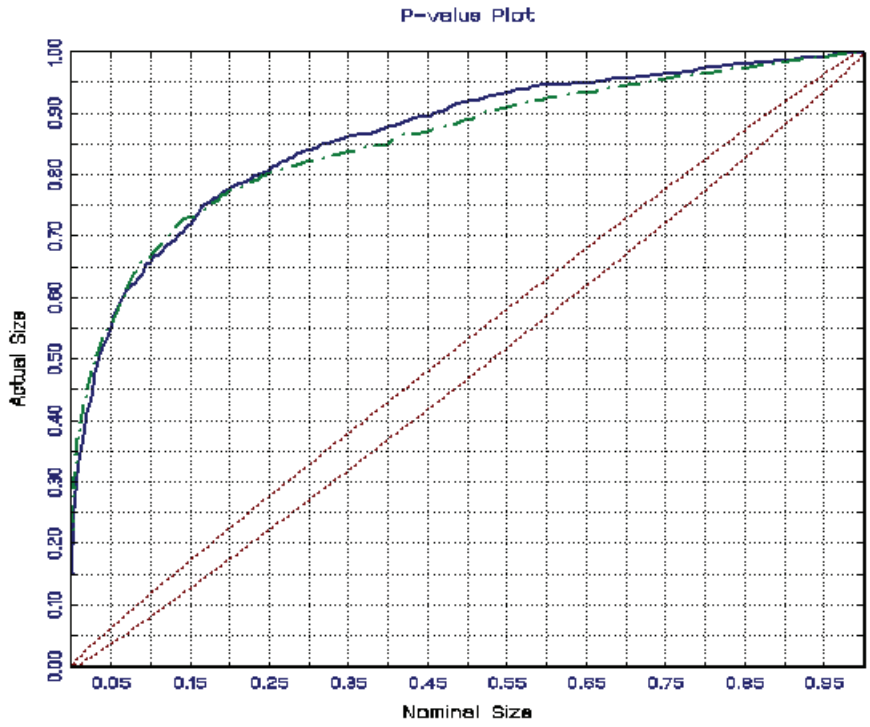


Figure 5b. 1000 observations



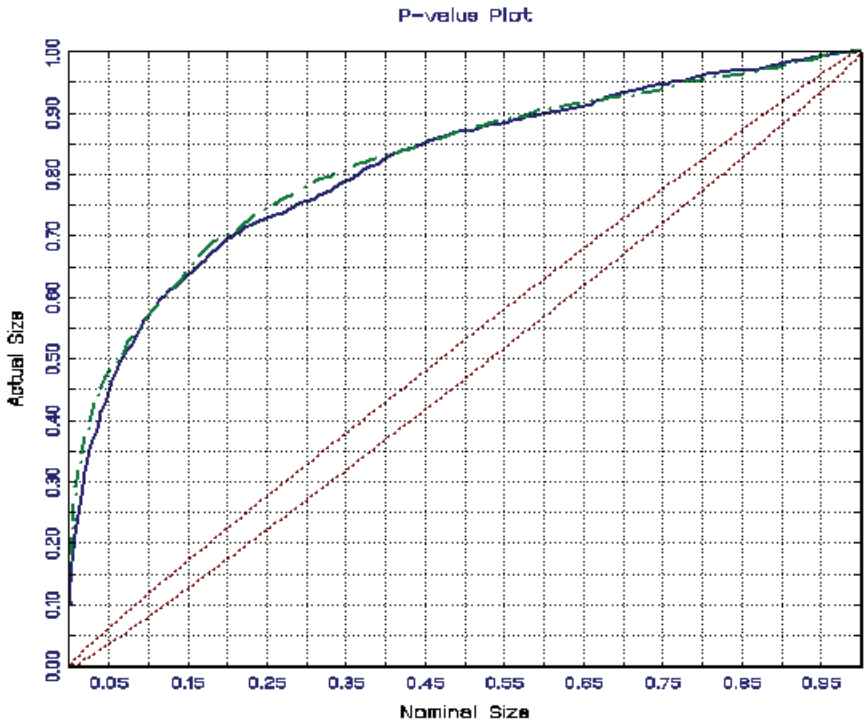

Solid lines = White noise. Dot dash line = GARCH. Dot lines = 95% confidence interval for nominal size.

Figure 6a. Power–Size plots of the Granger-causality test for 50 observations
VAR(1)




Solid lines = White noise. Dash line = GARCH. Dot lines = 95% confidence interval for nominal size.

Figure 6b Power–Size plots of the Granger-causality test for 50 observations
VAR(2)




Solid lines = White noise. Dash line = GARCH. Dot lines = 95% confidence interval for nominal size.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 487-491
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Large Deviations Techniques for Error Exponents to Multiple Hypotheses LAO Testing

Leader Navaei
Yerevan State University

In this article the problem of multiple hypotheses testing using a theory of large deviations is studied. The reliability matrix of Logarithmically Asymptotically Optimal (LAO) tests is introduced and described, and the conditions for the positive of all its elements are indicated.

Key words: hypotheses testing, empirical distributions, the method of types, reliability matrix, Sanov's theorem.

## Introduction

Many studies have been devoted to the study of exponential decrease, as the sample size $N$ goes to infinity, of the error probabilities $\alpha_1^N = \alpha_1$. For example Stain's lemma determines the exponential rate of convergence to zero of the error probability of the second kind $\alpha_2^N$ as $N$ goes to infinity. Perez (1984) considered independent identically distributed observations and different asymptotical aspects of two hypotheses, as the interdependence of exponents.

Csiszar and Shields (2004) considered independent identically distributed observations different asymptotical aspects of the two hypotheses testing via the theory of large deviations. This article is based on Haroutunian (1990), and provides a proof based on Sanov's theorem.

Leader Navaei is Assistant Professor in the Faculty of Mathematics. His research interests are in Markov chains, large deviations techniques, and applied information theory in multiple hypotheses testing. Address correspondence to Yerevan State University, St. Alex Manoogian 1,Yerevan 375049, Republic of Armenia. Email: ashkan_l1380@yahoo.com, or l_navaei@ysu.am

## Preliminaries

Let $\chi = \{1,2,....,K\}$ be the finite set of size $K$. The set of all probability distributions by $(PD's)$ on $\chi$ is denoted by $P(\chi)$. For $PD's$, $P$ and $Q$, $H(P)$ denotes entropy and $D(P \| Q)$ denotes information divergence (or the Kullback-Leibler distance).

$$H(P) \equiv -\sum_{x \in \chi} P(x) \log P(x),$$

$$D(P \| Q) \equiv \sum_{x \in \chi} P(x) \log \frac{P(x)}{Q(x)}.$$

In this article, exps and logs are used at base 2. Also considered are the standard conventions that $0 \log 0 = 0$, $0 \log \frac{0}{0} = 0$, $P \log \frac{P}{0} = \infty$ if $P > 0$.

The type of a vector $X = (x_1, x_2, ....., x_N) \in \chi^N$ is the empirical distribution given by $Q(x) = \frac{N(x \mid X)}{N}$ for all $x \in \chi$, where $N(x \mid X)$ denotes the number of occurrences of $x$ in $X$ or $Q(x) = (\frac{N_1}{N}, \frac{N_2}{N}, ....., \frac{N_N}{N}) \in \chi^N$ that $N_i \equiv$ number of times out $N$ trials that the

random variables $x_1, x_2, ...., x_N$ occurrences in $\chi$.

The subset of $P(\chi)$ consisting of the possible types of sequences $X \in \chi^N$ is denoted by $P_N(\chi)$. For $Q \in P_N(\chi)$, the set of sequences of type class $Q$ will be denoted by $T_Q^N$.

The probability that $N$ independent drawings from a $PD$, $P \in P(\chi)$ give $X \in \chi^N$, is denoted by $P^N(X)$. If $X \in T_Q^N$ then:

$$P^N(X) = \exp\{-N(H(Q) + D(Q \,||\, P)\}.$$

Lemma

The number of types for sequences of length $N$ grows at most polynomially with $N$:

$$| P_N(\chi) | < (N+1)^{|\chi|},$$

For any type $Q \in P_N(\chi)$:

$$(N+1)^{-|\chi|} \exp\{NH(Q)\} \leq | T_Q^N | \leq \exp\{NH(Q)\},$$

For any $PD, P \in P(\chi)$:

$$\frac{P^N(X)}{Q^N(X)} = \exp\{-ND(Q \,||\, P)\}, \text{ If } X \in T_Q^N,$$

and

$$(N+1)^{-|\chi|} \exp\{-ND(Q \,||\, P)\} \leq | T_Q^N | \leq \exp\{NH(Q)\},$$

Theorem 1(Sanov's theorem (Csiszar & Shields, 2004, Dembo & Zeitoni, 1993)

Let $A$ be a set of distributions from $P(\chi)$ such that its closure is equal to the closure of its interior, then for the empirical distribution $Q_X$ of a vector $X$ from a strictly positive distribution $P$ on $\chi$:

$$\lim_{N\to\infty}(-\frac{1}{N}\log P^N(X : Q_X \in A)) = \inf_{Q_X \in A}(D(Q_X \,||\, P)).$$

Problem Statement and Formulation of Results

The problem of multiple hypotheses testing is the following. Let $\chi = \{1, 2, ....., K\}$ be the finite set such that $M$ incompatible hypotheses $H_1, H_2, ....., H_M$ consist in that the random variable $X$ taking values on $\chi$ has one of $M$ distributions $P_1, P_2, ....., P_M$. For decision making $N$ independent experiences are carried out. When $H_m$ is true, the sample $X = \{x_1, x_2, ...., x_N\}$ of the experiments results has the probability

$$P_m^N(X) = \prod_{i=1}^{N} P_m(x_i), \; m = \overline{1, M}.$$

By means of non-randomized test $\varphi_N(X)$ on the basis of a sample of length $N$ one of the hypotheses must be accepted. For this aim one can divide the sample space $\chi^N$ on $M$ disjoint subsets,

$$\omega_m^N \equiv \{X : \varphi_N(X) = m\}, \; m = \overline{1, M}.$$

The probability of the erroneous acceptance of hypothesis $H_l$ provided that hypothesis $H_m$ is true, for $m \neq l$ is denoted:

$$\alpha_{m|l}^N(\varphi_N) \equiv P_m^N(\omega_l^N) = \sum_{X \in \omega_l^N} P_m^N(X).$$

For $m = l$ denote by $\alpha_{m|m}^N(\varphi_N)$ the probability to reject $H_m$ when it is true and this is:

$$\alpha_{m|m}^N(\varphi_N) \equiv \sum_{m \neq l} \alpha_{m|l}^N(\varphi_N). \quad (1)$$

The matrix $\omega(\varphi_N) \equiv \{ \alpha_{m|l}^N(\varphi_N)\}$ is called power of the test. Take into consideration the

rates of exponential decrease of the error probabilities and call them reliabilities:

$$E_{m|l}(\varphi) \equiv \overline{\lim_{N\to\infty}}(-\frac{1}{N}\log\alpha_{m|l}(\varphi_N)) \quad (2)$$

According to (1) and (2)

$$E_{m|m} = \min_{m\neq l} E_{m|l} \quad (3)$$

can be derived because

$$E_{m|m} = \lim_{N\to\infty}\frac{-1}{N}\log\alpha_{m|m}(\varphi)$$
$$= \lim_{N\to\infty}\frac{-1}{N}\log\sum_{m\neq l}\alpha_{m|l}(\varphi)$$
$$= \lim_{N\to\infty}\frac{-1}{N}[\log(Max\alpha_{m|l}\{\frac{\sum_{m\neq l}\alpha_{m|l}}{Max\alpha_{m|l}}+1\})]$$
$$= \lim_{N\to\infty}\frac{-1}{N}\log(Max_{\,m\neq l}\alpha_{m|l})+0$$
$$= \min_{m\neq l}\lim_{N\to\infty}\frac{-1}{N}\log(\alpha_{m|l})$$
$$= \min_{m\neq l} E_{m|l}(\varphi)$$

The matrix $E(\varphi)=\{E_{m|l}(\varphi)\}$ is called the reliability matrix of the tests sequences $\varphi$.

$$E(\varphi)=\begin{bmatrix} E_{1|1}.....E_{1|l}.....E_{1|M} \\ \vdots \\ E_{m|1}.....E_{m|l}.....E_{m|M} \\ \vdots \\ E_{M|1}.....E_{M|l}.....E_{M|M} \end{bmatrix}.$$

The problem is to find the matrix $E(\varphi)$ with largest elements, which can be achieved by tests when a part of elements of the matrix $E(\varphi)$ is fixed.

Definition

The test sequence $\varphi^* = (\varphi_1, \varphi_2, .....)$ is called LAO if for given values of the elements $E_{1|1}, E_{2|2}, ...., E_{M-1|M-1}$ it provides maximal values for all other elements of $E(\varphi^*)$.

Consider for a given positive and finite $E_{1|1}, E_{2|2}, ...., E_{M-1|M-1}$ the following family of regions:

$$\Re_l \equiv \{Q : D(Q \,||\, P_l) \le E_{l|l}\}, \quad (4.a)$$
$$l = \overline{1, M-1}$$
$$\Re_M \equiv \{Q : D(Q \,||\, P_l) > E_{l|l},\ l = \overline{1, M-1}\} \quad (4.b)$$
$$\Re_l^N = \Re_l \bigcap P_N(\chi), \quad (4.c)$$
$$l = \overline{1, M}$$

and introduce the functions:

$$E^*_{l|l} = E^*_{l|l}(E_{l|l}) \equiv E_{l|l},\ l = \overline{1, M-1}, \quad (5.d)$$

$$E^*_{m|l} = E^*_{m|l}(E_{l|l}) \equiv \inf_{Q\in\Re_l}(D(Q \,||\, P_m)), \quad (5.b)$$
$$l = \overline{1, M-1}\ , m = \overline{1, M}\ , m \neq l.$$

$$E^*_{m|M} = E^*_{m|l}(E_{1|1}, E_{2|2}, ....., E_{M-1|M-1}) \equiv \quad (5.c)$$
$$\inf_{Q\in\Re_M}(D(Q \,||\, P_m)),\ m = \overline{1, M-1},$$

$$E^*_{M|M} = E^*_{M|M}(E_{1|1}, E_{2|2}, ....., E_{M-1|M-1}) \equiv \quad (5.d)$$
$$\min_{l=1,M-1} E_{M|l}.$$

With the assumption $A = \Re_l$, $P = P_m$ in Sanov's theorem for conditions (4), (5) there is :

$$\lim_{N\to\infty}(-\frac{1}{N}\log\alpha^N_{m|l}(\varphi^*_N)$$
$$= \lim_{N\to\infty}(-\frac{1}{N}\log P^N_m(\Re_l) \quad (6)$$
$$= \inf_{Q_X\in\Re_l}(D(Q_X \,||\, P_m))$$

The notation $y_1^N \approx y_2^N$ can be used when $g(y_1^N) = g(y_2^N) + \varepsilon_N$, where $\varepsilon_N \to 0$, for $N \to \infty$. Using (6)

$$E_{m|l}(\varphi^*) = \inf_{Q_X \in \Re_l} (D(Q \parallel P_m)). \quad (7)$$

Therefore the value of $\alpha_{m|l}^N(\varphi_N^*)$ is equal to

$$\alpha_{m|l}^N(\varphi_N^*) \approx \exp(-N \inf_{Q_X \in \Re_l} (D(Q \parallel P_m)) \quad (8)$$
$$\approx \exp(-NE_{m|l}(\varphi_N{}^*)).$$

In fact, the error probability $\alpha_{m|l}^N(\varphi_N^*)$ goes to zero with exponential rate $\inf_{Q_X \in \Re_l} (D(Q \parallel P_m))$ for $P_m$ not in the set of $\Re_l$.

Theorem 2

For fixed on finite set $\chi$ family of distributions $P_1, P_2, ....., P_M$ the following two statements hold: If the positive finite numbers $E_{1|1}, E_{2|2}, ....., E_{M-1|M-1}$ satisfy conditions:

$$E_{l|l} \le \min_{l=\overline{2,M}} D(P_l \parallel P_1), \quad (9)$$
$$E_{M|M} < \min\left[ \min_{l=\overline{1,m-1}} E_{m|l}^*(E_{l|l}), \min_{l=\overline{m+1,M}} D(P_l \parallel P_m) \right],$$

Hence:

a) There exists a LAO sequence of tests $\varphi_N^*$, the reliability matrix of which $E^* = \{E_{m|l}^*(\varphi^*)\}$ is defined in (5), and all elements of it are positive.

b) Even if one of conditions (9) is violated, then the reliability matrix of an arbitrary test necessarily has an element equal to zero, (the corresponding error probability does not tend exponentially to zero).

Proof: At first it is remarked that $D(P_l \parallel P_m) > 0$, for $m \ne l$, because all measures $P_l$, $l = \overline{1,M}$, are distinct. Now for the proof of the sufficiency of the conditions (9). Consider the following sequence of tests $\varphi^*$ given by the sets

$$B_l^N = \bigcup_{Q \in \Re_l} T_Q^N, \; l = \overline{1,M}. \quad (10)$$

The sets $B_l^N$, $l = \overline{1,M}$, satisfies conditions to give test, by means:

$$B_l^N \bigcap B_m^N = \phi, \; l \ne m,$$

and

$$\bigcup_{l=1}^{M} B_l^N = \chi^N.$$

The following shows that exponent $E_{m|m}(\varphi^*)$ for sequence of tests $\varphi^*$ defined in (10) is not less than $E_{m|m}$. The following is known from lemma,

$$|T_Q^N| \approx \exp\{NH(Q)\}$$

and

$$P^N(T_Q^N) \approx \exp\{-N(D(Q \parallel P)\}, \; m = \overline{1,M}$$

and also by using the result of theorem 1 there is:

$$\alpha_{m|m}^N(\varphi^*) \approx \exp\{-NE_{m|m}\},$$

and

$$\alpha_{m|l}^N(\varphi^*) \approx \exp\{-NE_{m|l}^*(E_{l|l})\}, \; l = \overline{1,M-1},$$
$$m = \overline{1,M}, \; m \ne l,$$

$$\alpha_{m|l}^N(\varphi^*) \approx \exp\{-NE_{m|M}^*(E_{1|1}, E_{2|2}, ....., E_{M-1|M-1})\},$$
$$m = \overline{1,M}.$$

Using (9) and (4 - 5), all $E_{m|l}^*$ are strictly positive. The proof of part (a) will be finished if one demonstrate that the sequence of the test $\varphi^*$ is LAO, that is, at given finite

$E_{1|1}, E_{2|2}, ....., E_{M-1|M-1}$ for any other sequence of tests $\varphi^{**}$

$$E^*_{m|l}(\varphi^{**}) \le E^*_{m|l}(\varphi^*), \ m,l = \overline{1,M}.$$

For this purpose it is sufficient to see that the sequence of tests asymptotically dose not became better if the sets $B^N_m$ will not be union of some number of whole types $T^N_Q$, in other words, if a test $\varphi^{**}$ is defined, for example, by sets $G^N_1, G^N_2, ...., G^N_M$ and, in addition, $Q$ is such that $0 < | G^N_l \bigcap T^N_Q | \approx | T^N_Q |$,

The test $\varphi^{**}$ will not became worse if instead of the set $G^N_l$ one takes $G^N_l \supset T^N_Q$, it $G^N_l$ nonempty intersection with $T^N_Q$. At last is is able to prove the necessity of the condition (9).

If the sequence of the tests is LAO, then it can be given by sets of (10) form. But, the non-fulfillment of the conditions (9) is equivalent either to violation of (3), or to equality to zero some of $E^*_{m|l}$ given in (9), and this again contradicts with (3) because $E_{m|m}$, $m = \overline{1, M-1,}$ must be positive.

Remark 1

From definition (5) and (9) it follows that:

$$E^*_{m|m} = E^*_{m|M}, \ m = \overline{1, M-1}$$

Remark 2

After the change of hypotheses enumeration the theorem remains valid with corresponding changes in conditions (9).

Remark 3

The maximal likelihood test accepts the hypotheses maximising the probability of sample $X$. In fact

$$r^* = \arg\max_r P^N(X).$$

But it follows from equality $P^N(X) = \exp\{-N[H(Q) + D(Q \,||\, P)]\}$ that at the same time $r^* = \arg\min_r D(Q \,||\, P)$. In fact the principle of maximum of likelihood is equivalent to the principle of minimum ok Kullback-Leibler distance.

Acknowledgement

I am grateful to Professor E. A. Haroutunian for his very helpful comments which substantially improved the presentation of the paper.

References

Csiszar I., & Korner J. (1981). *Information theory: coding theorem for discrete memoryless systems.* NY: Academic Press.

Csiszar I., & Longo G. (1971). *On the error exponent for source coding and for testing simple statistical hypotheses, Studio sc. Mathem. Hungarica*, *6*, 181-191

Csiszar I., & Shields P. (2004). *Information theory and statistics*: *Fundamentals and trends in communications and information theory*. Hanover: MA.

Dembo A., &Zeitoni O. (1993). *Large deviations techniques and applications*, London: Jons and Bartlet .

Haroutunian, E. A. (1990). Logarithmically asymptotically optimal testing of multiple statistical hypotheses, *Problems of Control and Information Theory*, *19*, 413-421.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, *36*, 369-401.

Tusnady, G. (1977). On asymptotically optimal tests, *Annals of Statistics*, *5*(2), 385-393.

Longo, G., & Sgarro A. (1980). The error exponent for the testing of simple statistical hypotheses, A combinatorial. approach. *Journal of Combinotorics, Information. and System Sciences*, *5*(1), 58-67.

Perez, A. (1984). *Second-type-error exponent given the first-type-error exponent in the testing statistical hypotheses by unfitted procedures.* Abstract of papers of the Sixth International Symposium on Information Theory. (Part 1). Tashkent, 277-279.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 492--502
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Semi Parametric Estimation Of Some Reliability Measures Of Geometric Distribution

Mathachan Pathiyil
Nirmala College

E. S. Jeevanand
Union Christian College

Semi parametric estimators of the survival function, the hazard function, and the mean residual life function of geometric distribution using uncensored and Type II censored samples are obtained. The accuracy of the estimators so obtained is investigated empirically using simulated samples. The results are applied to a real life data set for illustration.

Key words: geometric distribution, hazard function, Kaplan-Meier estimator, least square estimation, mean residual life function, survival function, Type II censoring.

## Introduction

During the past twenty years, manufacturing industries have gone through a revolution in the use of statistical methods to improve product quality. Due to global competition, the industry faces immense pressure for shorter product-cycle times, stringent cost constraints, and higher customer expectations for quality and reliability. A natural extension of the revolution in product quality is to focus on product reliability, which is defined as quality over time. Reliability can be defined as the probability that a unit will perform its intended function until a specified point in time under *encountered* use conditions. The environment in which a product operates is a critical factor in evaluating the reliability of a product. The design for reliability requires careful consideration of product (process) failure modes. Broadly, failure modes can be classified as those that are anticipated and those that are unanticipated. Generally, engineers focus only on the anticipated failure modes. The main focus, however, of the statistician is in the unanticipated failures and it plays a crucial role in product reliability.

Reliability analysis of devices through failure time data when time is treated as discrete is a recently emerging area of research. Kemp (2004) provided a good discussion on the importance and applications of discrete life distributions. The sophisticated equipment used in the manufacturing process requires accurate measuring devices to record their failures in continuous time. In situations where such measuring instruments are very costly or their availability cannot be ensured, it may be desirable to go in for failure times that are in completed units of time (Xekalaki (1983)). The latter procedure is more desirable, provided the loss of accuracy in replacements of continuous measurements with discrete ones is more than compensated by the gain in terms of other considerations such as money, ease of analysis and time saved etc. Discrete distributions naturally arise when records are taken in completed units of time. The fact that many of the discrete distributions can be closely approximated by continuous distributions adds

E. S. Jeevanand received the Ph. D. from the Cochin University of Science and Technology, Cochin, Kerala, India. He is Reader in Statistics**.** His research interests include Bayes estimation, Inference, Reliability and Computational Statistics. He is Associate Editor of the *Half Yearly Discourses***.** Mathachan Pathiyil received his M. Phil from the University of Kerala, Kerala, India. He is Lecturer, Selection Grade, in the department of Statistics, Muvattupuzhaa, Kerala. His research interests include Bayes estimation, Inference and Reliability Analysis.

to the utility of the former as models of life length. Also, there is a well developed methodology to separately find the distribution of the integer parts and fractional parts of continuous random variables. This methodology often permits inference on parameters based on count data to be translated to those based on continuous measurements with a reasonable estimate of the margin of error on account of the translation. The geometric distribution owing to its lack of memory property is widely used to model such systems.

Reliability measures of the geometric distribution

An important property of a product or system is its ability to fulfill the intended purpose without failure for a specified period of time under stated conditions. Reliability is a yardstick of the capability of a component to operate without failure when put into service. The survival function, hazard function and mean residual life function are three important notions used extensively for characterizing life distributions.

Let X denote a discrete random variable in the support of $I^+ = \{0,1,...\}$ denoting the time to failure of a component. Defining

$$S(x) = P(X \geq x) \tag{2.1}$$

the survival function of $X$ and $f(x)$, the probability mass function of $X$, the hazard rate of $X$ is defined as

$$h(x) = P(X = x | X \geq x) = \frac{f(x)}{S(x)}. \tag{2.2}$$

and the mean residual life is defined as

$$r(x) = E(X - x | X > x) = \frac{1}{S(x+1)} \sum_{y=x+1}^{\infty} S(y). \tag{2.3}$$

Suppose that the life span $X$ of the component under observation follows a geometric distribution with probability mass function

$$f(x) = \theta(1-\theta)^x,\ 0 < \theta < 1,\ x = 0,1,... \tag{2.4}$$

Then

$$S(x) = (1-\theta)^x,\ 0 < \theta < 1, \tag{2.5}$$

$$h(x) = \theta,\ 0 < \theta < 1 \tag{2.6}$$

and

$$r(x) = \frac{1}{\theta},\ 0 < \theta < 1. \tag{2.7}$$

Estimation of the geometric parameter and the reliability measures using uncensored data

Wu (2001) and Faucher and Tyson (1988) proposed semi parametric estimation of the parameters of exponential and Pareto distributions using the empirical distribution function based on complete samples. The results are further extended, to the study of geometric distribution, by Mathachan and Jeevanand (2005). From (2.5)

$$\ln\left(S\left(x_{\{i\}}\right)\right) = x_{(i)} \ln(1-\theta). \tag{3.1}$$

Equation (3.1) can be written in the form, $Y_i = AX_i$, where $A = \ln(1-\theta)$, $X_i = x_{(i)}$ and $Y_{(i)} = \ln S(x_{(i)})$. By least square procedure, the estimator of $A$ is

$$\hat{A} = \frac{\sum_{i=1}^{n} \ln\left(S\left(x_{(i)}\right)\right)}{\sum_{i=1}^{n} x_{(i)}} \tag{3.2}$$

An estimate of the survival function $S\left(x_{(i)}\right)$ is $[1 - \hat{F}(x_{i:n}; q)]$ where $x_{i:n}$ is the $i^{th}$ order statistic and $\hat{F}(x_{i:n} : q) = \frac{i}{n}$, the empirical distribution function. In order to avoid $\log(0)$,

D'Agostino and Stephans (1986) suggested that, $\hat{F}(x_{i:n}:q)$ can be approximated by $\frac{i-c}{n-2c+1}$, $i=1,2,...,n$ where $0 \le c < 1$, generally. In this article three popular values for c are taken and considered in Wu (2001), Faucher and Tyson (1988), viz. $c = 0, 0.3$ and $0.5$. Then (3.2) becomes

$$\hat{A}_{FS} = \frac{\sum_{i=1}^{n} \ln\left(\frac{n+1-c-i}{n+1-2c}\right)}{\sum_{i=1}^{n} x_{(i)}} \qquad (3.3)$$

The estimated asymptotic variance of $\hat{A}$ is

$$\hat{V}\left(\hat{A}_{FS}\right) = \frac{\left(S_{yy} - \frac{S_{xy}}{S_{xx}}\right)}{n-2} \qquad (3.4)$$

where

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2, S_{yy} = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \text{ and}$$

$$S_{xy} = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}\ .$$

Estimation of the parameter of the geometric model

An estimator of the parameter of the model (2.4) using the uncensored sample is

$$\hat{\theta}_{FS} = 1 - e^{\hat{A}_{FS}} \qquad (3.5)$$

with estimated asymptotic variance

$$\hat{V}\left(\hat{\theta}_{FS}\right) = e^{2\hat{A}_{FS}} \hat{V}\left(\hat{A}_{FS}\right). \qquad (3.6)$$

Estimation of the survival function

The corresponding estimator of the survival function is

$$\hat{S}_{FS}(x) = Exp\left(x\,\hat{A}_{FS}\right) \qquad (3.7)$$

and the estimated asymptotic variance is

$$\hat{V}_{FS}\left(\hat{S}_{FS}(x)\right) = \hat{A}_{FS}\, e^{2x\,\hat{A}_{FS}} \hat{V}\left(\hat{A}_{FS}\right). \qquad (3.8)$$

Estimation of the hazard function

The estimator of the hazard function of the model (2.4) is

$$\hat{h}_{FS}(x) = \hat{\theta}_{FS} \qquad (3.9)$$

with estimated asymptotic variance

$$\hat{V}\left(\hat{h}_{FS}(x)\right) = e^{2\hat{A}_{FS}} \hat{V}\left(\hat{A}_{FS}\right). \qquad (3.10)$$

Estimation of the mean residual life function

The estimator of the mean residual life function of the model (2.4) is

$$\hat{r}_{FS}(x) = \left(1 - e^{\hat{A}_{FS}}\right)^{-1} \qquad (3.11)$$

with estimated asymptotic variance

$$\hat{V}\left(\hat{r}_{FS}\right) = \frac{e^{2\hat{A}_{FS}}}{\left(1 - e^{\hat{A}_{FS}}\right)^2} \hat{V}\left(\hat{A}_{FS}\right) \qquad (3.12)$$

Estimation of the geometric parameter and reliability measures using Type II censored data

With the high reliability products that are common today, testing under normal conditions is time consuming and even expensive. Thus, in life testing experiments, it is a common practice to cease testing before all the components under observation have failed. The resulting sample is a censored sample. Censored data occur frequently in medical research, and estimation of the reliability measures viz. survival function, hazard function and mean residual life function has been an attractive topic when the data are censored. Estimation of hazard function and mean residual life function has drawn less attention than that of the survival function. For the survival function, the Kaplan-Meier estimator (1958) is a widely used non-parametric estimator. It is strongly consistent and is asymptotically normal (see, Kim, et al., 2005, and Jan, et al., 2005). The focus of this our discussion is estimation of the survival function, hazard function and mean residual life function

of geometric distribution under Type II censoring.

Estimation of the parameter of the geometric model

A least square estimator is proposed for the parameter of the geometric distribution with survival function (2.5) when the data is censored at a pre defined time T. Suppose n components with geometric life times are put on test and observed the number of components failed at each time point $t,\ t+k,\ t+2k,\ldots$ up to the time T. Define, $n_j$ as the number of components still functioning at the time $t_j = t + jk,\ j = 0,1,\ldots,t_j \le T$ and $d_j$ as the number of components whose failures occur in the time interval $(t_{j-1}, t_j)$. Then, the Kaplan-Meier estimator of the survival function $S(t)$ (see Jan *et al.* (2005)) for a given $t$ is

$$S^*(t) = \prod_{j:t_{(j)} \le t} \left( \frac{n_j - d_j}{n_j} \right). \tag{4.1}$$

For $t < t_{(1)}$, $S^*(t) = 1$. From the survival function (2.5), the following may be written:

$$\ln\left(S^*(t)\right) = t \ln(1-\theta). \tag{4.2}$$

Now, (4.2) is of the form $y = At$, with $y = \ln\left(S^*(t)\right)$ and $A = \ln(1-\theta)$. By least square procedures, there is

$$\hat{A}_{CS} = \frac{\sum_{j=1}^{n} \ln\left(S^*(t_j)\right)}{\sum_{j=1}^{n} t_j}$$

and the estimated asymptotic variance of $\hat{A}$ is

$$\hat{V}\left(\hat{A}_{CS}\right) = \frac{\left(S_{yy} - \frac{S_{xy}}{S_{xx}}\right)}{n-2}.$$

Consequently, a least square estimator of the parameter $\theta$ is

$$\hat{\theta}_{CS} = 1 - e^{\hat{A}_{CS}} \tag{4.3}$$

with estimated asymptotic variance

$$\hat{V}(\theta_{CS}) = e^{2\hat{A}_{CS}} \hat{V}\left(\hat{A}_{CS}\right). \tag{4.4}$$

Estimation of the survival function

Substituting (4.3) in (2.5), a least square estimator of the survival function of the geometric distribution considered in (2.4) under Type II censoring scheme is obtained as

$$\hat{S}_{CS}(x) = e^{x\hat{A}_{CS}} \tag{4.5}$$

with asymptotic variance

$$\hat{V}\left(\hat{S}_{CS}(x)\right) = \hat{A}_{CS} e^{2x\hat{A}_{CS}} \hat{V}\left(\hat{A}_{CS}\right). \tag{4.6}$$

Estimation of the hazard function

Substituting (4.3) in (2.6), an estimator of the hazard function under the Type II censoring scheme is obtained as

$$\hat{h}_{CS} = \hat{\theta}_{CS} \tag{4.7}$$

and the asymptotic variance of the estimator is

$$\hat{V}\left(\hat{h}_{CS}\right) = e^{2\hat{A}_{CS}} \hat{V}\left(\hat{A}_{CS}\right). \tag{4.8}$$

Estimation of the mean residual life function

Substituting (4.3) in (2.7), an estimator of the mean residual life function under this scheme is obtained as

$$\hat{r}_{CS} = \frac{1}{\hat{\theta}_{CS}} \tag{4.9}$$

with asymptotic variance

$$\hat{\mathrm{V}}\left(\hat{\mathrm{r}}_{\mathrm{CS}}(\mathrm{x})\right) = \frac{e^{2\hat{A}_{CS}}}{\left(1 - e^{\hat{A}_{CS}}\right)^2} \hat{V}\left(\hat{A}_{CS}\right). \tag{4.10}$$

Simulation Study

These procedures are assessed by a numerical study based on simulated samples with different values of the parameters of the model. Performances of the proposed estimators are examined empirically by generating samples from the geometric distribution. In the study, the bias of the estimator is defined as

Bias = Average value of the estimate – Parameter value

and the mean square error (MSE) of an estimator is determined as

MSE = Variance of the estimator + (Bias) $^2$

The simulated absolute bias, $SD$ and $MSE$ of the estimators proposed for the reliability measures using uncensored samples of sizes 20, 50 and 100 for 1000 replications corresponding to different choices of the parameter values are given in Tables 1 – 3, respectively.

The simulated absolute bias, $SD$ and $MSE$ of the estimators proposed for the reliability measures using Type II censored samples of sizes 20, 50 and 100 with different censoring schemes (i. e., for different choices of the censoring time ) for 1,000 replications are given in following tables, Table 4 – 6, respectively. These values are computed for different values of the parameters.

A few features observed from Table 1 – 6:

1. For smaller values of $\theta$, the estimators under the two methods proposed have lesser bias and mean square error and a reverse trend is seen for larger values of $\theta$.
2. It seems that the bias and mean square error of all the proposed estimators become smaller as the sample size (or censoring times) increases for a given $\theta$.

Example for assessing the estimators with real data

The pattern of natal dispersal in vertebrate animals is an important factor affecting the genetic and demographic processes within and between populations. The geometric probability distribution is a common way to model the frequency distribution of vertebrate dispersal distances (Porter and Dooley (1993), Greenwood et al., (1979)). They define $X$ as the number of units (home ranges, habitat, nest sites, territories etc. with a fixed diameter) moved before stopping (settling and / or dying) and $\theta$, the probability of stopping while crossing any one unit of habitat before moving to an additional home- range diameter.

For an illustration of the present study, used is the data about the dispersal distance (in units of 200 meters diameter) from natal site to first year breeding site for different categories of 117 one-year-old male great tits given in page 141, Appendix I, Greenwood et al. (1979). The estimators of the geometric parameter and the reliability measures based on the censored and uncensored samples are given in Table 7.

The estimators are computed using Type II censored samples with different choices of the censoring time . The classical estimator of the geometric parameter under the maximum likelihood method of estimation ( $MLE$ ) is

$$\hat{\theta}_{MLE} = \frac{n}{n + \sum_{i=1}^{n} x_i}.$$

For the above data, the value 0.2566 is obtained. Table 7 suggests that the new semi-parametric estimates suggested are close to the $MLE$ in most cases.

Acknowledgements

The first author is thankful to the University Grants Commission, New Delhi, India, for the financial support.

Table 1: Estimators of $S(x)$ using uncensored samples when $x = 5$

| $\theta$ | $S(x)$ | $n$ | $c$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|---|
| **0.1** | 0.59049 | 20 | 0 | 0.0085 | 0.0764 | 0.0059 |
| | | | 0.3 | 0.0182 | 0.0774 | 0.0063 |
| | | | 0.5 | 0.0266 | 0.0782 | 0.0068 |
| | | 50 | 0 | 0.0141 | 0.0463 | 0.0023 |
| | | | 0.3 | 0.0179 | 0.0466 | 0.0025 |
| | | | 0.5 | 0.0224 | 0.0469 | 0.0027 |
| | | 100 | 0 | 0.0127 | 0.0336 | 0.0013 |
| | | | 0.3 | 0.0150 | 0.0337 | 0.0014 |
| | | | 0.5 | 0.0176 | 0.0338 | 0.0015 |
| **0.3** | 0.16807 | 20 | 0 | 0.0339 | 0.0659 | 0.0055 |
| | | | 0.3 | 0.0417 | 0.0639 | 0.0058 |
| | | | 0.5 | 0.0481 | 0.0622 | 0.0062 |
| | | 50 | 0 | 0.0394 | 0.0419 | 0.0033 |
| | | | 0.3 | 0.0439 | 0.0412 | 0.0036 |
| | | | 0.5 | 0.0474 | 0.0406 | 0.0039 |
| | | 100 | 0 | 0.0443 | 0.0301 | 0.0029 |
| | | | 0.3 | 0.0470 | 0.0297 | 0.0031 |
| | | | 0.5 | 0.0491 | 0.0295 | 0.0033 |
| **0.5** | 0.03125 | 20 | 0 | 0.0171 | 0.0178 | 0.0006 |
| | | | 0.3 | 0.0186 | 0.0164 | 0.0006 |
| | | | 0.5 | 0.0198 | 0.0152 | 0.0006 |
| | | 50 | 0 | 0.0212 | 0.0086 | 0.0005 |
| | | | 0.3 | 0.0219 | 0.0081 | 0.0005 |
| | | | 0.5 | 0.0225 | 0.0078 | 0.0006 |
| | | 100 | 0 | 0.0229 | 0.0059 | 0.0006 |
| | | | 0.3 | 0.0233 | 0.0056 | 0.0006 |
| | | | 0.5 | 0.0236 | 0.0055 | 0.0006 |
| **0.7** | 0.00243 | 20 | 0 | 0.0020 | 0.0019 | 0.0000 |
| | | | 0.3 | 0.0021 | 0.0016 | 0.0000 |
| | | | 0.5 | 0.0021 | 0.0015 | 0.0000 |
| | | 50 | 0 | 0.0023 | 0.0004 | 0.0000 |
| | | | 0.3 | 0.0023 | 0.0003 | 0.0000 |
| | | | 0.5 | 0.0023 | 0.0003 | 0.0000 |
| | | 100 | 0 | 0.0024 | 0.0001 | 0.0000 |
| | | | 0.3 | 0.0024 | 0.0001 | 0.0000 |
| | | | 0.5 | 0.0024 | 0.0001 | 0.0000 |

Table 2: Estimators of $h$ using uncensored samples

| $\theta$ | $h$ | $n$ | $c$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|---|
| **0.1** | 0.1 | 20 | 0 | 0.0039 | 0.0247 | 0.0006 |
| | | | 0.3 | 0.0070 | 0.0254 | 0.0007 |
| | | | 0.5 | 0.0097 | 0.0260 | 0.0008 |
| | | 50 | 0 | 0.0048 | 0.0145 | 0.0002 |
| | | | 0.3 | 0.0060 | 0.0147 | 0.0003 |
| | | | 0.5 | 0.0074 | 0.0149 | 0.0003 |
| | | 100 | 0 | 0.0041 | 0.0105 | 0.0001 |
| | | | 0.3 | 0.0049 | 0.0106 | 0.0001 |
| | | | 0.5 | 0.0057 | 0.0107 | 0.0001 |
| **0.3** | 0.3 | 20 | 0 | 0.0458 | 0.0735 | 0.0075 |
| | | | 0.3 | 0.0543 | 0.0747 | 0.0085 |
| | | | 0.5 | 0.0618 | 0.0757 | 0.0096 |
| | | 50 | 0 | 0.0425 | 0.0456 | 0.0039 |
| | | | 0.3 | 0.0473 | 0.0460 | 0.0044 |
| | | | 0.5 | 0.0512 | 0.0464 | 0.0048 |
| | | 100 | 0 | 0.0448 | 0.0328 | 0.0031 |
| | | | 0.3 | 0.0478 | 0.0330 | 0.0034 |
| | | | 0.5 | 0.0501 | 0.0332 | 0.0036 |
| **0.5** | 0.5 | 20 | 0 | 0.1315 | 0.1199 | 0.0316 |
| | | | 0.3 | 0.1422 | 0.1193 | 0.0344 |
| | | | 0.5 | 0.1514 | 0.1187 | 0.0370 |
| | | 50 | 0 | 0.1259 | 0.0716 | 0.0210 |
| | | | 0.3 | 0.1322 | 0.0716 | 0.0226 |
| | | | 0.5 | 0.1373 | 0.0715 | 0.0240 |
| | | 100 | 0 | 0.1297 | 0.0526 | 0.0196 |
| | | | 0.3 | 0.1336 | 0.0526 | 0.0206 |
| | | | 0.5 | 0.1367 | 0.0526 | 0.0214 |
| **0.7** | 0.7 | 20 | 0 | 0.1849 | 0.0882 | 0.0420 |
| | | | 0.3 | 0.1915 | 0.0853 | 0.0439 |
| | | | 0.5 | 0.1969 | 0.0827 | 0.0456 |
| | | 50 | 0 | 0.1907 | 0.0607 | 0.0401 |
| | | | 0.3 | 0.1946 | 0.0594 | 0.0414 |
| | | | 0.5 | 0.1977 | 0.0584 | 0.0425 |
| | | 100 | 0 | 0.1956 | 0.0427 | 0.0401 |
| | | | 0.3 | 0.1980 | 0.0422 | 0.0410 |
| | | | 0.5 | 0.1999 | 0.0417 | 0.0417 |

Table 3: Estimators of $r$ using uncensored samples

| $\theta$ | $r$ | $n$ | $c$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|---|
| **0.1** | 10 | 20 | 0 | 0.1385 | 2.2796 | 5.2156 |
| | | | 0.3 | 0.1556 | 2.2097 | 4.9072 |
| | | | 0.5 | 0.4002 | 2.1517 | 4.7898 |
| | | 50 | 0 | 0.2734 | 1.3644 | 1.9362 |
| | | | 0.3 | 0.3830 | 1.3481 | 1.9640 |
| | | | 0.5 | 0.5105 | 1.3292 | 2.0273 |
| | | 100 | 0 | 0.3013 | 0.9592 | 1.0108 |
| | | | 0.3 | 0.3680 | 0.9522 | 1.0422 |
| | | | 0.5 | 0.4443 | 0.9442 | 1.0889 |
| **0.3** | 3.3333 | 20 | 0 | 0.3114 | 0.6412 | 0.5082 |
| | | | 0.3 | 0.3863 | 0.6211 | 0.5350 |
| | | | 0.5 | 0.4485 | 0.6043 | 0.5663 |
| | | 50 | 0 | 0.3625 | 0.3920 | 0.2851 |
| | | | 0.3 | 0.4039 | 0.3851 | 0.3114 |
| | | | 0.5 | 0.4370 | 0.3795 | 0.3350 |
| | | 100 | 0 | 0.4069 | 0.2778 | 0.2428 |
| | | | 0.3 | 0.4319 | 0.2748 | 0.2620 |
| | | | 0.5 | 0.4514 | 0.2724 | 0.2780 |
| **0.5** | 2 | 20 | 0 | 0.3576 | 0.3206 | 0.2306 |
| | | | 0.3 | 0.3873 | 0.3092 | 0.2456 |
| | | | 0.5 | 0.4119 | 0.2996 | 0.2594 |
| | | 50 | 0 | 0.3812 | 0.1869 | 0.1802 |
| | | | 0.3 | 0.3977 | 0.1831 | 0.1917 |
| | | | 0.5 | 0.4108 | 0.1801 | 0.2012 |
| | | 100 | 0 | 0.4008 | 0.1356 | 0.1791 |
| | | | 0.3 | 0.4108 | 0.1339 | 0.1867 |
| | | | 0.5 | 0.4185 | 0.1326 | 0.1927 |
| **0.7** | 1.42857 | 20 | 0 | 0.2857 | 0.1309 | 0.0988 |
| | | | 0.3 | 0.2951 | 0.1246 | 0.1026 |
| | | | 0.5 | 0.3028 | 0.1193 | 0.1059 |
| | | 50 | 0 | 0.3004 | 0.0819 | 0.0969 |
| | | | 0.3 | 0.3055 | 0.0795 | 0.0997 |
| | | | 0.5 | 0.3096 | 0.0776 | 0.1019 |
| | | 100 | 0 | 0.3094 | 0.0547 | 0.0987 |
| | | | 0.3 | 0.3125 | 0.0537 | 0.1005 |
| | | | 0.5 | 0.3148 | 0.0529 | 0.1019 |

Table 4: Estimators of $S(x)$ using censored samples when $x = 5$

| $\theta$ | $S(x)$ | $t$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|
| **0.1** | 0.5905 | 5 | 0.0991 | 0.0118 | 0.0100 |
| | | 10 | 0.0506 | 0.0143 | 0.0028 |
| | | 20 | 0.0240 | 0.0065 | 0.0006 |
| | | 30 | 0.0126 | 0.0108 | 0.0003 |
| **0.3** | 0.1681 | 5 | 0.0701 | 0.0059 | 0.0049 |
| | | 10 | 0.0334 | 0.009 | 0.0012 |
| | | 15 | 0.0261 | 0.0066 | 0.0007 |
| **0.5** | 0.0313 | 5 | 0.0216 | 0.0023 | 0.0005 |
| | | 7 | 0.0165 | 0.0045 | 0.0003 |
| | | 9 | 0.0114 | 0.0056 | 0.0002 |
| **0.7** | 0.0024 | 5 | 0.0019 | 0.0002 | 0.0000 |
| | | 7 | 0.0005 | 0.0004 | 0.0000 |

Table 5: Estimators of $h$ using censored samples

| $\theta$ | $h$ | $t$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|
| **0.1** | 0.1 | 5 | 0.0325 | 0.0042 | 0.0011 |
| | | 10 | 0.0160 | 0.0047 | 0.0003 |
| | | 20 | 0.0074 | 0.0020 | 0.0001 |
| | | 30 | 0.0039 | 0.0033 | 0.0000 |
| **0.3** | 0.3 | 5 | 0.0718 | 0.0076 | 0.0052 |
| | | 10 | 0.0305 | 0.0088 | 0.0010 |
| | | 15 | 0.0233 | 0.0062 | 0.0006 |
| **0.5** | 0.5 | 5 | 0.1060 | 0.0203 | 0.0117 |
| | | 7 | 0.0721 | 0.0262 | 0.0059 |
| | | 9 | 0.0454 | 0.0241 | 0.0026 |
| **0.7** | 0.7 | 5 | 0.0772 | 0.0169 | 0.0062 |
| | | 7 | 0.0132 | 0.0117 | 0.0003 |

Table 6: Estimators of $r$ using censored samples

| $\theta$ | $r$ | $t$ | *Bias* | *SD* | *MSE* |
|---|---|---|---|---|---|
| **0.1** | 10 | 5 | 2.4462 | 0.2343 | 6.039 |
| | | 10 | 1.3706 | 0.3536 | 2.0036 |
| | | 20 | 0.6898 | 0.1782 | 0.5076 |
| | | 30 | 0.3674 | 0.3133 | 0.2332 |
| **0.3** | 3.3333 | 5 | 0.6425 | 0.0549 | 0.4159 |
| | | 10 | 0.3060 | 0.0826 | 0.1005 |
| | | 15 | 0.2393 | 0.0605 | 0.0609 |
| **0.5** | 2 | 5 | 0.3485 | 0.054 | 0.1243 |
| | | 7 | 0.2492 | 0.0804 | 0.0686 |
| | | 9 | 0.1635 | 0.0836 | 0.0337 |
| **0.7** | 1.4286 | 5 | 0.1414 | 0.0278 | 0.0208 |
| | | 7 | 0.0261 | 0.0229 | 0.0012 |

Table 7: Estimators of the parameter and the reliability measures using the real data when $x = 5$

| Censoring time $t$ | Estimate of $\theta$ | Estimate of $S(x)$ | Estimate of $h$ | Estimate of $r$ |
|---|---|---|---|---|
| **5** | 0.3345 | 0.1305 | 0.3345 | 2.9892 |
| **10** | 0.3229 | 0.1424 | 0.3229 | 3.0973 |
| **15** | 0.3102 | 0.1562 | 0.3102 | 3.2298 |
| Uncensored data | 0.2912 | 0.1790 | 0.2912 | 3.4347 |

References

D'Agostino, R. B. & Stephans, M. A. (1986). *Goodness of Fit Techniques.* Marcel Dekker, Inc., New York.

Faucher, B. & Tyson, W.R. (1988). On the determination of Weibull parameters. *Journal of Material Science Letters, 7*, 1199-1203.

Greenwood, P. J. Harvey, P. H. & Perrins, C. M. (1979). The role of dispersal in the Great Tit (Parus Major): The causes, consequences and heritability of natal dispersal. *Journal of Animal Ecology, 48,* 123-142.

Jan, B., Shah, S.W.A., Shah, S. & Qadir, M. F. (2005). Weighted Kaplan - Meier estimation of survival function in heavy censoring. *Pakistan Journal of Statistics, 21 (1),* 55- 83.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457- 481.

Kemp, A.W. (2004). Classes of discrete life distributions. *Communications in Statistics-Theory and Methods, 33(12),* 3069-3093.

Kim, C., Bae, W., Cho, H. & Park, B. U. (2005). Nonparametric hazard function estimation using the Kaplan – Meier estimator. *Nonparametric Statistics, 17 (8),* 937 – 948.

Mathachan Pathiyil & Jeevanand, E. S. (2005). Estimation of some reliability measures of geometric distribution by least square procedure. *Recent Advances in Statistical Theory and Applications,* 1, 117 – 125.

Porter, J.H & Dooley J.L (1993). Animal dispersal patterns: A reassessment of simple mathematical models. *Ecology, 74(8),* 2436-2443.

Wu, J.W. (2001). A note on determining the number of outliers in exponential sample by least square procedure. *Statistical papers,* 42*(4),* 489-503.

Xekalaki, E. (1983). Hazard functions and life distributions in discrete time. *Communications in Statistics - Theory and Methods*, *12,* 2503-2509.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 503- 516
Copyright © 2007 JMASM, Inc.
1538 – 9472/05/$95.00

# Inference on Overlapping Coefficients in Two Exponential Populations

Mohammad F. Al-Saleh
Yarmouk University

Hani M. Samawi
Georgia Southern University

Three measures of overlap, namely Matusita's measure $\rho$, Morisita's measure $\lambda$ and Weitzman's measure $\Delta$ are investigated in this article for two exponential populations with different means. It is well that the estimators of those measures of overlap are biased. The bias is of these estimators depends on the unknown overlap parameters. There are no closed-form, exact formulas, for those estimators variances or their exact sampling distributions. Monte Carlo evaluations are used to study the bias and precision of the proposed overlap measures. Bootstrap method and Taylor series approximation are used to construct confidence intervals for the overlap measures.

Key words: Bootstrap method; Matusita's measure; Morisita's measure; overlap coefficients; Taylor expansion; Weitzman's measure.

## Introduction

Overlap measure are commonly used in reliability analysis to estimate the proportion of machines or electronic devices that have similar range of failure time. The machines may come from two different sources or may be under different stress, which implies different probability densities of failure time. This proportion can be measured by the overlap coefficients of the two densities.

There are three overlap coefficients (OVL), (Matusita's measure $\rho$, Morisita's measure $\lambda$ and Weitzman's measure $\Delta$). However, the most commonly used overlap coefficient is the Weitzman's measure $\Delta$. OVL measure is defined to be the area intersected by the graphs of two probability density functions. It measures the similarity, the agreement or the closeness of the two probability distributions.

Mohammad Fraiwan Al-Saleh is Associate Professor, College of Science, Department of Statistics, Yarmouk University, Irbid-Jordan. Email: m-saleh@yu.edu.jo. Hani M. Samawi is Associate Professor of Biostatistics at Georgia Southern University. Email: hsamawi@georgiasouthern.edu.

The OVL measure $\Delta$ was originally introduced by Weitzman (1970). Recently, many authors considered this measure, see Bradley and Piantadosi (1982), Inman and Bradley(1989), Clemons (1996), Reiser and Faraggi (1999), Clemons and Bradley (2000) and Mulekar and Mishra (2000).

For other applications of $\Delta$, see Ichikawa (1993) (for the probability of failure in the stress-strength models of reliability analysis), Fedeer et al. (1963) (for estimating of the proportion of genetic deviates in segregating populations and Sneath (1977) (as a measure of distinctness of clusters). For additional references of such methodology applications in ecology and other fields, see Mulekar and Mishra (1994 and 2000). Inman and Bradley (1989) summarized the history of such procedures.

Let $f_1(x)$ and $f_2(x)$ be two probability density functions. Assume samples of observations are drawn from continuous distributions (Slobdchikoff and Schulz, 1980; Harner and Whiytmorte, 1997; MacArthur, 1972). The overlap measures are defined as follows:

Matusita's Measure (1955):

$$\rho = \int \sqrt{f_1(x) f_2(x)}\, dx,$$

Morisita's Measure (1959):

$$\lambda = \frac{2\int f_1(x) f_2(x)\, dx}{\int [f_1(x)]^2\, dx + \int [f_2(x)]^2\, dx},$$

and

Weitzman's Measure (1970):

$$\Delta = \int \min\{f_1(x), f_2(x)\}\, dx.$$

These measures can be directly applied to discrete distributions by replacing the integrals with summations and also can be generalized to multivariate distributions. All three overlap measures of two densities are measured on the scale of 0 to 1. Note that the overlap value close to 0 indicates extreme inequality of the two density functions, and the overlap value of 1 indicates exact equality.

Smith (1982) derived formulas for estimating the mean and the variance of the discrete version of Weizman's measure using delta method. Mishra et al. (1986) gave some properties of the sampling distributions for a function of the $\Delta$ estimator, under the assumption of homogeneity of variances for the case of two normal distributions. Mulekar and Mishra (1994) simulated the sampling distribution of estimators of the overlap measures for normal densities with equal means and obtained the approximate expressions for the bias and variance of their estimators. Lu et al. (1989) investigated the sampling variability of some estimators of these measures using simulation.

Dixon (1993) described the use of the bootstrap and jackknife techniques for Gini coefficient of size hierarchy and Jaccard index of community similarity. Mulekar and Mishra (2000) addressed the problem of making inferences about the overlap coefficients for two normal densities with equal means using jackknife, bootstrap, transformation and Taylor series approximation. Reiser and Faraggi (1999) considered the problem of making inference about the overlap coefficient $\Delta$, as a measure of bioequivalence, under the name, proportion of similar responses, for normal densities with the equal variances, based on the non-central *t*- and *F*- distributions. The sampling behavior of a nonparametric estimator of $\Delta$ was examined by Clemons and Bradley (2000), using Monte Carlo and bootstrap techniques. Finally, AL-Saidy et al. (2005) consider the problem of drawing inference about the three overlap measures under the Weibul distribution function with equal shape parameter.

Although, the exponential distribution is a special case of the Weibul distribution, this article considers the three proposed measures of overlap ($\rho$, $\lambda$ and $\Delta$) for two exponential distributions with different means. This special case provides some neat and closed form results. Exponential distributions are primarily used in reliability applications. They are used to model data with a constant failure rate (indicated by the hazard plot which is simply equal to a constant). Exponential distributions are the most commonly used life distribution models (see Mann et al. 1974.)

A random variable *X* follows the exponential (denotes by *EXP($\theta$)*) if it has the cdf and pdf given by:

$$F(x) = 1 - \exp\left\{-\frac{x}{\theta}\right\} \quad \text{for } x > 0, \quad (1.1)$$

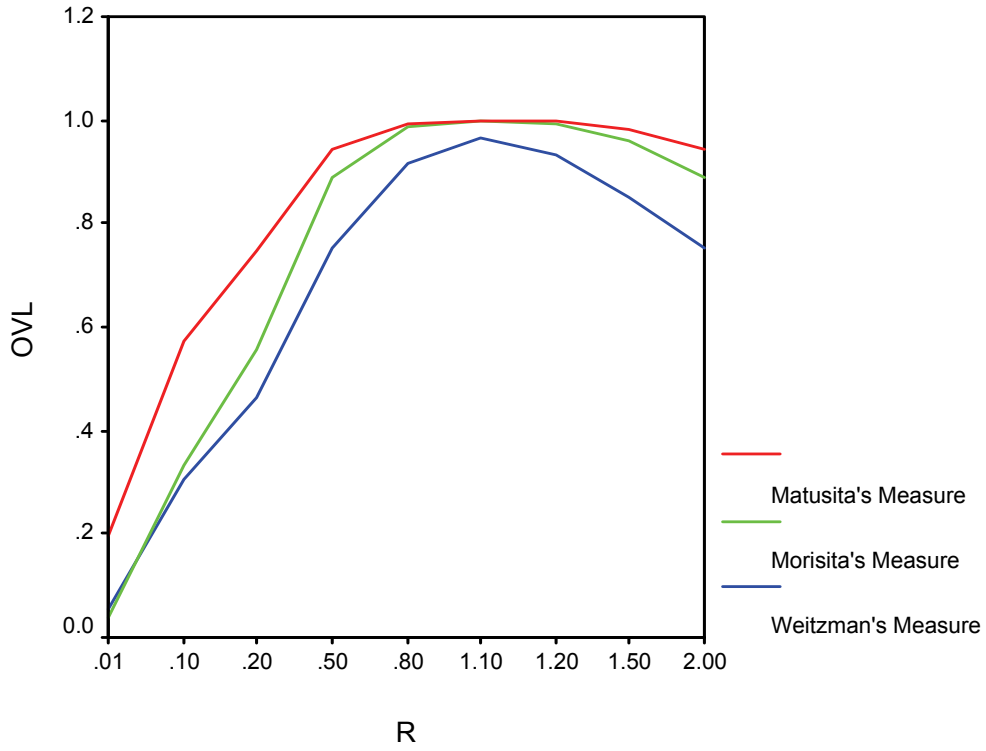
and

$$f(x) = \frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\} \quad \text{for } x > 0 \quad (1.2)$$

respectively, where $\theta > 0$ .

Overlap measures (OVL) for Exponential Distribution

Suppose $f_1(x)$ and $f_2(x)$ represent the exponential densities with means $\theta_1$ and $\theta_2$ respectively. Letting $R = \frac{\theta_1}{\theta_2}$, then the continuous version of the three proposed overlap measures can be expressed as a function of *R* as




follows (the derivation of the three overlap measures is straightforward and it is omitted from the content of this article):

$$\rho = \frac{2\sqrt{R}}{1+R}, \qquad (2.1)$$

$$\lambda = \frac{4R}{(1+R)^2} \qquad (2.2)$$

and

$$\Delta = 1 - R^{\frac{1}{1-R}} \left| 1 - \frac{1}{R} \right|, \quad R \neq 1. \qquad (2.3)$$

Figure 1 shows curves of the three overlap measures. All three measures are not monotone for all $R>0$. Similar to Mulekar and Mishra (2000), $\rho$, $\lambda$ and $\Delta$ have nice properties, such as, symmetry in $R$, i.e. OVL($R$)=OVL($1/R$) and invariance under linear transformation, $Y=aX+b$, $a \neq 0$. They all attain the maximum value of 1 at $R$=1.

Statistical Inference

Estimation

The OVL measures $\rho$, $\lambda$, and $\Delta$ are functions of $\theta_1$ and $\theta_2$. In order to draw any inference about the OVL measures, one first needs to get estimates of $\theta_1$ and $\theta_2$. Suppose that $(X_{11}, X_{12}, ..., X_{1n_1})$ and $(X_{21}, X_{22}, ..., X_{2n_1})$ are two independent random samples drawn from $f_1(x)$ and $f_2(x)$ respectively, where

$$f_1(x) = \frac{1}{\theta_1} \exp\left\{ -\frac{x}{\theta_1} \right\} \quad \text{for } x > 0$$

and

$$f_2(x) = \frac{1}{\theta_2} \exp\left\{ -\frac{x}{\theta_2} \right\} \quad \text{for } x > 0$$

The maximum likelihood estimators (MLEs) based on the two samples are given by:

1) From the first sample:

$$\hat{\theta}_1 = \overline{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}. \quad (3.1)$$

2) From the second sample

$$\hat{\theta}_2 = \overline{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}. \quad (3.2)$$

Note that, it is easy to show that $\hat{\theta}_1 \sim \mathrm{G}(n_1, \frac{\theta_1}{n_1})$ and $\hat{\theta}_2 \sim \mathrm{G}(n_2, \frac{\theta_2}{n_2})$, where G(., .) stands for the gamma distribution function. Hence, the variances of those MLE's are respectively $Var(\hat{\theta}_1) = \frac{\theta_1^2}{n_1}$ and $\mathrm{Var}(\hat{\theta}_2) = \frac{\theta_2^2}{n_2}$. Also, the MLE of $R$ is $\hat{R} = \frac{\hat{\theta}_1}{\hat{\theta}_2}$. Therefore, using the relationship between Gamma distribution and Chi-square distribution and the fact that the two samples are independent, it is easy to show that $\frac{\theta_2}{\theta_1}\hat{R}$ has F-distribution{i.e., $F(2n_1, 2n_2)$}. Hence, the variance of $\hat{R}$ is $Var(\hat{R}) = R^2 \frac{n_2^2(n_1+n_2-1)}{n_1(n_2-1)^2(n_2-2)}$. Also, an unbiased estimate of $R$ is given by $\hat{R}^* = \frac{\hat{\theta}_1}{\hat{\theta}_2}\frac{(n_2-1)}{n_2} = \frac{(n_2-1)}{n_2}\hat{R}$ with $Var(\hat{R}^*) = R^2 \frac{(n_1+n_2-1)}{n_1(n_2-2)}$. Clearly, $\hat{R}^*$ has less variance than $\hat{R}$.

The OVL measures considered here are functions of $R$, therefore, based on the MLE estimate of $R$, the OVL coefficients can be estimated by

$$\hat{\rho} = \frac{2\sqrt{\hat{R}^*}}{1+\hat{R}^*}, \quad (3.3)$$

$$\hat{\lambda} = \frac{4\hat{R}^*}{(1+\hat{R}^*)^2}, \quad (3.4)$$

and

$$\hat{\Delta} = 1 - (\hat{R}^*)^{\frac{1}{1-\hat{R}^*}}\left|1 - \frac{1}{\hat{R}^*}\right|. \quad (3.5)$$

Asymptotic Properties

Let OVL$= g(R)$, then $\widehat{OVL} = g(\hat{R}^*)$. Thus using the well-known delta method (Taylor series expansion) the approximate sampling variance of the OVL measures can be obtained as follows:

$$\mathrm{Var}(\hat{\rho}) = \sigma_{\hat{\rho}}^2 \approx \frac{R(1-R)^2(n_1+n_2-1)}{n_1(n_2-2)(1+R)^4}, \quad (3.6)$$

$$\mathrm{Var}(\hat{\lambda}) = \sigma_{\hat{\lambda}}^2 \approx \frac{16R^2(1-R)^2(n_1+n_2-1)}{n_1(n_2-2)(1+R)^6}, \quad (3.7)$$

and

$$\mathrm{Var}(\hat{\Delta}) = \sigma_{\hat{\Delta}}^2 \approx \frac{(n_1+n_2-1)(R)^{\frac{2}{1-R}}(\ln R)^2}{n_1(n_2-2)(1-R)^2}. \quad (3.8)$$

It is known that the estimators of those OVL coefficients are biased. Approximations for the biases of the OVL coefficients estimates, using Taylor series expansion, are as follow:

1. $Bias(\hat{\rho}^*) = \frac{(n_1+n_2-1)\sqrt{R}}{n_1(n_2-2)}\ \frac{3R(R-2)-1}{2(R+1)^3}$

2. $Bias(\hat{\lambda}^*) =$

$$\frac{(n_1+n_2-1)}{n_1(n_2-2)}\ \frac{8R^2(R-2)}{(R+1)^4}$$

3. $Bias(\hat{\Delta}^*) =$

$$\begin{Bmatrix} \frac{(n_1+n_2-1)R^2}{n_1(n_2-2)} \\ \frac{R^{\frac{2R-1}{1-R}}[R(2R-Ln(R)-2)Ln(R)-(R-1)^2]}{(R-1)^3} \text{ if } R>1 \\ \frac{(n_1+n_2-1)R^2}{n_1(n_2-2)} \\ \frac{R^{\frac{2R-1}{1-R}}[R(2R-Ln(R)-2)Ln(R)-(R-1)^2]}{(1-R)^3} \text{ if } R<1 \end{Bmatrix}.$$

Reasonable estimates for the above variances and the biases can be obtained by substituting $R$ by $\hat{R}^*$ in the above formulas.

Interval estimation

Transformation Technique

From Section 3.1, $\frac{\theta_2}{\theta_1}\hat{R} \sim F(2n_1, 2n_2)$, then $\frac{\theta_2(n_2)}{\theta_1(n_2-1)}\hat{R}^* \sim F(2n_1, 2n_2)$. Let $L$ and $U$ be the lower and upper confidence limits respectively of $R$, corresponding to the inclusion probability $1-\alpha$. Thus $L$ and $U$ can be determined by solving for $R$ the equation $P(F^{\alpha/2}_{(2n_1,2n_2)} < \frac{\theta_2}{\theta_1}\hat{R} < F^{1-\alpha/2}_{(2n_1,2n_2)}) = 1-\alpha$, where $F^{\alpha/2}_{(2n_1,2n_2)}$ and $F^{1-\alpha/2}_{(2n_1,2n_2)}$ are the lower and the upper $\alpha/2$ quantile of the $F(2n_1, 2n_2)$ distribution respectively. Thus $L = \frac{\hat{R}}{F^{1-\alpha/2}_{(2n_1,2n_2)}}$ and $U = \frac{\hat{R}}{F^{\alpha/2}_{(2n_1,2n_2)}}$. However, the OVL coefficients are not monotone functions of $R$ therefore, the $100(1-\alpha)$%confidence intervals for the OVL coefficients can be obtained using the transformation technique as follows:

1. $\{Min(\frac{2\sqrt{L}}{(L+1)}, \frac{2\sqrt{U}}{(U+1)}) \le \rho \le Max(\frac{2\sqrt{L}}{(L+1)}, \frac{2\sqrt{U}}{(U+1)})\}$

2. $\{Min(\frac{4L}{(L+1)^2}, \frac{4U}{(U+1)^2}) \le \lambda \le Max(\frac{4L}{(L+1)^2}, \frac{4U}{(U+1)^2})\}$

3. $\{Min(1-L^{\frac{1}{1-L}}|1-\frac{1}{L}|, 1-U^{\frac{1}{1-U}}|1-\frac{1}{U}|) \le \Delta \le Max(1-L^{\frac{1}{1-L}}|1-\frac{1}{L}|, 1-U^{\frac{1}{1-U}}|1-\frac{1}{U}|)\}$

Asymptotic technique

Normal approximation to the sampling distribution, using Delta-method, work fairly well for large sample because of the nice asymptotic properties of the MLE estimates of the exponential distribution. Therefore, the $100(1-\alpha)$% confidence intervals for the OVL coefficients can be computed easily as $\{O\hat{V}L - Z_{1-\alpha/2}\hat{\sigma}_{O\hat{V}L}, O\hat{V}L + Z_{1-\alpha/2}\hat{\sigma}_{O\hat{V}L}\}$, where $Z_{1-\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution.

These confidence intervals are not the best because of the bias involved in OVL coefficients estimates, however, for large samples they work fairly well. In Section 3.2, approximate the bias of those OVL coefficients. Using these approximations, the bias corrected interval can be computed as

$$\{(O\hat{V}L - Bias(O\hat{V}L) - Z_{1-\alpha/2}\hat{\sigma}_{O\hat{V}L}, (O\hat{V}L - Bias(O\hat{V}L)) + Z_{1-\alpha/2}\hat{\sigma}_{O\hat{V}L}\}.$$

Bootstrap Interference

Bootstrap methods are computer intensive which involves simulated data sets. Uniform (ordinary) bootstrap resampling by Efron (1979) is based on resampling with replacement from the observed sample according to a rule which places equal probabilities on sample values. Uniform bootstrap resampling as described by Efron(1979) and others is an assumption-free method that can be used for some inferential problems. However, it is designed for complete and continuous set of observations. For two-sample case the uniform resampling rules will apply to each sample separately and independently (see Ibrahim, 1991; Samawi et al., 1996; Samawi et al., 1998).

Suppose $\aleph_1 = (X_{11}, X_{12}, ..., X_{1n_1})$ and $\aleph_2^* = (X_{21}^*, X_{22}^*, ..., X_{2n_2}^*)$ are two independent random samples drawn from $f_1(x)$ and $f_2(x)$ respectively. Assume that the parameter of interest is the OVL coefficient, say $\theta$. Let $S$ be an estimate based on the random samples $\aleph_1$ and $\aleph_2$ i.e., $S=S(\aleph_1, \aleph_2)$. Furthermore, assume $S$ is a smooth function of the samples. Assume that $U$ is a function of $S$ i.e., $U = U(S)$. Write $U^*$ for the same function of the data but in resamples $\aleph_1^* = (X_{11}^*, X_{12}^*, ..., X_{1n_1}^*)$ and $\aleph_2^* = (X_{21}^*, X_{22}^*, ..., X_{2n_2}^*)$ which are drawn from $\aleph_1$ and $\aleph_2$ according to the rules which places probability $\frac{1}{n_1}$ on each sample value of $\aleph_1$ and probability $\frac{1}{n_2}$ on each sample value of $\aleph_2$. Let $u = E(U)$ then the bootstrap estimate (say $\hat{u}$) of $u$ is given by

$$\hat{u} = E(U^* \mid \aleph_1, \aleph_2) \qquad (3.17)$$

This expected value is often not computable.

Uniform Resampling Approximation for Bootstrap Estimate

Assume that the probability of selecting $X_{1i}$ in a resample is

$$P(X_1^* = X_{1i} \mid \aleph_1) = \frac{1}{n_1} \qquad (3.18)$$

and probability of selecting $X_{2i}$ in a resample is

$$P(X_2^* = X_{2i} \mid \aleph_2) = \frac{1}{n_2} \qquad (3.19)$$

Let $\aleph_{11}^*$, $\aleph_{12}^*$, ...., $\aleph_{1B}^*$ and $\aleph_{21}^*$, $\aleph_{22}^*$, ...., $\aleph_{2B}^*$ denote two independent resamples sets of size $B$ each drawn from $\aleph_1$ and $\aleph_2$ respectively. To obtain a Monte Carlo approximation to $\hat{u}$ using uniform resampling, let $U_b^*$ denote $U$ computed from $\aleph_{1b}^*$ and $\aleph_{2b}^*$. Then, the uniform resampling approximation to the bootstrap estimate $\hat{u}$ is given by

$$\hat{u}_B^* = B^{-1} \sum_{b=1}^{B} \left(U_b^*\right) \qquad (3.20)$$

Do and Hall (1991) showed that $\hat{u}_B^*$ is an unbiased approximation to $\hat{u}$, in the sense that $E(\hat{u}_B^* \mid \aleph_1, \aleph_2) = \hat{u}$. Moreover, an approximation of the bootstrap bias of $u$ can be obtained by $\hat{bias}^* = |\hat{u}_B^* - \hat{u}|$, and an approximation of the bootstrap MSE can be obtained by $\hat{MSE}^* = B^{-1} \sum_{b=1}^{B} \left(U_b^* - \hat{u}\right)^2$.

Estimation of distributions function and quantiles

Bootstrap method for calculating confidence limits, distribution function or a problem in testing hypothesis involves estimation of probabilities of the form

$$p = P(S \le d_p) \qquad (3.21)$$

Using the bootstrap estimation conditioned on the original samples one can estimate $p$ by $\hat{p}$ where

$$\hat{p} = P(S^* \le d_p) \qquad (3.21)$$

and $S^* = S(\aleph_1^*, \aleph_2^*)$. Note that (3.22) can be approximated by using empirical frequencies such as the proportion of $B$ simulated samples for which $S^* \le d_p$. In the literature, the problem is solved by defining a smooth transformation $h$ of $S$ viz., $T=h(S)$ with the property that the distribution of $T$ is approximately normal, see Hall (1992).

Adopting the notation of Section 3.3.1, the definitions of $U$ and $U^*$ for this problem become $U = I(S \le d_p)$ and $U^* = I(S^* \le d_p)$ respectively, where $I$ is the indicator function. Let $S_1^*, S_2^*, ...., S_B^*$ be the resampling realization of $S$. Then, the uniform resampling approximation to the bootstrap estimate $\hat{p}$ is

$$\hat{p}_B^* = B^{-1} \sum_{b=1}^{B} \left( I(S_b^* \le d_b) \right) \qquad (3.22)$$

To find the uniform resample approximation of the *p-th* quantile of the bootstrap distribution of $S$, say $\hat{q}_b$, let $S_{(1)}^*, S_{(2)}^*, ...., S_{(B)}^*$ be the order statistics of $S_1^*, S_2^*, ...., S_B^*$. Define $K$=Int($B$*$P$).

Then uniform resampling approximation of the lower limit is $\hat{q}_p^* = \dfrac{S_{(K)}^* + S_{(K+1)}^*}{2}$.

Simulation Study

A Monte Carlo simulation study was conducted for $R$=0.2, 0.5 and 0.8, $(n_1, n_2)$=(20,20), (20,30), (50, 50), (50, 100) and $\alpha$=0.05. All the 1000 simulated sets of observations were generated under the assumption that both densities have exponential distribution with the different means. A bootstrap approximation, based on 1000 resamples, was used.

Tables 1-3 indicate that the bias of the proposed OVL estimators are negligible and |bias| decreases as the sample sizes are increased. With respect to the coverage probability (1-$\alpha$), Taylor series approximation method seem to work well, except for $R$ close to one and very small sample sizes.

The coverage probability for all three OVL coefficients are getting closer to the nominal value when the sample sizes are increased. Bootstrap methods coverage probability work fairly good and increases when $R$ increases close to one. However, Transformation method, which is the easiest to be used, works very well when $R$ <0.5 and for small sample sizes. Also, transformation method is the best for all three OVL coefficients, with respect to the length of the confidence interval, except when the sample sizes are (50, 50).

Illustration: Survival Time from Dinse (1982)

In most of medical studies the progress of the patients is often monitored for a limited time after treatment. Dinse (1982) gives data for survival times in weeks for 10 patients with symptomatic lymphocytic non-Hodgkin's lymphoma and 28 asymptomatic patients. The precise survival time is not known for one patient in the symptomatic group and 12 patients in the asymptomatic group. They were alive when the study was terminated. Therefore, those patients were excluded from our illustration. Table 4 contains the survival time of the symptomatic and the asymptomatic group. The aim of this illustration to estimate the percentage of similarity in the range of survival time in the two groups.

Figure 2 and 3 indicate that the data for both groups (symptomatic and asymptomatic) can be accepted as exponential data. The MLE estimates for the scale parameters are respectively $\hat{\theta}_1$=138.22 and $\hat{\theta}_2 = 207.13$.

From Table 5, all three methods gave reasonable point and confidence interval estimates for the proposed OVL coefficients. However, $\Delta$ have the lowest asymptotic bias but the largest asymptotic variance. The confidence interval based on Taylor series

Table 1: Bias, length of interval (L.), and the coverage probability (Cov.) for *R*=0.20. Exact OVL coefficients: $\rho$ =0.745, $\lambda$ =0.556 and $\Delta$ =0.465.

| | | Taylor Series | | | Bootstrap | | | Transformation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n_1, n_2)$ | | Bias | L. | Cov. | Bias | L. | Cov | Bias* | L. | Cov. |
| (20,20) | $\rho$ | -0.028 | 0.314 | 0.933 | -0.016 | 0.279 | 0.936 | -0.015 | 0.296 | 0.959 |
| | $\lambda$ | -0.029 | 0.457 | 0.928 | -0.014 | 0.404 | 0.936 | -0.017 | 0.428 | 0.959 |
| | $\Delta$ | -0.018 | 0.337 | 0.949 | -0.006 | 0.311 | 0.936 | -0.004 | 0.330 | 0.959 |
| (20,30) | $\rho$ | -0.023 | 0.283 | 0.949 | -0.013 | 0.259 | 0.941 | 0.003 | 0.270 | 0.957 |
| | $\lambda$ | -0.024 | 0.415 | 0.947 | -0.010 | 0.376 | 0.941 | -0.011 | 0.396 | 0.957 |
| | $\Delta$ | -0.015 | 0.305 | 0.950 | -0.005 | 0.286 | 0.941 | -0.002 | 0.303 | 0.957 |
| (50, 50) | $\rho$ | -0.011 | 0.196 | 0.937 | -0.005 | 0.185 | 0.934 | -0.005 | 0.086 | 0.036 |
| | $\lambda$ | -0.014 | 0.290 | 0.937 | -0.002 | 0.274 | 0.934 | -0.004 | 0.109 | 0.036 |
| | $\Delta$ | -0.007 | 0.212 | 0.944 | -0.000 | 0.204 | 0.934 | -0.002 | 0.078 | 0.036 |
| (50, 100) | $\rho$ | -0.008 | 0.169 | 0.945 | -0.006 | 0.162 | 0.937 | -0.004 | 0.125 | 0.868 |
| | $\lambda$ | -0.008 | 0.250 | 0.945 | -0.005 | 0.240 | 0.937 | -0.004 | 0.185 | 0.868 |
| | $\Delta$ | -0.005 | 0.182 | 0.949 | -0.002 | 0.177 | 0.937 | -0.003 | 0.137 | 0.868 |

* Estimated bias using Monte Carlo simulation methods

Table 2. Bias, length of interval (L.), and the coverage probability (Cov.) for $R$=0.50. Exact OVL coefficients: $\rho$=0.943, $\lambda$=0.0.889 and $\Delta$=0.75.

| | | Taylor Series | | | Bootstrap | | | Transformation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n_1, n_2)$ | | Bias | L. | Cov. | Bias | L. | Cov | Bias* | L. | Cov. |
| (20,20) | $\rho$ | -0.035 | 0.203 | 0.921 | -0.024 | 0.179 | 0.944 | -0.016 | 0.186 | 0.957 |
| | $\lambda$ | -0.059 | 0.369 | 0.919 | -0.039 | 0.316 | 0.944 | -0.027 | 0.327 | 0.957 |
| | $\Delta$ | -0.029 | 0.430 | 0.917 | -0.016 | 0.369 | 0.944 | -0.013 | 0.365 | 0.957 |
| (20,30) | $\rho$ | -0.029 | 0.186 | 0.915 | -0.025 | 0.171 | 0.930 | -0.017 | 0.171 | 0.943 |
| | $\lambda$ | -0.048 | 0.339 | 0.915 | -0.042 | 0.304 | 0.930 | -0.030 | 0.303 | 0.943 |
| | $\Delta$ | -0.024 | 0.395 | 0.926 | -0.021 | 0.318 | 0.930 | 0.003 | 0.347 | 0.943 |
| (50, 50) | $\rho$ | -0.014 | 0.125 | 0.930 | -0.010 | 0.118 | 0.931 | -0.007 | 0.453 | 0.034 |
| | $\lambda$ | -0.024 | 0.232 | 0.930 | -0.017 | 0.215 | 0.931 | -0.014 | 0.583 | 0.034 |
| | $\Delta$ | -0.012 | 0.271 | 0.925 | -0.005 | 0.260 | 0.931 | -0.005 | 0.443 | 0.034 |
| (50, 100) | $\rho$ | -0.010 | 0.107 | 0.951 | -0.008 | 0.104 | 0.946 | -0.005 | 0.078 | 0.850 |
| | $\lambda$ | -0.018 | 0.200 | 0.948 | -0.014 | 0.191 | 0.946 | -0.008 | 0.144 | 0.850 |
| | $\Delta$ | -0.009 | 0.234 | 0.942 | -0.004 | 0.228 | 0.946 | 0.0009 | 0.175 | 0.850 |

* Estimated bias using Monte Carlo simulation methods

Table 3. Bias, length of interval (L.), and the coverage probability (Cov.) for $R$=0.80. Exact OVL coefficients: $\rho$=0.994, $\lambda$=0.988 and $\Delta$=0.918.

| | | Taylor Series | | | Bootstrap | | | Transformation | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n_1, n_2)$ | | Bias | L. | Cov. | Bias | L. | Cov | Bias* | L. | Cov. |
| (20,20) | $\rho$ | -0.031 | 0.106 | 0.712 | -0.026 | 0.108 | 0.955 | -0.016 | 0.096 | 0.337 |
| | $\lambda$ | -0.059 | 0.204 | 0.714 | -0.049 | 0.201 | 0.955 | -0.030 | 0.178 | 0.337 |
| | $\Delta$ | -0.020 | 0.500 | 0.953 | -0.071 | 0.320 | 0.955 | -0.046 | 0.226 | 0.337 |
| (20,30) | $\rho$ | -0.025 | 0.087 | 0.720 | -0.022 | 0.094 | 0.952 | -0.012 | 0.078 | 0.333 |
| | $\lambda$ | -0.048 | 0.168 | 0.721 | -0.041 | 0.176 | 0.952 | -0.023 | 0.145 | 0.333 |
| | $\Delta$ | -0.018 | 0.539 | 0.940 | -0.059 | 0.297 | 0.952 | -0.033 | 0.199 | 0.333 |
| (50, 50) | $\rho$ | -0.012 | 0.051 | 0.894 | -0.011 | 0.053 | 0.958 | -0.006 | 0.616 | 0.023 |
| | $\lambda$ | -0.024 | 0.10 | 0.892 | -0.020 | 0.103 | 0.958 | -0.011 | 0.794 | 0.023 |
| | $\Delta$ | -0.011 | 0.320 | 0.881 | -0.028 | 0.220 | 0.958 | -0.015 | 0.652 | 0.023 |
| (50, 100) | $\rho$ | -0.009 | 0.043 | 0.943 | -0.009 | 0.045 | 0.945 | -0.005 | 0.030 | 0.844 |
| | $\lambda$ | -0.017 | 0.083 | 0.943 | -0.017 | 0.088 | 0.945 | -0.009 | 0.058 | 0.844 |
| | $\Delta$ | -0.008 | 0.269 | 0.858 | -0.023 | 0.020 | 0.945 | -0.012 | 0.135 | 0.844 |

* Estimated bias using Monte Carlo simulation methods

Table 4. Survival time of symptomatic and asymptomatic lymphocytic patients by Dinse (1982)

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Symptomatic | 49 | 58 | 75 | 110 | 112 | 132 | 151 | 276 | 281 | | | | | | | |
| Asymptomatic | 50 | 58 | 96 | 139 | 152 | 159 | 189 | 225 | 239 | 242 | 257 | 262 | 292 | 294 | 301 | 359 |

Table 5. Results based on the real data of Dinse (1982)

| | | | **Asymptotic Inference** | | **Transformation Technique** | | **Bootstrap Inference based on 1000 resamples** | |
|---|---|---|---|---|---|---|---|---|
| | | | 95% confidence | | 95% confidence | | 95% confidence | |
| | | Asymptotic | Interval limits | | interval limits | | interval limits | |
| Coeff | MLEs (bias) | variance | Lower | Upper | Lower | upper | Lower | Upper |
| $\rho$ | 0.973(-0.063) | 0.0018 | 0.815 | 1.000 | 0.860 | 0.990 | 0.904 | 0.999 |
| $\lambda$ | 0.947(-0.117) | 0.0070 | 0.643 | 1.000 | 0.740 | 0.981 | 0.817 | 0.999 |
| $\Delta$ | 0.829(-0.060) | 0.0247 | 0.654 | 0.883 | 0.606 | 0.898 | 0.675 | 0.976 |

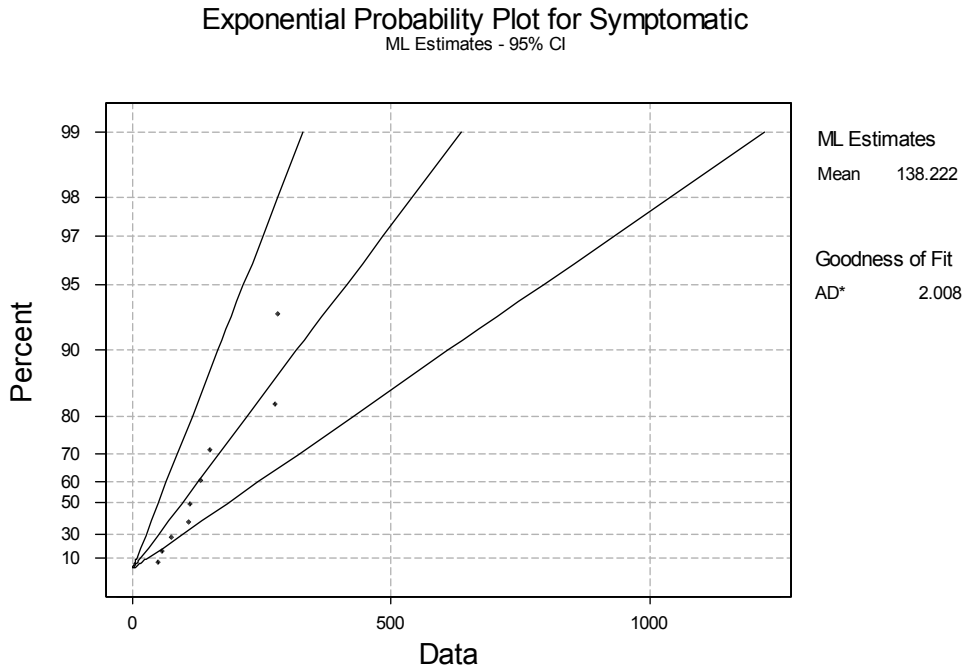



Figure 2. Exponential probability plot for symptomatic patients

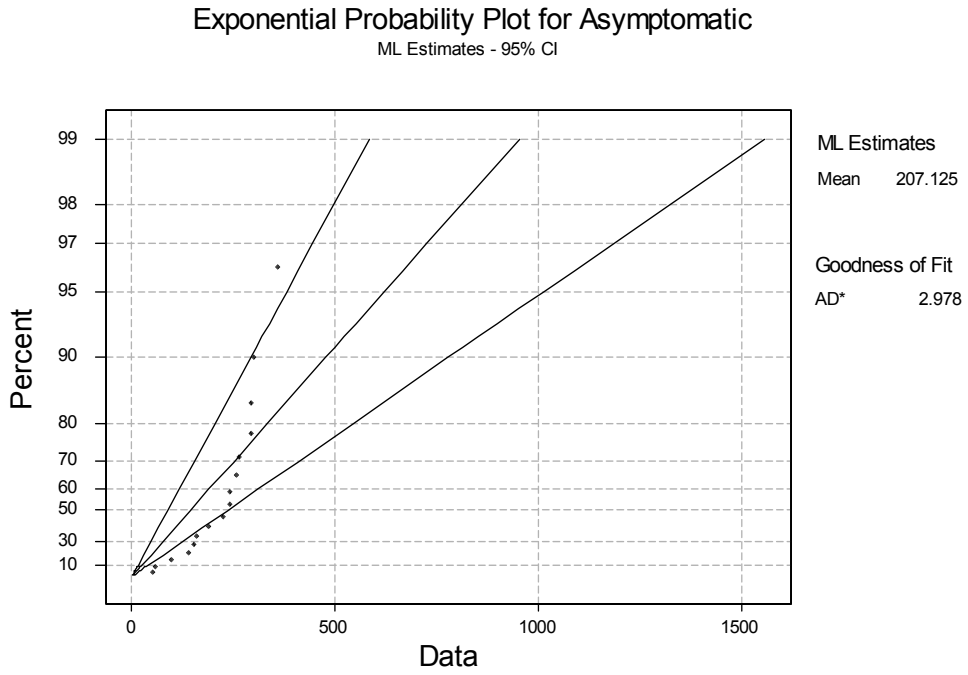



Figure 3: Exponential probability plot for asymptomatic patients.

approximation gave the shortest confidence for $\Delta$.

In conclusion, it seems that there is no best method in all situations. Therefore, when the sample size is small and $R$<0.5, transformation method is recommended. If computers are available, bootstrap method can be used. Taylor series approximation is recommended for larger sample sizes and $R$<0.8.

## Results


Al-Saidy, O., Samawi, H. M., & Al-Saleh, M. F. (2005). Inference on overlap coefficients under the Weibul distribution: Equal Shape Parameter. *ESAM*: *PS*, *9*, 206-219.

Bradley, E. L., & Piantadosi, S. (1982). *The overlapping coefficient as a measure of agreement between distributions.* Technical Report, Department of Biostatistics and Biomathematics, University of Alabama at Birmingham, Birmingham, AL.

Clemons. T. E. (1996). *The overlapping coefficient for two normal probability functions with unequal variances*. Unpublished Thesis, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL.

Clemons, T. E., & Bradley Jr. (2000). A nonparametric measure of the overlapping coefficient. *Comp. Statist. And Data Analysis*, *34*, 51-61.

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure Data. *Biometrics*, *38*, 417-431.

Dixon, P. M., (1993). The Bootstrap and the Jackknife: describing the precision of ecological Indices. In: Scheiner, S.M., Gurevitch J. (Eds.), *Design and Analysis of Ecological Experiments*. NY: Chapman & Hall, p. 209-318.

Do, K. N., & Hall, P. (1991). On importance resampling for the bootstrap. *Biometrika*, *78*, 161-167.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Biometrika* 73, 555-566.

Federer, W. T., Powers, L. R., & Payne, M. G. (1963). Studies on statistical procedures applied to chemical genetic data from sugar beets. *Technical Bulletin* 77, Agricultural Experimentation Station, Colorado State University.

Harner, E. J., & Whitmore, R.C. (1977). Multivariate measures of niche overlap using discriminant analysis. *Theoret. Pobulation Biol.*, *12*, 21-36.

Ibrahim, H. I. (1991). Evaluating the power of the Mann-Whitney test using the bootstrap method. *Commun. Statist. Theory Meth*., *20*, 2919- 2931.

Ichikawa, M. (1993). A meaning of the overlapped area under probability density curves of stress and strength. *Reliab. Eng. System Safety, 41*, 203-204.

Inman, H. F. , & Bradley, E. L. (1989). The Overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Comm. Statist. Theory and Methods*, *18*, 3851-3874.

Lu, R., Smith, E. P., & Good, I. J. (1989). Multivariate measures of similarity and niche overlap. *Theoret. Population Ecol.*, *35*, 1-21.

MacArthur, R.H. (1972). *Geographical ecology*. NY: Harper and Row.

Mann, N. R., Schafer, R. E., & Singpurwalla, N. D. (1974), *Methods For Statistical Analysis Of Reliability & Life Data*, NY: John Wiley & Sons.

Matusita, K. (1955). Decision rules based on the distance for problem of fir, two samples, and Estimation. *Ann. Math. Statist*., *26*, 631-640.

Mishra, S. N., Shah, A. K., & Lefante, J. J. (1986). Overlapping coeffecient: the generalized *t* approach. *Commun. Statist.-Theory and Methods*, *15*, 123-128.

Morisita, M. (1959). *Measuring interspecific association and similarity between communities*. [Incomplete reference available]

Memoirs of the faculty of Kyushu University. *Series E. Biology , 3*, 36-80.

Mulekar, M. S., & Mishra, S. N. (1994). Overlap Coefficient of two normal densities: equal means case. *J. Japan Statist. Soc.*, *24*, 169-180.

Mulekar, M. S., & Mishra, S. N. (2000). Confidence interval estimation of overlap: equal means case. Comp. *Statist .and Data Analysis,* 34, 121-137.

Reiser, B. and Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal equal variance case. *The statistician*, 48, Part 3, 413-418.

Samawi, H. M., Woodworth, G. G, & Al-Saleh M. F. (1996). Two-Sample importance resampling for the bootstrap. *Metron*, *LIV*, n. 3-4.

Samawi, H. M. (1998). Power estimation for two-sample tests using importance and antithetic resampling. *Biometrical Journal*. *40*, 3, 341-354.

Slobodchikoff, C. N., & Schulz, W. C. (1980). Measures of niche overlap. *Ecology*, *61*, 1051-1055.

Smith, E. P. (1982). Niche breadth, resource availability, and inference. *Ecology*, *63*, 1675-1681.

Sneath, P. H. A. (1977). A method for testing the distinctness of clusters: a test of the disjunction of two clusters in Euclidean space as measured by their overlap. *Math. Geol*., *9*, 123-143.

Weitzman, M. S. (1970). Measures of overlap of income distributions of white and Negro families in the United States. *Technical paper No*. 22, Departement of Commerce, Bureau of Census, Washington, D. C.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 517-529
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# The Correlation Coefficients

Rudy A. Gideon
University of Montana

A generalized method of defining and interpreting correlation coefficients is given. Seven correlation coefficients are defined — three for continuous data and four on the ranks of the data. A quick calculation of the rank based correlation coefficients using a 0-1 graph-matrix is shown. Examples and comparisons are given.

Key words: Pearson, Spearman, Kendall, Gini, Greatest Deviation, median, absolute value, nonparametrics, correlation, tied values

## Introduction

### Definitions

This article introduces a system of estimation that has numerous advantages over current practice. Among these advantages is the global tied value procedure for nonparametric or rank based correlation coefficients making estimation functional over all data and advanced statistical methods, such as multiple regression; the currently used local tied value procedure is very restrictive. This system has produced a way of viewing correlation that has allowed other correlation coefficients to be defined. In particular, the new continuous absolute value and median correlation coefficients should be used for L1 methods or the MAD scale estimate. It is general and provides a robust estimation procedure in correlation analysis and in advanced statistical procedures if robust correlation is used (www.math.umt.edu/gideon).

Rudy Gideon received the Ph.D. in Statistics in 1970 under John Gurland at the University of Wisconsin. His academic career began in the Department of Mathematical Sciences at the University of Montana in 1970; he retired from the Department in June of 2005. He has worked extensively with Masters and Doctoral students as well as on a multitude of various applied statistical projects. His prime goal in retirement is to disseminate his original correlation estimation system that encompasses basic statistical methods.

To make the definitions of the correlation coefficients more natural, Pearson's r is reformulated. This re-expression of r also makes possible a natural definition of parametric and nonparametric correlation coefficients based on absolute values and medians. Let CC and NP stand for correlation coefficient and for nonparametric. Some NPCCs are defined based on counting techniques. A 0-1 graph-matrix is used to establish relationships. Finally, some data is analyzed to examine the relative robustness of the NPCCs .

Let $(x_i, y_i), i = 1,2,\ldots n$ be a bivariate data set. The usual mean notation will be used and $x_i^* = x_i - \bar{x}, y_i^* = y_i - \bar{y}, i = 1,2,\ldots n$ are the centered data. The sample covariance is proportional to $\sum x_i^* y_i^*$ . To prepare for later definitions, this covariance is rewritten as

$$\sum x_i^* y_i^* = \left(\sum (x_i^* + y_i^*)^2 - \sum (x_i^* - y_i^*)^2\right)/4 .$$

In the uncentered notation, this can be written as

$$\left(\sum (x_i - \bar{x} + y_i - \bar{y})^2 - \sum (x_i - \bar{x} - y_i + \bar{y})^2 \right)/4 .$$

This form of the covariance function appeared as an interpretation of Pearson's r in Rodgers and Nicewater (1988), when their rescaled variance interpretations were added. Heuristic motivation

for this form as a measure of the relationship between the x-y data is now given and it holds for all CCs that are to be defined.

When there is positive correlation the terms $(x_i^* + y_i^*)^2 = (x_i - \bar{x} + y_i - \bar{y})^2$ will tend to be large, because the two deviations will tend to be in the same direction. The distance from a negative relationship is large, so the correlation would be positive. The terms $(x_i^* - y_i^*)^2, i = 1,2,\ldots n$ will have some canceling effect, so they will tend to be small. The net effect is that the covariance will be large. The distance from a positive relationship is small so that the correlation would be positive. When x and y are independent variables, a similar amount of canceling occurs in both terms and the covariance will fluctuate around zero. When there is negative correlation the distance from positive correlation will be large as the $(x_i^* - y_i^*)^2, i = 1,2,\ldots n$ terms will tend to be large, but cancellation will be occurring in the $(x_i^* + y_i^*)^2, i = 1,2,\ldots n$ terms, so the distance from negative correlation is small. Throughout this article the term distance does not mean just Euclidean distance, but is meant to describe the numerical measures of deviations from perfect positive or negative correlation.

These concepts are next elaborated in Euclidean n-space. For this paragraph *x* and *y* are the n-dimensional vectors of the centered data, normalized so that each has Euclidean length one, $\|x\| = \|y\| = 1$. Consider the vector x + *y* in n-space; the farther this vector is from the origin (for this vector the origin represents perfect negative correlation) the more positive is the correlation. For perfect positive correlation, cos(*x*, *y*) = 1 and $\|x + y\| = 2$; that is, distance from the origin is maximum. Consider the vector *x* – *y*. The closer this vector is to the origin, the more positive the correlation. For *x* – *y*, the origin represents perfect positive correlation and hence, $\|x - y\|$ small means distance from perfect positive correlation is small. Throughout this article the term distance does not mean just Euclidean distance, but is meant to describe the numerical measures of deviations from perfect positive or negative correlation.

To restate, for *x* – *y* the surface of the centered n-dimensional ball of radius 2 represents perfect negative correlation, so $\|x - y\|$ large means distance from perfect positive correlation is large. For perfect negative correlation, cos(*x*, *y* ) = –1, and $\|x - y\| = 0$, so the distance from the ball of radius 2 is a maximum.

Another way to express this, in terms of parameters, is that there is positive correlation when V(X+Y) > V(X–Y) and negative correlation when the inequality goes in the other direction. The connection between distance away from negative correlation and V(X+Y) and also for distance away from positive correlation and V(X–Y) is now illustrated for a bivariate normal distribution.

Let $Z_1$ and $Z_2$ be standardized normal random variables with CC $\rho$. Note that $E(Z_1Z_2) = \rho = [V(Z_1+Z_2) - V(Z_1-Z_2)] / 4$. The term $V(Z_1+Z_2)$ equals distance from perfect negative correlation and is a linear function of $\rho$, namely $2 + 2\rho$. For $\rho = -1$ this distance is zero but for $\rho = +1$, this distance is 4. Similarly, $V(Z_1-Z_2)$ is distance from perfect positive correlation and it is $2 - 2\rho$. For $\rho = -1$, this distance is 4, but for $\rho = +1$, this distance is 0. Note that these distances are monotonic functions of $\rho$ and the overall correlation $V(Z_1+Z_2) - V(Z_1-Z_2)$ combines to equal $4\rho$. However, for some of the other correlation coefficients this combining of the distance measures does not simplify. Also note that in the case of Fisher's normal transformation,

$$\frac{1}{2}\ln\frac{V(Z_1+Z_2)}{V(Z_1-Z_2)} = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} =$$

$$\tanh^{-1}\rho = \ln\frac{\sqrt{V(Z_1+Z_2)}}{\sqrt{V(Z_1-Z_2)}}.$$

It is possible that a similar normalizing concept would work for other correlation coefficients.

Additionally, a correlation coefficient could be based on the ratio, V(X+Y) / V(X–Y), which would be less than one for negative correlation, one for independent random variables, and greater than one for positive correlation.

Pearson's r and other correlation coefficients based on absolute values and medians can now be defined. Let $SS_X$ stand for a centered sum of squares and $SA_X$ stand for the sum of absolute values about the mean; i.e.,

$$SA_x = \sum |x_i - \bar{x}| .$$

Continuous correlation coefficients
Definition 1: Pearson's r

$$r(x,y) = \frac{1}{4}\left( \sum \left(\frac{x_i^*}{\sqrt{SS_x}} + \frac{y_i^*}{\sqrt{SS_y}}\right)^2 - \sum \left(\frac{x_i^*}{\sqrt{SS_x}} - \frac{y_i^*}{\sqrt{SS_y}}\right)^2 \right) \quad (1)$$

= {(standardized distance from perfect negative correlation) – (standardized distance from perfect positive correlation)} divided by a constant, that puts the value between –1 and +1.

Definition 2: An absolute value CC, $r_{av}$

$$r_{av}(x,y) = \frac{1}{2}\left( \sum \left| \frac{x_i^*}{SA_x} + \frac{y_i^*}{SA_y} \right| - \sum \left| \frac{x_i^*}{SA_x} - \frac{y_i^*}{SA_y} \right| \right) \quad (2)$$

where y,i.e. $\sum \left| \frac{x_i^*}{SA_x} \right| + \sum \left| \frac{y_i^*}{SA_y} \right| = 2$

Definition 3: The Median Absolute Deviation correlation coefficient.

For the final continuous correlation, a correlation analog of the MAD, median absolute deviation estimate of variation, is given and denoted by $r_{mad}$. For a random sample, define $MAD_x = med|x_i - med(x_i)|$ and similarly for the data from Y. A median-type correlation coefficient is defined as

$$r_{mad} = \frac{1}{2}\left( med\left| \frac{x_i - med(x_i)}{MAD_x} + \frac{y_i - med(y_i)}{MAD_y} \right| - med\left| \frac{x_i - med(x_i)}{MAD_x} - \frac{y_i - med(y_i)}{MAD_y} \right| \right). \quad (3)$$

It is not true that $|r_{mad}| \le 1$. Let $x_i^* = \frac{x_i - med(x)}{MAD_x}$, and similarly for $y_i^*$. Now, $med|x_i^*| = med|y_i^*| = 1$.

The proof that $|r_{mad}| \le 1$ breaks down is because the median of the sum of two sets of nonnegative numbers is not always less than the sum of the medians. It would be true if the following equation held for $r_{mad}$.

$$med|x_i^* + y_i^*| \le med\left(|x_i^*| + |y_i^*|\right) \le med|x_i^*| + med|y_i^*| = 2$$

However, the second inequality does not always hold. The computer package S+ has been used to examine $r_{mad}$, and values slightly greater than one were occasionally obtained. Simulation studies of $r_{mad}$ show it to behave very much like other correlation coefficients even with the anomaly of occasionally being greater than one. The spread of the distribution is very close to other correlations, and only when the population correlation is very near one can $r_{mad}$ become slightly greater than one. In the case when X, Y have a bivariate normal distribution with parameters, $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$, the population value is known to be $\rho_{mad} = \sqrt{\frac{1+\rho}{2}} - \sqrt{\frac{1-\rho}{2}}$. Substitute y for x in formula (3) and essentially MAD is recovered. Note that the same heuristic motivation for Pearson's r holds for this absolute value CC.

Rank based correlation coefficients

The first NPCC based on absolute values is now defined. In the same way that Spearman's CC is motivated from Pearson's r by using direct substitution of ranks, so is this new

correlation coefficient obtained from Definition 2 by substitution of ranks. An interesting historical note is that the NPCC in Definition 4 was found first and $r_{av}$ determined from it.

First rewrite

$$(x_i - \bar{x}) + (y_i - \bar{y}) \text{ as } x_i + y_i - (\bar{x} + \bar{y}).$$

Replacing the data by their ranks and ordering the bivariate data by the x data, gives the data in rank form. this is just before the $(i, p_i), i = 1,2,\ldots n$. Thus $p_i$ equals the rank of the $y_i$ for the x with rank i. The means of the ranked data are $\frac{n+1}{2}$, so $\bar{x} + \bar{y}$ becomes $n+1$. The ranks $p_i$ are here assumed distinct; tied values will be handled later. In Definition 2, with ranks substituted, the terms $SA_x$ and $SA_y$ are equal and can be factored from expression (2). Their value is

$$SA_x = SA_y = \sum \left| p_i - \frac{n+1}{2} \right| = \sum \left| i - \frac{n+1}{2} \right| = \sum \left| \frac{n+1-2i}{2} \right| .$$

For n odd, $\sum |n+1-2i|$ can be shown to be $\frac{n^2-1}{2}$ and for n even it becomes $\frac{n^2}{2}$; for either even or odd n, it is $\left[ \frac{n^2}{2} \right]$, the greatest integer in $\frac{n^2}{2}$. Thus the denominator in (2) becomes

$$2SA_x = \sum |n+1-2i| = \left[ \frac{n^2}{2} \right].$$

Definition 4: Spearman's modified footrule correlation coefficient, Gini (1914), Betro (1993)

$$r_{mf}(x,y) = \left( \sum |n+1-p_i-i| - \sum |p_i - i| \right) / \left[ \frac{n^2}{2} \right] \quad (4)$$

The attempt by Spearman (1906) to make an absolute value rank CC was also documented in Kendall and Gibbons (1990). Spearman tried to make a computationally simple and robust CC and based it on one summation. The idea in this article is that all or at least most correlations should be a difference of two functions that measure distance from positive and negative correlation, which contrasts with Kendall's method in Chapter 2 in Kendall and Gibbons (1990). There, Kendall advanced the idea that some type of inner product should be used to define all CCs. The above two absolute value CCs cannot be defined using Kendall's inner product concept. This difference of two functions gives the necessary symmetry to a CC. The denominator arises from the absolute value of the numerator which occurs when $p_i = i$ (correlation = +1), or when $p_i = n+1-i$ (correlation = –1). Note again that the same heuristic motivation applies. The forumulation of Spearman's correlation coefficient based on Definition 1 is:

Definition 5: Spearman's correlation coefficient, Spearman (1906)

$$r_s(x,y) = \frac{n(n^2-1)}{3} \left( \sum (n+1-p_i-i)^2 - \sum (p_i - i)^2 \right) \quad (5)$$

$$= 1 - \frac{6}{n(n^2-1)} \sum (p_i - i)^2 .$$

The linear restriction that allows $r_S$ to simplify as shown does not hold for $r_{mf}$. Two more CCs are to be defined — Kendall's, for which a linear restriction does allow a simplification of the defining formula and one based on maximum or greatest deviations for which no simplification occurs. Again the natural definitions are based on the difference of

two functions that measure distance from perfect positive and negative correlation and makes the distribution of the CCs symmetric about zero for the case when x and y are independent, i.e. the null case. It will also be shown that $r_{mf}$ can be computed from the quantities defined for the numerator of the Greatest Deviation CC.

Both Kendall's CC ($r_k$), usually called Tau, and the one based on greatest deviations ($r_{gd}$) use a counting technique that can be defined with an indicator function. Let

$$I(.) = \begin{cases} 1 \text{ if the argument is true} \\ 0 \text{ if false} \end{cases}$$

Recall that the data are assumed ordered by the x data and for the $i^{th}$ largest element of x, the rank of the corresponding y data is $p_i$. For Kendall's correlation coefficient, let

$$\sum_{j=i+1}^{n} I(p_j > p_i) = n_{c,i}$$

count the number of concordances and

$$\sum_{j=i+1}^{n} I(p_j < p_i) = n_{d,i}$$

count the number of discordances at position i (recall that no tied values are yet allowed). The larger the number of concordances the smaller the number of discordances. Let $n_c$ and $n_d$ be the sum over i, i=1,2,..., n-1 of the concordances and discordances, respectively. The concordance function, $n_c$, is a counting function that measures distance of the ranked data from a perfect negative monotone relationship, whereas $n_d$ is a similar discrete measure of the ranked data from a perfect positive monotone relationship.

Definition 6: Kendall's $r_k$ correlation coefficient, see e.g. Kendall and Gibbons (1990)

$$r_k(x,y) = \left( \sum_{i=1}^{n-1} n_{c,i} - \sum_{i=1}^{n-1} n_{d,i} \right) / \binom{n}{2} \qquad (6)$$

$$= (n_c - n_d) / \binom{n}{2}$$

$$= (4n_c / (n(n-1))) - 1 =$$

$$1 - (4n_d / (n(n-1))),$$

because $n_c + n_d = \binom{n}{2}$.

The quantity $\binom{n}{2}$ means n choose 2. This summation of $n_c$ and $n_d$ will be shown in the next section to be n choose 2 using a 0-1 graph-matrix formulation of the calculation of $r_k$.

For the Greatest Deviation CC let $d_i^+ = \sum_{j=1}^{i} I(p_j > i)$, a function that is large when there is negative correlation and small if not; that is, the measure is large if distance from positive correlation is great. Let

$$d_i^- = \sum_{j=1}^{i} I(n+1-p_j > i).$$

This is a measure that is large if distance from negative correlation is great.

Definition 7: The Greatest Deviation correlation coefficient, $r_{gd}$; Gideon and Hollister (1987) and in Gideon, Prentice, and Pyke (1989)

$$r_{gd}(x,y) = (\max_{1 \le i \le n} d_i^- - \max_{1 \le i \le n} d_i^+) / \left[ \frac{n}{2} \right] \qquad (7)$$

where $\left[ \frac{n}{2} \right]$ is the greatest integer in n/2; it's value is the maximum value of the difference in the numerator.

This completes the definitions of the correlation coefficients under consideration. The next section gives some insightful examples; the work is considerably eased using a computational aid that allows computations of the four nonparametric correlation coefficients

from an augmented plot of the data with a 0-1 matrix, called a graph-matrix.

## Methodology

Computations using the graph-matrix

The data in rank form are $(i, p_i), i = 1,2,\ldots n$. Let $e$ = (1,2,....,n) and $p$ = $(p_1, p_2, \ldots, p_n)$ be the data in vector form. The graph of the ranked data will have $e$ plotted on the horizontal axis and $p$ plotted on the vertical axis.

The YMCA basketball data that were used in illustrating the Greatest Deviation CC (Gideon & Hollister, 1987) is used here again. These data occurred as ranks and they will now be used to calculate all four of the NPCCs that have been defined. The $e$ contains the ranks of the won-lost records of the 16 teams that were in the fifth grade league in Missoula, Montana in 1980. Rank one is the team with the best record. Throughout the season, after each game, each coach was asked to rate the sportsmanship of the opposing team and at the end of the season the cumulative ratings were presented in rank form with rank one being the team with the highest rated sportsmanship. These ranks were $p$ = $(14,11,16,2,12,13,7,9,10,3,8,1,15,6,4,5)$.

Note that in general the teams with the best won-lost records had the lower sportsmanship ratings. The correlation coefficients put a measure on the relationship between winning and sportsmanship.

The graph-matrix appears in the middle of Figure 1 surrounded by auxiliary information. The two leftmost and the two rightmost columns as well as the two bottom rows are intermediate calculations explained below. Bordering the data plot are the axes labels. The *s indicate the plotted points $(i, p_i), i = 1,2,\ldots n$ and unlike a scatterplot, the Cartesian product, $e$ x $e$, on the graph is filled in with 0s above each of the plotted points and 1s below. The combination of these *s, 0s, and 1s are used to calculate all four NPCCs which appear on the three borders.

Although the definitions of the correlation coefficients may seem unwieldy, the counting technique is easy and quick to use. It is really more convenient to use the method if the diagonals to the data plot are drawn in, which is easier done by hand. The line of slope one is denoted $sl^{+1}$; this is the line through $(i,i), i = 1,2,\ldots,n$. The line of slope minus one, $sl^{-1}$, goes through points $(i, n+1-i), i = 1,2,\ldots,n$.

Immediately below the graph are two rows that give the values necessary to calculate the Spearman and Absolute Value CCs. The upper row counts from the * to the line $sl^{-1}$ with a minus sign if the * is below $sl^{-1}$. The lower row counts from the * to the line $sl^{+1}$ again with a minus sign if the * is below the line. It is readily apparent that this counting technique directly corresponds to the summands in the formulas of Definitions 4 and 5. The sum of the absolute values of these two rows are given just to the right of them (56, 106), followed by the sum of squares of them (348, 1012).

To the right of the graph-matrix are two columns that give the individual concordances and discordances in Kendall's Tau as given in Definition 6. Starting at a * in position $(i, p_i)$, a 0 appears in column j>i (to the right of the *) if and only if the rank of that column $p_j$ is in discordance ($p_i > p_j$) and a 1 appears in a column to the right of the * if and only if the rank of that column is in concordance ($p_i < p_j$). To obtain the discordances, count the 0s to the right of the * in each column, and to obtain the concordances count the 1s to the right of each * in each column. These results appear in the two columns to the right of the graph. The sums of the two columns, the total numbers of con- and discordances, are given below the columns as (38, 82). Note that the ordering within the two columns does not match the standard algorithm used to calculate Kendall's Tau, $r_k$.

To the left of the graph are two columns headed by $d_i^+$ and $d_i^-$. They label the values for which the maximums need to be taken in Definition 7 of the Greatest Deviation correlation coefficient. For each element in the $d_i^-$ column count all the 0's on and to the left of

YMCA basketball data: correlation computations

left: Greatest Deviation bottom: Spearman and Absolute Value right: Kendall

| $d_i^+$ | $d_i^-$ | vertical axis: sportsmanship rankings<br>horizontal axis: won and lost standings | | | | | | | | | | | | | | | | | $n_c$ | $n_d$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 16 | 0 | 0 | * | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 |
| 1 | 2 | 15 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | * | 0 | 0 | 0 | 0 | 3 |
| 2 | 1 | 14 | * | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 13 |
| 3 | 2 | 13 | 1 | 0 | 1 | 0 | 0 | * | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 9 |
| 3 | 2 | 12 | 1 | 0 | 1 | 0 | * | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 9 |
| 4 | 1 | 11 | 1 | * | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 4 | 10 |
| 5 | 2 | 10 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | * | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 6 |
| 6 | 2 | 9 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | * | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 6 |
| 6 | 2 | 8 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | * | 0 | 1 | 0 | 0 | 0 | 1 | 4 |
| 5 | 2 | 7 | 1 | 1 | 1 | 0 | 1 | 1 | * | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 5 |
| 5 | 2 | 6 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | * | 0 | 0 | 0 | 2 |
| 4 | 3 | 5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | * | 0 | 0 |
| 3 | 3 | 4 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | * | 1 | 1 | 0 |
| 3 | 2 | 3 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | * | 1 | 0 | 1 | 1 | 1 | 1 | 5 | 1 |
| 2 | 1 | 2 | 1 | 1 | 1 | * | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 11 | 1 |
| 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | * | 1 | 1 | 1 | 1 | 4 | 0 |
| 6 | 3 | gd | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 38 | 82 |
| 53 | 28 | mf | -2 | -4 | 2 | -11 | 0 | 2 | -3 | 0 | 2 | -4 | 2 | -4 | 11 | 3 | 2 | 4 | 56 | 348 |
| | | | 13 | 9 | 13 | -2 | 7 | 7 | 0 | 1 | 1 | -7 | -3 | -11 | 2 | -8 | -11 | -11 | 106 | 1012 |

Figure 1.

the $sl^{-1}$ line. To obtain each element in the $d_i^+$ column count all the 1's on and to the left of the $sl^{+1}$ line. For example, $d_7^+ = \sum_{j=1}^{7} I(p_j > 7) = 5$, because exactly $p_1, p_2, p_3, p_5, p_6$ are greater than 7.

Using the graph, there are exactly 5 1s on or to the left of $sl^{+1}$ in row 7, corresponding precisely to the five $p_i's$ mentioned above, because in that part of the plane, the second coordinate exceeds the first. Similarly, $d_7^- = \sum_{j=1}^{7} I(p_j < 17 - 7 = 10) = 2$ because only $p_4$ and $p_7$ are less than 10. Now for $d_i^-$, the term $n+1-p_j > i$ in the indicator function means $p_j < n+1-i$; that is, count all the zeroes at $n+1-i$ on the vertical axis on and to the left of the $sl^{-1}$ line. So for i=7, count all the zeroes at 17-7=10 on the vertical axis on and to the left of $sl^{-1}$; the 0s appear only in columns 4 and 7 corresponding to $p_4$ and $p_7$ being less than 10.

Just below the $d_i^-$ and $d_i^+$ columns are the maximums for $r_{gd}$ and the below them are the sums of these two columns. It will be shown that these sums can be used to compute $r_{mf}$.

Note that twice 53 is 106 and twice 28 is 56, the numbers needed for $r_{mf}$.

From the statistics given in Figure 1, the differences in the numerators of the four correlation coefficients can be obtained and the denominators are

$$\left[\frac{n^2}{2}\right] = 128,\ n(n^2-1)/3 = 1360,\ \binom{n}{2} = 120,$$

$$\left[\frac{n}{2}\right] = 8.$$

$$r_{mf} = \frac{56-106}{128} = \frac{-25}{64} = -0.3906,$$

$$r_s = \frac{348-1012}{1360} = \frac{-83}{170} = -0.4882$$

$$r_k = \frac{38-82}{120} = \frac{-11}{30} = -0.3667,$$

$$r_{gd} = \frac{3-6}{8} = \frac{-3}{8} = -0.3750.$$

Note that the two numbers in the numerator for $r_S$ and $r_K$ add to the denominator ( $r_S$: 348 + 1012 = 1360, and $r_K$: 38 + 82 = 120), the well-known linear restriction, but this does not occur for $r_{gd}$ and $r_{mf}$ as $r_{gd}$: 6 + 3 = 9 > 8, and $r_{mf}$: 56 + 106 = 162 > 128.

Special form for calculation of $r_{gd}$

If only $r_{gd}$ is desired, there is a convenient algorithm to compute the $d_i^-$ and $d_i^+$ values. Write down for i = 1,2, ..., n the three rows vectors $(i, p_i, n+1-p_i)$. Compute $d_i^+$ by placing a marker just to the right of the $i^{th}$ position and count left in the $p_i$ row and note all the ranks greater than i. Compute $d_i^-$ by keeping the same marker, but counting left in the $n+1-p_i$ row noting all the ranks greater than i. This is done in Table 2. Note that $d_i^-$ in Figure 1 and Table 2 appear in the same order whereas, the $d_i^+$ values are reversed.

Three theorems are given below which show some additional usefulness of this graph-matrix approach. The first shows the relationship between the statistics used in $r_{gd}$ and $r_{mf}$.

Theorem 1: $2\sum d_i^+ = \sum |p_i - i|$ and $2\sum d_i^- = \sum |n+1-p_i-i|$, all sums from 1 to n.

Proof: First the $d_i^+$ relationship is established. Clearly $\sum_{i=1}^{n}(p_i - i) = 0$; that is, the sum of the deviations about the $sl^{+1}$ is zero. Thus, $-\sum_{p_i<i}(p_i - i) = \sum_{p_i>i}(p_i - i)$. Now $\sum_{p_i>i}(p_i - i)$ just counts all the 1s on or above the $sl^{+1}$ line. But, $d_i^+ = \sum_{j=1}^{i} I(p_j > i)$ counts all the 1s in row I that are on or above the $sl^{+1}$ line so that $\sum d_i^+ = \sum_{p_i>i}(p_i - i) = \sum_{p_i<i}(i - p_i)$ or

$$2\sum d_i^+ = 2\sum_{p_i>i}(p_i - i) = \sum_{i=1}^{n}|p_i - i|.$$

These equalities are demonstrated in Figure 1. The bottom two rows carry signs to allow these equalities to be easily seen. The proof of the $d_i^-$ relationship follows in a similar manner.

Theorem 2: The number of 1s on or to the right of the $sl^{-1}$ line in row i-1 equals the number of 0s on or to the left of $sl^{-1}$ in row i, i=2,3,...,n. The number of 0s on or to the right of the $sl^{+1}$ line in row i equals the number of 1s on or to the left of the $sl^{+1}$ line in row i-1, i=2, 3,...,n. (In this theorem row i refers to the vertical axis, which are ranks; e.g. row 1 corresponds to the bottom row of the 0-1graph-matrix.) Figure 1 provides a guideline for the proof.

The symmetry displayed in this theorem shows that the Greatest Deviation CC could have been equivalently defined in a right-handed fashion; i.e. instead of counting 0s and 1s from the left to the diagonal lines, counting could have been done from the right with a suitable adjustment.

Theorem 3: For Kendall's CC, $n_c + n_d = \binom{n}{2}$.

Proof: If the data positions (*s) fell on the diagonal of the graph-matrix it is clear that there would be a total of $n^2 - n$ 0s and 1s with complete anti-symmetry. The permutation of the columns to depict the actual data does not change this total and hence, the total number of 0s and 1s to the left of the *s must equal the total number to the right. Thus, $n_c + n_d = \frac{n^2 - n}{2} = \binom{n}{2}$. Further, the number of 1s to the right (38 in Figure 1) equals the number of 0s to the left and the number of 0s to the right (82 in Figure 1) equals number of 1s to the left.

## Results

Which correlation coefficients are outlier resistant? In this section two examples are given to illustrate that the four NPCCs can have quite different values on the same data. The maximum differences between $r_k$ and $r_s$ appear on page 34 of Kendall and Gibbons (1990). The examples below suggest that $r_{gd}$ and $r_{mf}$ are the most robust, $r_k$ next, but that Spearman's $r_s$ is not very robust. Let $e$ and $p$ be the rank vectors. The calculation of the correlation coefficients is left to the reader. The values of the NPCCs for n = 10 and $p = (5,4,3,2,1,10,9,8,7,6)$ are

$$r_{mf} = \frac{26}{50} = 0.5200,\ r_s = \frac{17}{33} = 0.5152,$$
$$r_k = \frac{1}{9} = 0.1111,\ r_{gd} = \frac{3}{5} = 0.6000.$$

The values of the CCs now with $p = (10,1,2,3,4,5,6,7,8,9)$ are

$$r_{mf} = \frac{14}{50} = 0.2800,\ r_s = \frac{6}{330} = 0.0182,$$
$$r_k = \frac{11}{45} = 0.2444,\ r_{gd} = \frac{3}{5} = 0.6000.$$

It is known that for the bivariate normal distribution, the NPCCs estimate a function that is less than the correlation parameter, $\rho$. When the CCs differ greatly, it suggests that there are strange observations in the data. Here, $r_{gd}$ and $r_{mf}$ give the largest indication of a positive

Table 2. Calculation of the Greatest Deviation CC

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_i$ | 14 | 11 | 16 | 2 | 12 | 13 | 7 | 9 | 10 | 3 | 8 | 1 | 15 | 6 | 4 | 5 | |
| $n+1-p_i$ | 3 | 6 | 1 | 15 | 5 | 4 | 10 | 8 | 7 | 14 | 9 | 16 | 2 | 11 | 13 | 12 | |
| $d^+_i$ | 1 | 2 | 3 | 3 | 4 | 5 | 5 | 6 | 6 | 5 | 4 | 3 | 3 | 2 | 1 | 0 | 6 |
| $d^-_i$ | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 1 | 0 | 3 |

Table 3 Spearman data

| Person | addition | sound |
|---|---|---|
| D | 1 | 3 |
| I | 2 | 2 |
| H | 3 | 1 |
| B | 4 | 4.5 |
| J | 5 | 4.5 |
| E | 6 | 11 |
| A | 7 | 6 |
| K | 8 | 9 |
| F | 9 | 8 |
| C | 10 | 10 |
| G | 11 | 7 |
| | (i) | ($p_i$) |

relationship for the strange data of these two examples. Hence, they may be the most resistant to outliers or to any unusual data. (Work in progress shows them more resilient.)

Probabilities and asymptotics for the rank correlation coefficients

Some aspects of the rank CCs will be compared by using an example from Spearman (1906) concerning the relationship between the ability of people to add numbers quickly and accurately and their ability to distinguish between two sound tones. Spearman used this example to illustrate his footrule CC. The data were for eleven students of psychology; Spearman ranked their ability in pitch discrimination and a second person ranked independently for addition ability. The data are ordered by the addition variable and note the two tied values with the usual convention used, which could be called a local convention as opposed to a more useful global definition given below.

Spearmans's footrule CC is

$$r_f = 1 - \frac{6\sum_{p_i > i}(p_i - i)}{n^2 - 1} = 1 - \frac{6(8.5)}{120} = 0.57 .$$

Because this footrule only involved distance from perfect positive correlation, it is not a valid correlation coefficient. It is interesting from a historical perspective. He compared this number to probable error (derived in his article) of 0.13 and concluded because 0.57/0.13 = 4.38, "the faculty of adding numbers and that of discriminating pitch is just about large enough to be beyond all reasonable suspicion of mere chance coincidence" (p. 96).

Spearman did not use a table of critical values but instead stated a heuristic value for the above ratio to be significant. The four nonparametric CCs and their corresponding probability values are now computed for this data. Referring to what is now known as the Spearman CC (the rank equivalent of Pearson's CC; i.e., $r_S$) Spearman said, "the effect of squaring is to give more weight to the extreme differences as compared with the median ones. This is probably a considerable advantage in most physical measurements. But in other fields of research, and perhaps above all in Psychology, these extreme cases are just the ones of most suspicious validity, so that the squaring is here more likely to do harm than good" (p. 99). Thus, Spearman wanted a robust CC for his data.

This example illustrates the definition of a rank CC when tied values are present. In

advanced work on the use of CCs in estimation, the current local methods of tied value calculations are not adequate and hence a global method first introduced in Hollister and Gideon (1987) is presented. In this method, the calculations are done twice: first when Person B is assigned rank 4 for sound and Person J is assigned rank 5 for sound, favoring positive correlation; in the second calculation ties are broken in the reverse direction to favor negative correlation. Note that $r_{gd}$ is the only CC without a change. Each CC can be defined uniquely by averaging the values of the two extreme correlation coefficients.

In Table 4, $r_{gd}$ remains at 0.6000 but $r_{mf}$ becomes (0.7333 + 0.7000)/2 = 0.7167. A general global definition for an alternative tied value procedure is now given.

Definition: The global values of rank CC when ties are present

Let $(x, y)$ be a set of data, and $(I, P^+)$ be the corresponding ranks which are assigned among the tied values in the way that most favor positive correlation, and let $(I, P^-)$ the corresponding ranks assigned among the tied values in the way to most favors negative correlation. $I$ becomes $e$ and $P^+$ and $P^-$ are permutations of $e$. Then a rank correlation coefficient, $r$, is defined uniquely from the two extremes, $P^+$ and $P^-$. Its value is

$$r(x, y) = (r(e, P^+) + r(e, P^-))/2. \quad (8)$$

The quantities $r(e, P^+)$ and $r(e, P^-)$ are abbreviated to $r^+$ and $r^-$, respectively. As an example, let $(x, y)$ = ((1,2,2,4,5), (1,1,2,1,3)). Then $P^+$ = (1,2,4,3,5) and $P^-$ = (3,4,2,1,5). Thus, for $r_{gd}$,

$$r_{gd} = \frac{r^+ + r^-}{2} = \frac{1/2 + (-1/2)}{2} = 0.$$

Return to the level of significance for the Spearman example. The numerators and values of the four NPCCs as computed by the 0-1 graph-matrix method are given in Table 4. The denominators are

$$\left[\frac{11}{2}\right] = 5, \binom{11}{2} = 55, \left[\frac{11^2}{2}\right] = 60,$$
$$\frac{11(11^2 - 1)}{3} = 440.$$

Tail probabilities are obtained from Neave (1978) for $r_k$ and $r_s$, from Gideon and Hollister (1987) for $r_{gd}$, and from Betro (1993) for $r_{mf}$. The table values are compared to the asymptotic values computed from the asymptotic distributions which are given in Kendall and Gibbons (1990) for $r_s$ and $r_k$ and in Gideon, Prentice, and Pyke (1989) for $r_{gd}$. The asymptotic null distributions ($\rho = 0$) of the four CCs are given first. These are

$$\sqrt{n-1}r_s \text{ is } N(0,1); \sqrt{n-1}r_k \text{ is } N(0,4/9);$$
$$\sqrt{n}r_{gd} \text{ is } N(0,1); \sqrt{n-1}r_{mf} \text{ is } N(0,2/3).$$

For completeness the exact variances of each CC is given; $V(r_s) = 1/(n-1)$; $V(r_k) = 2(2n+5)/(9n(n-1))$; $V(r_{gd})$ is unknown; $V(r_{mf}) = 2(n^2+2)/(3n^2(n-1))$ for n even and $2(n^2+3)/(3(n-1)(n^2-1))$ for n odd. The one tie is neglected and the data for the most correlation case, $P^+$, is used. First, from tables,

$$0.001 \le P(r_s \ge 0.7636) \le 0.005;$$
$$0.01 \le P(r_k \ge 0.5636) \le 0.025;$$
$$0.01 \le P(r_{gd} \ge 0.6000) \le 0.05;$$
$$P(r_{mf} \ge 11/15 = 0.7333) = 0.0013 \text{ and}$$
$$P(r_{mf} \ge 7/10 = 0.7000) = 0.0024.$$

Thus, all of the CCs are significant with $r_s$ and $r_{mf}$ being the most significant. These results are now compared to the asymptotic approximations using the notation of $Z$ as N(0,1).

Table 4.

Spearman's 1906 Data and Correlations
The pairs of numbers in the numerators show distances from – and + correlation
Correlations are in second row of the named correlation

| | most + | most - | average |
|---|---|---|---|
| $r_{gd}$ | 5-2 | 5-2 | |
| | 0.6000 | 0.6000 | 0.6000 |
| $r_{mf}$ | 60-16 | 60-18 | |
| | 0.7333 | 0.7000 | 0.7167 |
| $r_k$ | 43-12 | 42-13 | |
| | 0.5636 | 0.5273 | 0.5455 |
| $r_s$ | 388-52 | 386-54 | |
| | 0.7636 | 0.7545 | 0.7591 |

Table 5: Some asymptotic comparisons

$$P(r_s \geq 0.7636) \cong P(Z \geq \sqrt{10}(0.7636) = 2.4147) = 0.0079$$

$$P(r_k \geq 0.5636) \cong P(Z \geq \frac{\sqrt{10}(0.5636)}{2/3} = 2.6734) = 0.0038$$

$$P(r_{gd} \geq 0.6000) \cong P(Z \geq \sqrt{11}(0.6000) = 1.9900) = 0.0233$$

$$P(r_{mf} \geq 0.7333) \cong P(Z \geq \frac{\sqrt{10}(0.7333)}{\sqrt{2/3}} = 2.8401) = 0.0023$$

All of these approximate results are reasonably good. All four correlations support Spearman's conclusion that his footrule CC gave. Spearman drew his conclusion by comparing his footrule value of 0.57 to the probable error, which he gave as 0.13. Thus, 0.57/0.13 = 4.38. This example is concluded by comparing the value of $r_{mf}$, the modified footrule CC, 0.7333, to

$$\sqrt{V(r_{mf})} = \sqrt{\frac{2(11^2+3)}{3(10)(11^2-1)}} = \sqrt{0.0689} = 0.2625 .$$

Now, 0.7333/0.2625 = 2.7937 and by Spearman's rule of "satisfactory demonstration" that this ratio be at least 4, had Spearman found the correct formulation, $r_{mf}$ , he would have drawn the opposite conclusion (p. 96).

Again for this example it should be pointed out that $r_S$ and $r_k$ have a linear restriction but $r_{mf}$ and $r_{gd}$ do not. Hence, the terms in the numerator, when added give the denominator for $r_S$ and $r_k$ but not for $r_{mf}$ and $r_{gd}$. For $r_S$: 388+52 = 440 and for $r_k$: 43+12 = 55 whereas for $r_{mf}$: 60+16 = 66 > 60 and for $r_{gd}$: 5+2 = 7 > 5.

## Conclusion

By viewing correlation broadly as the difference between measures of distance from perfect negative and perfect positive correlation, many new formulations of correlation may be defined. Two new continuous correlation coefficients are based on absolute values and medians. The median one is an extension of the MAD scale measurement and the absolute value one produces Gini's CC when data ranks are substituted. A 0-1 graph-matrix was introduced as an extension to the plot of the bivariate rank data and used to compute all four nonparametric correlation coefficients and exhibit some relationships. Several examples suggest which of the correlations are most robust: the Greatest Deviation and Gini. A data set from Spearman was used to demonstrate the application of the asymptotic distributions, to compare the correlations on the same data, and to illustrate a global tied value procedure. This procedure does not seem critical here, but for later developments on the use of correlation coefficients in estimation it is essential. Several times the normal distribution was selected to set up notation but this is not necessary, as any distribution from the class of bivariate t distributions would suffice. The four nonparametric correlation coefficients would be distribution-free on this class of bivariate distributions with elliptical shaped contours, including the Cauchy distribution.

.

## References


Betro, B. (1993), On the Distribution of Gini's Rank Correlation Association Coefficient, *Communications in Statistics: Simulation and Computation*, *22*, No. 2, 497-505.

Gideon, R. A. & Hollister, R. A. (1987), A Rank Correlation Coefficient Resistant to Outliers, *Journal of the American Statistical Association 82*, no.398, 656-666.

Gideon, R. A., Prentice, M. J., & Pyke, R. (1989). The Limiting Distribution of the Rank Correlation Coefficient $r_{gd}$, appears in *Contributions to probability and statistics* (Essays in Honor of Ingram Olkin) edited by Gleser, L. J., Perlman, M. D., Press, S. J., & Sampson, A. R. NY: Springer-Verlang, p 217-226.

Gideon, R. A. www.math.umt.edu/gideon.

Gini, C. (1914), L'Ammontare c la Composizione della Ricchezza della Nazioni, Bocca, Torino.

Kendall, M. G. & Gibbons, J. D. (1990), *Rank correlation methods*, 5th ed. Oxford University Press, or also Kendall, M. G. (1962), *Rank correlation methods*, 3rd ed. GB: Hafner Publ. Co.

Neave, H. R. (1978), *Statistical tables*, London: George Allen & Unwin Publishers, Ltd/.

Rodgers, J. L. & Nicewater, W. A. (1988), Thirteen Ways to Look at the Correlation Coefficient, *The American Statistician*, *42*, no. 1, 59-66.

Rousseeuw, P. J. & Croux, C. (1993), Alternatives to the Median Absolute Deviation, *Journal of the American Statistical Association*, *88*, 1273-1283.

Scarsini, M. (1984), On Measures of Condordance, *Stochastica*, *8*, No. 3, 201-218.

Schweizer, B. & Wolfe, E. F. (1981), On Nonparametric Measures of Dependence for Random Variables, *The Annals of Statistics*, *9*, 879-885.

Spearman, C. (1906), 'Footrule' for Measuring Correlations, *British Journal of Psychology*, *2*, 89-108.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 530-536
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data

L. W. Huson
Biostatistics Group, F.Hoffman-La Roche

Zero-clustered data occur widely in medical research and are characterised by the presence of a group of observations of value zero in a distribution of otherwise continuous non-negative responses. A simulation study was conducted to investigate the properties of a number of correlation coefficients applied to samples of zero-clustered data.

Key words: zero-clustered data, Pearson correlation, Spearman correlation, weighted rank correlation.

## Introduction

The defining characteristic of zero-clustered data is the presence of a group of observations of value zero in a distribution of otherwise continuous non-negative responses. This type of data is regularly encountered in a wide variety of medical and clinical applications (see e.g. Lachenbruch 1976; 2001a, 2001b, 2002).

Delucchi and Bostrom (2004) discussed a number of endpoints often used in psychiatric studies which typically exhibit zero-clustering, and Berk (2002) gave, as further examples of zero-clustered data, the antibody response to a vaccine, levels of alcohol consumption, severity rating of side-effects, and intensity of pain during labour. In the field of Health Economics, Buntin and Zaslavsky (2004) commented on the "spike of zero values" that is often seen in otherwise non-negative observations in data on health care costs or resource usage, and Chang and Pocock (2002) discussed a specific example of such data in their analysis of numbers of hours of personal care services received by a group of elderly patients. Other terms which have appeared in the literature to describe this type of data are semi-continuous (e.g. Schafer & Olsen 1999) and zero-inflated (e.g. Tu, 2002). Specifically excluded from consideration here are zero-inflated count data, which constitute a separate and widely studied phenomenon.

Some authors note that in the analysis of zero-clustered data, it may be appropriate to bear in mind the different possible origins of the zero values. Zeros may arise, for instance, by the deliberate censoring of any negative values and the setting of such values to zero. An example of such an endpoint is the ACRn score widely used in studies of rheumatoid arthritis (van Riel & van Gestel, 2000). Alternatively the zeros may arise from an unintentional censoring process, such as an imprecise or insensitive measuring device, where small values of an endpoint cannot be detected and response is therefore recorded as zero (see e.g. Moulton & Curriero, 2002 ). Finally, the zeros may be genuine and accurate values properly representing a patient's response (e.g. Chang & Pocock, 2002).

The proportion of zero values seen in practice in this type of data is variable from one type of endpoint to another. Delucchi and Bostrom (2004), for example, analysing data on addiction severity scores, reported proportions of zeros in different data sets ranging from 6% to 77%. Tu and Zhou (1999) cited data on

Dr. Les Huson has worked as a Consultant Medical Statistician in the pharmaceutical and biotechnology industries for 25 years, specializing in the design and analysis of controlled clinical trials. He obtained his PhD in Statistics at Imperial College, London, and currently holds an appointment as a Visiting Consultant Medical Statistician with the Statistical Advisory Service, Imperial College, London.

hospital in-patient charges in which approximately 75% of the values are zero. In many applications, however, the proportion of zeros would be expected to be smaller – Lachenbruch (2001a), for example, studied cases in which 10% or 20% of the values were zeros.

A further characteristic of zero-clustered data is that the distribution of non-zero part of the data is often skewed, with a long tail of high values. Models often suggested as appropriate for the non-zero part of the data are the lognormal or log-gamma distributions (see e.g. Lachenbruch, 2001a; Moulton & Curriero, 2002).

Although methods of analysis of zero-clustered data have been studied in the literature (see e.g. Lachenbruch 1976; 2001a, 2001b, 2002), the problem of measuring the degree of correlation between two samples of zero-clustered data has not previously been investigated. This article describes the results of a simulation study designed specifically to examine the performance of a number of different measures of correlation when applied to zero-clustered data. The study reported here was split into two parts. In the first simulation study the performance of two conventional correlation measures – the Pearson and Spearman correlation coefficients – was studied in the context of application to zero-clustered data. The second simulation study investigated the performance of three little known weighted rank correlation coefficients when applied to the same data structure.

## Methodology I

### Generating Samples of Correlated Zero-Clustered Data

Two different models were used to generate zero-clustered data for the simulation study – the binomial-lognormal model and the truncated lognormal model (Lachenbruch, 2001a; Moulton & Curriero, 2002). The first model assumes that the zero-clustered data arise from combination of binary and lognormal responses, and the second that the zeros arise from a process of truncation of lognormal data. These models are described in more detail below.

Samples sizes of 25, 50, 100, 200 and 1000 were used in the simulation study, with correlations in the data specified to be 0.30, 0.60 or 0.90, representing low, medium and high correlations respectively. The proportion of zeros in the generated samples was 10%, 20% or 30% in different series of simulations. For each of these combinations of parameters, 10000 simulated datasets were generated, and the value of each of the chosen correlation coefficients was calculated for each generated sample of data.

### Binomial-Lognormal Model

For these simulations, samples of zero-clustered data were generated as a mixture of binary responses and lognormal responses, with the same correlation applied to both components of the data. This gives samples of paired, correlated data which follow the binomial-lognormal model (Lachenbruch, 2001a).

The correlated binary components were generated using the algorithm described by Kang and Jung (2001), and the correlated lognormal components were generated using the methods described by Saucier (2000). The method of Kang and Jung permits the generation of pairs of binary observations - values (0,0), (0,1), (1,0) and (1,1) - with specified probabilities and correlations. For each sample size studied, a full set of such correlated binary pairs was generated, and also a full set of correlated lognormal responses. The final correlated zero-clustered binomial-lognormal dataset was then derived simply by multiplying these two sets of values together. Thus, a binary pair (0,0) and a lognormal pair $(X_1,X_2)$ when multiplied together yield the pair (0,0), a binary pair (0,1) and a lognormal pair $(X_3,X_4)$ when multiplied together yield the pair $(0,X_4)$, and similarly for other combinations.

### Truncated Lognormal Model

In this series of simulations, zero-clustered data were generated by truncating correlated lognormal data. To do this, correlated lognormal data were first generated, using the methods described by Saucier (2000), then, to generate a sample containing a given proportion

Table 1. Mean Value of Pearson and Spearman Correlation Coefficient Estimates
[Binomial-Lognormal Model - 10000 Simulations]

| Sample Size | Coefficient | True Correlation=0.3 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.6 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.9 Proportion of Zeros 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | Pearson | 0.30 | 0.28 | 0.27 | 0.56 | 0.54 | 0.53 | 0.85 | 0.85 | 0.84 |
| | Spearman | 0.32 | 0.30 | 0.29 | 0.56 | 0.56 | 0.56 | 0.81 | 0.83 | 0.85 |
| 50 | Pearson | 0.29 | 0.27 | 0.26 | 0.57 | 0.55 | 0.53 | 0.86 | 0.85 | 0.85 |
| | Spearman | 0.33 | 0.31 | 0.29 | 0.57 | 0.56 | 0.57 | 0.82 | 0.84 | 0.85 |
| 100 | Pearson | 0.29 | 0.27 | 0.26 | 0.57 | 0.55 | 0.53 | 0.87 | 0.86 | 0.86 |
| | Spearman | 0.33 | 0.31 | 0.30 | 0.57 | 0.56 | 0.57 | 0.82 | 0.84 | 0.86 |
| 200 | Pearson | 0.29 | 0.27 | 0.25 | 0.57 | 0.55 | 0.54 | 0.88 | 0.87 | 0.86 |
| | Spearman | 0.33 | 0.31 | 0.30 | 0.57 | 0.57 | 0.57 | 0.83 | 0.84 | 0.86 |
| 1000 | Pearson | 0.28 | 0.26 | 0.25 | 0.57 | 0.55 | 0.54 | 0.89 | 0.88 | 0.87 |
| | Spearman | 0.33 | 0.31 | 0.30 | 0.57 | 0.57 | 0.57 | 0.83 | 0.84 | 0.86 |

Table 2. Mean Value of Pearson and Spearman Correlation Coefficient Estimates
[Truncated Lognormal Model - 10000 Simulations]

| Sample Size | Coefficient | True Correlation=0.3 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.6 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.9 Proportion of Zeros 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | Pearson | 0.33 | 0.33 | 0.33 | 0.59 | 0.59 | 0.59 | 0.86 | 0.86 | 0.85 |
| | Spearman | 0.36 | 0.36 | 0.36 | 0.57 | 0.57 | 0.57 | 0.79 | 0.79 | 0.79 |
| 50 | Pearson | 0.32 | 0.32 | 0.32 | 0.59 | 0.59 | 0.59 | 0.87 | 0.87 | 0.86 |
| | Spearman | 0.36 | 0.36 | 0.36 | 0.58 | 0.58 | 0.58 | 0.80 | 0.80 | 0.80 |
| 100 | Pearson | 0.31 | 0.31 | 0.32 | 0.59 | 0.59 | 0.59 | 0.88 | 0.88 | 0.87 |
| | Spearman | 0.37 | 0.37 | 0.37 | 0.58 | 0.58 | 0.58 | 0.81 | 0.80 | 0.80 |
| 200 | Pearson | 0.31 | 0.31 | 0.31 | 0.60 | 0.60 | 0.60 | 0.89 | 0.89 | 0.88 |
| | Spearman | 0.37 | 0.37 | 0.37 | 0.58 | 0.58 | 0.58 | 0.81 | 0.81 | 0.81 |
| 1000 | Pearson | 0.30 | 0.30 | 0.31 | 0.60 | 0.60 | 0.60 | 0.90 | 0.89 | 0.89 |
| | Spearman | 0.37 | 0.37 | 0.37 | 0.58 | 0.58 | 0.59 | 0.81 | 0.81 | 0.81 |

p of zero data, any lognormal value lower than exp(probit(p)) to was set to zero.

## Results I

### First Simulation Study

#### Pearson and Spearman Correlations applied to the Binomial-Lognormal Model

Table 1 shows the results of the simulation study of the performance of the Pearson and Spearman correlation coefficients, when applied to zero-clustered data generated using the binomial-lognormal model.

The most obvious finding is that, under this data model, both the Pearson and Spearman coefficients on average slightly underestimate the true correlation in most simulated scenarios. The bias is relatively small, but persists across all sample sizes and for low, medium and high correlations. A second finding is that the bias of the Pearson correlation increases slightly as the proportion of zeros in the data increases. In contrast, with the Spearman estimate, the bias either remains much the same as the proportion of zeros increases, or diminishes slightly. The other interesting feature of the results is that the Spearman estimate is generally more accurate for low and medium correlations, across all sample sizes, while the Pearson estimate performs better for high correlations.

#### Pearson and Spearman Correlations applied to the Truncated Lognormal Model

Table 2 shows the results of the simulation study of the performance of the Pearson and Spearman correlation coefficients with correlated zero-clustered data generated using the truncated lognormal model. Under this data model, both the Pearson and Spearman coefficients tend to underestimate the true correlation for medium and high correlations, but tend to overestimate the true value when the true correlation is 0.3. When the true correlation is high, the bias of the Pearson correlation increases slightly as the proportion of zeros in the data increases, whereas with the Spearman estimate, the bias either remains much the same as the proportion of zeros increases, or diminishes slightly. Under this data model, the Pearson correlation performs better than the Spearman for most scenarios.

## Methodology II

### Weighted Rank Correlation Coefficients

#### Introduction

For the second simulation study correlation estimates were selected that were (a) based on ranks or functions of ranks, and (b) were defined in a way which allows lower weights to be attached to the zero values in the data, and higher weights to the non-zero values. These were considered likely to be properties which would result in better estimation of correlation in the presence of data containing many zeros.

Three weighted rank correlation coefficients which have these properties are described in the literature and are easily computed, but they are little known and little used in practice. They are the "top-down" correlation, the Blest-Genest-Plante correlation, and the Costa-Soares correlation. The second part of the simulation study investigated the properties of these three coefficients when applied to correlated zero-clustered data.

#### Top Down Correlation

Iman and Conover (1987) described a correlation estimate which they termed the "top down" correlation. This coefficient places emphasis on the higher ranked data in a sample (i.e. assigns lower weights to low-ranked zero values) by computing the correlation using Savage scores derived from the ranked data. Savage scores are defined as follows:

$$S_i = \sum_{j=1}^{n} 1/j \qquad (1)$$

where i is the rank assigned to the ith order statistic in a sample of size n. For example, with n = 3, the three Savage scores are $S_1 = 1 + 1/2 + 1/3$ , $S_2 = 1/2 + 1/3$, and $S_3 = 1/3$. The top-down coefficient is calculated as:

$$r_{td} = \left( \sum_{j=1}^{n} S_{Ri} S_{Qi} - n \right) / (n - S_1) \qquad (2)$$

where S indicates the Savage score, the $R_i$ and $Q_i$ are the ranks of the data in the two samples, and n is the sample size. A full description of the properties of this coefficient is given by Iman and Conover (1987).

Blest-Genest-Plante correlation

Blest (2000) also defined a rank correlation coefficient which allows lower weights to be assigned to lower ranked values in a dataset. This coefficient, whilst having some desirable properties, suffers from the disadvantage that in its original form it is not symmetric (i.e. corr[X,Y] does not equal corr [Y,X]). However, the Blest estimate was later modified by Genest and Plante (2003) to a symmetrical form, and this symmetric version for the simulation study reported here. The coefficient is calculated as:

$$r_{bgp} = -((4n+5)/(n-1)) + ( 6/(n^3 - n) \sum_{j=1}^{n} R_i Q_i (4- (R_i + Q_i)/ n+1))) \quad (3)$$

where the $R_i$ and $Q_i$ are the ranks of the data in the two samples, and n is the sample size. The detailed properties of the original Blest coefficient and its symmetrical generalization are described by Genest and Plante (2003).

Costa-Soares correlation

Costa and Soares (2005) also defined a rank correlation coefficient which, like the top-down correlation and the Blest-Genest-Plante correlation, allows lower weights to be assigned to lower ranked values in a dataset, and hence in this application allows lower weights to be assigned to the zero values in the zero-clustered data. The coefficient takes the form:

$$r_{cs} = 1 - 6 \sum_{j=1}^{n} (R_i - Q_i)^2 / (n^3 - n) \quad (4)$$

where the $R_i$ and $Q_i$ are the ranks of the data in the two samples, and n is the sample size. The properties of this coefficient, and in particular a comparison of the properties with those of the Blest correlation, are described by Costa and Soares (2005).

## Results

Second Simulation Study

Weighted Correlations with the Binomial-Lognormal Model

Table 3 shows the results of the simulation study of the performance of the weighted correlation coefficients with zero-clustered data generated using the binomial-lognormal model. These weighted correlation coefficients all slightly underestimate the true correlation in the data when the true correlation is medium or high, and overestimate the value when it is low, Their performance generally is as good as or better than that of the Spearman estimate.

Weighted Correlations with the Truncated Lognormal Model

Table 4 shows the results of the simulation study of the performance of the weighted correlation coefficients with zero-clustered data generated using the truncated lognormal model. As with the Pearson and Spearman coefficients, the general tendency of the estimates under this data model is that low correlations are overestimated and medium and high correlations are underestimated. Again under most conditions the weighted coefficients perform on average at least as well or better than the Spearman estimates.

## Conclusion

The literature contains no recommendations on an appropriate choice of correlation coefficient for use with zero-clustered data, but Delucchi & Bostrom (2004) reported the results of an informal survey showing that 22 of 35 articles reported analyses of zero-clustered data that used standard normal theory methods, despite the clear non-normality of such data. Hence it seems likely that some practitioners, in the absence of any specific alternative, might choose to apply commonly-used correlation measures - such as Pearson's correlation or Spearman's rank correlation – to zero-clustered data.

Table 3 Mean Value of Some Weighted Rank Correlation Coefficients
[Binomial-Lognormal Model - 10000 Simulations]

| Sample Size | Coefficient | True Correlation=0.3 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.6 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.9 Proportion of Zeros 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | Top-Down | 0.31 | 0.28 | 0.27 | 0.55 | 0.54 | 0.53 | 0.83 | 0.82 | 0.83 |
| | Blest-Genest-P | 0.33 | 0.31 | 0.31 | 0.56 | 0.56 | 0.57 | 0.82 | 0.83 | 0.86 |
| | Costa-Soares | 0.33 | 0.30 | 0.29 | 0.56 | 0.55 | 0.56 | 0.82 | 0.83 | 0.84 |
| 50 | Top-Down | 0.31 | 0.29 | 0.27 | 0.56 | 0.55 | 0.54 | 0.84 | 0.84 | 0.84 |
| | Blest-Genest-P | 0.33 | 0.31 | 0.31 | 0.57 | 0.56 | 0.57 | 0.82 | 0.84 | 0.86 |
| | Costa-Soares | 0.33 | 0.31 | 0.29 | 0.57 | 0.56 | 0.56 | 0.82 | 0.83 | 0.85 |
| 100 | Top-Down | 0.31 | 0.29 | 0.28 | 0.57 | 0.55 | 0.54 | 0.85 | 0.85 | 0.85 |
| | Blest-Genest-P | 0.33 | 0.31 | 0.31 | 0.57 | 0.57 | 0.58 | 0.83 | 0.84 | 0.86 |
| | Costa-Soares | 0.33 | 0.31 | 0.29 | 0.57 | 0.56 | 0.56 | 0.83 | 0.84 | 0.85 |
| 200 | Top-Down | 0.32 | 0.29 | 0.28 | 0.58 | 0.56 | 0.55 | 0.86 | 0.85 | 0.85 |
| | Blest-Genest-P | 0.34 | 0.31 | 0.31 | 0.57 | 0.57 | 0.58 | 0.83 | 0.84 | 0.87 |
| | Costa-Soares | 0.33 | 0.31 | 0.29 | 0.57 | 0.56 | 0.56 | 0.83 | 0.84 | 0.85 |
| 1000 | Top-Down | 0.32 | 0.30 | 0.28 | 0.58 | 0.56 | 0.55 | 0.86 | 0.86 | 0.86 |
| | Blest-Genest-P | 0.34 | 0.32 | 0.31 | 0.58 | 0.57 | 0.58 | 0.83 | 0.85 | 0.87 |
| | Costa-Soares | 0.34 | 0.31 | 0.30 | 0.58 | 0.57 | 0.56 | 0.83 | 0.84 | 0.85 |

Table 4. Mean Value of Some Differentially Weighted Correlation Coefficients
[Truncated Lognormal Model - 10000 Simulations]

| Sample Size | Coefficient | True Correlation=0.3 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.6 Proportion of Zeros 0.1 | 0.2 | 0.3 | True Correlation=0.9 Proportion of Zeros 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | Top-Down | 0.33 | 0.33 | 0.33 | 0.57 | 0.57 | 0.57 | 0.83 | 0.83 | 0.83 |
| | Blest-Genest-P | 0.36 | 0.36 | 0.37 | 0.58 | 0.58 | 0.59 | 0.81 | 0.81 | 0.81 |
| | Costa-Soares | 0.36 | 0.36 | 0.36 | 0.58 | 0.58 | 0.58 | 0.81 | 0.81 | 0.81 |
| 25 | Top-Down | 0.33 | 0.33 | 0.33 | 0.58 | 0.58 | 0.58 | 0.84 | 0.84 | 0.84 |
| | Blest-Genest-P | 0.36 | 0.37 | 0.37 | 0.59 | 0.59 | 0.59 | 0.82 | 0.82 | 0.82 |
| | Costa-Soares | 0.36 | 0.36 | 0.36 | 0.58 | 0.58 | 0.58 | 0.82 | 0.82 | 0.81 |
| 100 | Top-Down | 0.34 | 0.34 | 0.34 | 0.59 | 0.59 | 0.59 | 0.86 | 0.86 | 0.86 |
| | Blest-Genest-P | 0.37 | 0.37 | 0.37 | 0.59 | 0.59 | 0.60 | 0.82 | 0.82 | 0.83 |
| | Costa-Soares | 0.37 | 0.37 | 0.37 | 0.59 | 0.59 | 0.59 | 0.82 | 0.82 | 0.82 |
| 200 | Top-Down | 0.34 | 0.34 | 0.34 | 0.60 | 0.60 | 0.60 | 0.86 | 0.86 | 0.86 |
| | Blest-Genest-P | 0.37 | 0.37 | 0.37 | 0.59 | 0.59 | 0.60 | 0.82 | 0.83 | 0.83 |
| | Costa-Soares | 0.37 | 0.37 | 0.37 | 0.59 | 0.59 | 0.59 | 0.82 | 0.82 | 0.82 |
| 1000 | Top-Down | 0.34 | 0.34 | 0.34 | 0.60 | 0.60 | 0.60 | 0.87 | 0.87 | 0.87 |
| | Blest-Genest-P | 0.37 | 0.37 | 0.38 | 0.60 | 0.60 | 0.60 | 0.83 | 0.83 | 0.83 |
| | Costa-Soares | 0.37 | 0.37 | 0.37 | 0.59 | 0.59 | 0.59 | 0.83 | 0.82 | 0.82 |

The first part of the simulation study reported here was designed to examine the performance of these common correlation coefficients when applied to this type of data. The second part of the study investigated the properties of three little-known weighted rank correlation coefficients. This summary suggests that, overall, the Pearson estimate is in fact, for most practical purposes, an adequate choice from the coefficients studied, and that among rank correlation coefficients, those allowing differential weighting of zero values generally perform better than the much more widely known Spearman coefficient.

.

## References


Berk, K.N. (2002) Repeated measures with zeros. *Statistical Methods in Medical Research* 11:303-316.

Blest, D.C. (2000). Rank correlation—an alternative measure. *Australian and New Zealand Journal of Statistics* 42:101—111.

Buntin, M.B., Zaslavsky, A.M. (2004) Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures. *Journal of Health Economics* 23: 525–542.

Chang, B-H., Pocock, S. (2000) Analyzing data with clumping at zero: an example demonstration. *Journal of Clinical Epidemiology* 53:1036–1043.

Costa, J., Soares, C. (2005) A weighted rank measure of correlation. *Australian and New Zealand Journal of Statistics* 47(4):515–529.

Delucchi, K.L., Bostrom, A. (2004) Methods for Analysis of Skewed Data Distributions in Psychiatric Clinical Studies: Working With Many Zero Values. *American Journal of Psychiatry* 161:1159–1168.

Genest, C., Plante, J-F. (2003) On Blest's measure of rank correlation. *The Canadian Journal of Statistics* 31(1): 1–18.

Iman, R.L., Conover, W.J. (1987) A Measure of Top-Down Correlation. *Technometrics* 29(3): 351-357.

Kang, S-H., Jung, S-H. (2001) Generating Correlated Binary Variables with Complete Specification of the Joint Distribution. *Biometrical Journal* 43(3): 263–269.

Lachenbruch, P.A. (1976) Analysis of data with clumping at zero. *Biometrische Zeitschrift* 18: 851-856.

Lachenbruch, P.A. (2001a) Comparisons of two-part models with competitors. *Statistics In Medicine* 20:1215–1234.

Lachenbruch, P.A. (2001b) Power and sample size requirements for two-part models. *Statistics In Medicine* 20:1235–1238.

Lachenbruch, P.A. (2002) Analysis of data with excess zeros. *Statistical Methods in Medical Research* 11: 297.

Moulton, L.H., Curriero, F.C. (2002) Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* 11: 317-325.

Saucier, R. (2000) Computer Generation of Statistical Distributions. U.S. Army Research Laboratory Technical Report 2168, March 2000.

Schafer, J.L., Olsen, M.K. (1999) Modeling and imputation of semicontinuous survey variables In *Proceedings of Federal Committee on Statistical Methodology (FCSM) Research Conference*, Nov. 1999. 15.

Tu,W., Zhou, X.H. (1999) A Wald Test Comparing Medical Costs Based on Log-Normal Distributions With Zero-Valued Costs. *Statistics In Medicine* 18: 2749-2761.

Tu, W. (2002) Zero-inflated Data. *In: Encyclopedia Of Environmetric, Eds: El-Shaarawi AH, Piegorsch WW John Wiley & Sons, Chichester.* Volume 4: 2387-2391.

Van Riel, P.L.C.M., van Gestel, A.M. (2000) Clinical outcome measures in rheumatoid arthritis. *Annals of Rheumatic Disease* 59 (supplement): 28-31

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 537-543
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# From Information Lost to Knowledge Gained: The Benefits of Analyzing All the Research Evidence

Joseph L. Balloun
Mercer University

Hilton Barrett
Elizabeth City State University

Data analyses should reveal truths about data. To the extent possible analyses should tell a complete picture. Data analyses should not inadvertently ignore phenomena that might be discovered in sample data sets. However, common univariate or multivariate data analysis methods tend to be based on only the means, standard deviations, and Pearson correlations. The result is that many important truths are discovered, but not the whole truth. This article illustrates in a sample data set that (a) data analyses of other properties of variables and groups are feasible and practical, and (b) such analyses may reveal important information not otherwise detectable. These extensions of common statistical methods are applicable to data analysis and interpretation issues in the social and behavioral sciences.

Key words: Data analysis strategy, skewness, kurtosis, survey

## Introduction

Research findings depend on what is analyzed and on what is not. In this sense, data do not speak for themselves. The data analyst chooses what methods will be used, and this choice shapes what interpretations can be made of the data. The purpose of this article is to show how conventional data analysis strategies may ignore important information, and to demonstrate a somewhat more comprehensive data analysis approach.

Outside of the methodological or statistical literature, many researchers describe univariate data primarily by variables' means, and secondarily by the variables' standard deviations or variances. Relationships among variables tend to be analyzed by a variety of calculations derived from linear correlations. The univariate means and standard deviations often are used to appropriately scale such relationships, as in multiple regression analyses.

Joseph L. Balloun, Ph.D., is Professor of Education at Mercer University. He has published extensively on social science research methods, statistical techniques, and developing information systems. Hilton Barrett, Hilton Barrett, D.B.A., is Associate Professor of Business at Elizabeth City State University. He has published in *Journal of Marketing Theory and Practice*, *Entrepreneurial Theory and Practice*, and *Marketing News*.

## Strengths of Classical Statistical Methods

There are important strengths in current data analysis strategies that tend to assume univariate and multivariate normality. Their greatest advantage is that researchers in many different fields have made very impressive discoveries and practical improvements by using such statistical methods. Most academic researchers have been educated in the appropriate use of the traditional statistical methods. Widely distributed and inexpensive statistical packages such as SPSS (2007) and NCSS (Number Cruncher Statistical System, 2007) have made these methods easily accessible and usable for seasoned besides new researchers.

The classical methods have strong technical virtues. Their simplifying assumptions make them parsimonious, easily understood, and analytically or computationally tractable. When

their assumptions are met (e.g., no outliers or unduly influential observations, minimal missing data, univariate and/or multivariate normality, homoscedasticity, uncorrelated residual errors, sampling from a single identifiable population), they are fully informative.

### Extensions of Common Statistical Data Analysis Methods

For many data sets, there are individual observations within two or more subgroups. Most researchers typically report or analyze subgroup statistics such as the sample size, mean, standard deviation and within group covariances or correlations where applicable. Yet, there is substantial evidence that many population distributions are not Gaussian (Micceri, 1989; Rousseau & Leroy, 1987).

Subgroup differences in variances or covariances, or non-normality might provide useful information. There is some evidence that researchers do not typically analyze subgroup departures from normality. For instance, we searched article abstracts for "skewness" or "kurtosis" in the *Journal of Marketing Research*, the *Academy of Management Journal*, and the *Psychological Bulletin* during calendar years 2004 - 2006. However, in these three journals for these calendar years, the words "skewness", or "kurtosis" never appeared in the abstracts. This suggests that researchers seldom consider skewness or kurtosis of primary importance when summarizing data analyses.

A possible alternative data analysis strategy is to consider these other characteristics of subgroup data as possibly informative. The analyst should analyze more than subgroup means and pooled group statistics. Possibly subgroup differences in standard deviations, skewness or kurtosis may also be informative. Moreover, with current computer power, it is practical to analyze more than subgroup means and statistics pooled over groups.

## Methodology

### An Illustrative Example Using a Real World Data Set

Barrett, Balloun & Weinstein (2004) gathered data on marketing and management factors related to performance of profit and non-profit organizations. The resulting snowball sample consisted of 696 usable individual responses within 60 organizations. Barrett, et al. evaluated how organizations' implementations of market orientation (MKT), learning orientation (LRN), entrepreneurial orientation (ENT), and organizational flexibility (ORG) were related to perceived organizational performance (PERF) in for-profit and nonprofit settings. Further details about the purposes, methods and conclusions of the study are reported in Barrett, et al. (2004).

One of their intriguing results was that variability within organizations was greater than the variability among organizations for most of the variables used in their study. This finding points to the possibility that besides the mean levels of postulated success factors for each organization, levels of within-organization variability might be related to organizational performance. But the standard deviation and the mean do not necessarily describe all the information about univariate distributions. Possibly the skewness or kurtosis of the distributions of variable scores within an organization also might be related to organizational performance.

There are analogous ideas that come from the social or behavioral sciences. For example, Yerkes & Dodson (1908) described how arousal levels could be curvilinearly related in an inverse U-shaped way to the rapidity of habit formation. Katz & Kahn (1966) discussed how variety of internal subdivisions in an organization should be adapted to the variety of organizational inputs. Groupthink ideas also seem to imply that some variety of viewpoints should be important in creating better organizational decisions. Newell & Hancock (1984) discussed how skewness and kurtosis could influence inferences in studies of motor tasks. There was sufficient prior knowledge to warrant an exploration of the possible relations of within-organization variability, skewness, or kurtosis of variables to organizational performance.

### Calculations of Statistics

The calculations were done with SPSS 14.0 (SPSS, 2004). Several subgroup statistics

were computed for each of the scales ENT, ORG, MKT, LRN and PERF for each of the 60 organizations included in the study (Barrett, et al., 2004). The means of each scale were computed in the usual way within each organization. The large sample formulas were used to estimate the sampling errors of the standard deviation, skewness or kurtosis.

The Standard Deviation (SD) was computed as the square root of the unbiased variance. The Skewness (SK) and the Kurtosis (KU) were computed by Fisher's g1 and g2 formulas respectively. Within each organization, the sampling distribution was treated as normally distributed for each statistic. The means of the standard deviation, skewness or kurtosis were supplied for each organization. The distributions of the sample statistics within each organization were assumed normal in the population. The standard deviations of simulated sample statistic observations within each organization were calculated so that they would yield the large sample standard error of the statistic if the simulated sample observations were raw data and the standard error of interest were that of the sample mean. With those sampling assumptions, sample data observations were simulated within each organization for each statistic.

## Results

### Do Organizations Differ in the Central Tendency of the Statistics?

Four statistical attributes were used to describe the distributions of each of the five scales included in this study. Each attribute of each scale differed among the sixty organizations. Table 1 summarizes the distribution of each attribute for each scale over the sixty organizations.

For each of the five scales, the organizations were compared to see whether they differed significantly in the central tendency of the several distribution attributes. Eta Squared from the analysis of variance (ANOVA) was used to index the magnitude of differences among organizations. The comparisons of the distribution attributes were repeated for each of the five scales. Table 2 summarizes the results of these analyses.

For each of the five scales, a critical question is whether the additional distribution attributes improve modeling of PERF. Researchers tend to model the expected or conditional mean of a dependent variable. However, there may be additional aspects of a dependent variable to be modeled. These might include its spread or shape. In the following analyses, the organizational PERF means, standard deviations, skewnesses and kurtoses were modeled from attributes of the other scales in the study.

### Maintaining Parsimonious Models

A version of hierarchical multiple regression was used in this study. The subgroup scale attributes are somewhat correlated with each other. The first step of the hierarchical regression involved forcing the lower order moments as applicable into the regression equation first. Within each hierarchical step, the significant independent variables were chosen stepwise. On subsequent regression steps, the simpler attributes of each independent variable were entered into the equation first, followed by progressively more complex independent variable attributes. The purpose of this hierarchical or sequential procedure was to ensure that the developed regression models remain as parsimonious as possible (Cohen, et al., 2003, Pp. 186-187). At each of these steps, the information gain from the addition of the more complex independent scale attributes was assessed by the significance test for the increase in the ordinary least squares sample $R^2$. Hierarchical regression models were developed for each of the dependent variable attributes. The results are summarized in Table 3.

Table 1
Basic Description of Differences among Organizations

| Scale | Attribute | Differences among Organizations | | | | | |
|---|---|---|---|---|---|---|---|
| | | Minimum | Maximum | Mean | SD[a] | SK[b] | KU[c] |
| | N Per Organization | 4.00 | 31.00 | 11.60 | 5.05 | .93 | 2.42 |
| ENT | MEAN | 2.74 | 5.83 | 4.04 | .68 | .18 | -.13 |
| | SD | .32 | 1.70 | .88 | .27 | .44 | .34 |
| | SK | -1.44 | 1.25 | -.05 | .63 | .07 | -.54 |
| | KU | -2.92 | 2.05 | -.09 | 1.09 | -.02 | -.13 |
| ORG | MEAN | 2.90 | 5.68 | 4.04 | .56 | .43 | .67 |
| | SD | .34 | 1.64 | .91 | .25 | .41 | .85 |
| | SK | -1.87 | 2.02 | -.08 | .77 | .17 | .24 |
| | KU | -4.32 | 5.44 | .36 | 1.62 | .49 | 1.79 |
| MKT | MEAN | 3.19 | 5.99 | 4.60 | .67 | -.25 | -.66 |
| | SD | .23 | 1.35 | .83 | .21 | -.15 | 1.23 |
| | SK | -1.78 | 1.52 | -.18 | .64 | .00 | .14 |
| | KU | -2.82 | 3.82 | -.02 | 1.35 | .65 | .22 |
| LRN | MEAN | 3.20 | 5.31 | 4.40 | .49 | -.43 | -.26 |
| | SD | .50 | 2.21 | 1.04 | .31 | 1.14 | 2.74 |
| | SK | -1.82 | 2.01 | -.27 | .73 | .40 | .64 |
| | KU | -5.00 | 4.29 | .09 | 1.60 | -.04 | 1.89 |
| PERF | MEAN | 3.50 | 6.75 | 5.06 | .73 | -.01 | -.55 |
| | SD | .27 | 1.45 | 1.01 | .26 | -.63 | .39 |
| | SK | -1.94 | 1.49 | -.21 | .79 | .03 | -.31 |
| | KU | -3.03 | 3.19 | .06 | 1.53 | .41 | -.31 |

Table 2
Univariate Scale Attribute Differences among Organizations[a]

| Scale | Scale Distribution Attribute | | | |
|---|---|---|---|---|
| | MEAN | SD | SK | KU |
| ENT | .38*** | .19*** | .12* | .16*** |
| MKT | .25*** | .14*** | .18*** | .17*** |
| ORG | .38*** | .18*** | .14*** | .14*** |
| LRN | .19*** | .18*** | .14*** | .16*** |
| PERF | .32*** | .56*** | .18*** | .16*** |

Table 3
Hierarchical Multiple Regression to Detect Significant Effects

| Independent Scales' Subgroup Attributes | PERF (Dependent Scale) Subgroup Attributes | | | |
|---|---|---|---|---|
| | MEAN | SD | SK | KU |
| Means | .34***[b] | .04* | .00 | .00 |
| Squares of Means | .08***[c] | .00 | .00 | .00 |
| SDs | .00[d] | .27*** | .03* | .08** |
| SKs | .00[e] | .00 | .00 | .00 |
| KUs | .05**[f] | .00 | .00 | .00 |
| Total PRESS $R^2$ | .47***[g] | .31*** | .03* | .08** |

## Conclusion

### Do the Scale Attributes Differ Among Organizations?

Table 1 reveals substantial differences in scale attributes among organizations. The results shown in Table 2 reveal that all the statistical attributes for the five scales are significantly different among organizations at or beyond the .05 level by the one-way ANOVA. Among the subgroup means, the Eta Squareds are sizeable for social science studies, and vary from .19 to .38 with a median of about .32. Eta Squareds for the standard deviations varied from .14 to .56 with a median of .18. The skewnesses varied from .12 to .18 with a median of .14. Moreover, kurtoses had Eta Squareds in the range from .14 to .17 with a median of about .16. These results support the conclusion that the scale attributes differ importantly among organizations.

### Do the Additional Scale Attributes Add Useful Information?

Table 3 shows the effects of using distribution attributes beyond the mean for each organization. There are statistically significant effects for each of the attributes of PERF. Table 3 shows that aspects of the independent scales beyond the mean scores of each subgroup may contribute importantly to improving regression models. For example, the kurtosis of the ORG scale accounts for a PRESS $R^2$ increment of 5% in the variance in PERF. When considered in the context of the prior PRESS $R^2$ of .42, this is a 12% improvement in variance accounted for.

Such incremental improvements in the forecasting accuracy of prediction models can

produce large economic gains. For example, where there are many job applicants for a single job and there is high variance among people in their predicted job performance, then a small increment in $R^2$ can result in large financial gains for an employer. Similarly, in choosing which products to bring to market, a small improvement in demand forecasting accuracy can create large financial gains when spread over several hundred thousand potential customers.

The obtained increments in the sample or PRESS $R^2$s were expected to decline as more abstract attributes of the distribution of the dependent variable were modeled. For every organization's PERF attribute the stringent reproducibility requirement, that the PRESS $R^2$ be statistically significant at or beyond the .05 level, was met. This suggests that many more such effects may be found when one looks for them. And the example data set has shown with a substantial sample size and a carefully collected (albeit necessarily "snowball") data base that it is certainly possible to explore such phenomena.

Do these effects matter?

At present such possible effects as predictability of variability seem to be ignored. But ignoring phenomena observable in data does a disservice to researchers and to the general progress of our sciences and allied disciplines. For example, in the social sciences moderators remain a popular topic. But most discussions of moderation assume that moderators only are interaction effects in the analysis of variance sense. Yet, moderation may connote at least two different things. First, it may be that the slope of the regression of a dependent variable on two or more independent scales depends on the levels of one or more other independent scales (IVs). This is equivalent to interaction effects in the analysis of variance sense

But there is another sense in which the term moderator has been used. Second, correlations, or the absolute size of model errors, among IVs and the dependent variable (DV) may differ depending on the levels of other IVs. This also implies that the multiple correlations among a subset of IVs and the DV may differ depending on the level of other IVs. This is not the same phenomenon as possible interaction effects. It is theoretically similar to suggestions that the absolute size of errors in a model may be a replicable function of one or more independent scales of predictability (Ghiselli, 1956). Ghiselli discussed several applications of his moderator idea in personnel selection. Modeling the conditional spread (standard error of prediction) is quite similar to this old idea of moderation. In an econometric context, such effects are called conditional variance, or are discussed under the topic of heteroscedasticity (e.g., Vytlacil, 2005). In econometrics researchers have also successfully modeled the conditional SK or conditional KU besides the conditional SD (e.g., Ahgiray, Booth, Hatem & Mustafa, 1991; Perez-Quiros & Timmermann, 2001).

The methods suggested here for modeling the spread of the dependent variable pose another strategy for dealing with this possible phenomenon. Moreover, by also modeling the conditional SK and KU of the DV, the methods suggested can lead to further extensions of moderation ideas. See also Sharma, et al. (1981) and Baron & Kenny (1986) for related ideas.

Some Cautions

In this article, it has been argued that data analysts should use more of the information available in a data set. The information gain made possible by expanding the data analyses has been demonstrated in this example data set. Yet reasonable caution should be exercised. Data analysts should tell the truth and the whole truth. But one should ensure that the data analysis tells only the truth. In statistical folklore the cautionary saying is "Torture the data and it will confess." In any practical application one should be cautious to not create artificial results or misleading interpretations from overly elaborate data analyses. There is a danger that using the methods suggested in this article might lead to unnecessarily complex models for a given purpose. Research is constrained by time and cost factors and expected payoffs from more complex analyses. That is certainly a valid point, and Ghiselli (1956) and others were aware of this some time ago (cf. Zikmund, 2003, p. 12).

What are the Implications of this Study?

If researchers do not look for distribution differences among subgroups other than central tendency then they are bound to not find them. The demonstration data set was chosen because the authors of the prior study made it available. The data set was not chosen because it was expected to reveal SD, SK or KU differences among organizations. Instead it was matter of strong suspicion that most data sets involve differences in spread and shape besides differences in central tendencies. Upon analysis, some of the suspected effects with higher order moments were revealed.

It is not known how large or important such effects from higher order subgroup moments may be. But in this study, when the subgroup variances or shapes of the independent variables were included, replicable gains in variance accounted for in attributes of the dependent variable were common. Other researchers should routinely examine their data to see whether subgroup SDs, SKs or KUs, as in this study, produce large and important information gains.

## References


Ahgiray, V., Booth, G. G., Hatem, J. J., & Mustafa, C. (1991). Conditional dependence in precious metal prices. *The Financial Review*, *26*, 367-386.

Baron, R. & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *31*, 1173-1182.

Barrett, H., Balloun, J. L. & Weinstein, A. (2004). Success factors for 21st century organizations: A multi-sector, multiple respondents approach. *The International Journal of Business, 15 (2),* 79-88.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). . *Applied multiple regression/ correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, *40*, 374-377.

Katz, D. & Kahn, R. (1966). *The social psychology of organizations*. New York: Wiley.

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

Number Cruncher Statistical System. (2007). *NCSS 2007 for Windows*. Released March 16, 2007. Kaysville, Utah: NCSS, Inc.

Newell, K. M., & Hancock, P. A. (1984). Forgotten moments: A note on skewness and kurtosis as influential factors in inferences extrapolated from response distributions. *Journal of Motor Behavior*, *16*, 320-335.

Perez-Quiros, G. & Timmermann, A. G. (2001). Business cycle asymmetries in stock returns: evidence from higher order moments and conditional densities. ECB Working Paper No. 58. Retrieved May 2, 2005 from http://ssrn.com/abstract=356061

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression and outlier detection*. Hoboken, New Jersey: Wiley.

Sharma, S., Durand, R. M., & Gur-Arie, O. (1981). Identification and analysis of moderator scales. *Journal of Marketing Research*, *18* , 291-300.

SPSS. (2007). SPSS 16 for Windows. Released 2007. Chicago: SPSS, Inc.

Vytlacil, E. (2005). 102B: Introduction to econometrics Stanford University, Winter 2005. Retrieved April 26, 2005 from Stanford University Website: http://www.stanford.edu/~vytlacil/_teaching/syllabus_102B_winter_2005.pdf.

Yerkes, R. M. & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459-482.

Zikmund, W. G. (2003). *Business research methods*. (7th edition). Mason, OH:South-Western.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 544-550
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Global Measure of the Deviation of a Wavelet Density Estimator

Kussiy K. Alyass
Lawrence Technological University

A wavelet estimator $f^*(x)$ of an unknown probability density function $f(x) \in \mathcal{L}^2(\mathbf{R})$ is considered. A conditional central limit theorem for martingales is used to show that $\int [f^*(x) - f(x)]^2 dx$ is asymptotically normally distributed. Results obtained can be used in a test of goodness-of-fit.

Key words: Wavelet estimator, martingale difference array, multiresolution analysis, asymptotic distribution.

## Introduction

The problem of finding the asymptotic distribution of the quadratic norm of the deviation of the probability density function $f(x)$ from its estimator $f^*(x)$ have been studied by many authors. Bickel and Rosenblatt (1973) obtained the asymptotic distribution of

$$T_n = n\,h(n)\int [f^*(x) - f(x)]^2\, a(x)\,dx$$

where $f^*(x)$ is the kernel estimator of $f(x)$ , $h(n) \to 0$, $n\,h(n) \to \infty$, and $a(x)$ is a weight function. The basic technique in obtaining the result consists in finding the asymptotic distribution of $T_n$ with $f^*(x)$ replaced by conveniently chosen Gaussian process and showing that two functionals converge to the same law. Viollaz (1980) considered orthogonal series estimators and Lii (1978) considered spline estimators, in both cases the above method is used to establish limit theorems for the quadratic norm of the deviation of the probability density function from its estimator. A method using a conditional central limit theorem for martingales due to Adnan (1981) was used by Ghorai (1980) to find the asymptotic distribution of the quadratic norm of the deviation of the orthogonal series estimator.

Rosenblatt (1975) used a method involving the Poissonization of the sample size to obtain the asymptotic distribution of the quadratic norm of the deviation of the two-dimensional kernel estimator. Alyass and Sun (1994) considered two-dimensional orthogonal series estimators; they used the method of Poissonization to establish a limit theorem for the properly normalized quadratic norm of the deviation of the estimator.

A wavelet estimator $f^*(x)$ is used here to estimate the probability density function $f(x)$. Then, a martingale central limit theorem is used to show that $\int [f^*(x) - f(x)]^2 \, dx$ is asymptotically normally distributed.

A brief review and a statement of a conditional central limit theorem for martingales will now be given. For further details, refer to Adnan (1981). Let $\{V_n, n \geq 1\}$ be a sequence of integrable random variables on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{B}_0 \subset \mathcal{B}_1 \subset \mathcal{B}_2 \subset \ldots\ldots$ be an increasing sequence of sub-σ-fields of $\mathcal{F}$. Suppose the sequence $\{(V_n, \mathcal{B}_n), n \geq 1\}$ is a martingale, then the sequence $\{(V_n - V_{n-1}, \mathcal{B}_n), n \geq 1\}$ is called a martingale difference. A double sequence $\{(W_{nj}, \mathcal{B}_{nj}), n \geq 1, j \geq 0\}$ is said to be a martingale difference array if it is a martingale difference for each n.

Suppose that $\{Y_n, n \geq 1\}$ is a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$. Let $\{\mathcal{F}_n, n \geq 1\}$ be a sequence of sub-σ-fields of $\mathcal{F}$. $Y_n \mid \mathcal{F}_n$ converges weakly to

Kussiy K. Alyass holds appointments in the Mathematics and Computer Sciences Department He can be contacted via e-mail at alyass@ltu.edu.

a random variable $Y$ defined on $(\Omega, \mathcal{F}, P)$ if and only if

$$E[f(Y_n)|\mathcal{F}_n] \to Ef(Y),$$

for every bounded continuous function $f$. This convergence will be denoted by

$$Y_n | \mathcal{F}_n \xrightarrow{d} Y .$$

Methodology

The following theorem due to Adnan (1981) will be used in the proof of the main result in this article.

Theorem 1:

Suppose $\{(W_{n\,j}, \mathcal{B}_{n\,j}), n \geq 1, j \geq 0\}$ is a martingale difference array. Assume that:

$(i)$ $\sup_n \sum_{j=1}^{\infty} EW_{nj}^2 < \infty,$

$(ii)$ $\sum_{j=1}^{\infty} W_{nj}^2 \xrightarrow{p} c^2$ for some positive constant $c$

$(iii)$ $\sup_j |W_{nj}| \xrightarrow{p} 0.$

Then, $\sum_{j=1}^{\infty} W_{nj} | \mathcal{B}_{n0} \xrightarrow{d} N(0, c^2)$.

Remark:

Let $\gamma$ denote the trivial $\sigma$-field. If $\gamma \subset \mathcal{B}_{n0}$ then the conditional convergence in the above theorem is equivalent to the usual unconditional convergence in distribution (see Adnan, 1981).

A multiresolution analysis $\ldots \subseteq A_{-2} \subseteq A_{-1} \subseteq A_0 \subseteq A_1 \subseteq A_2 \subseteq \ldots$ of $\mathcal{L}^2(\mathbf{R})$ is an increasing sequence of subspaces $A_j, j \in \mathbf{Z}$, of $\mathcal{L}^2(\mathbf{R})$ satisfying the following conditions:

(M1) $\bigcup_{j \in \mathbf{Z}} A_j$ is dense in $\mathcal{L}^2(\mathbf{R})$,

(M2) $\bigcap_{j \in \mathbf{Z}} A_j = \{0\}$,

(M3) $g(x) \in A_j$ if and only if $g(2^{-j}x) \in A_0$,

(M4) there exists a function $\varphi(x)$ in $A_0$ such that $\{\varphi(x-k)\}_{k \in \mathbf{Z}}$ is an orthonormal basis for $A_0$.

Remarks:

(i) It follows that

$$\left\{ 2^{j/2} \varphi(2^j x - k) \right\}_{k \in \mathbf{Z}}$$

forms an orthonormal basis for $A_j$.

(ii) Assume that $\varphi$ is integrable and $\int \varphi(x) dx \neq 0$ because if $\int \varphi(x) dx = 0$ then the same is true for all functions in all $A_j$, and one would not expect to have condition (M1). In fact one can show that if $\varphi$ has compact support and $\int \varphi(x)\, dx = 1$ then condition (M1) holds (see Strichartz (1993)).

In order to construct the wavelets, let $B_j$ be the orthogonal complement of $A_j$ in $A_{j+1}$, thus $A_{j+1} = A_j \oplus B_j$. There exists a function $\psi(x)$ called the wavelet such that the family $\{\varphi(x-k), \psi(x-k)\}_{k \in \mathbf{Z}}$ is an orthonormal basis for $A_1$. This implies that $\{\psi(x-k)\}_{k \in \mathbf{Z}}$ is an orthonormal basis for $B_0$. The space $\mathcal{L}^2(\mathbf{R})$ is represented as a direct sum

$$\mathcal{L}^2(\mathbf{R}) = \bigoplus_{j \in Z} B_j .$$

Also

$$\left\{ 2^{j/2} \psi(2^j x - k) \right\}_{k \in \mathbf{Z}}$$

is an orthonormal basis for $B_j$ and that the spaces $B_j$ are all mutually orthogonal. Therefore, it is possible to combine all the orthonormal bases for $B_j$ into one orthonormal basis:

$$\left\{ 2^{j/2}\ \varphi\left(2^j x-k\right)\right\}_{\mathrm{j}\in\mathbf{Z},\,\mathrm{k}\in\mathbf{Z}}$$

for $\mathcal{L}^2(\mathbf{R})$. Because the following decomposition of $\mathcal{L}^2(\mathbf{R})$ is also true

$$\mathcal{L}^2(\mathbf{R}) = A_q \oplus \left[ \bigoplus_{j=q}^{\infty} B_j \right], \qquad q \in \mathbf{Z},$$

then one can combine the basis

$$\left\{ 2^{q/2}\ \varphi\left(2^q x-k\right)\right\}_{k\in\mathbf{Z}}$$

for $A_q$ with the bases

$$\left\{ 2^{j/2}\ \psi\left(2^j x-k\right)\right\}_{\mathrm{k}\in\mathbf{Z}}$$

for $B_j$ with $j \geq q$ to obtain an orthonormal basis for $\mathcal{L}^2(\mathbf{R})$. Then, if

$$\varphi_{j,k}(x) = 2^{j/2}\ \varphi\left(2^j x-k\right)$$

and

$$\psi_{j,k}(x) = 2^{j/2}\ \psi\left(2^j x-k\right),$$

the family

$$\left\{ \varphi_{q,k}(x)\,, \psi_{j,k}(x) \right\}_{j\geq q\,,\,k\in\mathbf{Z}}$$

forms an orthonormal basis for $\mathcal{L}^2(\mathbf{R})$. Thus, for any $f(x)\in\ \mathcal{L}^2(\mathbf{R})$. there is

$$f(x) = \qquad (1)$$
$$\sum_{k=-\infty}^{\infty} \alpha_{q,k}\ \varphi_{q,k} + \sum_{j=q}^{\infty}\ \sum_{k=-\infty}^{\infty} \beta_{j,k}\ \psi_{j,k}(x),$$

where

$$\alpha_{q,k} = \int f(x)\varphi_{q,k}(x)dx,$$
$$\beta_{j,k} = \int f(x)\psi_{j,k}(x)dx, \qquad j \geq q.$$

For more detailed account of the subject of multi-resolution analysis and wavelets see Meyer (1990) and Daubechies (1992).

Suppose $X_1, X_2, \ldots, X_n$ are independent, identically distributed, real-valued random variables with common, but unknown, continuous probability density function $f(x)\in\ \mathcal{L}^2(\mathbf{R})$. Estimate $f(x)$ by

$$f^*(x) = \sum_{k=-\infty}^{\infty}\ \hat{\alpha}_{q(n),k}\varphi_{q(n),k}(x),$$

where (2)

$$\hat{\alpha}_{q(n),k} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{q(n),k}(x_i).$$

Some of the properties of this estimator may be found in Doukhan and Leon (1990) and Kerkyacharian and Picard (1992). Throughout the remainder of this article, assume the function $\varphi$ is compactly supported in the interval $[s, t]$. This will ensure that, in (2), only finite random number of coefficients $\hat{\alpha}_{q(n),k}$ are non-zero. To simplify the calculations, assume that $s, t\in\mathbf{Z}$. Under these assumptions,

$$f^*(x) = \qquad (3)$$
$$\frac{1}{n}\sum_{i=1}^{n}\ \sum_{k=1-t}^{-s}\ \varphi_{q(n),k}(x-[x_i])\ \varphi_{q(n),k}(x_i-[x_i]),$$

where $[x]$ denotes the largest integer that is less than or equal to $x$.

Let

$$Y_i = X_i - [x_i],\ \ i = 1, 2, \ldots\ldots, n,$$
$$\eta_{q(n),k} = E\varphi_{q(n),k}(Y)$$

and

$$\theta_{q(n),k}(y) = \varphi_{q(n),k}(y) - \eta_{q(n),k}.$$

Using (1) and (3) the result is

$$\int\left[f^*(x)-f(x)\right]^2 dx = \qquad (4)$$
$$\frac{2}{n^2}\sum_{j=2}^{n}\ \sum_{i=1}^{j-1} U_{ij}(n) + \frac{1}{n^2}\sum_{i=1}^{n}\ \sum_{k=1-t}^{-s}\ \theta^2_{q(n),k}(y_i)$$

$$+\frac{1}{n^2}\sum_{i\neq j} Z_{ij}(n)+\sum_{j=q(n)}^{\infty}\sum_{k=-\infty}^{\infty}\beta_{j,k}^2,$$

where

$$Z_{ij}(n)= \quad (5)$$

$$\sum_{k=1-t}^{-s}\Big[\sum_{l=1-t}^{-s}\varphi_{q(n),k}(Y_i)\,\varphi_{q(n),l}(Y_j)\,\delta_{k+2^{q(n)}[x_i],\,l+2^{q(n)}[x_j]}-\varphi_{q(n),k}(Y_i)\,\varphi_{q(n),k}(Y_j)\Big],$$

and

$$U_{ij}(n)=\sum_{k=1-t}^{-s}\theta_{q(n),k}(Y_i)\,\theta_{q(n),k}(Y_j)$$

Now put

$$c_{k\,l}=\operatorname{cov}\left(\varphi_{q(n),k}(Y),\quad \varphi_{q(n),l}(Y)\right), \quad (6)$$

$$\mu_n=\frac{1}{n}\sum_{k=1-t}^{-s} c_{k\,k},\ \sigma_n^2=\sum_{k=1-t}^{-s}\sum_{l=1-t}^{-s} c_{k\,l}^2,$$

$$W_{n\,j}=\begin{cases}\dfrac{\sqrt{2}}{n\sigma_n}\displaystyle\sum_{i=1}^{j-1} U_{ij}(n), & j=2,3,....,n,\\[2ex] 0, & j=0,1 \ \text{ and } \ j>n,\end{cases}$$

$$V_{n\,j}=\sum_{i=1}^{j} W_{ni} \ \text{ for all } n \text{ and } j,$$

and let $\mathfrak{B}_{nj}$ be the σ-field generated by $Y_1, Y_2, ....,Y_j$, and $\mathfrak{B}_{n0}$ be the trivial σ-field. The sequence $\{(V_{nj}, \mathfrak{B}_{nj}), j \geq 0\}$ is a martingale for each $n \geq 1$. Because $W_{nj} = V_{nj} - V_{n,j-1}$, $\{(W_{nj}, \mathfrak{B}_{nj}), n > 0, j \geq 0\}$ is a martingale difference array.

## Results

Now, to state and prove the main theorem:

Theorem 2: Assume that

$$\frac{n}{\sigma_n}\sum_{k=1-t}^{-s}\left[E\left|\varphi_{q(n),k}(Y)\right|\right]^2\to 0 \quad (7)$$

$$\text{as } n\to\infty$$

$$\frac{n}{\sigma_n}\sum_{j=q(n)}^{\infty}\sum_{k=-\infty}^{\infty}\beta_{j,k}^2\to 0 \ \text{ as } \ n\to\infty \quad (8)$$

$$\sigma_n^2\to\infty \quad \text{as} \quad n\to\infty, \quad (9)$$

$$\sup_k E\varphi_{q(n),k}^4(Y)\leq M. \quad (10)$$

for some constant $M$

It follows that if $t - s < 1+2^{q(n)}$, then

$$\frac{n}{\sqrt{2}\,\sigma_n}\left(\int\left[f^*(x)-f(x)\right]^2 dx-\mu_n\right)\xrightarrow{d} N(0,1).$$

Proof: Using (4) and (6) gives

$$\frac{n}{\sqrt{2}\sigma_n}\left(\int\left[f^*(x)-f(x)\right]^2 dx-\mu_n\right)=$$

$$\sum_{j=1}^{\infty} W_{nj}+\frac{1}{\sqrt{2}n\sigma_n}\sum_{i=1}^{n}\sum_{k=1-t}^{-s}\left[\theta_{q(n),k}^2(Y_i)-c_{kk}\right]+$$

$$\frac{1}{\sqrt{2}n\sigma_n}\sum_{i\neq j} Z_{ij}$$

$$+\sum_{j=q(n)}^{\infty}\sum_{k=-\infty}^{\infty}\beta_{j,k}^2=H_1+H_2+H_3+H_4\,.$$

By assumption (8), $H_4\to 0$ as $n\to\infty$. Let

$$A=\left\{\begin{array}{l}(k,l):\ k-l=2^{q(n)}\,([X_j]-\\ [x_i]),\ k,l=1-t,.....,-s\end{array}\right\}$$

and

$$\mathrm{B}=\left\{\begin{array}{c}(\mathrm{i,j}):[X_j]-[X_i]>\dfrac{(t-s)-1}{2^{q(n)}},\\ i\neq j,\quad i,j=1,2,.....,n\end{array}\right\}$$

From (5) it follows that

$$H_3 = \frac{1}{\sqrt{2}n\sigma_n}$$
$$\left\{ \sum_B \left[ \sum_A \varphi_{q(n),k}(Y_i)\varphi_{q(n),1}(Y_j) - \sum_{k=1-t}^{-s} \varphi_{q(n),k}(Y_i)\varphi_{q(n),k}(Y_j) \right] \right\}$$
$$+ \sum_{B^c} \left[ \sum_A \varphi_{q(n),k}(Y_i)\varphi_{q(n),1}(Y_j) - \sum_{k=1-t}^{-s} \varphi_{q(n),k}(Y_i)\varphi_{q(n),k}(Y_j) \right]$$

The second term in the above formula is equal to zero because $\frac{(t-s)-1}{2^{q(n)}} < 1$ forces $[X_j] = [X_i]$. Also for $(i,j) \in B$ $A = \varnothing$. Therefore

$$H_3 = -\frac{1}{\sqrt{2}n\sigma_n} \sum_B \sum_{k=1-t}^{-s} \varphi_{q(n),k}(Y_i)\varphi_{q(n),k}(Y_j),$$

$$\operatorname{var}(H_3) \le \frac{1}{2n^2\sigma_n^2}$$

$$\sum_{i\neq j} \sum_{i'\neq j'} \sum_{k=1-t}^{-s} \sum_{l=1-t}^{-s} E\left|\varphi_{q(n),k}(Y_i)\varphi_{q(n),k}(Y_j)\varphi_{q(n),l}(Y_i')\varphi_{q(n),l}(Y_j')\right|$$
$$= \frac{n(n-1)}{2n^2\sigma_n^2} \sum_{k=1-t}^{-s} \sum_{l=1-t}^{-s} E\left|\varphi_{q(n),k}(Y)\varphi_{q(n),l}(Y)\right| \left[ E\left|\varphi_{q(n),k}(Y)\varphi_{q(n),l}(Y)\right| \right.$$
$$\left. + 2E\left|\varphi_{q(n),k}(Y)\right| E\left|\varphi_{q(n),l}(Y)\right| \right] + \frac{n(n-1)(n^2-n-3)}{2n^2\sigma_n^2}$$
$$\left[ \sum_{k=1-t}^{-s} \left\{ E\left|\varphi_{q(n),k}(Y)\right| \right\}^2 \right]^2.$$

Hence, if (7), (9) and (10) hold, then $\operatorname{var}(H_3) \to 0$ as $n \to \infty$. Next, observe under assumption (10)

$$\operatorname{var}(H_2) = \frac{1}{2n\sigma_n^2} E\left[ \sum_{k=1-t}^{-s} \left( \theta_{q(n),k}^2(Y) - c_{kk} \right) \right]^2$$
$$= O\left( \frac{1}{n\sigma_n^2} \right).$$

Consequently, $\operatorname{var}(H_2) \to 0$ as $n \to \infty$. if assumption (9) holds.

Therefore, to complete the proof of the theorem, it is sufficient to show that $H_1 \xrightarrow{d} N(0,1)$. To prove this, observe:

$$EW_{nj}^2 = \frac{2}{n^2\sigma_n^2} E\left( \sum_{i=1}^{j-1} U_{ij} \right)^2 = \frac{2(j-1)}{n^2}$$

Therefore,

$$\sum_{j=1}^{\infty} EW_{nj}^2 = \frac{2}{n^2\sigma_n^2} E\left( \sum_{i=1}^{j-1} U_{ij} \right)^2 \qquad (11)$$
$$= \frac{2(j-1)}{n^2} \qquad \text{for all } n.$$

Next to be shown is

$$\sum_{j=1}^{\infty} W_{nj}^2 \xrightarrow{p} 1. \qquad (12)$$

In order to establish (12), it is enough, in view of Chebychev's inequality and (11), to show that

$$E\left( \sum_{j=1}^{n} W_{nj}^2 \right)^2 = \qquad (13)$$
$$\sum_{j=1}^{n} EW_{nj}^4 + 2\sum_{j<j'}^{n} EW_{nj}^2 W_{nj'}^2 \longrightarrow 1$$
$$\text{as } n \longrightarrow \infty.$$

Using Holder's inequality,

$$W_{nj}^4 \le \frac{4(t-s)^3}{n^4\sigma_n^4}$$
$$E\left[ \sum_{k=1-t}^{-s} \left( \sum_{i=1}^{j-1} \theta_{q(n),k}(Y_i) \right)^4 \theta_{q(n),k}^4(Y_j) \right]$$
$$= \frac{4(t-s)^3}{n^4\sigma_n^4}$$
$$\sum_{k=1-t}^{-s} E\theta_{q(n),k}^4(Y) \left[ (j-1)E\theta_{q(n),k}^4(Y) + 3(j-1)(j-2)c_{kk}^2 \right]$$

By summing over $j$,

$$\sum_{j=1}^{n} EW_{nj}^4 \le \frac{2(n-1)(t-s)^3}{n^3\sigma_n^4}$$

$$\sum_{k=1-t}^{-s} E\theta_{q(n),k}^4(Y)\left[E\theta_{q(n),k}^4(Y)+2(n-s)c_{kk}^2\right]$$

Therefore, under assumptions (9) and (10)

$$\sum_{j=1}^{n} EW_{nj}^4 \to 0 \text{ as } n\to\infty. \qquad (14)$$

Now,

$$2\sum_{j<j'} EW_{nj}^2 W_{nj'}^2 = \frac{8}{n^4\sigma_n^4}\sum_{j<j'} E\left[\left(\sum_{i=1}^{j-1} U_{ij}^2\right)\left(\sum_{r=1}^{j'-1} U_{rj'}^2\right)+\left(\sum_{i=1}^{j-1} U_{ij}^2\right)\left(\sum_{\substack{r=1\\ r\neq r'}}^{j'-1}\sum_{r'=1}^{j'-1} U_{rj'}U_{r'j'}\right)+\left(\sum_{r=1}^{j'-1} U_{rj'}^2\right)\left(\sum_{\substack{i=1\\ i\neq i'}}^{j-1}\sum_{i'=1}^{j-1} U_{ij}U_{i'j}\right)+\left(\sum_{\substack{i=1\\ i\neq i'}}^{j-1}\sum_{i'=1}^{j-1} U_{ij}U_{i'j}\right)\left(\sum_{\substack{r=1\\ r\neq r'}}^{j'-1}\sum_{r'=1}^{j'-1} U_{rj'}U_{r'j'}\right)\right]$$
$$= G_1+G_2+G_3+G_4.$$

Note that

$$G_1 = \frac{8}{n^4\sigma_n^4}\sum_{j<j'}\left[2(j-1)EU_{12}^2U_{23}^2+(j-1)(j'-3)\left\{EU_{12}^2\right\}^2\right]$$
$$= \frac{8(n-1)(n-2)}{3n^3\sigma_n^4}\sum_{k}\sum_{l}\sum_{k'}\sum_{l'} c_{kl}c_{k'l'}E\theta_{q(n),k}\,\theta_{q(n),l}\,\theta_{q(n),k'}\,\theta_{q(n),l'} + \frac{(n-1)(n-2)(n-3)}{n^3}.$$

Thus,

$$G_1 \to 1 \quad \text{as} \quad n\to\infty. \qquad (16)$$

If assumptions (9) and (10) hold,

$$G_3 = 0 \quad \text{and } G_2 \to 0 \quad \text{as} \quad n\to\infty. \ (17)$$

Also, computations (see Ghorai, 1980) show

$$G_4 \to 0 \text{ as } n\to\infty. \qquad (18)$$

Therefore (15) together with (16),(17) and (18) gives

$$2\sum_{j<j'} EW_{nj}^2\,W_{nj'}^2 \to 1 \text{ as } n\to\infty. \ (19)$$

Relation (12) now follows by combining (13), (14) and (19). Finally,

$$P\left(\sup_j |W_{nj}| > \epsilon\right) \le \sum_{j=1}^{n} P\left(W_{nj}^2 > \epsilon^2\right) \le \frac{1}{\epsilon^4}\sum_{j=1}^{n} EW_{nj}^4 .$$

By using (14)

$$\sup_j |W_{nj}| \xrightarrow{\ p\ } 0. \qquad (20)$$

may be deduced. The theorem now follows by combining Theorem 1 with (11), (12) and (20).

## Conclusion

Tests of goodness-of-fit can be obtained as a direct application to Theorem 2. In fact, a test may be constructed for the hypothesis H: $f(x) = f_0(x)$ at a given level $\alpha$, where $f_0(x)$ is a given function. To do this, the statistic

$$R_n = \int \left[ f^*(x) - f(x) \right]^2 dx$$

is to be computed for $f(x) = f_0(x)$ and the hypothesis is to be rejected if $R_n \ge d_n(\alpha)$ where by Theorem 2

$$d_n(\alpha) = \mu_n + \frac{\sqrt{2}\sigma_n}{n}\Phi^{-1}(1-\alpha),$$

where

$$\Phi(z) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

References

Adnan, M. A. (1981). Conditional central limit theorem for martingales and reversed martingales, *Sankhya*: *The Indian Journal of Statistics*, Series A , 43 , No. 1 , 100-106.

Alyass, K. K. & Sun, T. C. (1994). Asymptotic distribution of the quadratic norms of the deviation of orthogonal series type density estimates, *Sankhya*: *The Indian Journal of Statistics*, Volume 56, Series A, Pt. 2, 238-264.

Bickel, P. L. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimators, *Ann. Statist*., 1071-1095.

Daubechies, I.(1992). Ten lectures on wavelets, *Society for Industrial and Applied Mathematics*, Philadelphia, Pa.

Doukhan, P. & Leon, J. (1990). Deviation quadratique d'estimateurs de densite par projections orthogonales, *Comptes Rendus de 1'Academic des sciences, Series*-1 *Mathematiquel* 310, No.6 425-430.

Ghorai, J. (1980). Asymptotic normality of quadratic measure of orthogonal series type density estimate, *Ann. Inst. Statist. Math*. 32, Part A, 341-350.

Kerkyacharian, G. & Picard, D. (1992). Density estimation in Besov space, *Statistics & Probability letters* 13, 15-24.

Lii, K. S. (1978). A global measure of a spline density estimate, *Ann. Statist*., Vol. 6 , No 5, 1138-1148

Meyer, Y. (1990). Ondelettes et operateurs, Hermann, Paris.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence, *Ann. Statist.* 3, 1-14.

Strichartz, R. S. (1993). How to make wavelets. Amer. *Math. Monthly,* 100(6), 539-556.

Viollaz, A. J. (1980). Asymptotic distribution of $L_2$ norms of the deviation of density function estimates, *Ann. Statist*. 8, 322-346.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 551-560
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Bayesian Subset Selection of Binomial Parameters Using Possibly Misclassified Data

James D. Stamey  Thomas L. Bratcher  Dean M. Young
Baylor University

Three Bayesian approaches are considered for the selection of binomial proportion parameters when data is subject to misclassification. The cases where the misclassification is non-differential and differential were considered, thus extending previous work which considered only non-differential misclassification. In this article, various selection criteria are applied to a simulated data set and a real data set.

Key words: Bayes, posterior approximation, Gibbs Sampler, binomial parameter subset selection

## Introduction

A decision maker is often interested in selecting the population from among several populations that will produce the largest or smallest parameter value. For example, an experimenter might be interested in determining which production technique gives the lowest percentage of defects; a crime analyst might consider which reporting district has the highest rate of violent crimes; a baseball fan might inquire about the best home run hitter of the twentieth century. In each case a selection of a population parameter must be made from a set of parameters using data from the populations of interest. This process is known as the subset-selection problem. Of course various procedures exist for selecting a subset that contains the best (largest or smallest) parameter. Here, the concern is with the Bayesian subset selection paradigm.

James Stamey is an Assistant Professor in the Department of Statistical Science at Baylor University. His research interests are in discrete data with misclassification errors and Bayesian analysis. Email: James_Stamey@baylor.edu. Tom Bratcher is a Professor of Statistics. His research interests are selection and ranking. Dean Young is Professor of Statistics. His research interests are multivariate analysis and linear models.

The concept of subset selection essentially began with an article by Gupta and Sobel (1957), who described a statistic that can be used in parameter ranking and selection for multiple populations. Early work on Bayesian subset selection was initiated by Bratcher & Bhalla (1974), who have used a constant loss function to derive a Bayesian subset selection procedure, and Govindarajulu & Harvey (1974). For other Bayesian subset-selection approaches and related topics, see Goel & Rubin (1977), Gupta & Hsu (1977), Berger (1979, 1980), Gupta & Yang (1985), Gupta & Liang (1987), Berger & Deely (1988), Dixon & Simon (1991, 1994), Schulter, Deely, & Nicholson (1997) and Deely & Smith (1998).

Examples abound where interest might be in selecting a subset of binomial proportion parameters using correctly classified and misclassified data. For example, Hanson, Johnson, & Gardner (2003) have considered the prevalence of the disease bovine brucellosis in cattle herds in twenty regions of Mexico. This application can be thought of as a type of quality control in which one wishes to determine a set of herds deemed most likely to develop bovine brucellosis or, conversely, perhaps a set of herds that could be considered least likely to have the disease. A second application of a subset-selection method for binomial proportion parameters using possibly misclassified data is auditing. For instance, Raats & Moors (2004) have estimated the proportion of errors in social security payments in the Netherlands combining

fallible and validation data. One could also compare or select a subset of the proportion parameters of errors in auditing across geographical regions, industries, or some other variable of interest.

In both of the above examples, one cannot reasonably assume the observed counts are infallible. Most diagnostic tests are well known to be fallible. That is, most diagnostic tests can indicate that subjects have a disease when they do not or that they are disease free when they are actually infected. An appropriate statistical model will adjust for the error rates of the fallible test. Joseph, Gyorkos, & Coupal (1995) and Dendukuri & Joseph (2001) considered the case of estimating the prevalence of one population with fallible data. Hanson et al. (2003) have extended this work to multiple populations. Hanson et al. (2003) assumed that the properties of the diagnostic tests remain constant across populations. This assumption is referred to as non-differential misclassification.

Two subset-selection criteria of Schluter et al. (1997) and a subset- selection criterion proposed by Stamey, Bratcher, & Young (2004) are applied here to the bovine brucellosis data found in Hanson et al. (2003). Also proposed is a method of extending the hierarchical model to allow for differential misclassification. Differential misclassification occurs when the false positive and false negative rates are different in each population. For this scenario it is assumed that an expensive error-free classifier is available for a small sample of units. A sample where both fallible and infallible observations are made is often called a validation sample. A simulated binomial parameter subset-selection problem with differential misclassification motivated by an auditing application in Raats & Moors (2004) is considered.

## Methodology

A parametric hierarchical model for binomial data with misclassification analogous to Hanson et al. (2003) is provided and a Bayesian extension is proposed for the case of differential misclassification. For the non-differential misclassification model, consider the case where only a single classifier is utilized; however, the method is easily extended to allow for two or more classifiers. The hierarchical model is

$$Z_i \mid \pi_i, \eta, \theta \sim \text{binomial}(n_i, p_i)$$

with

$$p_i = \pi_i\eta + (1 - \pi_i)(1 - \theta),$$

where $p_i$ is the population proportion of observable occurrences in population $i = 1, \ldots, m$. The parameter $\pi_i$ is the true probability of a positive response for population $i$ and is assumed to vary across populations. The parameter $\eta$ = 1 – P(false negative) is the sensitivity, or probability that a true positive is observed and is assumed to be the same for all populations. The parameter $\theta$ = 1 – P(false positive) is the specificity, or probability that a true negative is labeled as a negative and is also assumed to be the same for all populations. The first-stage priors of the Bayesian hierarchical model are

$$\pi_i \sim \text{beta}(\alpha, \beta),$$

$$\eta \sim \text{beta}(\alpha_\eta, \beta_\eta),$$

and

$$\theta \sim \text{beta}(\alpha_\theta, \beta_\theta).$$

The beta prior is the usual first-stage prior for hierarchical binomial models and is consistent with the models of Hanson et al. (2003). One can elicit priors for the sensitivity and specificity by using the approaches of Chaloner (1996) and Kadane & Wolfson (1996).

To model the heterogeneity of the prevalences, the parametric prior of Hanson et al. (2003) is used for both its convenience and ease of interpretation. Here, $\alpha = \mu\gamma$ and $\beta = \gamma$, where the parameter $\mu$ is the grand mean of the population prevalences and $\gamma$ controls the heterogeneity of the prevalences since the variance is $\frac{\mu(1-\mu)}{1+\gamma}$. Specifically, the larger the value of $\gamma$, the tighter the distribution of the prevalences. To finish the hierarchy, assume $\mu \sim \text{beta}(\alpha_\mu, \beta_\mu)$ and $\gamma \sim \text{gamma}(\alpha_\gamma, \beta_\gamma)$, where $\alpha_\mu$, $\beta_\mu$, $\alpha_\gamma$, and $\beta_\gamma$ are hyperpriors specified by the

experimenter. The joint posterior of all parameters is proportional to

$$p(\pi_i,\theta,\eta,\mu,\gamma\,|\,\mathbf{d}) \propto \prod_{i=1}^{r}\pi_i^{\mu\gamma-1}(1-\pi_i)^{\gamma-1}\theta^{\alpha_\theta-1}(1-\theta)^{\beta_\theta-1}\eta^{\alpha_\eta-1}(1-\eta)^{\beta_\eta-1}p_i^{z_i}(1-p_i)^{n_i-z_i}.$$

Hanson et al. (2003) provided a method for eliciting values for the parameters of the priors. However, in the analyses diffuse non-informative priors are used. No apparent closed-form posterior distributions exist, but the parameters can be estimated using either Monte Carlo integration or Markov Chain Monte Carlo methods. The free software WinBugs is used to approximate the posterior densities that is used. These WinBugs software programs are available from the first author.

The assumption that the sensitivity and specificity do not vary across populations is quite strong and often fails in practice. Here the model of Hanson et al. (2003) is extended to the case where the sensitivity and specificity are not the same across populations. If it is believed that the misclassification parameters vary across populations, it is recommended to use one of the following approaches. If the number of populations is not large, an expert to elicit prior parameters for each specificity and sensitivity can be used, using methods detailed in Chaloner (1996) and Kadane & Wolfson (1996). This approach results in the following change in the hierarchical model: $\eta_i \sim \text{beta}(\alpha_{\eta i}, \beta_{\eta i})$ and $\theta_i \sim \text{beta}(\alpha_{\theta i}, \beta_{\theta i})$.

However, if expert opinion is not available for each of the sensitivities and specificities, another method is needed. One possibility is to use validation data for each population. For instance, Raats & Moors (2004) have assumed that a large sample of accounts is audited by a fallible auditor, and then a small random sample of these accounts is double checked by an infallible expert. Suppose in each population $r_i$ units are classified by both the fallible and infallible procedure. The validation data adds the following binomial likelihoods to the experiment likelihood:

$$T_i\,|\,\pi_i \sim \text{binomial}(r_i, \pi_i),$$

$$X_i\,|\,t_i, \eta_i \sim \text{binomial}(t_i, \eta_i),$$

and

$$Y_i\,|\,t_i, \theta_i \sim \text{binomial}(r_i - t_i, \theta_i).$$

Here, $T_i$ is the number of positive responses determined by the infallible classifier, $X_i$ is the number of true positive responses correctly labeled as positive by the fallible classifier, and $Y_i$ is the number of true negative responses labeled as negative by the fallible classifier. Then, a hierarchical structure for the sensitivity and specificity parameters similar to that used on the prevalences is used. That is, $\eta_i \sim \text{beta}(\alpha_\eta, \beta_\eta)$ and $\theta_i \sim \text{beta}(\alpha_\theta, \beta_\theta)$ and define $\alpha_\eta = \mu_\eta\gamma_\eta$, $\beta_\eta = \gamma_\eta$, $\alpha_\theta = \mu_\theta\gamma_\theta$, and $\beta_\theta = \gamma_\theta$. The hierarchy is completed with the priors

$$\mu_\eta \sim \text{beta}(\alpha_{\mu\eta}, \beta_{\mu\eta}),$$
$$\gamma_\eta \sim \text{gamma}(\alpha_{\gamma\eta}, \beta_{\gamma\eta}),$$
$$\mu_\theta \sim \text{beta}(\alpha_{\mu\theta}, \beta_{\mu\theta}),$$

and

$$\gamma_\theta \sim \text{gamma}(\alpha_{\gamma\theta}, \beta_{\gamma\theta}).$$

The WinBugs computer programs used to approximate the posterior distributions are available from the first author.

## Three Subset Selection Procedures

Reviewed next are two subset selection criteria from Schluter et al. (1997) and a decision theoretic subset selection criterion from Stamey, Bratcher, & Young (2004) and extend them to apply to the binomial parameter case using possibly misclassified data.

### A Posterior Probabilities Approach (Schluter et al. (1997))

The first subset-selection procedure that is considered uses the posterior probability that a site has the largest prevalence or is largest by a multiple of, say, $\nu$. That is,

$$p_i(\nu) = P(\pi_i > \nu\pi_j, \forall\, j \neq i\,|\,\mathbf{z}) \qquad (1)$$

where $\mathbf{z}$ represents the vector of observed data. The probability (1) does not have a closed form; however, MCMC methods make (1) trivial to calculate. Suppose that after an initial burn-in, the Gibbs sampler is run $B$ iterations. One can

approximate the posterior probability (1) by counting the number of times

$$\pi_{ik} = \max(\nu\pi_{1k}, ..., \nu\pi_{i-1,k}, \pi_{ik}, \nu\pi_{i+1,k}, ..., \nu\pi_{mk}),$$

where $k = 1, \ldots, B$. Specifically, probability (1) is approximated as

$$p_i(\nu) \approx \frac{\#(\pi_{ik} = \max(\nu\pi_{1k}, ..., \nu\pi_{i-1,k}, \pi_{ik}, \nu\pi_{i+1,k}, ..., \nu\pi_{mk}))}{B}$$

where $\#(\cdot)$ denotes the number of elements in a set. In this case count the number of Gibbs sampler iterations such that the prevalence of interest is the largest. Schluter et al. (1997) have remarked that if $\nu = 1$, then (1) simply becomes the probability that $\pi_i$ is the largest prevalence. The populations can be ranked via

i) the use of the posterior probability (1),
ii) the use of some probability threshold chosen such that the groups selected are the smallest subset where the sum of the $p_i(\nu)$ probabilities exceed the threshold, or
iii) the choice of $r < m$ largest probabilities to be included in the superior set.

A Predictive Probabilities Approach (Schluter, et al., 1997)

A second criterion is based on the predictive number of future occurrences in a future sample. The criterion is based on the probability that a future number of true positives, say $W_i$, exceeds some experimenter-chosen quantity, say $w^*$, or

$$pd_i(w^*) = P(W_i > w^* \mid \mathbf{z}, n_0) \qquad (2)$$

where $n_0$ represents the future sample size. To compute probability (2) with the Gibbs sampler, add the variables $W_i \mid \pi_i \sim$ binomial($n_0$, $\pi_i$) for $i = 1, 2, \ldots, m$, to the likelihood. The approximation

$$pd_i(w_i) \approx \frac{\#(W_i \geq w^*)}{B}$$

is then straightforward to calculate. One can rank the populations via probability (2) and then either include the top $r$ of them in a superior set or select all populations whose predictive probability (2) is greater than some user-specified value $P_0$. Difficulties with this criterion include determining a meaningful future sample size $n_0$ and defining a meaningful comparison number $w^*$.

A Decision Theoretic Approach (Bratcher & Bhalla (1974))

Stamey et al. (2004) used a constant loss function for Poisson parameters with misclassified data. Here a similar loss function for the binomial data case is utilized,

$$L(\pi) = \begin{cases} c_1 \#(S) + c_2 & \text{if } \pi_{\max} \notin S \\ c_1 \left[\#(S) - 1\right] & \text{if } \pi_{\max} \in S \end{cases},$$

where $S$ denotes the superior set, $\#(S)$ denotes the number of parameters in the superior set, and $\pi_{\max} \in S$ represents placing the actual maximum proportion in the superior set. The corresponding risk is a linear combination of the expected size of the superior set and the probability of correct selection. Formally, the risk is

$$R(\pi) = c_1 E[\#(S)] + (c_1 + c_2)(1 - P(CS)) - c_1$$

where $P(CS)$ denotes probability of correct selection, i.e., $\pi_{\max}$ is selected. The Bayes threshold for inclusion is

$$p(\pi_i = \pi_{\max}) \geq 1/(c+1), \qquad (3)$$

where $c = c_2/c_1$. This loss ratio represents the relative seriousness of the two types of mistakes: leaving the largest parameter out of the superior set and putting a parameter in the superior set that is not the largest. Additionally, $c + 1$ may be considered the rate of change in $E[\#(S)]$ with respect to $P(CS)$. To guarantee that at least one parameter is placed in the superior set $S$, it is required that $c \geq m - 1$. The left side of (3) is approximated identically to (1) when $\nu = 1$. The estimated probabilities are then compared to $1/(c + 1)$, and the parameter $\pi_k$ is placed in the superior set $S$ when

$$p_i(\nu) \approx \frac{\#(\pi_{ik} = \max(\nu\pi_{1k}, ..., \nu\pi_{i-1,k}, \pi_{ik}, \nu\pi_{i+1,k}, ..., \nu\pi_{mk}))}{B}$$

$$> 1/(c + 1).$$

## Results

The methods discussed are now applied to real data originally found in Hanson et al. (2003). Twenty cow herds in an area of Mexico where the disease is known to occur are sampled and tested with the buffered acidified plate agglutination (BAPA) serologic test. The BAPA is known to be imperfect, and its properties are discussed in Stemshorn et al. (1985). Point estimates of the sensitivity and specificity are .75 and .97, respectively. As in Hanson et al. (2003), this article used an equivalent sample size of 20 for the beta priors based on the prior means of .75 and .97, respectively. That is, seek beta priors with means of .75 and .97 where the sum of the parameters is 20; thus, $\eta \sim$ beta(15,5) and $\theta \sim$ beta(19.4, .6) was used. Had an equivalent sample size of 40 been used, it would have been assumed that $\eta \sim$ beta(30, 10) and $\theta \sim$ beta(38.8, 1.2). Interestingly, virtually identical inferences resulted from the two sets of priors with only a slight decrease in posterior variation. For this example the non-informative priors $\mu \sim$ beta(1, 1) and $\gamma \sim$ gamma(.001, .001) were used for the hierarchical parameters in the model for the prevalences.

WinBugs was used to approximate the posterior distributions. We show a plot of the approximate posterior densities for the bovine brucellosis prevalences in Figure 1. For this data one can visually see that clear differences exist among the posterior densities. The posterior distributions for the prevalences $\pi_{15}$ and $\pi_7$ are centered at considerably larger values than the posterior distributions of the other prevalences.

Using (1), the posterior probabilities that each prevalence was the largest were calculated. Table 1 gives results for the posterior probabilities approach of selecting the largest prevalence for values of $\nu$ of 1, 1.1, and 1.25. In the table there are sites and corresponding posterior probabilities where $P(\pi_i > \nu\pi_j, \forall\, j \neq i \,|\, \mathbf{z})$ exceed 0.01 when $\nu = 1$. If one use criterion ii) with a threshold of .9 in conjunction with the posterior probability criterion, one can see from Table 1 that the two prevalences $\pi_{15}$ and $\pi_7$ were the only elements contained in the superior set $S$ using the posterior probability criterion. If the threshold had been increased to .99, then the prevalences $\pi_{14}$ and $\pi_{19}$ would be added to the superior set $S$. If one increases $\nu$ to 1.1 and 1.25, then it becomes evident from Table 1 that $\pi_{15}$ is the sole choice for the largest prevalence.

Next, the predictino approach criterion is applied to the bovine brucellosis data. It was assumed a future sample size of $n_0 = 10$ and provided the probabilities for various values of $w^*$. Figure 2 is a plot of the results for values of $w^*$ ranging from 0 to 5. For illustrative purposes supposed $w^* = 3$ and $P_0 = .8$, then placed a rectangle or box in the area of Figure 2 where the prediction criterion holds. All curves that fall inside the box, which in this case corresponded to the prevalences $\pi_7$, $\pi_{14}$, and $\pi_{15}$, satisfied the prediction criterion. The graph could easily be changed to allow for different values of $P_0$ and $w^*$.

Consider the decision theoretic approach to selecting herds with the largest bovine brucellosis prevalence. Only the prevalences $\pi_{15}$, $\pi_7$, and $\pi_{14}$ would be selected at the boundary for the rate of change, $(c + 1) = 20$, which gave a critical probability of $1/(c + 1) = .05$. Thus, for this example we assumed it is 19 times more serious to leave the largest prevalence out of the superior set than to include a prevalence in the superior set $S$ that is not the largest. If it were to be considered to be 99 more times serious to leave the largest prevalence out of the superior set than to include a prevalence in the superior set $S$ that is not the largest, the critical probability would decrease to .01, and the prevalences $\pi_{15}$, $\pi_7$, $\pi_{14}$, and $\pi_9$ would be included in the superior set.

## Auditing Application

As a second example, data were simulated similar to that found in Raats & Moors (2004). Suppose we wish to compare 15 locations in terms of the proportion of errors in accounts. As in Raats & Moors (2004), we assumed that the initial audit is fallible, that is, some accounts that are in error could be missed

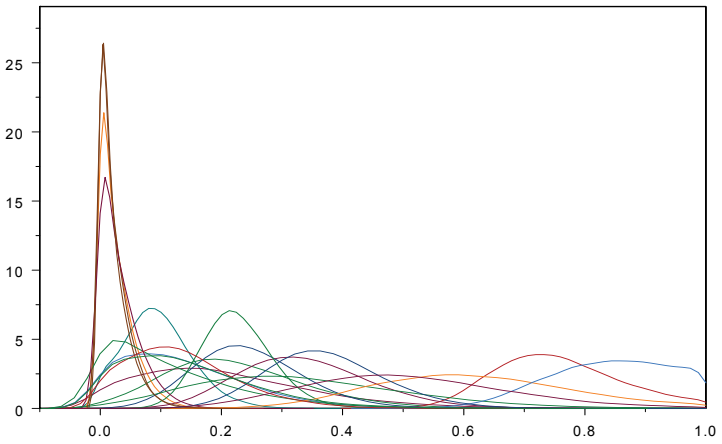



Figure 1. Posterior densities of prevalences for bovine data

Table 1. Posterior probabilities of having the largest prevalence

| ν | Herd 15 | Herd 7 | Herd 14 | Herd 19 | Others |
|---|---|---|---|---|---|
| 1 | .773 | .158 | .052 | .013 | .004 |
| 1.1 | .469 | .032 | .000 | .000 | .000 |
| 1.25 | .123 | .000 | .000 | .000 | .000 |

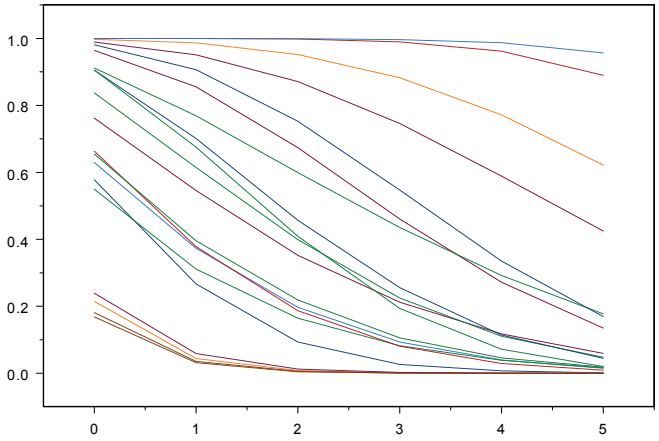



Figure 2. Predictive probabilities for bovine brucellosis data. The rectangle includes herds that satisfy a predictive probability of 3 or more events with probability greater than .8.

and some accounts that are correct could be labeled as in error. For each of the 15 locations, the parameters of the populations with the following distributions: $\pi_i \sim$ beta(2, 18), $\eta_i \sim$ beta(12, 8), and $\theta_i \sim$ beta(19, 1) were generated. These distributions are consistent with Raats & Moors (2004) in the sense that the overall proportion of errors is small with a mean of 10%, the sensitivity is moderate with a mean of 60%, and the specificity is high with a mean of 95%. For each site the following was generated $z_i \sim$ binomial(500, $p_i$), $t_i \sim$ binomial(60, $\pi_i$), $x_i \sim$ binomial($t_i$, $\eta_i$), and $y_i \sim$ binomial(60 – $t_i$, $\theta_i$), where $p_i = \pi_i\eta_i + (1 - \pi_i)(1 - \theta_i)$.

For the hierarchical model, allow for differential misclassification by using diffuse priors for all hyperprior distributions. Specifically, let $\mu \sim$ beta(1, 1), $\gamma \sim$ gamma(.001, .001), $\mu_\eta \sim$ beta(1, 1), $\gamma_\eta \sim$ gamma(.001, .001), $\mu_\theta \sim$ beta(1, 1), and $\gamma_\theta \sim$ gamma(.001, .001).

Two competing models were considered. The first was an independence-based model where each of the 15 sites was modeled independently and, thus, no information-sharing occurred among the sites. For the independence models beta(1, 1) priors were used for all parameters. Also considered was the hierarchical model of Hanson et al. (2003), previously used on the first example, where all the specificities and sensitivities were constant. For this non-differential misclassification model, the actual distributions from which the sensitivities and specificities were generated are used as the prior distributions. That is, the priors $\eta \sim$ beta(12, 8) and $\theta \sim$ beta(19, 1) were assumed. The generated proportions, posterior means of the validation data hierarchical model, and 95% intervals for all three models are provided in Table 3.

Note that the 95% intervals for the hierarchical model and the independence model both contained the true parameter values in all cases while the non-differential misclassification model missed two of the parameters. Also, the hierarchical model had the narrowest intervals, thus supporting the use of this model.

Table 4 gives the sites and corresponding posterior probabilities of having the largest prevalence for parameters where $P(\pi_i > \nu\pi_j, \forall\, j \neq i \mid \mathbf{z})$ exceed 0.01 when $\nu = 1$. Probabilities are provided for the case where $\nu =$ 1 and 1.1. Assuming criterion ii) with a probability threshold of .9, it was determined that the proportions $\pi_8$, $\pi_1$, and $\pi_3$, were included in the superior set because the sum of their probabilities is .923. In Table 4 are the three largest proportions used to generate the data in order from largest to smallest are $\pi_8$, $\pi_3$, and $\pi_1$. Thus, the posterior probability procedure included the three largest proportions in this example. If the threshold was increased to .99, then the proportions $\pi_8$, $\pi_1$, $\pi_3$, $\pi_{10}$, $\pi_7$, $\pi_9$, and $\pi_2$ would all be included in the superior set *S*.

If non-differential misclassification is incorrectly assumed, then one would have incorrectly concluded that $\pi_{13}$ was the largest proportion with a corresponding posterior probability of .865 of being the largest proportion. Also, if the incorrect non-differential misclassification model were applied, one would have determined that the second largest proportion was $\pi_{10}$ with a posterior probability of .106 of being the largest proportion. In this case the non-differential misclassification assumption leads to incorrect inferences because neither site 13 nor 10 was actually among the three largest proportions.

For this same data the prediction subset-selection criterion was applied. It was assumed a future sample size of 50. For the validation-data model with differential misclassification, the plot for all 15 sites for values of $w^*$ from 0 to 6 is given in Figure 3. Included is a decision box for $w^* = 2$ and $P_0 = .6$. It was found that sites 1, 3, 7, 8, and 10 satisfied this particular configuration and, therefore, $\pi_1$, $\pi_3$, $\pi_7$, $\pi_8$, and $\pi_{10}$ would be placed in the superior set. Recall that $\pi_1$, $\pi_3$, and $\pi_8$ were the largest three proportions so, again, this proposed prediction subset-selection criterion yielded very reasonable results.

For the decision theoretic approach, this article again considered *c*'s of 19 and 99 that yielded critical probabilities of .05 and .99. For a critical probability of .05, the proportions $\pi_8$, $\pi_1$, and $\pi_3$ were included in the superior set *S*.

Table 3. Posterior means and intervals for simulated auditing example
(Intervals that failed to cover the true parameter are bolded.)

| Site | True value | Posterior mean differential hierarchical | 95% Interval differential hierarchical | 95% Interval independence | 95% Interval non-differential hierarchical |
|---|---|---|---|---|---|
| 1 | .141 | 0.157 | (.080, .244) | (.096, .290) | (.117, .465) |
| 2 | .076 | 0.102 | (.051, .161) | (.049, .175) | (.000, .230) |
| 3 | .148 | 0.137 | (.089, .202) | (.099, .258) | (.000, .162) |
| 4 | .017 | 0.047 | (.012, .097) | (.004, .087) | (.000, 162) |
| 5 | .055 | 0.044 | (.010, .093) | (.004, .083) | (.000, .174) |
| 6 | .131 | 0.100 | (.055, .146) | (.054, .163) | (.000, .214) |
| 7 | .126 | 0.118 | (.067, .180) | (.073, .229) | (.000, .205) |
| 8 | .201 | 0.190 | (.128, .262) | (.145, .295) | (.122, .480) |
| 9 | .103 | 0.119 | (.068, .175) | (.070, .183) | (.003, .267) |
| 10 | .101 | 0.120 | (.063, .190) | (.067, .221) | **(.150, .542)** |
| 11 | .059 | 0.063 | (.023, .108) | (.017, .117) | (.000, .219) |
| 12 | .092 | 0.090 | (.046, .145) | (.037, .149) | (.000, .214) |
| 13 | .070 | 0.061 | (.017, .115) | (.011, .113) | **(.200, .658)** |
| 14 | .119 | 0.079 | (.037, .135) | (.029, .142) | (.072, .372) |
| 15 | .089 | 0.090 | (.038, .153) | (.028, .163) | (.065, .361) |

Table 4. Posterior probabilities of having the largest proportion of errors

| $\nu$ | $\pi_8$ | $\pi_1$ | $\pi_3$ | $\pi_{10}$ | $\pi_7$ | $\pi_9$ |
|---|---|---|---|---|---|---|
| 1 | .618 | .246 | .059 | .026 | .021 | .015 |
| 1.1 | .452 | .142 | .023 | .010 | .007 | .005 |

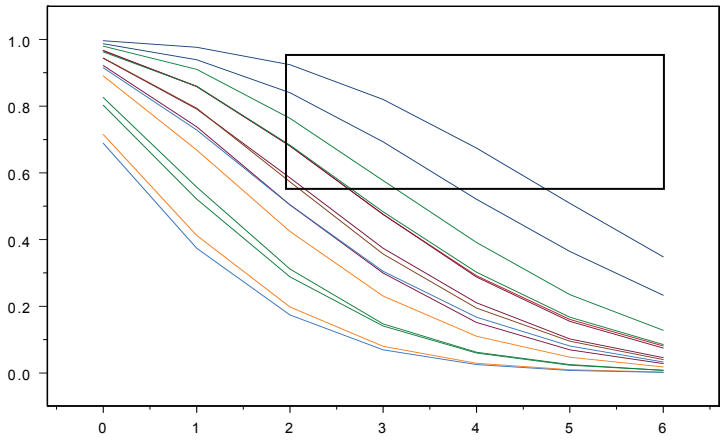



Figure 3. Predictive probabilities for auditing example. The rectangle includes populations that satisfied a predictive probability of 2 or more events with probability greater than .6.

For a critical probability of .01, the proportions $\pi_{10}$, $\pi_7$, and $\pi_9$ also entered the superior set. For the decision theoretic approach, this article again considered *c*'s of 19 and 99 that yielded critical probabilities of .05 and .99. For a critical probability of .05, the proportions $\pi_8$, $\pi_1$, and $\pi_3$ were included in the superior set *S*. For a critical probability of .01, the proportions $\pi_{10}$, $\pi_7$, and $\pi_9$ also entered the superior set.

## Conclusion

In this article, three ranking criteria were applied to a hierarchical binomial model with misclassification first proposed in Hanson et al. (2003). These criteria are easy to use and understand and are computationally practical because of currently available statistical software. This has also extended the non-differential misclassification model of Hanson et al. (2003) to allow for differential misclassification. The example using simulated audit data with misclassified observations illustrates the importance of appropriately incorporating differential misclassification in the analysis. It is note that the Bayesian binomial parameter selection methods proposed here could also apply to psychology and medical subset-selection problems, where interest might lie in comparing various treatments when a fallible diagnostic test is used to assess presence of a particular psychological or medical condition. Finally, the computations in this article have been performed using WinBugs, which is a free statistical computing package available on the Internet.

## References

Berger, R. L. (1980), Minimax subset selection for the multinomial distribution, *Journal of Statistical Planning and Inference 4,* 391-402.

Berger, R. L. (1979), Minimax subset selection for loss measured by subset size, *Annals of Statistics 7,* 1333-1338.

Berger, J. O., & Deely, J. (1988), A Bayesian approach to ranking and selection of related means with alternatives to analysis-of-variance methodology, *Journal of the American Statistical Association 83,* 364-373.

Bratcher, T.L., & Bhalla, P. (1974), On the properties of an optimal selection procedure, *Communications in Statistics - Theory and Methods 3,* 191-196.

Chaloner, K. (1996), Elicitation of prior distributions, *Bayesian Biostatistics*, 41-156.

Deely, J. J., & Smith, A. F. (1998), Quantitative refinements for comparison of institutional performance, *Journal of the Royal Statistical Society, A, 161*, 5-12.

Dendukuri, N., & Joseph, L. (2001), Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests, *Biometrics 57,* 208-217.

Dixon, D. O., & Simon, R. (1991), Bayesian subset analysis**,** *Biometrics 47*, 871-881.

Dixon, D. O., & Simon, R. (1994), Corrections: Bayesian subset analysis**,** *Biometrics 50*, 322.

Goel, P. K., & Rubin, H. (1977), On selecting a subset containing the best population-a Bayesian approach, *Annals of Statistics 5,* 969-983

Govindarajulu, Z., & Harvey, C. (1974), Bayesian procedures for ranking and selection problems, *Annals of the Institute of Statistical Mathematics 26*, 35-53.

Gupta, S. S., & Hsu, J. C. (1977), On the monotonicity of Bayes subset selection procedures, *Proceedings of the 41st Session of the International Statistical Institute 47*, New Delhi, India, 208–211.

Gupta, S. S., & Liang, T. (1987), On some Bayes and empirical Bayes selection procedures, *Probability and Bayesian Statistics* (Ed., R. Viertl), New York: Plenum Publishing Corporation, 233–246.

Gupta, S. S., & Sobel, M. (1957), On a statistic which arises in ranking and selection problems, *Annals of Mathematical Statistics 28*, 957-967.

Gupta, S. S., & Yang, H.-M. (1985), Bayes-$P^{*}$ subset selection procedures, *Journal of Statistical Planning and Inference* 12, 213-233.

Hanson, T., Johnson, W., & Gardner, I. (2003), Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard, *Journal of Agricultural, Biological, and Environmental Statistics 8,* 223-239.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995), Bayesian Estimation of Disease Prevalence and Parameters of Diagnostic Tests in the Absence of a Gold Standard, *American Journal of Epidemiology 141,* 263-272.

Kadane, J. B., & Wolfson, L. J. (1996), Priors for the design and analysis of clinical trials, *Bayesian Biostatistics*, 157-184.

Raats, V. A., & Moors, J. J. A. (2004), Double-checking auditors: a Bayesian approach, *Journal of the Royal Statistical Society, D*, *52,* 351-366.

Schluter, D. C., Deely, J. J., & Nicholson, D.G. (1997), Ranking and Selecting Motor Vehicle Accident Sites Using a Hierarchical Bayesian Model, *The Statistician 46,* 293-316.

Stamey, J. D., Bratcher, T. L., & Young, D. M. (2004), Parameter subset selection and multiple comparisons of Poisson rate parameters with misclassified data, *Computational Statistics and Data Analysis 45,* 467479.

Stemshorn, B. W., Forbes, L. B., Eaglesome, M. D., Nielsen, K. H., Robertson, F. J., & Samagh, B. S. (1985), A comparison of standard serologic tests for the diagnosis of bovine brucellosis in Canada, *Canadian Journal of Comparative Medicine 49,* 391–394.

Journal of Modern Applied Statistical Methods
March, 2008, Vol. 6, No. 2, 561-572
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Covariate Dependent Markov Models for Analysis of Repeated Binary Outcomes

M. A. Islam
Department of Statistics
University of Dhaka

R. I. Chowdhury
Health Science Center
Kuwait University

K. P. Singh
University of North Texas
Health Science Center

The covariate dependence in a higher order Markov models is examined. First order Markov models with covariate dependence are discussed and are generalized for higher order. A simple alternative is also proposed. The estimation procedure is discussed for higher order with a number of covariates. The proposed model takes into account the past transitions. Transitions are fitted and are tested in order to examine their influence on the most recent transitions. Applications are illustrated using maternal morbidity during pregnancy. The binary outcome at each visit during pregnancy is observed for each subject and then the covariate dependent Markov models are fitted. The results indicate that the proposed model can be employed for analyzing repeated observations conveniently.

Key words**:** Markov models, higher order, covariate dependence, repeated observations, transitions

## Introduction

Markov chain models can be used in analyzing longitudinal data. There are several discrete time Markov chain models proposed for analyzing repeated categorical data over decades. A model for estimating odds ratio from a two state transition matrix was proposed by Regier (1968). Prentice and Gloeckler (1978) proposed a grouped data version of the proportional hazards regression model for estimating computationally feasible estimators of the relative risk function. Korn and Whittemore (1979) proposed a model to incorporate role of previous state as a covariate to analyze the probability of occupying the current state.

To analyze the binary sequence of presence or absence of diseases, Muenz and Rubinstein (1985) introduced a discrete time Markov chain for expressing the transition probabilities in terms of covariates. The technique proposed by them is applicable for first order Markov model but they provided hints that the approach can be extended for second-order Markov chains. For analyzing sequences of ordinal data from relapsing and remitting of a disease, Albert (1994) developed a finite Markov chain model. In addition, Albert and Waclawiw (1998) developed a class of quasi-likelihood models for a two state Markov chain with stationary transition probabilities for heterogeneous transitional data. Raftery (1985), Raftery and Tavare (1994) proposed a higher order Markov chain model with dependence on contribution of the past transitions. Islam and Chowdhury (2006) presented a higher order version of the covariate dependent Markov model.

For analyzing repeated observations, there is a renewed interest in the development of multivariate models based on Markov chains. These models can be employed for analyzing data generated from meteorology, epidemiology and survival analysis, reliability, econometric analysis, biological concerns, etc. Muenz and Rubinstein (1985) employed logistic regression models to analyze the transition probabilities from one state to another. The estimation for first-order Markov models is quite straight forward, but still there is serious lack of generalization in estimation and testing for models applicability for higher order Markov chains. Islam and Chowdhury (2006) provided a further generalization for covariate dependent

Email Professor M. A. Islam at: mataharul@yahoo.com. Email R. I. Chowdhury, M.Sc., at: rafiq@hsc.edu.kw. Email Professor and Chair, K. P. Singh at ksingh@hsc.unt.edu.

higher order models. This paper makes an attempt to present a simplified version of the covariate dependent higher order Markov models.

A parallel stream of development is observed in analyzing transition models with serial dependence of the first or higher orders on the basis of the marginal mean regression structure models. Azzalini (1994) introduced a stochastic model, more specifically, first order Markov model, to examine the influence of time-dependent covariates on the marginal distribution of the of the binary outcome variables in serially correlated binary data. Markov chains are expressed in transitional form rather than marginally and the solutions are obtained such that covariates relate only to the mean value of the process, independent of association parameters. Following Azzalini (1994), Heagerty and Zeger (2000) presented a class of marginalized transition models (MTM) and Heagerty (2002) proposed a class of generalized MTMs to allow serial dependence of first or higher order. These models are computationally tedious and the form of serial dependence is quite restricted. If the regression parameters are strongly influenced by inaccurate modeling for serial correlation then the MTMs can result in misleading conclusions. Heagerty (2002) provided derivatives for score and information computations.

Transition models are used here for Markov chain regression for binary responses proposed by Diggle et al. (2002). This type of models takes into account the potential impact of explanatory variables depending on the order of the underlying Markov model. This class of models has the flexibility to address a wide range of possible situations, ranging from only main effects to main effects and all possible interactions that emerge from different past transitions of the underlying Markov model. Some hypothesized situations are considered with main effects and some potential interactions emerging from past transitions of the process. In addition, a simple alternative is suggested to test for the order of Markov model.

Covariate Dependent Higher Order Model

As the first serious attempt to analyze covariate dependence of transition probabilities in a Markov model was proposed by Muenz and Rubinstein (1985), a brief review of the model provides a useful background for the proposed model for higher order.

Consider a two-state Markov chain for a discrete time binary sequence as follows:

$$\pi = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix} \qquad (2.1)$$

where $\pi_{00} = 1 - \pi_{01}$ and $\pi_{10} = 1 - \pi_{11}$. Here, 0 and 1 are two possible outcomes of a dependent variable, Y. Each row of the above transition probability matrix provides a model on the basis of conditional probabilities. For instance, the probability of a transition from 0 at time $t_{j-1}$ to 1 at time $t_j$ is $\pi_{01} = P(Y_j = 1 \mid Y_{j-1} = 0)$ and similarly the probability of a transition from 1 at time $t_{j-1}$ to 1 at time $t_j$ is $\pi_{11} = P(Y_j = 1 \mid Y_{j-1} = 1)$. It is evident that $\pi_{00} + \pi_{01} = 1$, and similarly $\pi_{10} + \pi_{11} = 1$.

The covariate dependent higher order models can be proposed by extending the model for first order Markov chain. To illustrate the extension, a second order Markov model is considered. The second order Markov model for time points $t_{j-2}$, $t_{j-1}$ and $t_j$ with corresponding binary outcomes $Y_{j-2} = s_2$, $Y_{j-1} = s_1$ and $Y_j = s_0$, respectively, is shown as follows:

| $Y_{j-2}$ | $Y_{j-1}$ | $Y_j$ | |
|---|---|---|---|
| | | 0 | 1 |
| 0 | 0 | $\pi_{000}$ | $\pi_{001}$ |
| 0 | 1 | $\pi_{010}$ | $\pi_{011}$ |
| 1 | 0 | $\pi_{100}$ | $\pi_{101}$ |
| 1 | 1 | $\pi_{110}$ | $\pi_{111}$ |

(2.2)

Following the outline of Diggle et al. (2002), the transition probabilities are defined as follows:

$$\pi_{s_2 s_1}(Y_j = 1 \mid Y_{j-2} = s_2, Y_{j-1} = s_1, X) = \frac{e^{\beta' X + s_1 \alpha_1' X + s_2 \alpha_2' X + s_1 s_2 \alpha_3' X}}{1 + e^{\beta' X + s_1 \alpha_1' X + s_2 \alpha_2' X + s_1 s_2 \alpha_3' X}} \quad (2.3)$$

The vector $X$ includes $X_0 = 1$ and p covariates such that $X = [X_0, X_1, ..., X_p]$.

The parameter vectors $\alpha_1$, $\alpha_2$ and $\alpha_3$ are defined as follows:

$$\beta' = [\beta_0, \beta_1, ....., \beta_p]$$
$$\alpha_1' = [\alpha_{10}, \alpha_{11}, ....., \alpha_{1p}]$$
$$\alpha_2' = [\alpha_{20}, \alpha_{21}, ....., \alpha_{2p}]$$
$$\alpha_3' = [\alpha_{30}, \alpha_{31}, ....., \alpha_{3p}]$$

and define

$$\beta_{00}' = \beta' \; \beta_{01}' = \beta' + \alpha_1' \; \beta_{10}' = \beta' + \alpha_2'$$
$$\beta_{11}' = \beta' + \alpha_1' + \alpha_2'$$

Equation 2.3 can be expressed more precisely as follows:

$$\pi_{s_2 s_1}(Y_j = 1 / Y_{j-2} = s_2, Y_{j-1} = s_1, X) = \frac{e^{(\beta' + \sum_{m=1}^{2^2} \lambda_m \alpha_m) X}}{1 + e^{(\beta' + \sum_{m=1}^{2^2} \lambda_m \alpha_m) X}} \quad (2.4)$$

where

$m = 1,2,...,2^2$, $\lambda_1 = 0, \lambda_2 = s_1, \lambda_3 = s_2, \lambda_4 = s_1.s_2$.

The third order Markov model for time points $t_{j-3}$, $t_{j-2}$, $t_{j-1}$ and $t_j$ with corresponding outcomes $Y_{j-3}$=s 3 , $Y_{j-2}$=s 2 , $Y_{j-1}$=s 1 and $Y_j = s_0$, respectively, is shown as follows:

| $Y_{j-3}$ | $Y_{j-2}$ | $Y_{j-1}$ | $Y_j$ | |
|---|---|---|---|---|
| | | | 0 | 1 |
| 0 | 0 | 0 | $\pi_{0000}$ | $\pi_{0001}$ |
| 0 | 0 | 1 | $\pi_{0010}$ | $\pi_{0011}$ |
| 0 | 1 | 0 | $\pi_{0100}$ | $\pi_{0101}$ |
| 0 | 1 | 1 | $\pi_{0110}$ | $\pi_{0111}$ |
| 1 | 0 | 0 | $\pi_{1000}$ | $\pi_{1001}$ |
| 1 | 0 | 1 | $\pi_{1010}$ | $\pi_{1011}$ |
| 1 | 1 | 0 | $\pi_{1100}$ | $\pi_{1101}$ |
| 1 | 1 | 1 | $\pi_{1110}$ | $\pi_{1111}$ |

For a Markov model of order three we can rewrite the transition probability as follows:

$$\pi_{s_3 s_2 s_1}(Y_j = 1 \mid Y_{j-3} = s_3, Y_{j-2} = s_2, Y_{j-1} = s_1, X) \quad (2.5)$$

where

$$\lambda_1 = 0, \lambda_2 = s_1, \lambda_3 = s_2, \lambda_4 = s_1.s_2, \lambda_5 = s_3, \lambda_6 = s_1.s_3,$$
$$\lambda_7 = s_2.s_3, \lambda_8 = s_1.s_2.s_3$$

To generalize this to the k-th order, consider $2^k$ sets of models. The transition probability matrix for the k-th order Markov model can be represented by binary outcomes at different time points $Y_{j-k} = s_k$, $Y_{j-(k-1)} = s_{k-1}, ....., Y_{j-1} = s_1, Y_j = s_0$ at time points $t_{j-k}, t_{j-(k-1)}, ....., t_{j-1}, t_j$, respectively, where $Y_j = 1$ for occurrence of the event and $Y_j = 0$ for non-occurrence of the event at time $t_j$. The transition probability is given by

$$\pi_{s_k \ldots s_1}(Y_j = 1 \mid Y_{j-k} = s_k, \ldots, Y_{j-1} = s_1, X) = \frac{e^{(\beta' + \sum_{m=1}^{2^k} \lambda_m \alpha_m) X}}{1 + e^{(\beta' + \sum_{m=1}^{2^k} \lambda_m \alpha_m) X}} \qquad (2.6)$$

The likelihood function is given by

$$L = \prod_{i=1}^{n_{s_k \ldots s_1}} \prod_{s_k=0}^{1} \cdots \prod_{s_1=1}^{1} \left[ \{\pi_{is_k \ldots s_1}\}^{\delta_{is_k \ldots s_1}} \{1 - \pi_{is_k \ldots s_1}\}^{1-\delta_{is_k \ldots s_1}} \right]$$

where $\delta_{is_k s_{k-1} \cdots s_1} = 1$ if the outcome at time $t_j$ is $Y_j = 1$ for individual $i$ and $\delta_{is_k s_{k-1} \cdots s_1} = 0$ if the outcome at time $t_j$ is $Y_j = 0$ for individual $i$ for the transition type $Y_{j-k} = s_k, Y_{j-(k-1)} = s_{k-1}, \ldots\ldots, Y_{j-1} = s_1$ prior to time $t_j$ and $n_{s_k \ldots s_1}$ denotes the number of subjects experiencing transition type $Y_{j-k} = s_k, Y_{j-(k-1)} = s_{k-1}, \ldots\ldots, Y_{j-1} = s_1$ prior to time $t_j$.

Then the parameters $\beta_{\ell\ s_k s_{k-1} \cdots s_1}$ and $\alpha_{\ell,m}$ can be obtained from the following equations

$$\frac{\partial \ln L}{\partial \beta_{\ell s_k s_{k-1} \cdots s_1}} = 0$$

and

$$\frac{\partial \ln L}{\partial \alpha_{\ell,m}} = 0$$

A Simple Model

In the previous model, the number of parameters increases exponentially with an increase in the order of the dependence, although, the proposed model provides more detailed information for each transition type. Another major limitation of such model is that it requires a large sample size to ensure adequate transitions for each transition type. To address such problems, a simple model is proposed in this section. In the model, the transition probability takes into account selected covariates and previous transitions are also incorporated as covariates for a $k$- order Markov model. The model is as follows:

$$\pi_{s_k s_{k-1} \cdots s_1}(Y_j = 1 \mid Y_{s_k}, Y_{s_k - 1}, \cdots, Y_{s1}, X) = \frac{e^{(\beta X + \theta_1 Y_{s1} + \cdots + \theta_k Y_{s_k})}}{1 + e^{(\beta X + \theta_1 Y_{s1} + \cdots + \theta_k Y_{s_k})}} \qquad (2.7)$$

$$\frac{\partial \ln L}{\partial \beta_\ell} = 0 \quad \ell = 1,2, \cdots, p$$

,

$$\frac{\partial \ln L}{\partial \theta_r} = 0 \quad r = 1,2, \cdots, k$$

,

Testing for the Significance of Parameters

The following vector shows the $2^k$ sets of parameters for the k-th order Markov model:

$$\beta'_G = \left[ \beta'_1, \beta'_2, \ldots\ldots\ldots, \beta'_{2^k} \right]$$

where

$$\beta'_m = \left[ \beta_{m1}, \ldots\ldots\ldots, \beta_{mp} \right], \quad \text{m=1,2,…..,} 2^k$$

To test the null hypothesis $H_0 : \beta = 0$, the usual likelihood ratio test is employed and is given by

$$-2[\ln L(\beta_0) - \ln L(\beta_G)] \sim \chi^2_{2^k p}$$

where

$$\beta'_{G0} = \left[ \beta'_{10}, \beta'_{20}, \ldots, \beta'_{2^k 0} \right]$$

and

$$\beta'_{m0} = \left[ \beta_{m1}, \ldots, \beta_{mp} \right], \quad \text{m=1,2,…..,} 2^k$$

To test the significance of the qth parameter of the m-th set of parameters, the null

hypothesis is $H_0 : \beta_{mq} = 0$ and the corresponding Wald-test is given by

$$W = \frac{\hat{\beta}_{mq}}{se(\hat{\beta}_{mq})}.$$

Test the order of the Markov model on the basis of the simple model (2.7) such that $H_0 : \theta_i = 0$ versus $H_1 : \theta_i \neq 0$ (i=1,2,…,k) that can identify the order is at least i if the null hypothesis is rejected. Use the test procedure discussed above for testing for the order of the Markov model as well.

The computer program employed in this paper is the modified version of the algorithm appeared in Chowdhury et al. (2005) for higher order covariate dependent Markov model.

An Application to Maternal Morbidity Data

Data are used from the survey on Maternal Morbidity in Bangladesh conducted by the Bangladesh Institute for Research for Promotion of Essential and Reproductive Health Technologies (BIRPERHT) during November 1992 to December 1993. The data were collected using both cross-sectional and prospective study designs. The study is based on the data from the prospective component of the survey. A multistage sampling design was used for collecting the data for this study. Districts were selected randomly in the first stage, one district from each Division. Then, Thanas were selected randomly in the second stage, one Thana from each of the selected Districts. A Thana is comprised of several Unions, while Union is the smallest administrative geographical unit in Bangladesh. At the third stage, two Unions were selected randomly from each selected Thana. The subjects comprised of pregnant women with less than 6 months in the selected Unions. The pregnant women from the selected Unions were followed on regular basis (roughly at an interval of one month) throughout the pregnancy. During the follow-up visits, pregnancy complications were recorded.

A total of 1020 pregnant women were interviewed in the follow-up component of the study. The survey collected information on socio-economic and demographic characteristics, pregnancy related care and practice, morbidity during the period of follow-up as well as in the past, information concerning complications at the time of delivery and during the post partum period. For the purpose of this study, 993 pregnant women were selected, with at least one antenatal follow-up. Table 1 shows the number of respondents at different follow-up visits during antenatal period. At the first follow-up 992 respondents were recorded (out of 993 respondents one was missing at the first follow-up but reported subsequently). The number dropped to 917 at the second follow-up and the rate of dropout increased sharply at subsequent follow-ups. The number of respondents observed at the third and the fourth follow-ups were 771 and 594, respectively. The following pregnancy complications are considered under the complications in this study: hemorrhage, edema, excessive vomiting, fits/convulsion. If one or more of these complications occurred to the respondents, they were considered as having complications.

The explanatory variables are: pregnancies prior to the index pregnancy (yes, no), education of respondent (no schooling, some schooling), economic status (low, high), age at marriage (less than 15 years, 15 years or more), involved with gainful employment (no, yes), index pregnancy was wanted or not (no, yes). The number of transitions for the first, second, and third order Markov chains are displayed in Table 2. The estimates of parameters of covariate dependent Markov models are presented in Table 3.

Two variables, economic status and whether the pregnancy was wanted, show significant association with transition from no complication in previous visit to complication in current visit during pregnancy (transition type $0 \rightarrow 1$). If the respondent has economically better status, she is expected to experience higher transition to pregnancy complications. On the other hand, if the index pregnancy is wanted, as compared to that of unwanted pregnancy, there is a decreased risk of transition to pregnancy complications during the current visit.

If the previous outcome was complication, three variables influence to the

transition to the same status at the time of current visit during pregnancy (transition type $1 \to 1$) which are whether the index pregnancy was wanted or not, gainful employment, and education. The desired pregnancies appear to have higher risk of pregnancy complications in consecutive follow-up visits. In other words, undesired pregnancies seem to result in higher risk of transition to complications but risk of complications at consecutive follow-up visits appears to be higher for desired pregnancies. The respondents who are involved with gainful employment have higher risk of transition to complication in consecutive visits during pregnancy but respondents with some education have reduced risk of continued complications in consecutive visits.

The second order model shows that there is a lower risk for desired pregnancies to make transition to the state of complications at current visit after two consecutive no complications status prior to the current visit (transition type $0 \to 0 \to 1$). There is no significant association between reverse transition of the type $1 \to 0 \to 1$ and the selected covariates. Like the transition type $1 \to 1$, transition type $0 \to 1 \to 1$ is observed to be positively associated with desired pregnancy and negatively associated with education. Similarly, similar to $1 \to 1$, desired pregnancy and gainful employment are positively associated with the complications at three consecutive visits (transition type $1 \to 1 \to 1$).

There are eight models for the third order Markov chain. Among those, some of the transition types do not show any clear association with the selected covariates (considered at p-value = 0.05) such as transition types $1 \to 0 \to 0 \to 1$, $0 \to 0 \to 1 \to 1$, $1 \to 1 \to 0 \to 1$, $0 \to 1 \to 1 \to 1$ and $1 \to 1 \to 1 \to 1$. For the transition type, $0 \to 0 \to 0 \to 1$, gainful employment appears to have positive association. There is a marginal positive association (p-value is observed to be little higher than 0.05) between age at marriage and transition type $0 \to 1 \to 0 \to 1$, where the complication is repeated during four follow-ups. Economic status is associated marginally and positively and previous pregnancies are associated negatively with transition type $1 \to 0 \to 1 \to 1$.

The global chi-square and likelihood ratio tests show good fit for all the first, second and third order models. Hence, in order to find the best selection, we have employed the AIC and the BIC procedures. The AIC and the BIC results indicate that the third order models provide the best fit as compared to the first and second order models.

The number of parameters increases geometrically with an increase in the order of Markov model. Hence, a simple alternative is employed to the same data. Table 4 presents the results for the simple model as an alternative to the hierarchical model for the higher order Markov chain. For the second order model, first order outcome is considered as a covariate. Similarly, for the third order model, first and second order outcomes are included as covariates in order to examine the impact of previous outcomes on the subsequent outcomes. In the first model, economic status, wanted pregnancy, age at marriage and education appear to be significantly associated with pregnancy complications. The first order outcome, $s_1$, is included as a covariate for the second order model and confirms that first order outcome exerts a positive influence on the second order outcome. Similarly, first and second order outcomes are also associated positively with the outcome for the third order. Hence, in these models, third order Markov model is expected to fit better. Economic status, gainful employment and occurrence of the complications at previous two follow-ups all are positively associated with current complications. This finding confirms the conclusion based on the results presented in Table 3. In other words, the simple model can be employed confidently if the detailed impact of covariates on the response variable is not needed for each transition type separately for policy purposes.

## Conclusion

In this article, the fitting of higher order covariate dependent Markov model is illustrated.

The method shown here is based on a suggestion provided by Diggle et al. (2002). The inference procedure is described for any higher order Markov model and the proposed method can be employed conveniently to identify the risk factors having significant impact on the repeated binary outcomes of interest at different time points. The proposed technique has been applied to a set of maternal morbidity data and the pregnancy complications at follow-up observations during pregnancy are analyzed. Some selected covariates are used to examine whether the transition probabilities for pregnancy complications at consecutive visits during pregnancy depend on the covariates. A simple alternative is also examined.

If the factors affecting different types of transitions depending on past transitions are not of much interest then we can use the simple alternative. However, the proposed model provides a detailed analysis of the factors affecting transitions of first or higher order Markov models. The detailed analysis may be considered to have useful interpretations for policymakers. On the other hand, the number of parameters in the simple model does not increase geometrically unlike the proposed model.

## Acknowledgments

The authors would like to thank Dr. Halida Hanum Akhter, Director of the Bangladesh Institute of Research for Promotion of Essential and Reproductive Health and Technologies (BIRPERHT) for her permission to use the data employed in this paper. The authors are greatly indebted to the Ford Foundation for funding the data collection of the maternal morbidity study.

## References

Albert, P. S. (1994) A Markov model for sequence of ordinal data from a relapsing-remitting disease. Biometrics 50, 51-60.

Albert, P. S. & Waclawiw, M. A. (1998) A two state Markov chain for heterogeneous transitional data: A quasilikelihood approach. Statistics in Medicine 17, 1481-1493.

Azzalini, A. (1994) Logistic regression for autocorrelated data with application to repeated measures. Biometrika 81, 767-775.

Chowdhury, R. I., Islam, M. A., Shah, M.A. & Al-Enezi,N. (2005). A computer program to estimate the parameters of covariate dependent higher order Markov model. Computer Methods and Program in Biomedicine 77: 175-181.

Diggle, P. J., Heagerty, P. J., Liang, K. Y. & Zeger, S. L. (2002) Analysis of Longitudinal Data (Second Edition). Oxford University Press, Oxford.

Heagerty, P. J. (2002) Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics 58, 342-351.

Heagerty, P. J. & Zeger, S. L. (2000) Marginalized multi-level models and likelihood inference (with Discussion). Statistical Science 15, 1-26.

Islam, M. A., & Chowdhury, R. I. (2006) A higher-order Markov model for analyzing covariate dependence. Applied Mathematical Modelling 30, 477-488.

Korn, E. L. & Whittemore, A. S. (1979) Methods of analyzing panel studies of acute health effects of air pollution. Biometrics 35, 795-802.

Muenz, L. R & Rubinstein, L. V. (1985) Markov models for covariate dependence of binary sequences. Biometrics 41, 91-101.

Prentice, R. & Gloeckler, L. (1978) Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57-67.

Raftery, A. & Tavare, S. (1994) Estimating and modeling repeated patterns in higher order Markov chains with the mixture transition distribution model. Applied Statistics 43(1), 179-199.

Raftery, A. E. (1985) A model for higher order Markov chains. Journal of Royal Statistical Society B 47, 1528-39.

Regier, M. H. (1968) A two state Markov model for behavior change. Journal of the American Statistical Association 63, 993-999.

## Appendix

Table 1: Number of Respondents at Different Follow-ups During Antenatal Period

| Follow-up Number | Frequency |
|---|---|
| 1 | 992 |
| 2 | 917 |
| 3 | 771 |
| 4 | 594 |
| 5 | 370 |
| 6 | 148 |

Table 2: Number of Transitions for Pregnancy Complications

| Transitions | **→0** | **→1** |
|---|---|---|
| First Order | | |
| **0→** | 1577 | 277 |
| **1→** | 366 | 614 |
| | | |
| Second Order | | |
| **0→ 0→** | 923 | 138 |
| **1→ 0→** | 176 | 79 |
| **0→ 1→** | 95 | 79 |
| **1→ 1→** | 110 | 295 |
| | | |
| Third Order | | |
| **0→ 0→0→** | 459 | 72 |
| **1→ 0→0→** | 91 | 30 |
| **0→ 1→0→** | 31 | 16 |
| **0→ 0→1→** | 49 | 27 |
| **1→ 1→0→** | 40 | 28 |
| **1→ 0→1→** | 26 | 20 |
| **0→ 1→1→** | 8 | 23 |
| **1→ 1→1→** | 44 | 132 |

Table 3: Estimates of Parameters of Covariate Dependent Markov Models for Analyzing Pregnancy Complications

| Variables | Estimates | Std. error | t-value | p-value |
|---|---|---|---|---|
| | First Order | | | |
| **0→ 1** | | | | |
| Constant | -1.5233 | 0.1559 | -9.7718 | 0.0000 |
| Economic Status (Good=1) | 0.4209 | 0.1699 | 2.4770 | 0.0132 |
| Wanted pregnancy (Yes=1) | -0.4217 | 0.1371 | -3.0762 | 0.0021 |
| Gainful employment (Yes=1) | 0.0827 | 0.1453 | 0.5693 | 0.5692 |
| Age at marriage (< 15 = 1) | -0.0353 | 0.1394 | -0.2535 | 0.7998 |
| Education (Yes=1) | -0.1161 | 0.1354 | -0.8576 | 0.3911 |
| Previous pregnancies (Yes=1) | 0.0482 | 0.1354 | 0.3561 | 0.7218 |
| **1→ 1** | | | | |
| Constant | 1.9102 | 0.1589 | 12.0175 | 0.0000 |
| Economic Status (Good=1) | 0.1378 | 0.1778 | 0.7749 | 0.4384 |
| Wanted pregnancy (Yes=1) | 0.6904 | 0.1406 | 4.9084 | 0.0000 |
| Gainful employment (Yes=1) | 0.3187 | 0.1506 | 2.1167 | 0.0343 |
| Age at marriage (< 15 = 1) | -0.1535 | 0.1483 | -1.0352 | 0.3006 |
| Education (Yes=1) | -0.5238 | 0.1430 | -3.6635 | 0.0002 |
| Previous pregnancies (Yes=1) | 0.0493 | 0.1411 | 0.3492 | 0.7269 |
| Global Chi-square | 1020.03596; d.f. = 14; p-value=0.00000 | | | |
| LRT | 1126.20664; d.f. = 14; p-value=0.00000 | | | |
| AIC | 2830.55158 | | | |
| BIC | 2898.37897 | | | |
| | Second Order | | | |
| **0→ 0→ 1** | | | | |
| Constant | -1.8003 | 0.2210 | -8.1447 | 0.0000 |
| Economic Status (Good=1) | 0.1830 | 0.2534 | 0.7222 | 0.4702 |
| Wanted pregnancy (Yes=1) | -0.4847 | 0.1919 | -2.5256 | 0.0115 |
| Gainful employment (Yes=1) | 0.2149 | 0.2005 | 1.0715 | 0.2840 |
| Age at marriage (< 15 = 1) | -0.1075 | 0.1983 | -0.5422 | 0.5877 |
| Education (Yes=1) | 0.1420 | 0.1884 | 0.7536 | 0.4511 |
| Previous pregnancies (Yes=1) | 0.1877 | 0.1899 | 0.9886 | 0.3228 |
| **1→ 0→ 1** | | | | |
| Constant | 0.9524 | 0.3091 | 3.0815 | 0.0021 |
| Economic Status (Good=1) | 0.3405 | 0.3523 | 0.9663 | 0.3339 |
| Wanted pregnancy (Yes=1) | 0.4857 | 0.2811 | 1.7283 | 0.0839 |
| Gainful employment (Yes=1) | -0.0526 | 0.3123 | -0.1686 | 0.8661 |
| Age at marriage (< 15 = 1) | 0.2394 | 0.2933 | 0.8163 | 0.4144 |
| Education (Yes=1) | -0.3707 | 0.2845 | -1.3030 | 0.1926 |
| Previous pregnancies (Yes=1) | -0.2887 | 0.2837 | -1.0177 | 0.3088 |
| **0→ 1→ 1** | | | | |
| Constant | 1.7303 | 0.3845 | 4.5000 | 0.0000 |
| Economic Status (Good=1) | 0.4831 | 0.4203 | 1.1494 | 0.2504 |
| Wanted pregnancy (Yes=1) | 0.7516 | 0.3348 | 2.2450 | 0.0248 |
| Gainful employment (Yes=1) | 0.1031 | 0.3511 | 0.2935 | 0.7691 |
| Age at marriage (< 15 = 1) | -0.2899 | 0.3596 | -0.8062 | 0.4201 |
| Education (Yes=1) | -1.0315 | 0.3404 | -3.0301 | 0.0024 |
| Previous pregnancies (Yes=1) | -0.1730 | 0.3359 | -0.5150 | 0.6066 |

**Table 3 Continued…**

| Variables | Estimates | Std. error | t-value | p-value |
|---|---|---|---|---|
| **1→ 1→ 1** | | | | |
| Constant | 2.1680 | 0.2742 | 7.9081 | 0.0000 |
| Economic Status (Good=1) | 0.2688 | 0.2989 | 0.8993 | 0.3685 |
| Wanted pregnancy (Yes=1) | 0.7322 | 0.2414 | 3.0332 | 0.0024 |
| Gainful employment (Yes=1) | 0.6301 | 0.2637 | 2.3891 | 0.0169 |
| Age at marriage (< 15 = 1) | 0.0853 | 0.2512 | 0.3395 | 0.7342 |
| Education (Yes=1) | -0.4259 | 0.2498 | -1.7049 | 0.0882 |
| Previous pregnancies (Yes=1) | 0.3097 | 0.2461 | 1.2581 | 0.2083 |
| Global Chi-square | 736.2494; d.f. = 28; p-value=0.00000 | | | |
| LRT | 819.7761; d.f. = 28; p-value=0.00000 | | | |
| AIC | 1863.2516 | | | |
| BIC | 1998.9065 | | | |
| | Third Order | | | |
| **0→ 0→ 0→ 1** | | | | |
| Constant | -2.0755 | 0.3272 | -6.3440 | 0.0000 |
| Economic Status (Good=1) | 0.3901 | 0.3502 | 1.1140 | 0.2653 |
| Wanted pregnancy (Yes=1) | -0.1337 | 0.2803 | -0.4771 | 0.6333 |
| Gainful employment (Yes=1) | 0.5960 | 0.2780 | 2.1435 | 0.0321 |
| Age at marriage (< 15 = 1) | -0.1715 | 0.2777 | -0.6174 | 0.5370 |
| Education (Yes=1) | -0.1050 | 0.2628 | -0.3997 | 0.6894 |
| Previous pregnancies (Yes=1) | 0.2914 | 0.2716 | 1.0730 | 0.2833 |
| **1→ 0→ 0→ 1** | | | | |
| Constant | 1.1866 | 0.4645 | 2.5547 | 0.0106 |
| Economic Status (Good=1) | -0.4560 | 0.6635 | -0.6874 | 0.4919 |
| Wanted pregnancy (Yes=1) | -0.7254 | 0.4440 | -1.6340 | 0.1023 |
| Gainful employment (Yes=1) | -0.1067 | 0.4893 | -0.2181 | 0.8273 |
| Age at marriage (< 15 = 1) | 0.8524 | 0.4898 | 1.7404 | 0.0818 |
| Education (Yes=1) | -0.4170 | 0.4657 | -0.8954 | 0.3706 |
| Previous pregnancies (Yes=1) | -0.0466 | 0.4411 | -0.1056 | 0.9159 |
| **0→1→ 0→ 1** | | | | |
| Constant | 0.2618 | 0.9645 | 0.2714 | 0.7861 |
| Economic Status (Good=1) | -1.5657 | 0.9903 | -1.5810 | 0.1139 |
| Wanted pregnancy (Yes=1) | 0.6496 | 0.7006 | 0.9271 | 0.3539 |
| Gainful employment (Yes=1) | 0.2875 | 0.7673 | 0.3747 | 0.7079 |
| Age at marriage (< 15 = 1) | 1.4826 | 0.7838 | 1.8916 | 0.0585 |
| Education (Yes=1) | -0.1186 | 0.7317 | -0.1621 | 0.8713 |
| Previous pregnancies (Yes=1) | 0.5271 | 0.7445 | 0.7080 | 0.4790 |
| **0→ 0→ 1→ 1** | | | | |
| Constant | 2.1936 | 0.5945 | 3.6901 | 0.0002 |
| Economic Status (Good=1) | -0.3760 | 0.7314 | -0.5140 | 0.6073 |
| Wanted pregnancy (Yes=1) | -0.4947 | 0.5163 | -0.9581 | 0.3380 |
| Gainful employment (Yes=1) | -0.6332 | 0.5392 | -1.1744 | 0.2402 |
| Age at marriage (< 15 = 1) | -0.3622 | 0.5999 | -0.6037 | 0.5460 |
| Education (Yes=1) | -0.4388 | 0.5363 | -0.8182 | 0.4132 |
| Previous pregnancies (Yes=1) | -0.1100 | 0.5295 | -0.2078 | 0.8354 |
| Gainful employment (Yes=1) | -0.3197 | 0.6776 | -0.4718 | 0.6371 |

**Table 3 Continued…**

| Variables | Estimates | Std. error | t-value | p-value |
|---|---|---|---|---|
| **1→ 1→ 0→ 1** | | | | |
| Constant | 2.1797 | 0.5997 | 3.6347 | 0.0003 |
| Economic Status (Good=1) | 0.8516 | 0.6716 | 1.2680 | 0.2048 |
| Wanted pregnancy (Yes=1) | -0.4881 | 0.5618 | -0.8689 | 0.3849 |
| Age at marriage (< 15 = 1) | -0.0976 | 0.5685 | -0.1718 | 0.8636 |
| Education (Yes=1) | -0.0570 | 0.5623 | -0.1014 | 0.9192 |
| Previous pregnancies (Yes=1) | -1.0848 | 0.5944 | -1.8251 | 0.0680 |
| **1→ 0→ 1→ 1** | | | | |
| Constant | 2.4045 | 0.7779 | 3.0909 | 0.0020 |
| Economic Status (Good=1) | 2.2623 | 1.1836 | 1.9114 | 0.0560 |
| Wanted pregnancy (Yes=1) | -0.2161 | 0.7619 | -0.2837 | 0.7767 |
| Gainful employment (Yes=1) | 0.8775 | 0.8714 | 1.0070 | 0.3139 |
| Age at marriage (< 15 = 1) | -0.5819 | 0.7777 | -0.7482 | 0.4543 |
| Education (Yes=1) | -1.1563 | 0.8015 | -1.4428 | 0.1491 |
| Previous pregnancies (Yes=1) | -1.8188 | 0.9178 | -1.9818 | 0.0475 |
| **0→ 1→ 1→ 1** | | | | |
| Constant | 3.1416 | 1.4192 | 2.2137 | 0.0268 |
| Economic Status (Good=1) | -0.4749 | 1.0472 | -0.4535 | 0.6502 |
| Wanted pregnancy (Yes=1) | -0.6499 | 1.2058 | -0.5390 | 0.5899 |
| Gainful employment (Yes=1) | -0.2974 | 1.0250 | -0.2902 | 0.7717 |
| Age at marriage (< 15 = 1) | 1.0829 | 1.2160 | 0.8905 | 0.3732 |
| Education (Yes=1) | 0.6433 | 1.2520 | 0.5138 | 0.6074 |
| Previous pregnancies (Yes=1) | 0.2593 | 1.0143 | 0.2556 | 0.7983 |
| **1→ 1→ 1→ 1** | | | | |
| Constant | 2.7444 | 0.4633 | 5.9233 | 0.0000 |
| Economic Status (Good=1) | 0.0139 | 0.4792 | 0.0291 | 0.9768 |
| Wanted pregnancy (Yes=1) | -0.1399 | 0.3928 | -0.3562 | 0.7217 |
| Gainful employment (Yes=1) | 0.0439 | 0.4056 | 0.1082 | 0.9138 |
| Age at marriage (< 15 = 1) | -0.2489 | 0.3807 | -0.6538 | 0.5132 |
| Education (Yes=1) | 0.3410 | 0.4179 | 0.8160 | 0.4145 |
| Previous pregnancies (Yes=1) | 0.6496 | 0.4144 | 1.5673 | 0.1170 |
| Global Chi-square | 418.79388; d.f. = 56; p-value=0.00000 | | | |
| LRT | 467.3335; d.f. = 56; p-value=0.00000 | | | |
| AIC | 1164.0451 | | | |
| BIC | 1435.3548 | | | |

Table 4: Estimates of Parameters of Simple Model for Higher Order Markov Chain

| Variables | Estimates | Std. error | t-value | p-value |
|---|---|---|---|---|
| Logistic regression for first order Markov model | | | | |
| Economic status (good=1) | .502 | .091 | 30.467 | .000 |
| Wanted pregnancy (Yes=1) | -.331 | .073 | 20.401 | .000 |
| Gainful employment (Yes=1) | .069 | .076 | .811 | .368 |
| Age at marriage (< 15=1) | -.222 | .076 | 8.662 | .003 |
| Education (Yes=1) | -.408 | .073 | 31.372 | .000 |
| Previous pregnancies (Yes=1) | -.041 | .072 | .317 | .573 |
| Constant | -.293 | .083 | 12.461 | .000 |
| Model Chi-square | | 86.92 (p=0.000) | | |
| Logistic regression for second order Markov model | | | | |
| Economic status (good=1) | .496 | .122 | 16.378 | .000 |
| Wanted pregnancy (Yes=1) | -.080 | .099 | .646 | .422 |
| Gainful employment (Yes=1) | .216 | .102 | 4.447 | .035 |
| Age at marriage (< 15=1) | -.117 | .102 | 1.323 | .250 |
| Education (Yes=1) | -.385 | .098 | 15.269 | .000 |
| Previous pregnancies (Yes=1) | .065 | .097 | .448 | .503 |
| $S_1$ | 2.223 | .094 | 562.140 | .000 |
| Constant | -1.669 | .125 | 177.689 | .000 |
| Model Chi-square | | 704.48 (p=0.000) | | |
| Logistic regression for third order Markov model | | | | |
| Economic status (good=1) | .402 | .155 | 6.768 | .009 |
| Wanted pregnancy (Yes=1) | -.084 | .123 | .469 | .494 |
| Gainful employment (Yes=1) | .361 | .127 | 8.037 | .005 |
| Age at marriage (< 15=1) | -.100 | .126 | .625 | .429 |
| Education (Yes=1) | -.209 | .123 | 2.898 | .089 |
| Previous pregnancies (Yes=1) | .152 | .122 | 1.556 | .212 |
| $S_1$ | 1.720 | .126 | 186.921 | .000 |
| $S_2$ | 1.127 | .125 | 81.547 | .000 |
| Constant | -1.984 | .160 | 152.904 | .000 |
| Model Chi-square | | 521.18 (p=0.000) | | |

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 573-588
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Operating Characteristics Of The DIF MIMIC Approach Using Jöreskog's Covariance Matrix With ML And WLS Estimation For Short Scales

Michaela N. Gelin Bruno D. Zumbo
University of British Columbia

Type I error rate of a structural equation modeling (SEM) approach for investigating differential item functioning (DIF) in short scales was studied. Muthén's SEM model for DIF was examined using a covariance matrix (Jöreskog, 2002). It is conditioned on the latent variable, while testing the effect of the grouping variable over-and-above the underlying latent variable. Thus, it is a multiple-indicators, multiple-causes (MIMIC) DIF model. Type I error rates were determined using data reflective of short scales with ordinal item response formats typically found in the social and behavioral sciences. Results indicate Type I error rates for the DIF MIMIC model, as implemented in LISREL, are inflated for both estimation methods for the design conditions examined.

Key words: Type I error, multiple-causes model for DIF, Monte Carlo simulation.

## Introduction

A variety of statistical methods have been developed over the years to aid the researcher in identifying DIF items for the purposes of (a) fairness and equity in testing, (b) evidence during litigation, (c) investigating whether item properties are changing over time, (d) dealing with a possible "threat to internal validity," and (e) trying to understand the (cognitive and/or psychosocial) processes of item responding and test performance, and investigating whether these processes are the same for different groups of individuals (Shimizu & Zumbo, 2005; Zumbo & Gelin, 2005; Zumbo & Hubley, 2003; Zumbo, 2007).

The statistical methods developed for analyzing DIF have primarily focused on educational ability and achievement tests that are typically quite long (i.e., tests containing many items). As a result, most DIF methods require test s that contain many items (e.g., greater than 30) for the results to be reliable (e.g., Fidalgo, Mellenbergh, & Muñiz, 2000). Measures used in educational, psychological, and more broadly social and health science research (e.g., Rosenberg's Self-Esteem Scale, RSE; Rosenberg, 1965; Center for Epidemiologic Studies Depression Scale, CESD; Radloff, 1977) tend to have relatively fewer items, typically ranging from 3 to 30 items.

Reliability decreases with shorter scales and hence measurement error increases. Observed score DIF methods, such as logistic regression (LogR) or Mantel-Haesnzel (MH) that match on the observed score (e.g., total score or corrected total score often called the rest score), which has measurement error, are of particular concern in short scales because of the lower reliability and error of measurement. A latent variable approach for investigating DIF with short scales is more appropriate compared to an observed score approach because one can condition on the measurement error free latent variable.

A latent variable approach is in line with the formal definition of DIF in which the underlying variable is the conditioning variable. In addition, a latent variable approach is recommended by Zwick (1990), Meredith (1993), Meredith and Millsap (1992), and Millsap and Meredith (1992), who argued that

Michaela N. Gelin is a Research Scientist at CTB/McGraw-Hill in Monterey, California. Bruno D. Zumbo is Professor of Measurement, Evaluation, and Research Methodology, and Department of Statistics. Email him at bruno.zumbo@ubc.ca

observed variable matching DIF methods such as the MH and LogR are not generally diagnostic of item bias. These observed score matching variable DIF methods use the manifest matching variable as a proxy for the latent matching variable and will only be appropriate when the two (manifest and latent) correspond.

This correspondence holds when the observed item responses are consistent with a Rasch (i.e., one-parameter logistic) item response theory model. Under the Rasch model, the observed total score is a sufficient statistic for the latent variable score – assuring the correspondence between the observed and latent matching variables.

Another situation where the observed and latent matching variables correspond is with long scales (a measure or scale with more than 30 items being combined into the composite score) in which all of the items are strong indicators (high factor loadings) of one underlying latent variable, assuming a one-dimensional scale. Shorter scales, containing up to 30 items, do not share this property even though they also may display uni-dimensionality. This rests partly on the notion that in item response theory modeling it is necessary to estimate the latent distribution, and that requires long scales for unbiased estimation and precision. The latent variable approach for investigating DIF in short scales rests on the structural equation modeling (SEM) multiple indicators multiple causes (MIMIC) method (Muthén, 1989).

In this study, Muthén's MIMIC DIF method was implemented using a relatively new covariance matrix available in LISREL for factor models for ordinal variables with covariate effects on the manifest and latent variables (Jöreskog, 2002; Moustaki, Jöreskog, & Mavridis, 2004).

Given that (a) short scales are typically found in the educational and psychological disciplines, (b) the SEM MIMIC method is the most appropriate method for investigating DIF in short scales, and (c) the increasing number of published articles using the MIMIC method suggests this approach is growing in popularity, the purpose of this study is to investigate the statistical properties of a relatively new covariance matrix for the SEM DIF MIMIC method. The proposed MIMIC methodology uses Muthén's (1989) SEM model computed via Jöreskog's covariance matrix. The Type I error rate of this DIF approach have not been investigated. The primary focus of this study is to examine the Type I error rate of the proposed DIF MIMIC approach by means of a simulation study under a variety of study conditions designed to reflect real responses to short scales with ordinal item formats typically found in the social and behavioral sciences.

A statistical test that maintains its Type I error rate is a valid test of the hypothesis. Type I error rates are often referred to as operating characteristics of a test. A Type I error rate, the probability of rejecting $H_0$ when in fact it is true, in detecting DIF refers to declaring an item as DIF when it is not a DIF item. Once the statistical null hypothesis is rejected and the conclusion is reached that an item functions differentially for different groups, further evaluation of the item is necessary in order to determine whether the DIF is attributable to item bias or item impact.

In the context of high stakes testing, for example, making a Type I error may be of great concern because of the matter of test fairness. The Type I error rate is also important in terms of the decisions being made about items flagged as showing DIF. As a result, the empirical Type I error rate of the DIF MIMIC model must be explored. If the Type I error rate is found to be within reason (e.g., 0.05; Bradley, 1978), the power of the DIF MIMIC model needs to be examined (i.e., power is not formally defined unless the statistical test protects the Type I error rate).

## DIF MIMIC model

Although technical descriptions of Muthén's approach can be found, the description below is intended to be less technical with a broader audience of researchers who may be interested in SEM but less familiar with the psychometrics of DIF. The DIF MIMIC model was first proposed by Muthén in 1989. In general, this method conditions on the latent variable while simultaneously testing the effect of group membership (e.g., gender) over-and-above the underlying latent variable of interest. This is a multiple-indicators, multiple-causes

(MIMIC) model which is akin to a latent variable ANCOVA. As Zumbo and Hubley (2003) noted, DIF methods are akin to ANCOVA or attribute-by-treatment interaction (ATI) methodologies.

The MIMIC model was introduced by Jöreskog and Goldberger (1975). It contains one or more latent variables that are simultaneously identified by both multiple endogenous item indicators, which comprise the scale under consideration, and by multiple exogenous causal variables such as background variables of gender or ethnicity. The MIMIC model allows the regression of latent variables on the background variables. Several uses of the MIMIC approach were described by Muthén (1989) and colleagues (e.g., Muthén, Tam, Muthén, Stolzenberg & Hollis, 1993).

One advantage of this approach is that it involves the inclusion of multiple relevant background variables that allow one to study the relative importance of the predictors. Including multiple exogenous variables provides extra information about the measurement, which is particularly useful in detecting population heterogeneity (see Mast & Lichtenberg, 2000) and provides information to help validate scales, permitting the testing of the factor structure of a measure (Zumbo, 2005). The MIMIC approach allows for the detection of item-level measurement non-invariance (i.e., DIF).

Muthén's (1989) modeling approach, the MIMIC model, can be thought of in the context of an example using a 10-item scale, in this case of depression. The MIMIC model consists of three components: (1) a measurement model, (2) a regression model, and (3) a direct effects estimate. Figure 1 is a conceptual, or path, diagram to assist in the description of each of the components of the MIMIC DIF model.

The measurement component refers to the hypothesized relationship between a latent variable and its indicators. The measurement model relates the observed indicators (items) to the continuous latent variable, representing 'depression'. The latent variable is defined for this analysis by the 10 items that form the 10-item scale measuring depression. The relationship between the latent variable and its indicators or factor loadings, which are associated with the endogenous measurement model, are represented by directional arrows that point from the latent conditioning variable to the 10 individual items. The measurement errors for the indicators of the endogenous variables or residuals are set free in this model. Similarly, the measurement errors for the endogenous latent factors are set free.

The regression model relates the latent variable depression to the covariate sex or gender. The effect of the grouping variable, assumed to influence the latent factor, on the underlying latent construct is represented by an arrow from the latent grouping variable, the covariate, to the latent variable depression. This single directional relationship is set free in this model. This is analogous to regression of a continuous outcome variable onto one or more covariates such as gender, marital status, and education level.

The interpretation of the regression coefficient for the grouping variable will depend, of course, on the coding. If, for example, the grouping variable denotes gender such that males are 0 and females are 1, a negative coefficient for the regression of the latent variable, depression, on gender would indicate that females have lower underlying depression than males. The third component is a direct effect estimate that detects measurement invariance in an item response associated with group membership. In other words, adding direct effects from the covariate(s) to the observed indicators, unmediated by the latent factor, incorporates DIF.

It is possible to have a directional arrow pointing from the grouping variable to the individual item being analyzed. This analysis is repeated for each individual item on the scale that one wishes to investigate DIF. More than one item could be tested at a time by specifying more than one direction arrow at a time. This path, or paths for more than one item at a time, represents a systematic difference in responses, controlling for the latent variable.

Having described the DIF MIMIC method there are numerous advantages for using this method: (1) follows the formal definition of DIF, (2) allows for multiple conditioning variables, (3) the combination of covariates (e.g., demographics, attitudes) indirectly represent group membership and hence group

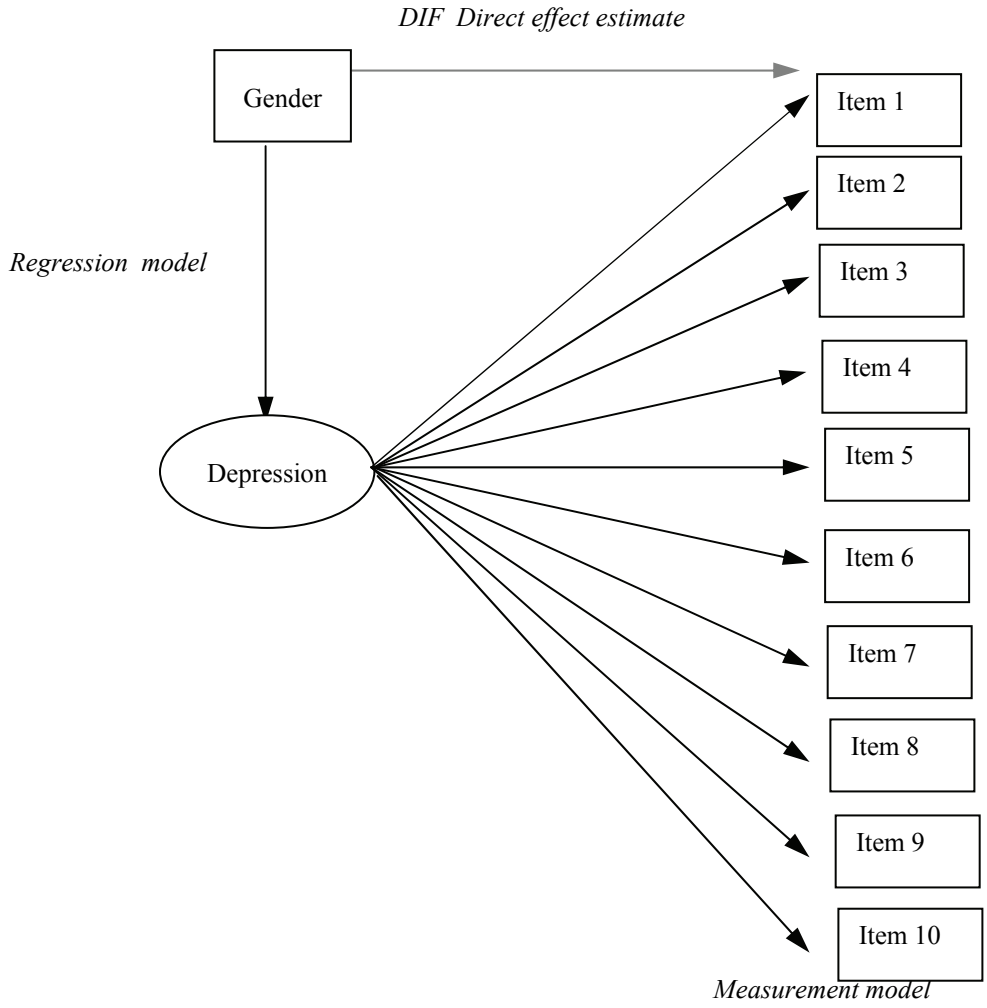



Figure 1. Conceptual (Path) diagram for the DIF MIMIC model for a 10-item scale.

membership does not have to be assigned a priori, (4) can be used with binary, ordinal, and mixed item formats, (5) can be used with multidimensional scales, (6) can model complex data structures involving complex item and test formats (testlets, item bundles, correlated errors), and (7) can be used with short scales. One limitation of this method is that it does not test for interactions (non-uniform DIF); it only investigates uniform DIF. The DIF MIMIC method only examines DIF that is attributable to differences in item difficulty (differences in thresholds). This method assumes the measurement model is the same in both groups (an implicit assumption in GLIM models such as LogR or MH, as well as conditional and unconditional DIF methods, see Zumbo & Hubley, 2003).

A Covariance Matrix for SEM DIF

Recently, Jöreskog (2002) and Moustaki, Jöreskog, and Mavridis (2004) described a new covariance matrix that takes

into consideration that one or more ordinal variables are observed jointly with a covariate(s) (possible explanatory variables). This covariance matrix makes it possible to implement Muthén's MIMIC DIF modeling approach in LISREL. The estimation problem comes down to constructing and estimating the correct covariance matrix of the grouping variable and item response variables for input into the structural equation model. For technical details see Jöreskog (2002) and Moustaki, Jöreskog, and Mavridis (2004). The description below is intended to be less technical with a broader audience of researchers in mind.

In order to understand the advantage of Jöreskog's (2002) covariance matrix, a psychometric problem with be clarified. For ordered discrete response data (ordinal data) the proper correlation measure is a polychoric (tetrachoric if ordered binary) correlation. For metric data (interval or ratio) the proper correlation is a Pearson correlation. It is also known from regression and correlation theory that for truly binary variables (e.g., grouping variables representing a contrast in a design matrix) the Pearson correlation can be used, and this models a difference in means for the continuous dependent or response variables in the model. The construction of a proper covariance matrix becomes a problem when there is a mix of ordinal and continuous data. Figure 2 illustrates this problem, in which items 1 through 3 are 4-point ordered discrete response categories, and the variables age and height are continuous (truly discrete binary variables such as gender are also treated as continuous in the specification of a design matrix representing group differences). The correct correlation between the test items in Figure 2, such as item1 and item2, is a polychoric correlation (ordinal: ordinal). Similarly, the correct correlation between the continuous variables age and height is a Pearson correlation (continuous: continuous). However, the correlation between an ordinal variable (item1) and a continuous variable (age) is problematic because of their different variable formats.

If the data contain mixed variable formats, as is the case shown in Figure 2 between the ordinal and continuous variables, and a Pearson correlation matrix is used, it will treat the ordinal item responses as interval or ratio, resulting in incorrect attenuated correlation values. This type of measurement error caused by using Pearson's correlation with ordinal data, such as Likert-type response formats, has long been debated in the literature (O'Brien, 1979; Bollen & Barb, 1981). As cited by Byrne (1998), Jöreskog and Sörbom (1993) noted that when the observed variables in SEM analyses are either all ordinal or a combination of ordinal and metric scales, the analyses should be *not* be based on Pearson product-moment correlation, but rather be based on either polychoric or polyserial correlations. If a polychoric (or tetrachoric for ordered binary) correlation matrix is used when data are of mixed formats, the continuous variables will be treated as ordinal, which they are not. The resulting correlation values will be incorrect.

Jöreskog's (2002) new method correctly treats the variables according to their variable type (see Figure 2). The ordinal item responses (items 1 through 3 in Figure 2) are correctly treated as ordinal variables, and the age and height variables are correctly treated as continuous covariates. This method allows computing the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables (this is done simultaneously). Given that one or more ordinal item response variables are jointly observed with one or more manifest (observed) variables, such as gender, that can be treated as covariates or predictor variables, one can estimate the effect of the predictor variables on the probability of responding to the ordered categorical (ordinal) variables using either a logistic or probit model. The joint covariance matrix may be computed for the predictor and the variables underlying each of the ordinal variables. This covariance matrix can then be used as input for any structural equation modeling and ML or WLS estimation can be correctly applied.

The statistical test of DIF is examined via (a) the t-statistic of the DIF direct effects coefficient, or (b) a Chi-squared difference test of two models, one with and a second without the DIF direct effects, wherein the nominal alpha of .05 is used in the test for DIF.

| | Item 1 | Item 2 | Item 3 | Age | Height |
|---|---|---|---|---|---|
| Item 1 | | ordinal: ordinal | ordinal: ordinal | ordinal: continuous | ordinal: continuous |
| Item 2 | | | ordinal: ordinal | ordinal: continuous | ordinal: continuous |
| Item 3 | | | | ordinal: continuous | ordinal: continuous |
| Age | | | | | continuous: continuous |
| Height | | | | | |

Figure 2. Example of a correlation matrix with mixed variable formats.

## Methodology

Monte Carlo methods were used to examine the Type I error rates of Muthén's (1989) DIF MIMIC methodology computed via Jöreskog's (2002) covariance matrix with ML and WLS estimation methods. To provide a realistic set of values within the various study design variables described below in the simulation study, real item response data using the 10 and 20 item versions of the Center for Epidemiologic Studies Depression scale (CESD; Radloff, 1977) was used. The CESD is a widely used self-report measure developed for use in studies exploring the epidemiology of depressive symptomology in the general population. Each item is rated on a four-point (0 - 3) Likert-type scale of which a total scale score is computed from the sum of the items. The real response data came from 600 community-dwelling adults living in northern British Columbia (290 females; 310 males) who completed the 20-item CESD scale. The item response data came from the Health and Health Care Survey carried out by the Institute for Social Research and Evaluation in the fall of 1998. The mean age of female participants was 42 years (SD = 13.4, range = 18 to 87 years), and the mean age of male participants was 46 years (SD = 12.1, range = 17 to 82 years). This same item response data was also used to represent the short 10-item CESD scale. See Figure 3 for the specific items that make-up the 20- and 10-item versions.

Data from the CESD scale was chosen because it is a commonly used measure and hence is reflective of measures used in the social and behavioral sciences. Moreover, the scale and item characteristics (unidimensionality, scale length and item format) were representative of data typically found in psychological measures. Specifically, the 10 item short form (CESD-10: Andresen, Malmgren, Carter, Patrick, 1994) and the original 20 item (CESD-20: Radloff, 1977) CESD scales are essentially unidimensional (Clark, Aneshensel, Frerichs & Morgan, 1981; Hertzog, Van Alstine, Usala, Hultsch & Dixon, 1990; Sheehan, Fifield, Reisine & Tennen, 1995; Zumbo, Gelin & Hubley, 2002), supporting the use of a single-factor model with both test lengths for this simulation.

The variables in this simulation study are seven sample size combinations (three equal and four unequal group combinations), two item response distributions (normal/symmetric and positively skewed), two scale lengths (10 and 20 items per scale), and two estimation methods (ML and WLS).

For ease of interpretation, this simulation study is divided into two sub-studies. The first sub-study (Part A) investigates the Type I error rates in which two groups have equal sample sizes (e.g., 200 simulees per

group). The second sub-study (Part B) investigates the Type I error rates in which two groups have unequal sample sizes (200 simulees in one group and 800 simulees in the second group). As a result, the first sub-study (Part A) has a 2 x 2 x 2 x 3 factorial design: two estimation methods by two item response distributions by two scale lengths by three sample size combinations. Similarly, the second sub-study has a 2 x 2 x 2 x 4 factorial design, of which the variables are the same as in Part A except there are four sample size combinations instead of three. Given that the simulation methodology is the same for both sub-studies, only the results section of this simulation study will be divided into the sub-studies.

Study design

Scale length and item format

Consistent with the CESD-10 and CESD-20 scales, data are simulated to represent 10 and 20 item scales, respectively. These two scale lengths are also chosen because they are representative of numerous short scales typically found in the social and behavioral sciences. As found in the CESD scales, all items are simulated to represent ordered categorical data with four categories. This number of rating scale points is also representative of item response formats typically encountered in psychological measures. Ordinal variables are commonly referred to as rating scale, or Likert variables, and thus these terms will be used interchangeably. As in numerous psychological, educational, and behavioral sciences, the ordinal variables used in this study are conceptualized as observed ordered-categorical variables, y, wherein the underlying variable, y*, is completely unobserved (latent) and continuous. As the normally distributed latent variable increases beyond threshold values, the observed variable takes on higher scores, referred to as scale points. Thus, a person endorsing one category has more of a characteristic than if he/she had chosen a lower category, but one does not know how much more.

Item response distribution

Following the simulation study by DiStefano (2002), two distributions are investigated: approximately normally distributed and non-normally distributed. To approximate Likert-type data with four ordered response categories, the generated continuous data are divided using three threshold values.

For the normal (symmetric) distribution, the three equal interval cut points (thresholds) used to categorize the continuous data into four ordered categories are chosen in accordance with the area under the normal curve. For the ordered categories 1 through 4, the item response thresholds (-1.67, 0, and 1.67) corresponded to approximately 5%, 45%, 45%, and 5% of the area under the normal curve. A check on the generated item-level characteristics revealed that the population data (i.e., all items for both the 10 and 20 item scales) are approximately normally distributed for both groups (Skewness approximately 0; Kurtosis approximately -0.2).

To determine the effect of skewness of the item response distribution on the DIF MIMIC method, the generated continuous data are divided into non-normally distributed four-category ordered categorical data with a targeted skewness of 1.7. This skewness level is chosen based on data from the CESD-20 in which skewness values ranged from 0.64 to 3.1, with an average positive skew of 1.7. This type (i.e., positive) and magnitude of skewness is also consistent with item characteristics of other psychological measures (e.g., Golding, 1988; Micceri, 1989; Olsson, 1979) and with other simulation studies (e.g., Babakus, Ferguson & Jöreskog, 1987). To create skewed ordered categorical data, the percentage of responses in each category is approximately 66, 22, 7, and 5 under the normal curve (as determined from real data using the CESD-20) for response categories 1 though 4, respectively (thresholds = 0.4, 1.16, 1.65). A check on the generated item-level data for both the 10 and 20 item scales show skewness and kurtosis values close to the target levels for both groups in the population data (skewness approx. 1.6, kurtosis approx. 1.8).

Sample size combinations

Building on simulation designs in the literature (De Champlain & Gessaroli, 1998; Curran, Bollen, Paxton, Kirby, & Chen, 2002; Muñiz, Hambleton & Xing, 2001; Muthén & Kaplan, 1992), as well as from published

INSTRUCTIONS: Using the scale below, please circle the number for each statement that best describes how often you felt or behaved this way during the past week.

0 = Rarely or none of the time (less than 1 day)
1 = Some or a little of the time (1-2 days)
2 = Occasionally or a moderate amount of time (3-4 days)
3 = Most or all of the time (5-7 days)

| *DURING THE PAST WEEK:* | Less than 1 day | 1-2 days | 3-4 days | 5-7 days | Factor Loadings 10 item scale | Factor Loadings 20 item scale |
|---|---|---|---|---|---|---|
| 1. **I was bothered by things that usually don't bother me.** | 0 | 1 | 2 | 3 | .669 | .698 |
| 2. I did not feel like eating; my appetite was poor. | 0 | 1 | 2 | 3 | -- | .533 |
| 3. I felt that I could not shake off the blues even with help from my family or friends. | 0 | 1 | 2 | 3 | -- | .918 |
| 4. I felt that I was just as good as other people. | 0 | 1 | 2 | 3 | -- | .462 |
| 5. **I had trouble keeping my mind on what I was doing.** | 0 | 1 | 2 | 3 | .744 | .692 |
| 6. **I felt depressed.** | 0 | 1 | 2 | 3 | .857 | .856 |
| 7. **I felt that everything I did was an effort.** | 0 | 1 | 2 | 3 | .743 | .697 |
| 8. **I felt hopeful about the future.** | 0 | 1 | 2 | 3 | .532 | .554 |
| 9. I thought my life had been a failure. | 0 | 1 | 2 | 3 | -- | .751 |
| 10. **I felt fearful.** | 0 | 1 | 2 | 3 | .653 | .658 |
| 11. **My sleep was restless.** | 0 | 1 | 2 | 3 | .597 | .584 |
| 12. **I was happy.** | 0 | 1 | 2 | 3 | .680 | .708 |
| 13. I talked less than usual. | 0 | 1 | 2 | 3 | -- | .671 |
| 14. **I felt lonely.** | 0 | 1 | 2 | 3 | .658 | .713 |
| 15. People were unfriendly. | 0 | 1 | 2 | 3 | -- | .505 |
| 16. I enjoyed life. | 0 | 1 | 2 | 3 | -- | .749 |
| 17. I had crying spells. | 0 | 1 | 2 | 3 | -- | .729 |
| 18. I felt sad. | 0 | 1 | 2 | 3 | -- | .853 |
| 19. I felt that people dislike me. | 0 | 1 | 2 | 3 | -- | .605 |
| 20. **I could not get "going".** | 0 | 1 | 2 | 3 | .775 | .734 |

*Note*: All 20 items are part of the CESD-20, whereas only the **bold** formatted items are part of the CESD-10. For the CESD-20 the items are summed after reverse scoring of items 4, 8, 12, and 16. Total CESD-20 scores range from 0-60, with higher scores indicating higher levels of general depression. For the CESD-10 the items are summed after reverse scoring items 8 and 12.

Figure 3. Center for Epidemiologic Studies Depression Scales: CESD-10 and CESD-20.

literature using the CESD between 2000 and 2004 (PsycINFO search), seven combinations of equal and unequal sample sizes are considered.

The first sub-study investigates the Type I error rates of the DIF MIMIC model when two groups have equal sample sizes. The equal sample size combinations included 1000, 500, and 200 simulees per group. The second sub-study investigates the Type I error rates in when the two groups have unequal sample sizes. For this sub-study, a total sample size of 1000 is used to avoid the problem of confounding the sample size with the per group size. By controlling the total sample size to be 1000 allows for the investigation of whether the Type I error rates are affected by differences in group sizes; if the total sample size was not held constant it would be difficult to distinguish whether or not the Type I error rate was affected by the difference in group sizes or the total sample size. Using a sample size of 1000, four different ratios are considered: 1:9, 2:8, 3:7, and 4:6. These ratios represent the size of Group 1 compared to the size of Group 2. For example, the ratio 1:9 indicates that there are 100 simulees in Group 1 and 900 simulees in Group 2. Overall, these sample size combinations reflect the range of sample sizes used in psychological and educational research (moderate-to-small-scale testing).

Estimation methods

Given that (i) the primary focus of this article is on short scales that are typically found in the educational and psychological disciplines of which often contain ordinal item formats (e.g., 4-point scale) and (ii) DIF often involves a truly binary variables (e.g., gender), Jöreskog's (2002) covariance matrix with ML (which involves the asymptotic covariance matrix and WLS estimation methods will be used. As previously described, Jöreskog's method was chosen because the LISREL software is widely used and it correctly treats the variables according to their variable type thereby allowing one to compute the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables. In turn, this covariance matrix can then be used as input for any structural equation modeling and ML or WLS estimation can be correctly applied.

Procedure / data generation

First, a population covariance matrix, $\Sigma$, as $\Sigma(y^*)_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g$ for two subgroups is created from pre-specified factor loadings. Unlike some simulation studies in which researchers choose factor loadings arbitrarily, the factor loadings (i.e., lambdas) from real data were used to reflect the range of item loadings commonly encountered in practice. Based on the real data described above, the factor loadings for simulating the 10 and 20 item scales are listed in Figure 3. Using the population correlation matrix among the variables, continuous item response data, y*, of a specified population size, with normally distributed but independent (i.e., uncorrelated) continuous scores are generated and saved for each of two groups. A grouping variable is created and saved in the data set. For Group 1 in the equal sample size condition, the specified sample size is 50 000. However, for the unequal sample size conditions, the specified sample sizes for Group 1 are either 10 000, 20 000, 30 000, or 40 000 which correspond to the data with sample size ratios of 1:9, 2:8, 3:7, and 4:6, respectively. For Group 2, the specified sample size is 50 000 for data representing an equal sample size condition. Conversely, the specified sample size for data representing the unequal sample size conditions with ratios of 1:9, 2:8, 3:7, and 4:6 are 90 000, 80 000, 70 000, and 60 000, respectively, for Group 2. These normally distributed scores represent the (typically unobserved) latent scores from which ordered responses are generated.

The generated continuous data are divided into four ordered categories by using three thresholds. Thus, the ordered responses are computed by recoding the continuous item response data into the appropriate thresholds for a 4-point scale: the thresholds for the symmetric data (i.e., equal latent thresholds) are -1.67, 0, and 1.67, and the thresholds for the skewed data (i.e., unequal latent thresholds) are 0.4, 1.16, and 1.65. The continuous scores are manipulated to mimic responses on a rating scale while simultaneously modifying the distributional shape of the data. Lastly, the data from Group 1 is appended to the data from Group 2 to create a population data set with a total of 100,000 simulees for the appropriate design cell.

Type I error is defined as the proportion of times that a null-DIF item was falsely rejected at the 0.05 level. In other words, the empirical Type I error rates are computed as the number of rejections divided by the number of replications. Based on Bradley's (1978) liberal criteria, an empirical Type I error rate exceeding 7.5% (i.e., > 0.075 level of significance) will be considered to be inflated. Bradley's liberal criterion for robustness of validity requires Type I error values of p to lie between 0.025 and 0.075. Note that both the t-test and the Chi-squared tests are investigated. The Chi-square test is a more general (i.e., omnibus) test that can be used to test several items at a time, whereas the t-test (t-value) is a one-degree of freedom test and can therefore only test one item at a time. In this case, however, because there is a large number of degrees of freedom the t-statistic "operates as a z-statistic in testing that the estimate is statistically different from zero" (Byrne, 1998, p. 104).

## Results

### Psychometric properties of the data

Before sampling from the population data files it is important to verify that the simulated data has the desired psychometric properties. A confirmatory factor analysis (CFA) with the polychoric correlation matrix and weighted least squares (WLS) estimation procedure with the asymptotic covariance matrix (Byrne, 1998) was computed using LISREL 8.54 (Jöreskog & Sörbom, 2003b). The goodness-of-fit statistics suggest that both the 10 ($\chi^2(35) = 110.82$, RMSEA = .06) and 20 ($\chi^2(170) = 442.47$, RMSEA = .052) item one-factor models have a reasonable fit to the data.

### Reliability of the data

Different population data sets were created for the equal and unequal sample size conditions. Four population data sets (two levels of the number of items in the scale by two levels of item distributions were created for the equal sample size conditions (Part A). For each of these population data files, the reliability, as computed using alpha, was as follows: the 10-item symmetric data $\alpha = .86$, the 10-item skewed data $\alpha = .85$, the 20-item symmetric data $\alpha = .92$, and the 20-item skewed data $\alpha = .92$. As expected, the longer scales (the 20-item scales) had better reliabilities.

### Monte Carlo

For each of the 1000 replications, the model fit and test statistics (t and $\chi^2$) were recorded. The asymptotic covariance matrix of the estimated coefficients is used for the WLS and ML estimation. More specifically, the computation of WLS takes the inverse of the asymptotic covariance matrix. If this matrix is not positive definite there is no inverse matrix and thus the computation either fails entirely or gives results that are statistically incorrect. This problem is identified by (1) a warning message in the LISREL software output file and (2) an examination of the results where incorrect statistical values are revealed (e.g., negative chi-square values are incorrect because squared values, by definition, must be positive).

There are a few simulation cells in which the first run of the simulation resulted in all of the replications being non-computable, as the results are not interpretable because they are statistically incorrect. For these cases the simulation was re-run, however, the results were the same – the solution was not valid. The solution was not valid because the matrix was not positive definite and therefore the inverse of the asymptotic covariance matrix could not be computed which is needed in order to implement the WLS method for covariance and correlation structures (for a discussion on not positive definite matrices see Wothke, 1993). The computation of ML, on the other hand, does not require the inverse of this matrix. To get ML estimates you maximize the likelihood of the parameters given the data; thus, it does not involve a direct inversion of the asymptotic covariance matrix. Hence, the results using ML, as shown below, were computable.

There are a number of reasons why the asymptotic covariance matrix is "not positive definite." One possible reason could be due to sampling variation. When sample size is small, a sample covariance or correlation matrix may be not positive definite due to mere sampling fluctuation (Anderson & Gerbing, 1984). A second reason could be due to poor parameter values at the start of the iteration process (Byrne,

1998). For example, if the start value is a positive number but the true estimated value is negative, the solution may be unable to continue iterations or may not converge. Thus, it is really a problem when there is a wide discrepancy between the start values and the true estimates. Another explanation "is that the model is empirically underidentified in the sense that the information matrix is nearly singular (i.e., it is close to be nonpositive definite)" (Byrne, 1998, p. 68). Given the problem of a not positive definite matrix, one limitation with this DIF MIMIC approach is that errors are inevitable. One should therefore be cautious and always check that the matrix being analyzed is correct. With this in mind, the following results for the equal sample size condition (Part A) and the unequal sample size condition (Part B) are presented below.

Part A: Equal sample size condition

Model fit

The overall model fit values over the 1000 replications for the DIF MIMIC models with ML estimation method for each cell of the 10- and 20-item scales fits at least adequately. For the 10 and 20-item scales, the RMSEA values are all less than .10 suggesting the data have a good fit to the model.

The mean fit statistics for the DIF MIMIC model conducted with WLS estimation method showed that for the 10-item skewed scale data with a sample size combination of 500:500 the fit values were not computed because the asymptotic covariance matrix was not positive definite. Similarly, the 20-item symmetrical and skewed 200:200 scale data with WLS estimation did not produce any valid data because of the not positive definite matrix. A further discussion of this problem is located at the end of the results section of this article. For the cells that had valid data, the RMSEA values were reasonable (i.e., less than .10). Given that the models fit adequately, the DIF MIMIC model is consistent with our use.

Type I error rates

The DIF MIMIC model was evaluated based on its ability to control Type I error rates under a variety of conditions. For the individual parameters, the chi-square values were examined since there is only one path (direct effects estimate) one is able to also test if the path is equal to zero via the *t* statistic. As expected, the *t* statistic is also inflated and follows the same patterns as the chi-square statistic reported in the results tables.

The chi-square value used for examining the Type I error rate is the difference in chi-squares between the MIMIC model with no group to the item path and the MIMIC model with the group to item path ($\lambda_{12}$ in Figure 1). Using this chi-square value, the proportion of rejections was counted, which represent the Type I error rates, based on the chi-square p-value, with p-values less than 0.05 leading to a decision not to reject the hypothesis. The chi-square rejection rates (Type I error rates) across estimation method, scale length, distributional condition, for the equal sample size combinations are shown in Table 1.

For the symmetrically distributed 10-item data using ML estimation, the Type I error rate was inflated (7.7% - 10.3%) for all three sample size conditions. Similarly, for the skewed 10-item data using ML estimation, the Type I error rate was also inflated (12.5% to 14.8%) for all sample size conditions. Table 1 also shows that the empirical Type I error rates for the symmetrically distributed 20-item data using ML estimation were also inflated (10.8% - 14.7%) for all three sample size conditions. As shown in the same table, the Type I error rates for the skewed 20-item data using ML estimation were even more inflated than the symmetrically distributed data and ranged from 11.6% to 16.3% for all sample size conditions.

In terms of the 10-item scale with WLS estimation (Table 1), the symmetrically distributed data showed inflated Type I error rates ranging from 9.9% to 23.5%. Likewise, the skewed data was also inflated (14.7% to 28.3%). It should also be noted that there were no valid cells for the 10-item scale with skewed data for the 500:500 sample size combination because the matrix was not positive definite.

The 20-item scale using WLS estimation (see Table 1) showed even higher Type I error rates ranging from 24.9% - 46.7%. As one can also see, there were no valid chi-square for the 200:200 sample sizes combinations due to the problem of a non-positive definite matrix.

Table 1. Empirical Type I error rates of the Chi-squared Test of the DIF MIMIC model across estimation method, scale length, distributional condition, for the *equal* sample size combination.

| Estimation method | Scale length | Distribution | | Sample size combinations | | |
|---|---|---|---|---|---|---|
| | | | | 200:200 | 500:500 | 1000:1000 |
| ML | 10-item | Symmetric | *Reject* | .103 | .093 | .077 |
| | | | *Valid reps* | 964 | 995 | 993 |
| | | Skewed | *Reject* | .126 | .148 | .125 |
| | | | *Valid reps* | 957 | 991 | 995 |
| | 20-item | Symmetric | *Reject* | .118 | .108 | .147 |
| | | | *Valid reps* | 626 | 508 | 470 |
| | | Skewed | *Reject* | .162 | .163 | .116 |
| | | | *Valid reps* | 660 | 575 | 481 |
| WLS | 10-item | Symmetric | *Reject* | .235 | .131 | .099 |
| | | | *Valid reps* | 948 | 996 | 997 |
| | | Skewed | *Reject* | .283 | Not | .147 |
| | | | *Valid reps* | 972 | computable | 991 |
| | 20-item | Symmetric | *Reject* | Not | .341 | .249 |
| | | | *Valid reps* | computable | 988 | 993 |
| | | Skewed | *Reject* | Not | .467 | .305 |
| | | | *Valid reps* | computable | 959 | 957 |

'*Valid reps*' is shorthand for the number of valid replications.

Part B: Unequal sample size condition

Model fit

The fit statistics for the DIF MIMIC model conducted with ML estimation suggest that the overall model for each cell of the 10- and 20-item scales fit adequately. For both the scale lengths, the RMSEA values are all <.5 suggesting the data fit the model very well. In addition, the RMSEA fit statistic for the DIF MIMIC models conducted with the WLS estimation also suggest that the data fit the model adequately.

Type I error rates

As in Part A, the chi-square values were examined and used to evaluate the Type I error rates of the DIF MIMIC model under a variety of conditions. The chi-square rejection rates (Type I error rates) for the unequal sample size conditions are shown in Table 2.

For the symmetrically distributed 10-item data using ML estimation, the Type I error rate was inflated (9% - 11.6%) for all four sample size conditions. Likewise, the skewed 10-item data using ML estimation also showed inflated Type I error rates (13.4% to 14.3%) for all sample size conditions.

the MIMIC method is the most appropriate

Table 2. Empirical Type I error rates of the Chi-squared test of the DIF MIMIC model across estimation method, scale length, distributional condition, for the *unequal* sample size combinations.

| Estimation method | Scale length | Distribution | | Sample size combinations 1:9 | 2:8 | 3:7 | 4:6 |
|---|---|---|---|---|---|---|---|
| ML | 10-item | Symmetric | *Reject* | .097 | .116 | .090 | .103 |
| | | | *Valid reps* | 982 | 988 | 996 | 996 |
| | | Skewed | *Reject* | .136 | .134 | .134 | .143 |
| | | | *Valid reps* | 974 | 983 | 991 | 994 |
| | 20-item | Symmetric | *Reject* | .123 | .105 | .098 | .124 |
| | | | *Valid reps* | 528 | 513 | 479 | 467 |
| | | Skewed | *Reject* | .159 | .113 | .143 | .163 |
| | | | *Valid reps* | 536 | 503 | 490 | 491 |
| WLS | 10-item | Symmetric | *Reject* | .115 | .126 | .114 | .125 |
| | | | *Valid reps* | 979 | 994 | 999 | 995 |
| | | Skewed | *Reject* | .138 | .162 | .171 | .178 |
| | | | *Valid reps* | 982 | 995 | 996 | 998 |
| | 20-item | Symmetric | *Reject* | .188 | .211 | .207 | .232 |
| | | | *Valid reps* | 903 | 966 | 998 | 999 |
| | | Skewed | *Reject* | .224 | .259 | .279 | .320 |
| | | | *Valid reps* | 991 | 999 | 999 | 982 |

For the symmetrically distributed 20-item data using ML estimation, the Type I error rate was also moderately inflated (9.8% - 12.4%) for all four sample size conditions. The Type I error rate for the skewed 20-item data using ML estimation was even more inflated than the symmetrically distributed data and ranged from 11.3% to 16.3% for all sample size conditions.

In terms of the 10-item scale with WLS estimation (see Table 2), the symmetrically distributed data showed inflated Type I error rates ranging from 11.4% to 12.5%. Likewise, the skewed data was also inflated (13.8% to 17.8%). The 20-item scale using WLS estimation (see Table 2) showed even higher Type I error rates for both the symmetrically distributed data (18.8% to 23.2%) and the skewed data (22.4% to 32%).

## Discussion

Given that short scales are typically found in the educational and psychological disciplines and method for investigating DIF in short scales, the primary purpose of this article was to investigate the Type I error rates for this DIF method as implemented using Jöreskog's (2002) covariance matrix with ML and WLS estimation methods. As mentioned in the introduction of this article, no previous study had examined the Type I error rates for the DIF MIMIC method let alone its implementation in the LISREL software. Accordingly, the primary focus of this article was to examine the Type I error rate of the proposed MIMIC approach under a variety of study conditions including seven sample size combinations, two item response distributions, two scale lengths, and two estimation methods.

The results of this study clearly show that the DIF MIMIC model has inflated Type I error rates with both the 10- and 20-item scales with ML and WLS estimation methods under all study design conditions. The Type I error rates were more inflated for the skewed data than the symmetric data and the Type I error rates were more inflated for WLS compared to

ML estimation. The results also illustrated that a limitation of the DIF MIMIC method with WLS estimation is that it produced not positive definite asymptotic covariance matrices. As discussed in the results section, the matter of a not positive definite matrix is problematic for WLS estimation (as opposed to ML) because the inverse of the asymptotic covariance matrix is needed in order to implement the method for covariance and correlation structure.

Based on the results from the current study we caution researchers against the use of the DIF MIMIC method with Jöreskog's methods in LISREL. Accordingly, given that this simulation study was motivated by practical contexts wherein the data were reflective of real test data and the design conditions were chosen based on practical contexts, this author recommends avoiding the DIF MIMIC approach currently available in LISREL. Moreover, for studies that have used this DIF MIMIC method (with the new covariance matrix described above), it is likely that too many DIF items were flagged as functioning differently between groups because of the inflated Type I error rate of this method. Thus, for these studies, it is difficult to determine which items are truly functioning differently from those items that are falsely flagged as functioning differently.

## References


Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155-73.

Andresen, E.M., Malmgren, J.A., Carter, W.B., & Patrick, D.L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventative Medicine, 10*, 77-84.

Babakus, E., Ferguson, C.E., & Jöreskog, K. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24*, 222-228.

Bollen, K.A., & Barb, K.H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Byrne, B.M. (1998). *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Clark, V.A., Aneshensel, C.S., Frerichs, R.R., & Morgan, T.M. (1981). Analysis of effects of sex and age in response to items on the CES-D scale. *Psychiatry Research, 5*, 171-181.

Curran, P.J., Bollen, K.A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37,* 1-36.

De Champlain, A., & Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11,* 231-253.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327-346.

Fidalog, A.M., Mellenbergh, G.J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3)*,* 1-11. Retrieved October 25, 2004, from http://www.mpr-online.de

French, A.W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Gallo, J.J., Anthony, J.C., & Muthén, B.O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology, 49,* 251-264.

Golding, J.M. (1988). Gender differences in depressive symptoms. *Psychology of Women Quarterly*, *12*, 61-74.

Hertzog, C., Van Alstine, J., Usala, P.D., Hultsch, D.F., & Dixon, R. (1990). Measurement properties of the Center for Epidemiological Studies Depression scale (CES-

D) in older populations. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 64-72.

Jöreskog, K.G. (2002, June). *Analysis of ordinal variables 5: Covariates*. Retrieved January 6, 2004 from http://www.ssicentral.com/lisrel/column11.htm.

Jöreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631-639.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (2003a). *PRELIS* (Version 2.51) [Computer software]. Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (2003b). *LISREL* (Version 8.54) [Computer software]. Chicago, IL: Scientific Software International.

Mantel, N., & Haenszel, W.M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mast, B.T., & Lichtenburg, P.A. (2000). Assessment of functional abilities among geriatric patients: A MIMIC model of the Functional Independence Measure. *Rehabilitation Psychology, 45*, 94-64.

Meredith, W. (1993). Measurement invariance, factor invariance and factorial invariance. *Psychometrika, 58*, 525-543.

Meredith, W., & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289–311.

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

Millsap, R., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16,* 389–402.

Moustaki, I., Jöreskog, K. G., & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling, 11,* 487-513.

Muñiz, J., Hambleon, R.K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135.

Muthén, B.O. (1989). Using item-specific instructional information in achievement modelling. *Psychometrika, 54*, 385-396.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19-30.

Muthén, B.O., Tam, W.Y., Muthén, L.K., Stolzenberg, R.M., & Hollis, M. (1993). Latent variable modeling in the LISCOMP Framework: Measurement of attitudes toward career choice. In D. Krebs & P. Schmidt (Eds.), *New directions in attitude measurement*, *Festschrift for Karl Schuessler* (pp. 277-290). Berlin, Germany: Walter de Gruyter.

O'Brien, R.M. (1979). The use of Pearson's *r* with ordinal data. *American Sociological Review, 44*, 851-857.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485-500.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 3*, 385-401.

Rigdon, E.E., & Ferguson, C.E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28,* 491-497.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sheehan, T.J., Fifield, J., Reisine, S., & Tennen, H. (1995). The measurement structure of the Center for Epidemiological Studies Depression scale. *Journal of Personality Assessment, 64*, 507-521.

Shimizu, Y., & Zumbo, B. D. (2005). A Logistic Regression for Differential Item Functioning Primer. *Japan Language Testing Association Journal, 7,* 110-124.

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 256-293). Newbury Park, CA: Sage.

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In Brian Everitt and David C. Howell, *Encyclopedia of Behavioral Statistics*, (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Zumbo, B. D., & Gelin, M.N. (2005). A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological / Community Moderated (or Mediated) Test and Item Bias. *Educational Research and Policy Studies, 4, 223-233.*

Zumbo, B.D., Gelin, M.N., & Hubley, A.M. (2002). The construction and use of psychological tests and measures. *Encyclopedia of Life Support Systems*. France: United Nations Educational, Scientific and Cultural Organization Publishing (UNESCO-EOLSS Publishing).

Zumbo, B.D., & Hubley, A.M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 505-509). Thousand Oaks, CA: Sage Press.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 589-595
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# A Simple Method For Finding Empirical Likelihood Type Intervals For The ROC Curve

Ayman Baklizi
Qatar University

Interval estimation of the ROC curve is considered using the empirical likelihood techniques. Suggested is a procedure that is very simple computationally and avoids the constrained optimization problems usually faced with empirical likelihood methods. Various modifications are suggested and the performance of the intervals is evaluated in terms of their coverage probability. The results show tat some of the suggested intervals compete well with other intervals known in the literature.

Key words: ROC curve, empirical likelihood, kernel estimators, bootstrap

## Introduction

The Receiver Operating Characteristic (ROC) curve is used to assess the accuracy of a diagnostic test in discriminating between healthy and diseased individuals. A threshold value $c$ is determined, and people with test measurements greater than $c$ are classified as diseased, otherwise as healthy. Let $X$ be a random variable representing the test score of a healthy individual and let $Y$ be the score of a diseased patients. Let $F$ and $G$ be the distribution functions of $X$ and $Y$ respectively. The sensitivity of the test is defined as $1-G(c)$. It is the probability that the test score of diseased patient is greater than $c$. The specificity of the test is defined as $F(c)$, it is the probability of correctly classifying a healthy individual. The receiver operating characteristic curve is defined as the plot of $1-F(c)$ against $1-G(c)$ as $c$ varies from $-\infty$ to $\infty$ or equivalently as the plot of $1-G\left(F^{-1}(1-t)\right)$ where $0 \le t \le 1$, (Hsieh and Turnbull, 1996).

Ayman Baklizi is an Associate Professor of statistics, and is at the department of mathematics and physics at Qatar University. He is an elected-member at the ISI. Email: baklizi1@yahoo.com

The estimation of the ROC curve has received considerable attention. The problem has been considered in parametric, nonparametric and semi-parametric situations. For example, see Hsieh and Turnbull (1996), Li et al. (1999), Hall et al., (2003).

Claeskens et al. (2003) developed empirical likelihood confidence regions for the ROC curve. Let $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ be two random samples from the distributions $F$ and $G$ respectively. Define the ROC curve as $R(t)=1-G\left(F^{-1}(1-t)\right)$ where $0 \le t \le 1$ and let $\theta = R(t)$, Claeskens et al. (2003) constructed confidence intervals for $\theta$ using the smoothed empirical likelihood function

$$L(\theta) = \sup_{(p,q,\eta)} \left( \prod_{i=1}^{n} p_i \right) \left( \prod_{j=1}^{m} q_j \right),$$

where $p' = (p_1,\ldots,p_n)$ and $q' = (q_1,\ldots,q_m)$ are probability vectors each summing to one and subject to certain constraints on the smoothed versions of the empirical distributions of $X$ and $Y$. They showed that the asymptotic distribution of the log-likelihood ratio $l(\theta) = -2\log L(\theta)$ is chi square with one degree of freedom and conducted some simulations to investigate the performance of their intervals and show that it performs better than some other asymptotic and bootstrap intervals.

## Purpose

An alternative procedure is suggested here based on the empirical likelihood which is very simple computationally, does not need numerical constrained optimization, and produces interval estimates that are, in some cases, about as accurate as those of Claeskens et al. (2003). This procedure and some modifications are described. A simulation experiment was conducted to investigate and compare the suggested procedure with other well known procedures.

## Empirical Likelihood Based Intervals

Assume that an interval estimator of $R(t^*) = 1 - G(F^{-1}(1-t^*))$ is desired where $t^*$ is some specific point in the unit interval. Proceed in two stages as follows; in the first stage obtain a point estimator for $F^{-1}(1-t^*)$. This is equivalent to estimating $x_{1-t^*}$: $(1-t^*)^{th}$ quantile of $F$ denote this estimator by $\hat{x}_{1-t^*}$. In the second stage obtain an interval estimator of $\overline{G}(\hat{x}_{1-t^*}) = 1 - G(\hat{x}_{1-t^*})$ which is the right tail probability of the random variable $Y$ having distribution function $G$.

In an empirical likelihood setup, the first stage amounts to estimating $x_{1-t^*}$ which may be done using interpolation between the values of the ordered statistics of the sample of the distribution of $X$. In the second stage consider the empirical likelihood function for quantiles (Owen, 2001) given by

$$R(p,q) = \max\left\{\prod_{i=1}^{m} n w_i \mid \sum_{i=0}^{m+1} w_i Z_i(p,q) = 0, w_i \geq 0, \sum_{i=0}^{m+1} w_i = 1\right\}$$

where $0 \leq p \leq 1$, $-\infty < q < \infty$ is the $p^{th}$ quantile and $Z_i(p,q) = I_{(X_i \leq q)} - p$. Substituting $\hat{x}_{1-t^*}$ for $q$ and $G(\hat{x}_{1-t^*})$ for $p$ and, conditional on $\hat{x}_{1-t^*}$ using the empirical likelihood function $R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*})$ one can construct confidence interval for $G(\hat{x}_{1-t^*})$ as

$$\left\{G(\hat{x}_{1-t^*}) \mid -2\log R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > \chi^2_{\alpha,1}\right\}$$

and then transform it to a confidence interval for $1 - G(\hat{x}_{1-t^*})$ this results in a confidence interval for $R(t^*)$ Call this interval the (EL) interval.

The chi square calibration used in the empirical likelihood interval may be replaced by the E-Calibration of Tsao (2004). This calibration is based on the quantiles of a new family of distributions arising from the normal distribution. It is derived using the finite sample similarity between the empirical and parametric likelihoods. Some quantiles $e_{\alpha,1,m}$ of that distribution are given in Tsao (2004). The E – calibration corrects for under coverage resulting from using the chi square calibration. The new interval (EC interval) based on this calibration is given by

$$\left\{G(\hat{x}_{1-t^*}) \mid -2\log R(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > e_{\alpha,1,m}\right\}$$

Another modification may be obtained by using the "smoother" version of the empirical likelihood function for quantiles introduced by Adimari (1998). In this modification the empirical likelihood is replaced by a smoother version which, when considered as a function of $G(\hat{x}_{1-t^*})$, may be written as

$$-2\log \tilde{R}\left(G\left(\hat{x}_{1-t^*}\right), \hat{x}_{1-t^*}\right) = 2m\left[\tilde{G}\left(\hat{x}_{1-t^*}\right)\log\left(\frac{\tilde{G}\left(\hat{x}_{1-t^*}\right)}{G\left(\hat{x}_{1-t^*}\right)}\right) + \left(1-\tilde{G}\left(\hat{x}_{1-t^*}\right)\log\left(\frac{1-\tilde{G}\left(\hat{x}_{1-t^*}\right)}{1-G\left(\hat{x}_{1-t^*}\right)}\right)\right)\right]$$

where

$$\tilde{G}(\hat{x}_{1-t^*}) = \begin{cases} G^*(\hat{x}_{1-t^*}) & \text{if } [Y_{(1)}, Y_{(m)}] \text{contains } \hat{x}_{1-t^*} \\ \hat{G}(\hat{x}_{1-t^*}) & \text{otherwise} \end{cases}$$

where $G^* = \dfrac{2i-1}{2n}$ on each $Y_{(i)}$ and is linear in each $[Y_{(i)}, Y_{(i+1)}]$, and where $Y_{(1)}, \ldots, Y_{(n)}$ are the

order statistics of the sample of $Y$ values. Adimari showed that the limiting distribution is also $\chi_1^2$. A $(1-\alpha)\%$ confidence interval for $G(\hat{x}_{1-t^*})$ (AD interval) is given by

$$\{G(\hat{x}_{1-t^*}) | -2\log \tilde{R}(G(\hat{x}_{1-t^*}), \hat{x}_{1-t^*}) > \chi^2_{1,\alpha}\}$$

Simulation

Simulation studies were conducted to assess the performance of the interval estimates based in the empirical likelihood. Also considered were the bootstrapped version of the empirical likelihood interval (BEL), and the bootstrapped version of the (AD) interval, the (BTAD) interval. A Bartlett type correction factor is obtained as the mean of the B bootstrap empirical log-likelihood ratios which in turn used to find the (BRT) interval. The simulation design used similar to those used by Claeskens et al.(2003) and Hall et. al. (2003). The coverage probability were investigated at values of $t$ = 0.1, 0.3, 0.5, 0.7 and 0.9 with sample size $(n,m) = (30,30)$, (50,50), (70, 70), (100,100), (50,70) and (70,50). In each case 2000 pair of samples is generated from

$$1\text{-}\ X \sim N(0,1),\ Y \sim N(1,1)$$

$$2\text{-}\ X \sim \Gamma(2),\ Y \sim \Gamma(3)$$

$$3\text{-}\ X \sim t(5),$$

$$Y \sim 0.2(t(5)-1) + 0.8(t(5)-1)$$

B = 500 is used in bootstrap calculations. The coverage probabilities of the intervals with nominal confidence levels $(1-\alpha) = 0.90$ and 0.95 are given in Tables 1-3.

Result

The results are given in tables 1 – 3 where the following abbreviations are used EL: The empirical likelihood interval based on the asymptotic $\chi^2$ approximation. BEL: The empirical likelihood interval based on bootstrap cririca values.

EC: The empirical likelihood interval based on Tsao's E-Calibration. BRT: The empirical likelihood interval with the bootstrap Bartlett type correction. AD: The empirical likelihood interval based on Adimari's modification.

BTAD: The empirical likelihood interval based on Adimari's modification and bootstrap critical values.

Conclusion

It appears that the coverage probabilities of the intervals are close to the nominals for small values of $t$. For larger values of $t$ most intervals tend to have an undercoverage problem. Exceptions are the bootstrapped empirical likelihood interval (BEL) and the corrected interval (BRT). These two intervals tend to be conservative for larger values of $t$. A drawback of the (BEL) interval is that it has a very low coverage probability for small values of $t$ when the sample sizes differ.

This is not the case with the (BRT) interval. These observations are also applicable to the results given in tables 2 and 3. The BRT in most cases have the closest coverage probability to nominal. Comparison of these results with Hall et. al. (2003) and Claeskens et al. (2003) shows that the (BRT) interval considered in this article competes very well with theirs in terms of its coverage probability. The simplicity of the methods discussed in this article and the avoidance of complicated restricted optimization problems or sophisticated bandwidth rules used for the construction of kernel based intervals may balance the slightly better performance of the Hall et al. (2003) or Claeskens et. al. (2003) intervals.

References

Adimari, G. (1998). An empirical likelihood statistic for quantiles. *Journal of Statistical Computation and Simulation*, *60*, 85 – 95.

Hall, P., Hyndman, R. & Fan, Y. (2003). Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, *91*, 3, 743 – 750.

Claeskens, G., Jing, B. Peng, L., & Zhou, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics*, *31*, 173 – 190.

Table 1. Coverage Probabilities of the Intervals, The Normal Distribution

| $n$ | $m$ | $t$ | EL | BEL | EC | BRT | AD | BTAD | EL | BEL | EC | BRT | AD | BTAD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | | $\alpha = 0.05$ | | | |
| 30 | 30 | 0.1 | 887 | 182 | 887 | 791 | 926 | 932 | 887 | 228 | 967 | 900 | 951 | 961 |
| | | 0.3 | 743 | 741 | 743 | 779 | 761 | 768 | 794 | 791 | 946 | 803 | 789 | 947 |
| | | 0.5 | 816 | 852 | 816 | 872 | 805 | 817 | 864 | 903 | 864 | 906 | 877 | 899 |
| | | 0.7 | 719 | 881 | 743 | 903 | 767 | 789 | 840 | 928 | 840 | 931 | 848 | 863 |
| | | 0.9 | 600 | 839 | 600 | 876 | 676 | 695 | 623 | 921 | 705 | 917 | 768 | 788 |
| 50 | 50 | 0.1 | 927 | 378 | 927 | 878 | 920 | 925 | 927 | 411 | 927 | 917 | 951 | 954 |
| | | 0.3 | 852 | 809 | 852 | 824 | 848 | 855 | 905 | 864 | 905 | 872 | 905 | 906 |
| | | 0.5 | 757 | 906 | 819 | 926 | 803 | 811 | 868 | 951 | 868 | 953 | 874 | 884 |
| | | 0.7 | 729 | 914 | 729 | 950 | 746 | 756 | 811 | 965 | 811 | 966 | 832 | 841 |
| | | 0.9 | 651 | 890 | 651 | 948 | 692 | 700 | 731 | 965 | 731 | 971 | 783 | 791 |
| 70 | 70 | 0.1 | 905 | 531 | 905 | 927 | 936 | 937 | 966 | 561 | 966 | 956 | 960 | 961 |
| | | 0.3 | 855 | 874 | 855 | 890 | 836 | 840 | 855 | 909 | 892 | 919 | 922 | 925 |
| | | 0.5 | 783 | 924 | 783 | 952 | 790 | 797 | 860 | 961 | 860 | 971 | 866 | 870 |
| | | 0.7 | 756 | 932 | 756 | 971 | 761 | 765 | 838 | 978 | 838 | 984 | 838 | 844 |
| | | 0.9 | 662 | 907 | 662 | 972 | 692 | 694 | 732 | 974 | 732 | 984 | 773 | 777 |
| 100 | 100 | 0.1 | 941 | 658 | 941 | 945 | 938 | 938 | 941 | 677 | 941 | 970 | 967 | 967 |
| | | 0.3 | 830 | 895 | 830 | 918 | 833 | 833 | 880 | 940 | 880 | 941 | 906 | 906 |
| | | 0.5 | 800 | 946 | 800 | 978 | 803 | 803 | 868 | 978 | 868 | 986 | 882 | 882 |
| | | 0.7 | 719 | 945 | 719 | 986 | 743 | 743 | 832 | 988 | 832 | 992 | 832 | 832 |
| | | 0.9 | 646 | 929 | 646 | 980 | 661 | 661 | 722 | 980 | 722 | 991 | 752 | 752 |
| 50 | 70 | 0.1 | 878 | 506 | 878 | 901 | 918 | 922 | 947 | 540 | 947 | 939 | 946 | 948 |
| | | 0.3 | 815 | 847 | 815 | 867 | 812 | 813 | 815 | 889 | 859 | 899 | 902 | 907 |
| | | 0.5 | 747 | 910 | 747 | 943 | 763 | 768 | 833 | 959 | 833 | 959 | 836 | 842 |
| | | 0.7 | 708 | 911 | 708 | 958 | 715 | 724 | 791 | 973 | 791 | 978 | 800 | 800 |
| | | 0.9 | 587 | 885 | 587 | 946 | 610 | 620 | 646 | 962 | 646 | 971 | 694 | 701 |
| 70 | 50 | 0.1 | 958 | 397 | 958 | 924 | 952 | 954 | 958 | 435 | 958 | 953 | 972 | 974 |
| | | 0.3 | 869 | 815 | 869 | 823 | 864 | 870 | 912 | 869 | 912 | 875 | 907 | 911 |
| | | 0.5 | 780 | 909 | 845 | 925 | 823 | 835 | 881 | 954 | 881 | 951 | 890 | 899 |
| | | 0.7 | 791 | 928 | 791 | 962 | 789 | 797 | 871 | 976 | 871 | 979 | 867 | 874 |
| | | 0.9 | 710 | 910 | 710 | 957 | 735 | 743 | 786 | 965 | 786 | 976 | 812 | 818 |

Table 2. Coverage Probabilities of the Intervals, Asymmetric Distributions Case

| $n$ | $m$ | $t$ | EL | BEL | EC | BRT | AD | BTAD | EL | BEL | EC | BRT | AD | BTAD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | | $\alpha = 0.05$ | | | |
| 30 | 30 | 0.1 | 922 | 292 | 922 | 844 | 919 | 925 | 922 | 346 | 922 | 884 | 950 | 955 |
| | | 0.3 | 761 | 767 | 840 | 779 | 822 | 833 | 840 | 828 | 840 | 825 | 867 | 878 |
| | | 0.5 | 769 | 871 | 769 | 892 | 754 | 769 | 820 | 918 | 868 | 920 | 844 | 863 |
| | | 0.7 | 704 | 878 | 704 | 913 | 725 | 748 | 735 | 933 | 813 | 937 | 812 | 834 |
| | | 0.9 | 561 | 828 | 561 | 853 | 698 | 711 | 649 | 905 | 649 | 895 | 776 | 802 |
| 50 | 50 | 0.1 | 857 | 516 | 857 | 885 | 914 | 918 | 943 | 546 | 943 | 932 | 950 | 951 |
| | | 0.3 | 796 | 840 | 796 | 874 | 807 | 816 | 850 | 898 | 850 | 899 | 875 | 881 |
| | | 0.5 | 749 | 905 | 749 | 949 | 764 | 775 | 842 | 967 | 842 | 964 | 841 | 850 |
| | | 0.7 | 691 | 928 | 691 | 959 | 741 | 749 | 767 | 970 | 780 | 978 | 820 | 826 |
| | | 0.9 | 601 | 881 | 675 | 920 | 713 | 721 | 740 | 945 | 757 | 949 | 787 | 796 |
| 70 | 70 | 0.1 | 906 | 648 | 906 | 921 | 912 | 915 | 906 | 681 | 906 | 950 | 946 | 949 |
| | | 0.3 | 769 | 878 | 769 | 904 | 790 | 799 | 873 | 926 | 873 | 928 | 874 | 884 |
| | | 0.5 | 723 | 929 | 762 | 973 | 760 | 765 | 804 | 980 | 832 | 986 | 840 | 849 |
| | | 0.7 | 689 | 934 | 736 | 978 | 732 | 737 | 820 | 979 | 820 | 990 | 810 | 816 |
| | | 0.9 | 638 | 900 | 638 | 947 | 677 | 684 | 723 | 965 | 739 | 965 | 765 | 770 |
| 100 | 100 | 0.1 | 733 | 703 | 733 | 752 | 729 | 729 | 923 | 762 | 923 | 969 | 948 | 948 |
| | | 0.3 | 805 | 924 | 805 | 952 | 806 | 806 | 875 | 961 | 875 | 967 | 875 | 875 |
| | | 0.5 | 758 | 943 | 758 | 984 | 761 | 761 | 850 | 984 | 850 | 991 | 839 | 839 |
| | | 0.7 | 729 | 945 | 729 | 988 | 715 | 715 | 774 | 988 | 774 | 994 | 807 | 807 |
| | | 0.9 | 631 | 926 | 631 | 971 | 680 | 680 | 732 | 978 | 732 | 985 | 768 | 767 |
| 50 | 70 | 0.1 | 874 | 659 | 874 | 895 | 900 | 901 | 874 | 687 | 874 | 924 | 939 | 943 |
| | | 0.3 | 738 | 872 | 738 | 898 | 761 | 766 | 844 | 924 | 844 | 927 | 837 | 841 |
| | | 0.5 | 690 | 920 | 727 | 958 | 726 | 732 | 769 | 967 | 805 | 976 | 814 | 819 |
| | | 0.7 | 647 | 924 | 697 | 971 | 711 | 717 | 772 | 973 | 772 | 981 | 794 | 799 |
| | | 0.9 | 566 | 860 | 566 | 922 | 641 | 643 | 632 | 940 | 645 | 949 | 719 | 728 |
| 70 | 50 | 0.1 | 880 | 532 | 880 | 898 | 927 | 930 | 960 | 566 | 960 | 938 | 958 | 960 |
| | | 0.3 | 824 | 854 | 824 | 866 | 817 | 824 | 870 | 894 | 870 | 898 | 887 | 898 |
| | | 0.5 | 795 | 921 | 795 | 956 | 797 | 806 | 876 | 961 | 876 | 968 | 875 | 885 |
| | | 0.7 | 746 | 939 | 746 | 980 | 784 | 792 | 830 | 981 | 843 | 988 | 866 | 874 |
| | | 0.9 | 662 | 899 | 729 | 937 | 737 | 747 | 789 | 957 | 811 | 961 | 818 | 827 |

Table 3: Coverage Probabilities of the Intervals, Mixture Distributions Case

| $n$ | $m$ | $t$ | EL | BEL | EC | BRT | AD | BTAD | EL | BEL | EC | BRT | AD | BTAD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | | $\alpha = 0.05$ | | | |
| 30 | 30 | 0.1 | 808 | 792 | 808 | 804 | 844 | 852 | 861 | 850 | 861 | 844 | 873 | 879 |
| | | 0.3 | 820 | 899 | 877 | 921 | 846 | 858 | 877 | 938 | 877 | 949 | 912 | 930 |
| | | 0.5 | 828 | 927 | 828 | 956 | 854 | 874 | 914 | 964 | 914 | 972 | 916 | 928 |
| | | 0.7 | 755 | 903 | 755 | 937 | 792 | 809 | 825 | 960 | 842 | 961 | 864 | 884 |
| | | 0.9 | 590 | 824 | 590 | 851 | 678 | 692 | 699 | 906 | 699 | 906 | 764 | 783 |
| 50 | 50 | 0.1 | 838 | 866 | 838 | 886 | 826 | 834 | 838 | 902 | 838 | 917 | 907 | 913 |
| | | 0.3 | 863 | 934 | 863 | 962 | 842 | 848 | 895 | 968 | 927 | 971 | 907 | 914 |
| | | 0.5 | 811 | 952 | 851 | 984 | 846 | 853 | 911 | 985 | 911 | 990 | 909 | 916 |
| | | 0.7 | 813 | 937 | 813 | 977 | 798 | 807 | 879 | 978 | 879 | 986 | 869 | 876 |
| | | 0.9 | 595 | 880 | 595 | 919 | 668 | 679 | 685 | 946 | 685 | 951 | 749 | 760 |
| 70 | 70 | 0.1 | 828 | 887 | 828 | 906 | 815 | 820 | 867 | 928 | 867 | 939 | 891 | 895 |
| | | 0.3 | 841 | 951 | 881 | 984 | 852 | 857 | 908 | 982 | 908 | 988 | 917 | 921 |
| | | 0.5 | 837 | 956 | 837 | 988 | 830 | 836 | 893 | 987 | 893 | 994 | 901 | 905 |
| | | 0.7 | 807 | 962 | 807 | 990 | 803 | 808 | 872 | 991 | 885 | 995 | 880 | 882 |
| | | 0.9 | 604 | 907 | 662 | 957 | 663 | 672 | 734 | 968 | 734 | 974 | 752 | 762 |
| 100 | 100 | 0.1 | 787 | 919 | 787 | 941 | 811 | 811 | 873 | 959 | 873 | 959 | 889 | 889 |
| | | 0.3 | 843 | 965 | 843 | 991 | 853 | 853 | 924 | 989 | 924 | 997 | 923 | 923 |
| | | 0.5 | 834 | 970 | 834 | 996 | 843 | 843 | 899 | 995 | 899 | 998 | 907 | 907 |
| | | 0.7 | 775 | 967 | 775 | 996 | 799 | 799 | 873 | 993 | 873 | 998 | 868 | 868 |
| | | 0.9 | 634 | 908 | 634 | 968 | 662 | 662 | 686 | 977 | 686 | 986 | 746 | 746 |
| 50 | 70 | 0.1 | 780 | 858 | 780 | 887 | 781 | 786 | 827 | 908 | 827 | 922 | 858 | 861 |
| | | 0.3 | 832 | 949 | 871 | 977 | 840 | 846 | 894 | 981 | 894 | 986 | 900 | 904 |
| | | 0.5 | 820 | 964 | 820 | 992 | 810 | 815 | 875 | 990 | 875 | 997 | 882 | 886 |
| | | 0.7 | 766 | 943 | 766 | 979 | 760 | 767 | 819 | 981 | 836 | 990 | 841 | 846 |
| | | 0.9 | 521 | 870 | 570 | 929 | 600 | 605 | 638 | 946 | 638 | 956 | 696 | 702 |
| 70 | 50 | 0.1 | 859 | 869 | 859 | 884 | 839 | 851 | 859 | 903 | 859 | 911 | 918 | 924 |
| | | 0.3 | 886 | 940 | 886 | 965 | 867 | 871 | 912 | 975 | 946 | 979 | 931 | 938 |
| | | 0.5 | 818 | 962 | 853 | 980 | 847 | 853 | 920 | 985 | 920 | 990 | 915 | 922 |
| | | 0.7 | 834 | 942 | 834 | 983 | 824 | 835 | 907 | 983 | 907 | 990 | 897 | 904 |
| | | 0.9 | 656 | 889 | 656 | 949 | 715 | 724 | 733 | 958 | 733 | 971 | 793 | 801 |

Hsieh, F., & Turnbull, B. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, *24*, *1*, 25 – 40.

Li, G., Tiwari, R., & Wells, M. (1999). Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curve. *Biometrika, 86, 3*, 487 – 502.

Owen, A. B. (2001). *Empirical Likelihood*. NY: Chapman and Hall.

Tsao, M. (2004). A new method of calibration for the empirical loglikelihood ratio. *Statistics and Probability Letters*, *68*, 305 – 314.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 596-607
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# A Modified $\bar{X}$ Control Chart for Samples Drawn from Finite Populations

Michael B. C. Khoo
Universiti Sains Malaysia

The $\bar{X}$ chart works well under the assumption of random sampling from infinite populations. However, many process monitoring scenarios may consist of random sampling from finite populations. A modified $\bar{X}$ chart is proposed in this article to solve the problems encountered by the standard $\bar{X}$ chart when samples are drawn from finite populations.

Key words: $\bar{X}$ control chart, finite population, infinite population, average run length (ARL), in-control, out-of-control (o.o.c.), upper control limit (*UCL*), lower control limit (*LCL*).

## Introduction

The Shewhart $\bar{X}$ control chart is widely used in manufacturing industries to monitor the stability of the mean of a process. Since its introduction in the late 1920's, numerous extensions and enhancements of the $\bar{X}$ chart have been suggested.

Nelson (1984) discussed eight types of runs rules which increase the sensitivity of the $\bar{X}$ chart for the detection of a shift in the mean of a normally distributed process. Wheeler (1983) provided tables of the power function of the $\bar{X}$ chart and the Type-I error probabilities of each of the four different sets of detection rules. False signal rates of the $\bar{X}$ chart incorporating each of the eight different runs rules are studied by Walker et al. (1991). Seven of these rules are discussed in Nelson (1984). Klein (2000) proposed two different runs rules for the chart, the 2-of-2 and 2-of-3 rules, based on a Markov chain approach in setting the limits of the chart. Using the same Markov chain approach. Khoo (2004a) extended the work of Klein (2000) by suggesting three additional rules, i.e., the 2-of-4, 3-of-3 and 3-of-4 rules.

Superior alternatives to the two rules of Klein (2000) are proposed by Khoo and Khotrun (2006) to enable a quicker detection of a big shift, while maintaining the same sensitivity towards a small shift. Shmueli and Cohen (2003) introduced a new method for computing the run length distribution of a Shewhart chart with runs and scans rules. Davis and Krehbiel (2002) compared the ARL performances of Shewhart charts with all possible combinations of supplementary runs rules and that of zone charts and found the latter to outperform the former.

The first optimum economic design of the $\bar{X}$ chart which considered statistical and cost considerations in the selection of design parameters, i.e., sample size, sampling intervals and location of control limits was proposed by Duncan (1956). Tagaras (1989) studied the statistical properties and the economic design of $\bar{X}$ charts with asymmetric control limits. Del Castillo et al. (1996) applied an interactive multicriteria nonlinear optimization algorithm to a model for the design of $\bar{X}$ charts where only the sampling cost needs to be specified while the cost of false alarms need not be specified. Jaraiedi and Zhuang (1991) presented a computer program to perform optimal cost-based design of $\bar{X}$ charts when multiple

Michael B. C. Khoo is an Associate Professor in the School of Mathematical Sciences, Universiti Sains Malaysia (USM). He earned his Ph.D. in Applied Statistics from USM in 2001. His research interest is Statistical Quality Control. He is a member of the editorial boards of *Quality Engineering*, *Quality Management Journal*, *Journal of Modern Applied Statistical Methods and International Journal of Statistics and Management System*.

assignable causes can shift the process to an out-of-control state. McWilliams et al. (2001) gave a FORTRAN program that can be used to jointly determine the parameters of $\bar{X}$ charts used in combination with either the $R$ or $S$ charts. Waheba and Nickerson (2005) developed a comprehensive cost model to incorporate two cost functions, i.e., the reactive and proactive functions for obtaining economically optimum designs of $\bar{X}$ charts for controlling the process mean. Keats et al. (1995) presented and analyzed a methodology for using average production length (APL) and sampling constraints to aid in the design of $\bar{X}$ control schemes.

Costa (1994) studied the properties of the variable sample size (VSS) $\bar{X}$ chart when the size of each sample depends on what is observed in the preceding sample and compared its performance with the other methods. Sim et al. (2004) considered the occurrence of double assignable causes in a process, adopted the Markov chain approach to investigate the statistical properties of the VSS $\bar{X}$ chart and suggested a procedure to compute the optimal sample size. Lin and Chou (2005a) proposed the variable sample size and control limit (VSSCL) $\bar{X}$ chart which was shown to have a lower false alarm rate and to be quicker than the VSS $\bar{X}$ chart in detecting small and moderate shifts in a process involving non-normal populations. Reynolds and Stoumbos (2001) showed that the variable sampling interval (VSI) $X$ chart which allows the sampling interval to be varied enables a substantial reduction in the expected times in detecting shifts in process parameters. Chen and Chiou (2005) developed an economic design of VSI $\bar{X}$ control charts. Lin and Chou (2005b) proposed two adaptive $\bar{X}$ charts, i.e., the variable sampling rate with sampling at fixed times (VSRFT) $\bar{X}$ chart and the variable parameters with sampling at fixed times (VPFT) chart.

Nedumaran and Pignatiello (2001) addressed the problem of estimating the $\bar{X}$ chart limits when the values of the process parameters are unknown. Nedumaran and Pignatiello (2005) also proposed the use of the analysis of means (ANOM) technique for constructing retrospective $\bar{X}$ control chart limits so as to control the overall probability of a false alarm at a desired level. Champ and Jones (2004) examined methods for obtaining probability limits of Phase-I $\bar{X}$ charts when the process mean and standard deviation are estimated.

Methods of making the $\bar{X}$ charts less influenced by extreme observations and hence more effective in the detection of outliers are the trimmed mean $\bar{X}$ and R charts proposed by Langenberg and Iglewicz (1986) and the robust $\bar{X}_Q$ and $R_Q$ charts based on the sample interquartile range estimator suggested by Rocke (1989 and 1992). Among the procedures of using the charts for skewed populations are those based on the weighted variance concept proposed by Bai and Choi (1995) and Chang and Bai (2001), as well as that using the skewness correction method suggested by Chan and Cui (2003).

Other extensions of the $\bar{X}$ chart are as follows: The estimation of the time of a change in the mean following an out-of-control signal using the maximum likelihood estimation technique was proposed by Samuel et al. (1998). Park and Park (2004) suggested a maximum likelihood joint estimator of the change point to identify the time of a change in the process mean or variance when $\bar{X}$ and $S$ control charts issue a signal. Daudin (1992) presented a double sampling $\bar{X}$ chart which offers better statistical efficiency than the standard $\bar{X}$ chart without increasing the sampling. Costa and Rahim (2004) suggested joint $\bar{X}$ and $R$ charts with a two stage sampling procedure which speeds up the detection of process disturbances. Del Castillo (1996) presented a C program for the computation of the run length distribution and average run length of $\bar{X}$ charts with unknown process variance. Khoo (2004b) reviewed and studied some commonly used performance measures for the $\bar{X}$ charts. Maragah and Woodall (1992) showed the effect of autocorrelation on the retrospective $X$ chart for individuals. Roes et al. (1993), Rigdon et al. (1994) and Trip and Wieringa (2006) showed that using the $X$ chart alone is as efficient as the combined $X$-$MR$ chart for detecting changes in the process variance. However, Rigdon et al. (1994) recommended that the limits on the

individuals $X$ chart be based on the moving range ($MR$) rather than the sample standard deviation. Rahardja (2005) found that adding the MR chart to an $X$ chart is not helpful for detecting independently and identically distributed (i.i.d.) departures from standard conditions, but is beneficial in detecting some non-i.i.d. conditions. Combined $\bar{X}$ and $S$ charts such as the semicircle and Max charts are proposed by Chao and Cheng (1996) and Chen and Cheng (1998) respectively.

A Modified $\bar{X}$ Control Chart

Suppose that a quality characteristic is normally distributed with mean μ and standard deviation σ, where both μ and σ are known. If $X_1, X_2, ..., X_n$ is a sample of size n, then the mean of this sample is

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} \qquad (1)$$

For sampling from infinite populations, which is usually assumed to be the case in process monitoring, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Thus, the ±3 sigma limits of the standard $\bar{X}$ chart are $UCL_S / LCL_S = \mu_0 \pm 3\sigma_{\bar{X}} = \mu_0 \pm \frac{3\sigma}{\sqrt{n}}$, where $\mu_0$ is the in-control mean of the process.

However, in some industrial settings, sampling is made from finite populations. Here, the use of the standard $\bar{X}$ chart's limits can lead to erroneous conclusions as it will cause an inflated Type-II error which will be discussed later. For sampling from finite populations (Bluman, 2004), the sample mean, $\bar{X}_i \sim N\left(\mu, \frac{\sigma^2(N_i - n_i)}{n_i(N_i - 1)}\right)$, $i$ =1, 2, …. Assuming that a manufacturing process is producing items at a steady rate such as in a conveyor belt system and that the number of items drawn for each sample are of equal size, then $\bar{X}_i \sim N\left(\mu, \frac{\sigma^2(N-n)}{n(N-1)}\right)$. The correction factor for the variance of $\bar{X}_i$, i.e., $\frac{N-n}{N-1}$ is necessary if relatively large samples are taken from a small population so that the sample mean will more accurately estimate the population mean and there will be less error in the estimation. The ±3 sigma limits of the modified $\bar{X}$ chart for samples drawn from finite populations are

$$UCL_M = \mu_0 + \frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (2a)$$

and

$$LCL_M = \mu_0 - \frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (2b)$$

If the process parameters $\mu_0$ and σ are unknown, they are estimated from $\bar{\bar{X}}$ and $\bar{R}/d_2$ or $\bar{S}/c_4$ respectively, where $\bar{\bar{X}}$ is the grand average, $\bar{R}$ is the average range and $\bar{S}$ is the average standard deviation. Here $\bar{\bar{X}}$, $\bar{R}$ and, $\bar{S}$ are computed from the following formulae:

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_m}{m} \qquad (3)$$

$$\bar{R} = \frac{R_1 + R_2 + ... + R_m}{m} \qquad (4)$$

and

$$\bar{S} = \frac{S_1 + S_2 + ... + S_m}{m} \qquad (5)$$

where $\bar{X}_1, \bar{X}_2, ..., \bar{X}_m$ denote the means of the $m$ samples, $R_1, R_2, ..., R_m$ the ranges of the $m$ samples while $S_1, S_2, ..., S_m$ the standard deviations of the $m$ samples. The $m$ samples from which $\bar{X}_i$, $R_i$, and $S_i$, $i$ = 1, 2, …, $m$, are computed are assumed to be taken from an in-control process.

If $\mu_0$ and σ are unknown, the limits in eqs. (2a) and (2b) when σ is estimated using $\bar{R}/d_2$ are

$$UCL_M = \bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \bar{\bar{X}} + A_2'\bar{R} \quad (6a)$$

and

$$LCL_M = \bar{\bar{X}} - \frac{3\bar{R}}{d_2\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \bar{\bar{X}} - A_2'\bar{R}, \quad (6b)$$

respectively, where $A_2' = \frac{3}{d_2\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

If $\bar{S}/c_4$ is used to estimate σ, then the limits in eqs. (2a) and (2b) become

$$UCL_M = \bar{\bar{X}} + \frac{3\bar{S}}{c_4\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \bar{\bar{X}} + A_3'\bar{S} \quad (7a)$$

and

$$\mathrm{LCL_M} = \bar{\bar{X}} - \frac{3\bar{S}}{c_4\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \bar{\bar{X}} - A_3'\bar{S}, \quad (7b)$$

respectively, where $A_3' = \frac{3}{c_4\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Generally, the estimator $\hat{\sigma} = \bar{R}/d_2$ is used for small sample sizes, say $n < 10$ while the estimator $\hat{\sigma} = \bar{S}/c_4$ is used for big sample sizes, say $n \geq 10$. Factors $A_2'$ and $A_3'$ based on various values of $n$ and $N$, for the construction of the limits of the modified $\bar{X}$ chart are given in Tables A1 and A2 respectively in the Appendix, where sample sizes of $n$ = 2, 3, …, 25 and selected population sizes of $N \leq 1000$ are considered. Note that $N > 1000$ is not considered because it will be shown via Monte Carlo simulation in this article that the results of the standard and modified $\bar{X}$ charts are about the same for $N > 1000$.

Formulae for Computing the Type-I and Type-II Errors of the Modified and Standard $\bar{X}$ Charts

This section deals with the derivation of formulae for computing the probabilities of Type-I, $\alpha$ and Type-II, $\beta$ errors of the modified and standard $\bar{X}$ charts. The exact in-control and out-of-control ARLs can be easily computed using formulae

$$\mathrm{ARL}_0 = \frac{1}{\alpha} \quad (8)$$

and

$$\mathrm{ARL}_1 = \frac{1}{1-\beta}, \quad (9)$$

respectively.

Assume that the out-of-control process mean is represented by $\mu = \mu_0 + \delta\sigma$, where $\mu_0$ denotes the in-control mean. Note that $\delta = 0$ shows that the process is in-control while $\delta > 0$ or $\delta < 0$ indicates that the process is out-of-control. For sampling from finite populations, it is known that $\bar{X} \sim N\left[\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right]$. The probability of a Type-I error of the modified $\bar{X}$ chart for sampling from finite populations is

$$\begin{aligned}\alpha_M &= P\left(\bar{X} > UCL_M \mid \mu = \mu_0\right) + P\left(\bar{X} < LCL_M \mid \mu = \mu_0\right) \\ &= P\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} > \frac{\mu_0 + \frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} - \mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) + \\ &\quad P\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0 - \frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} - \mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &= P(Z > 3) + P(Z < -3) \end{aligned} \quad (10)$$

while the corresponding probability of a Type-I

error of the standard $\bar{X}$ chart is

$$\begin{aligned}\alpha_S &= P\left(\bar{X} > UCL_S \mid \mu = \mu_0\right) + P\left(\bar{X} < LCL_S \mid \mu = \mu_0\right) \\ &= P\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} > \frac{\mu_0+\frac{3\sigma}{\sqrt{n}}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) + \\ &\quad P\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0-\frac{3\sigma}{\sqrt{n}}-\mu_0}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &= P\left(Z > 3\sqrt{\frac{N-1}{N-n}}\right) + \\ &\quad P\left(Z < -3\sqrt{\frac{N-1}{N-n}}\right)\end{aligned} \tag{11}$$

The probability of a Type-II error of the modified $\bar{X}$ chart for sampling from finite populations is computed as follows:

$$\begin{aligned}\beta_M &= P\left(LCL_M < \bar{X} < UCL_M \mid \mu = \mu_0 + \delta\sigma\right) \\ &= P\left(\bar{X} < UCL_M \mid \mu = \mu_0 + \delta\sigma\right) - \\ &\quad P\left(\bar{X} < LCL_M \mid \mu = \mu_0 + \delta\sigma\right) \\ &= P\left(\frac{\bar{X}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0+\frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &\quad - P\left(\frac{\bar{X}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0-\frac{3\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &= P\left(Z < 3-\delta\sqrt{\frac{n(N-1)}{N-n}}\right) \\ &\quad - P\left(Z < -3-\delta\sqrt{\frac{n(N-1)}{N-n}}\right)\end{aligned} \tag{12}$$

while that of the standard $\bar{X}$ chart is

$$\begin{aligned}\beta_S &= P\left(LCL_S < \bar{X} < UCL_S \mid \mu = \mu_0 + \delta\sigma\right) \\ &= P\left(\bar{X} < UCL_S \mid \mu = \mu_0 + \delta\sigma\right) \\ &\quad - P\left(\bar{X} < LCL_S \mid \mu = \mu_0 + \delta\sigma\right) \\ &= P\left(\frac{\bar{X}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0+\frac{3\sigma}{\sqrt{n}}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &\quad - P\left(\frac{\bar{X}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}} < \frac{\mu_0-\frac{3\sigma}{\sqrt{n}}-\mu_0-\delta\sigma}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}\right) \\ &= P\left[Z < \sqrt{\frac{N-1}{N-n}}\left(3-\delta\sqrt{n}\right)\right] \\ &\quad - P\left[Z < -\sqrt{\frac{N-1}{N-n}}\left(3+\delta\sqrt{n}\right)\right]\end{aligned} \tag{13}$$

A Comparison of the ARL Performances of the Modified and Standard $\bar{X}$ charts

The ARL profiles of the modified $\bar{X}$ chart can be easily computed using eqs. (8), (9), (10) and (12) while that of the standard $\bar{X}$ chart from eqs. (8), (9), (11) and (13). SAS version 9 is used in the computation of the ARL values. For ease of computation, the in-control process is assumed to follow a standard normal, N(0,1) distribution while the out-of-control process a normal, N(δ,1) distribution so that the out-of-control mean is $\mu = \mu_0 + \delta\sigma$ where $\mu_0 = 0$ and $\sigma = 1$. Values of $\delta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 1.2, 1.4, 1.6, 1.8, 2\}$ are used so that a positive shift is considered. Due to the symmetrical limits of the modified and standard $\bar{X}$ charts, similar ARL profiles will be obtained for positive and negative values of δ. The sample sizes, $n \in \{1, 2, 5\}$ and population sizes, $N \in \{10, 25, 50, 100, 500, 1000, 2500, 5000, 7500, 10000\}$ are considered. Tables 1 and 2 give the ARL results of the modified and standard $\bar{X}$ charts respectively.

When $n = 1$, both the modified and standard $\bar{X}$ charts are reduced to the individuals $X$ charts. From eqs. (10) and (11), it is observed

that $\alpha_M = \alpha_S$ for $n$ = 1 and similarly, from eqs. (12) and (13), $\beta_M = \beta_S$ for $n$ = 1. Thus, the ARL profiles of the two charts in Tables 1 and 2 are exactly the same when $n$ =1, where $ARL_0$ = 370.4 irrespective of the population size, $N$. Note that the results in Tables 1 and 2 for n = 1 are also similar to that of the standard $\bar{X}$ chart when samples are drawn from infinite populations because it can be shown easily that $\alpha_M = \alpha_S = \alpha$ and $\beta_M = \beta_S = \beta$, where α and β are the probabilities of the Type-I and Type-II errors of the standard $\bar{X}$ chart for sampling from infinite populations.

For bigger sample sizes of n = 2 or 5, it is observed that the modified $\bar{X}$ chart gives reliable results (see Table 1) compared to that of the standard $\bar{X}$ chart (see Table 2). The $ARL_0$ values of the modified $\bar{X}$ chart for $n$ = 2 and 5 are all 370.4, irrespective of the value of $N$, i.e., similar to the case of the standard $\bar{X}$ chart when sampling is made from infinite populations. On the contrary, the $ARL_0$ values of the standard $\bar{X}$ chart for $n$ = 2 and 5 are greatly larger than 370.4 for small values of $N$, which are more pronounced for $n$ = 5. For example, when $n$ = 5, $ARL_0$ = 17545.7, 985.2, 573.0, 455.6 and 385.4 for N = 10, 25, 50, 100 and 500 respectively. The $ARL_0$ values of the standard $\bar{X}$ chart in Table 2 for $n$ = 2 and 5 decreases as $N$ increases and approximates 370.4 when $N$ > 1000. The $ARL_1$ values of the standard $\bar{X}$ chart in Table 2 for $n$ = 2 and 5 involving small values of $N$ and δ are greatly larger than the corresponding values in Table 1. For example, when $n$ = 5, $N$ = 10 and $\delta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ the $ARL_1$ values of the standard $\bar{X}$ chart are 9495.2, 3232.5, 1124, 422.9, and 172.8 respectively, while that of the modified $\bar{X}$ chart are significantly smaller at 253.1, 119.7, 55.8, 27.8 and 15 respectively. Thus, using the standard $\bar{X}$ chart in the detection of process shifts when sampling is made from finite populations where $N$ is small or of moderate size can lead to a significant delay in the detection of small shifts in the mean. The $ARL_1$ value of the standard $\bar{X}$ chart that corresponds to a fixed small value of δ for $n$ = 2 and 5, say δ = 0.1 decreases as $N$ increases and approximates that of the modified chart when $N$ > 1000. From the above discussion, it is found that the use of the standard $\bar{X}$ chart can lead to erroneous conclusion and wrong understanding of the probabilities of Type-I and Type-II errors of the chart if sampling is made from finite populations of N < 1000. The use of the modified chart is justified in that it produces reliable in-control and out-of-control ARL values which are somewhat close to that of the standard $\bar{X}$ chart where sampling is made from infinite populations.

## Conclusion

The standard $\bar{X}$ chart caters only for the case involving sampling from infinite populations. This article identifies the problems faced by the standard $\bar{X}$ chart when it is used in the monitoring of processes for samples drawn from finite populations or if the population which is supposedly assumed to be infinite consists of less than $N$ = 1000 items of a certain part. As highlighted above, the problems arise include $ARL_0$s for $n \geq 2$ and $N$ < 1000 are greatly larger than the target value of approximately 370 and the corresponding $ARL_1$s involving small shifts in the mean are also greatly larger than that of the modified $\bar{X}$ chart. In an industrial setting if the assumption of an infinite population size where sampling is made cannot be met, the modified $\bar{X}$ chart should be used in place of the standard $\bar{X}$ chart. Tables A1 and A2 in the Appendix provide factors $A_2'$ and $A_3'$ used in the computation of the control limits of the modified $\bar{X}$ chart if the process parameters need to be estimated from a preliminary set of data of in-control subgroups. These factors simplify the computation of the limits of the modified $\bar{X}$ chart.

## References

Bai, D. S., & Choi, I. S. (1995). $\bar{X}$ and $R$ control charts for skewed populations. *Journal of Quality Technology, 27*, 120 – 131.

Bluman, A. G. (2004). *Elementary Statistics*, 5th ed. McGraw-Hill, New York.

Table 1. ARL profiles of the modified $\bar{\bar{X}}$ chart

| $n$ | δ | $N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10 | 25 | 50 | 100 | 500 | 1000 | 2500 | 5000 | 7500 | 10000 |
| | 0 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 |
| | 0.1 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 |
| | 0.2 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 |
| | 0.3 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 |
| | 0.4 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 |
| | 0.5 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 |
| 1 | 0.75 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 |
| | 1 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 |
| | 1.2 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 |
| | 1.4 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 |
| | 1.6 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 |
| | 1.8 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 |
| | 2 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 |
| | 0 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 |
| | 0.1 | 333.1 | 335.6 | 336.2 | 336.6 | 336.8 | 336.9 | 336.9 | 336.9 | 336.9 | 336.9 |
| | 0.2 | 253.1 | 259.3 | 261.1 | 261.9 | 262.6 | 262.7 | 262.7 | 262.7 | 262.7 | 262.7 |
| | 0.3 | 176.5 | 184.1 | 186.3 | 187.3 | 188.1 | 188.2 | 188.3 | 188.3 | 188.3 | 188.3 |
| | 0.4 | 119.7 | 126.8 | 128.9 | 129.9 | 130.7 | 130.8 | 130.8 | 130.8 | 130.9 | 130.9 |
| | 0.5 | 81.2 | 87.2 | 89 | 89.8 | 90.5 | 90.6 | 90.6 | 90.6 | 90.6 | 90.6 |
| 2 | 0.75 | 32.9 | 36.1 | 37.1 | 37.6 | 38 | 38 | 38.1 | 38.1 | 38.1 | 38.1 |
| | 1 | 15 | 16.7 | 17.2 | 17.5 | 17.7 | 17.7 | 17.7 | 17.7 | 17.7 | 17.7 |
| | 1.2 | 8.7 | 9.7 | 10.1 | 10.2 | 10.4 | 10.4 | 10.4 | 10.4 | 10.4 | 10.4 |
| | 1.4 | 5.4 | 6.1 | 6.3 | 6.4 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | 1.6 | 3.6 | 4.1 | 4.2 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 |
| | 1.8 | 2.6 | 2.9 | 3 | 3 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 |
| | 2 | 2 | 2.2 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 |
| | 0 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 |
| | 0.1 | 253.1 | 284 | 290.4 | 293.2 | 295.3 | 295.5 | 295.7 | 295.7 | 295.7 | 295.7 |
| | 0.2 | 119.7 | 159.3 | 169.1 | 173.6 | 176.9 | 177.3 | 177.6 | 177.7 | 177.7 | 177.7 |
| | 0.3 | 55.8 | 84.4 | 92.3 | 96 | 98.9 | 99.2 | 99.4 | 99.5 | 99.5 | 99.5 |
| | 0.4 | 27.8 | 46 | 51.5 | 54.1 | 56.1 | 56.3 | 56.5 | 56.5 | 56.6 | 56.6 |
| | 0.5 | 15 | 26.4 | 29.9 | 31.7 | 33.1 | 33.2 | 33.3 | 33.4 | 33.4 | 33.4 |
| 5 | 0.75 | 4.4 | 8.2 | 9.5 | 10.1 | 10.6 | 10.7 | 10.7 | 10.7 | 10.8 | 10.8 |
| | 1 | 2 | 3.4 | 4 | 4.2 | 4.4 | 4.5 | 4.5 | 4.5 | 4.5 | 4.5 |
| | 1.2 | 1.4 | 2.1 | 2.4 | 2.5 | 2.6 | 2.6 | 2.7 | 2.7 | 2.7 | 2.7 |
| | 1.4 | 1.1 | 1.5 | 1.7 | 1.7 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 |
| | 1.6 | 1 | 1.2 | 1.3 | 1.3 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 |
| | 1.8 | 1 | 1.1 | 1.1 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| | 2 | 1 | 1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |

Table 2. ARL profiles of the standard $\overline{X}$ chart

| $N$ | δ | $N$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 10 | 25 | 50 | 100 | 500 | 1000 | 2500 | 5000 | 7500 | 10000 |
| | 0 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 | 370.4 |
| | 0.1 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 | 352.9 |
| | 0.2 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 | 308.4 |
| | 0.3 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 | 253.1 |
| | 0.4 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 | 200.1 |
| | 0.5 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 | 155.2 |
| 1 | 0.75 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 | 81.2 |
| | 1 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 | 43.9 |
| | 1.2 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 | 27.8 |
| | 1.4 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 | 18.2 |
| | 1.6 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 | 12.4 |
| | 1.8 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 |
| | 2 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 | 6.3 |
| | 0 | 683.7 | 458.7 | 410.4 | 389.5 | 374.1 | 372.2 | 371.1 | 370.8 | 370.6 | 370.6 |
| | 0.1 | 607.6 | 413.9 | 371.9 | 353.6 | 340.1 | 338.5 | 337.5 | 337.2 | 337.1 | 337 |
| | 0.2 | 449.5 | 317.1 | 287.5 | 274.6 | 265 | 263.9 | 263.2 | 263 | 262.9 | 262.9 |
| | 0.3 | 304.1 | 222.7 | 204.1 | 195.9 | 189.8 | 189.1 | 188.6 | 188.5 | 188.4 | 188.4 |
| | 0.4 | 200.4 | 151.9 | 140.6 | 135.5 | 131.8 | 131.3 | 131 | 131 | 130.9 | 130.9 |
| | 0.5 | 132.4 | 103.5 | 96.6 | 93.5 | 91.2 | 90.9 | 90.8 | 90.7 | 90.7 | 90.7 |
| 2 | 0.75 | 50.4 | 42 | 39.9 | 39 | 38.3 | 38.2 | 38.1 | 38.1 | 38.1 | 38.1 |
| | 1 | 21.6 | 19 | 18.3 | 18 | 17.8 | 17.8 | 17.7 | 17.7 | 17.7 | 17.7 |
| | 1.2 | 12 | 10.9 | 10.6 | 10.5 | 10.4 | 10.4 | 10.4 | 10.4 | 10.4 | 10.4 |
| | 1.4 | 7.2 | 6.7 | 6.6 | 6.6 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 | 6.5 |
| | 1.6 | 4.6 | 4.4 | 4.4 | 4.4 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 |
| | 1.8 | 3.2 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 | 3.1 |
| | 2 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 |
| | 0 | 17545.7 | 985.2 | 573.0 | 455.6 | 385.4 | 377.8 | 373.3 | 371.9 | 371.4 | 371.1 |
| | 0.1 | 9495.2 | 722.4 | 441.1 | 357.6 | 306.8 | 301.2 | 297.9 | 296.8 | 296.5 | 296.3 |
| | 0.2 | 3232.5 | 375.5 | 248.5 | 208.5 | 183.3 | 180.5 | 178.8 | 178.3 | 178.1 | 178 |
| | 0.3 | 1124 | 185.5 | 131.5 | 113.6 | 102.1 | 100.8 | 100.1 | 99.8 | 99.7 | 99.7 |
| | 0.4 | 422.9 | 94.8 | 71.3 | 63.2 | 57.8 | 57.2 | 56.8 | 56.7 | 56.7 | 56.7 |
| | 0.5 | 172.8 | 51.0 | 40.3 | 36.5 | 34.0 | 33.7 | 33.5 | 33.5 | 33.4 | 33.4 |
| 5 | 0.75 | 26.3 | 13.6 | 11.9 | 11.3 | 10.9 | 10.8 | 10.8 | 10.8 | 10.8 | 10.8 |
| | 1 | 6.5 | 5 | 4.7 | 4.6 | 4.5 | 4.5 | 4.5 | 4.5 | 4.5 | 4.5 |
| | 1.2 | 3 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 | 2.7 |
| | 1.4 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 |
| | 1.6 | 1.3 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 1.4 |
| | 1.8 | 1.1 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| | 2 | 1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |

Chan, L. K., & Cui, H. J. (2003). Skewness correction $\bar{X}$ and *R* charts for skewed distributions. *Naval Research Logistics, 50*, 555 – 573.

Champ, C. W., & Jones, L. A. (2004). Designing phase I $\bar{X}$ charts with small sample sizes. *Quality and Reliability Engineering International, 20*, 497 – 510.

Chang, Y. S., & Bai, D. S. (2001). Control charts for positively-skewed populations with weighted standard deviations. *Quality and Reliability Engineering International, 17*, 397 – 406.

Chao, M. T., & Cheng, S. W. (1996). Semicircle control chart for variables data. *Quality Engineering, 8*, 441 – 446.

Chen, G., & Cheng, S. W. (1998). Max chart: Combining *X*-bar chart and *S* chart. *Statistica Sinica, 8*, 263 – 271.

Chen, Y. K., & Chiou, K. C. (2005). Optimal design of VSI $\bar{X}$ control charts for monitoring correlated samples. *Quality and Reliability Engineering International, 21*, 757 – 768.

Costa, A. F. B. (1994). $\bar{X}$ charts with variable sample size. *Journal of Quality Technology, 26*, 155 – 163.

Costa, A. F. B., & Rahim, M. A. (2004). Joint $\bar{X}$ and *R* charts with two-stage samplings. *Quality and Reliability Engineering International, 20*, 699 – 708.

Daudin, J. J. (1992). Double sampling $\bar{X}$ charts. *Journal of Quality Technology, 24*, 78 – 87.

Davis, R. B., & Krehbiel, T. C. (2002). Shewhart and zone control chart performance under linear trend. *Communications in Statistics: Simulation and Computation, 31*, 91 – 96.

Del Castillo, E., Mackin, P., & Montgomery, D. C. (1996a). Multiple-criteria optimal design of $\bar{X}$ control charts. *IIE Transactions, 28*, 467 – 474.

Del Castillo, E. (1996b). Evaluation of run length distribution for $\bar{X}$ charts with unknown variance. *Journal of Quality Technology, 28*, 116 – 122.

Duncan, A. J. (1956). The economic design of $\bar{X}$ charts used to maintain current control of a process. *Journal of the American Statistical Association, 51,* 228 – 242.

Jaraiedi, M. & Zhuang, Z. (1991). Determination of optimal design parameters of $\bar{X}$ charts when there is a multiplicity of assignable causes. *Journal of Quality Technology, 23*, 253 – 258.

Keats, J. B., Miskulin, J. D., & Runger, G. C. (1995). Statistical process control scheme design. *Journal of Quality Technology, 27*, 214 – 225.

Khoo, M. B. C. (2004). Design of runs rules schemes. *Quality Engineering, 16*, 27 – 43.

Khoo, M. B. C. (2004). Performance measures for the Shewhart $\bar{X}$ control chart. *Quality Engineering, 16*, 585 – 590.

Khoo, M. B. C., & Khotrun, N. A. (2006). Two improved runs rules for the Shewhart $\bar{X}$ control chart. *Quality Engineering, 18,* 173 – 178.

Klein, M. (2000). Two alternatives to the Shewhart $\bar{X}$ control chart. *Journal of Quality Technology, 32*, 427 – 431.

Langenberg, P., & Iglewicz, B. (1986). Trimmed mean $\bar{X}$ and *R* charts. *Journal of Quality Technology, 18*, 152 – 161.

Lin, Y. C., & Chou, C. Y. (2005a). Robustness of the variable sample size and control limit $\bar{X}$ chart to non normality. *Communications in Statistics: Theory and Methods, 34*, 721 – 743.

Lin, Y. C., & Chou, C. Y. (2005b). Adaptive $\bar{X}$ control charts with sampling at fixed times. *Quality and Reliability Engineering International, 21*, 163 – 175.

Maragah, H. D., & Woodall, W. H. (1992). The effect of autocorrelation on the retrospective *X*-chart. *Journal of Statistical Computation and Simulation, 40*, 29 – 42.

McWilliams, T. P., Saniga, E. M., & Davis, D. J. (2001). Economic-statistical design of $\bar{X}$ and *R* or $\bar{X}$ and *S* charts. *Journal of Quality Technology, 33*, 234 – 241.

Nedumaran, G., & Pignatiello, J. J., Jr. (2001). On estimating $\bar{X}$ control chart limits. *Journal of Quality Technology, 33*, 206 – 212.

Nedumaran, G., & Pignatiello, J. J., Jr. (2005). On constructing retrospective $\bar{X}$ control chart limits. *Quality and Reliability Engineering International, 21*, 81 – 89.

Nelson, L. S. (1984). The Shewhart control chart: Tests for special causes. *Journal of Quality Technology, 16*, 237 – 239.

Park, J., & Park, S. (2004). Estimation of the change point in the $\bar{X}$ and $S$ control charts. *Communications in Statistics: Simulation and Computation, 33*, 1115 – 1132.

Rahardja, D. (2005). $X$-charts versus $X$/$MR$ chart combinations: IID cases and non-iid cases. *Quality Engineering, 17*, 189 – 196.

Reynolds, M. R., Jr., & Stoumbos, Z. G. (2001). Monitoring the process mean and variance using individual observations and variable sampling intervals. *Journal of Quality Technology, 33*, 181 – 205.

Rigdon, S. E., Cruthis, E. N., & Champ, C. W. (1994). Design strategies for individuals and moving range control charts. *Journal of Quality Technology, 26*, 274 – 287.

Rocke, D. M. (1989). Robust control charts. *Technometrics, 31*, 173 – 184.

Rocke, D. M. (1992). $\bar{X}_Q$ and $R_Q$ charts: Robust control charts. *The Statistician, 41*, 97 – 104.

Roes, K. C. B., Does, R. J. M. M., & Schurink, Y. (1993). *Journal of Quality Technology, 25*, 188 – 198.

Samuel, T. R., Pignatiello, J. J., Jr., & Calvin, J. A. (1998). Identifying the time of a step change with $\bar{X}$ control charts. *Quality Engineering, 10*, 521 – 527.

Shmueli, G., & Cohen, A. (2003). Run-length distribution for control charts with runs and scans rules. *Communications in Statistics: Theory and Methods, 32*, 475 – 495.

Sim, S. B., Kang, C. W. & Xie, M. (2004). On variable sample size $\bar{X}$ chart for processes with double assignable causes. *International Journal of Reliability, Quality and Safety Engineering, 11*, 47 – 58.

Tagaras, G. (1989). Economic $\bar{X}$ charts with asymmetric control limits. *Journal of Quality Technology, 21*, 147 – 154.

Trip, A., & Wieringa, J. E. (2006). Individuals charts and additional tests for changes in spread. *Quality and Reliability Engineering International, 22*, 239 – 249.

Waheba, G. S., & Nickerson, D. M. (2005). The economic design of $\bar{X}$ charts: A proactive approach. *Quality and Reliability Engineering International, 21*, 91 – 104.

Walker, E., Philpot, J. W., & Clement, J. (1991). False signal rates for the Shewhart control chart with supplementary runs rules. *Journal of Quality Technology, 23*, 247 – 252.

Wheeler, D. J. (1983). Detecting a shift in process average: Table of the power function for $\bar{X}$ charts. *Journal of Quality Technology, 15*, 155 – 169.

Acknowledgement

This material is funded by Research Grant no. 03-385 RG/MATHS/AS from the Third World Academy of Sciences (TWAS), Trieste, Italy.

## Appendix

Table A1. Values of factor, $A_2'$ for the Modified $\bar{X}$ chart

| Sample Size, $n$ | Population size, $N$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| 2 | 1.773 | 1.841 | 1.861 | 1.871 | 1.876 | 1.877 | 1.878 | 1.879 | 1.879 | 1.879 | 1.879 | 1.880 | 1.880 |
| 3 | 0.902 | 0.980 | 1.002 | 1.013 | 1.018 | 1.020 | 1.021 | 1.021 | 1.021 | 1.022 | 1.022 | 1.022 | 1.022 |
| 4 | 0.595 | 0.681 | 0.706 | 0.717 | 0.723 | 0.725 | 0.726 | 0.726 | 0.727 | 0.727 | 0.727 | 0.727 | 0.727 |
| 5 | 0.430 | 0.527 | 0.553 | 0.565 | 0.571 | 0.573 | 0.574 | 0.574 | 0.575 | 0.575 | 0.575 | 0.576 | 0.576 |
| 6 | 0.322 | 0.430 | 0.458 | 0.471 | 0.477 | 0.479 | 0.480 | 0.481 | 0.481 | 0.482 | 0.482 | 0.482 | 0.482 |
| 7 | 0.242 | 0.363 | 0.393 | 0.406 | 0.413 | 0.415 | 0.416 | 0.417 | 0.417 | 0.418 | 0.418 | 0.418 | 0.418 |
| 8 | 0.176 | 0.316 | 0.345 | 0.359 | 0.366 | 0.368 | 0.369 | 0.370 | 0.370 | 0.371 | 0.371 | 0.371 | 0.371 |
| 9 | 0.112 | 0.275 | 0.308 | 0.323 | 0.330 | 0.332 | 0.333 | 0.334 | 0.334 | 0.335 | 0.335 | 0.335 | 0.335 |
| 10 | | 0.244 | 0.278 | 0.294 | 0.301 | 0.304 | 0.305 | 0.305 | 0.306 | 0.306 | 0.306 | 0.307 | 0.307 |
| 11 | | 0.218 | 0.254 | 0.270 | 0.278 | 0.280 | 0.281 | 0.282 | 0.283 | 0.283 | 0.283 | 0.283 | 0.284 |
| 12 | | 0.196 | 0.234 | 0.251 | 0.258 | 0.261 | 0.262 | 0.263 | 0263 | 0.264 | 0.264 | 0.264 | 0.264 |
| 13 | | 0.176 | 0.217 | 0.234 | 0.242 | 0.244 | 0.246 | 0.246 | 0.247 | 0.247 | 0.248 | 0.248 | 0.248 |
| 14 | | 0.159 | 0.202 | 0.219 | 0.228 | 0.230 | 0.231 | 0.232 | 0.233 | 0.233 | 0.233 | 0.234 | 0.234 |
| 15 | | 0.144 | 0.189 | 0.207 | 0.215 | 0.218 | 0.219 | 0.220 | 0.220 | 0.221 | 0.221 | 0.221 | 0.222 |
| 16 | | 0.130 | 0.177 | 0.196 | 0.204 | 0.207 | 0.208 | 0.209 | 0.210 | 0.210 | 0.210 | 0.211 | 0.211 |
| 17 | | 0.117 | 0.166 | 0.186 | 0.194 | 0.197 | 0.199 | 0.200 | 0.200 | 0.200 | 0.201 | 0.201 | 0.201 |
| 18 | | 0.105 | 0.157 | 0.177 | 0.186 | 0.189 | 0.190 | 0.191 | 0.191 | 0.192 | 0.192 | 0.192 | 0.193 |
| 19 | | 0.093 | 0.148 | 0.169 | 0.178 | 0.181 | 0.182 | 0.183 | 0.184 | 0.184 | 0.184 | 0.185 | 0.185 |
| 20 | | 0.082 | 0.141 | 0.161 | 0.171 | 0.174 | 0.175 | 0.176 | 0.177 | 0.177 | 0.177 | 0.178 | 0.178 |
| 21 | | 0.071 | 0.133 | 0.155 | 0.164 | 0.167 | 0.169 | 0.170 | 0.170 | 0.171 | 0.171 | 0.171 | 0.172 |
| 22 | | 0.059 | 0.127 | 0.149 | 0.158 | 0.161 | 0.163 | 0.164 | 0.165 | 0.165 | 0.165 | 0.166 | 0.166 |
| 23 | | 0.047 | 0.120 | 0.143 | 0.153 | 0.156 | 0.158 | 0.159 | 0.159 | 0.160 | 0.160 | 0.160 | 0.160 |
| 24 | | 0.032 | 0.115 | 0.138 | 0.148 | 0.151 | 0.153 | 0.154 | 0.154 | 0.155 | 0.155 | 0.155 | 0.155 |
| 25 | | | 0.109 | 0.133 | 0.143 | 0.146 | 0.148 | 0.149 | 0.150 | 0.150 | 0.150 | 0.151 | 0.151 |

Table A2. Values of factor, $A_3'$ for the Modified $\bar{\bar{X}}$ chart

| Sample Size, $n$ | Population size, $N$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 25 | 50 | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| 2 | 2.507 | 2.603 | 2.631 | 2.645 | 2.652 | 2.654 | 2.655 | 2.656 | 2.656 | 2.657 | 2.657 | 2.657 | 2.657 |
| 3 | 1.724 | 1.871 | 1.914 | 1.935 | 1.945 | 1.948 | 1.950 | 1.951 | 1.951 | 1.952 | 1.952 | 1.952 | 1.953 |
| 4 | 1.329 | 1.523 | 1.578 | 1.603 | 1.616 | 1.620 | 1.622 | 1.623 | 1.624 | 1.625 | 1.625 | 1.625 | 1.626 |
| 5 | 1.064 | 1.303 | 1.368 | 1.398 | 1.413 | 1.418 | 1.420 | 1.422 | 1.423 | 1.423 | 1.424 | 1.424 | 1.424 |
| 6 | 0.858 | 1.145 | 1.220 | 1.254 | 1.271 | 1.276 | 1.279 | 1.281 | 1.282 | 1.283 | 1.283 | 1.284 | 1.284 |
| 7 | 0.682 | 1.024 | 1.107 | 1.146 | 1.164 | 1.170 | 1.173 | 1.175 | 1.176 | 1.177 | 1.177 | 1.178 | 1.178 |
| 8 | 0.518 | 0.925 | 1.018 | 1.060 | 1.080 | 1.086 | 1.089 | 1.091 | 1.093 | 1.094 | 1.094 | 1.095 | 1.095 |
| 9 | 0.344 | 0.842 | 0.944 | 0.989 | 1.011 | 1.018 | 1.021 | 1.023 | 1.025 | 1.026 | 1.026 | 1.027 | 1.028 |
| 10 | | 0.771 | 0.881 | 0.930 | 0.953 | 0.961 | 0.964 | 0.966 | 0.968 | 0.969 | 0.970 | 0.970 | 0.971 |
| 11 | | 0.708 | 0.827 | 0.879 | 0.904 | 0.912 | 0.916 | 0.918 | 0.920 | 0.921 | 0.922 | 0.922 | 0.923 |
| 12 | | 0.652 | 0.780 | 0.835 | 0.861 | 0.869 | 0.874 | 0.876 | 0.878 | 0.879 | 0.880 | 0.880 | 0.881 |
| 13 | | 0.601 | 0.738 | 0.796 | 0.824 | 0.832 | 0.837 | 0.839 | 0.841 | 0.842 | 0.843 | 0.844 | 0.844 |
| 14 | | 0.553 | 0.701 | 0.762 | 0.790 | 0.799 | 0.804 | 0.807 | 0.808 | 0.810 | 0.811 | 0.811 | 0.812 |
| 15 | | 0.509 | 0.666 | 0.731 | 0.760 | 0.770 | 0.775 | 0.777 | 0.779 | 0.781 | 0.782 | 0.782 | 0.783 |
| 16 | | 0.467 | 0.635 | 0.702 | 0.733 | 0.743 | 0.748 | 0.751 | 0.753 | 0.754 | 0.755 | 0.756 | 0.757 |
| 17 | | 0.427 | 0.607 | 0.677 | 0.709 | 0.719 | 0.724 | 0.727 | 0.729 | 0.731 | 0.732 | 0.732 | 0.733 |
| 18 | | 0.388 | 0.580 | 0.653 | 0.686 | 0.697 | 0.702 | 0.705 | 0.707 | 0.709 | 0.710 | 0.711 | 0.711 |
| 19 | | 0.349 | 0.555 | 0.631 | 0.666 | 0.677 | 0.682 | 0.685 | 0.687 | 0.689 | 0.690 | 0.691 | 0.692 |
| 20 | | 0.310 | 0.532 | 0.611 | 0.646 | 0.658 | 0.663 | 0.667 | 0.669 | 0.670 | 0.672 | 0.673 | 0.673 |
| 21 | | 0.271 | 0.510 | 0.592 | 0.629 | 0.640 | 0.646 | 0.649 | 0.652 | 0.653 | 0.655 | 0.655 | 0.656 |
| 22 | | 0.229 | 0.489 | 0.575 | 0.612 | 0.624 | 0.630 | 0.633 | 0.636 | 0.637 | 0.639 | 0.640 | 0.640 |
| 23 | | 0.183 | 0.470 | 0.558 | 0.597 | 0.609 | 0.615 | 0.619 | 0.621 | 0.623 | 0.624 | 0.625 | 0.626 |
| 24 | | 0.126 | 0.451 | 0.542 | 0.582 | 0.595 | 0.601 | 0.605 | 0.607 | 0.609 | 0.610 | 0.611 | 0.612 |
| 25 | | | 0.433 | 0.528 | 0.569 | 0.581 | 0.588 | 0.592 | 0.594 | 0.596 | 0.597 | 0.598 | 0.599 |

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 608-618
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Generalized Linear Mixed-Effects Models for the Analysis of Odor Detection Data

Sandra Hall Matthew S. Mayo
University of Kansas Medical Center

Xu-Feng Niu James C. Walker
Florida State University

Olfactory detection has become a science of interest. Seven individuals' odor detection abilities are explored and an attempt is made to characterize all subjects with one generalized linear mixed effects model. Two methods of fitting the models were used and simulations were conducted to discover which method yielded the best results.

Key words: olfactory, conditional distribution, Metropolis Algorithm**,** Monte Carlo Newton Raphson Method, random effects, detectability, odor, human, sensitivity.

## Introduction

The quality of indoor air is one of the least understood health problems that industry faces today. A major problem that poor indoor air quality causes is Sick Building Syndrome (EPA, 1989). This occurs when a substantial proportion of a building's occupants experience discomfort and health effects that are relieved upon leaving the building. It has been reported that sick buildings cause an estimated loss of between ten and one hundred billion dollars a year for non-medical aspects of diminished indoor air quality, excluding medical events such as asthmatic attacks (Fisk & Rosenfeld, 1997). Human symptoms of Sick Building Syndrome range from repetitive office headaches and common cold-like symptoms to serious ailments such as respiratory infections, asthma and allergies. A 1996 Cornell University study found that**,** in each of 35 buildings surveyed, at least 20% of the occupants had experienced symptoms associated with Sick Building Syndrome (Mann, 1998). Odor threshold is the point at which the probability of odor detection becomes greater than chance. Threshold is the most basic measure of sensory function. To understand higher order capabilities (e.g. odor discrimination, odor identification, identification of target in mixtures, and perception of odor quality), it is necessary to take into account the sensitivity of each individual to each chemical. Thus, it is important to have a valid way to quantify sensitivity. One example of why odor threshold might be studied is to gain a better understanding of issues related to olfaction such as Sick Building Syndrome. Another is that it has been hypothesized that early stages of Alzheimer's disease can be detected by a loss of odor detectability (Devanand et. al., 2000).

To help researchers understand this concept of accurately quantifying odor detection ability, a study was conducted at the Florida State University Sensory Research Institute's (SRI). Subjects received stimuli via a facemask that covered the person's mouth and nose, although the stimuli were taken in through only the nose. The subject then responded using a computer mouse and monitor screen as to whether or not an odor was detected. By using this olfactometer, the subject was given a precise concentration of the chemical (Walker et. al., 2003).

Sandra Hall is an independent statistical consultant. Her research interests include applied statistical analysis, brain imaging data analysis and computation statistics. Email: sandy_b_h@hotmail.com. Xu-Feng Niu is Professor of Statistics. His research interests include time series analysis and spatial statistics, niu@stat.fsu.edu. Matthew Mayo is Professor and Chair of Biostatistics and his research interests include clinical trials and experimental design, MMAYO@kumc.edu. James Walker's interests are in quantifying capabilities of the olfactory systems, jwalker@psy.fsu.edu.

For that study, seven subjects were recruited. They were selected so as to have a variety of different ages as well as subjects of each and both genders. As the subjects responded to the posters via phone calls they were asked routine questions to determine if they had a prior history of nasal defect. The researchers desired both a male and female subject in each of the following age categories: 18-20, 21-30, 31-44 and > 45 years. After three weeks of recruitment, no male subject was found in the 31-44 year- old group and the experimenters elected to continue the study without a male subject from this age group. Each subject completed 12 to 14 sessions over the course of 3 to 4 months. Each session consisted of 75 trials (15 trials of clean air in addition to 15 trials at each of 4 different concentrations of amyl acetate) each lasting approximately 18 seconds and separated by 90-second intervals. Hence, a typical session ran for 2 hours and 15 minutes.

A trial consists of the subject being asked to come to the mask, where they breathed the stimulus. The subject then used a mouse to click whether they detected an odor or not. The method of stimulus presentation allowed for very precise control. Before and during stimulus presentation, breathing was measured. After several seconds of pre-stimulus sampling of respiratory behavior was stored, the next exhalation onset triggered operation of the flow valve that (unless a clean air trial is scheduled) sent odorant to the mask. This approach essentially eliminated the vexing problem of a stimulus rise time, because the concentration reached its asymptotic value during the interval from an exhalation onset to the next inhalation onset (Prah, Sears & Walker, 1995). The specific concentrations and corresponding yes's (y's) and no's (n's) from the subject for the session were recorded on the same computer that randomized the concentrations to be given.

Traditionally, longitudinal data might have been analyzed using a generalized linear model (GLM) for each subject. However, this method does not accommodate a population based model, which is ultimately desired. Thus, generalized linear mixed models (GLMMs) were used to address the problem.

The class of functions known as GLMMs extends GLMs by adding random effects to the linear predictor(s). The benefit of this model is that it allows for responses that are correlated and non-normally distributed, which can frequently occur in actual problems. By including the random effects, the GLMMs can model correlated errors, smooth regression relationships and model dependence among variables that occurs in repeated measure designs. Many problems involve multiple sources of variation such as analysis of data that has a hierarchical structure like clinical trial data. The GLMM can be used to model such data. In this particular study, the model needed to account for the randomness of the session. This random nature is not considered in the traditional generalized linear model which is initially used to describe the data analyzed in this study.

A natural alternative to this approach is to utilize generalized estimating equations (GEE). The GEE approach is attractive because it allows for a weighted estimate of the regression parameters and correctly adjusts for correlated data. The problems with GEE are that a) it provides only a population model of the data and b) it requires a large amount of subjects for the large sample distribution properties to provide correct standard errors for inference (hypothesis testing and confidence intervals). Since only seven subjects were available, the GEE approach would not be an appropriate choice.

GLMMs are useful as an alternative to GEE and might be an approach that is useful in small sample sizes. For example, SAS has a procedure called "GLIMMIX" that is promising. The problem is that GLIMMIX has not been completely assessed for its usefulness in small sample sizes.

GLMMs provide insight into the behavior, but accurately estimating the model can be quite difficult. Because GLMMs are an extension of GLMs, one might logically try to fit the model using maximum likelihood, the common method to fit GLMs. The maximum likelihood method will only work for very simple GLMMs due to the need to numerically evaluate high dimensional integrals that are

irreducible. Thus, statisticians have looked for other methods to fit these models that do not involve the difficulties of the numerically complicated integration. Many different methods have been proposed to fit generalized linear mixed models. The model and two specific previously proposed methods (one being the SAS- GLIMMIX approach) will be discussed. Shown next will be results of the simulation study comparing these two methods for the data, fit the model that was deemed best in the simulation study and summarize the work.

## Methodology

### Model

Let $Y_{ij}$ be the jth response for subject i, with j = 1 to $n_i$ and i = 1 to m where m is the number of subjects and $n_i$ is the number of observations per subject. Let $X_{ijk}$ be the $j^{th}$ value of the $k^{th}$ fixed effect for subject i, with k = 1 to p and i and j as described previously. Thus, the traditional generalized linear model is

$$g(\mu_{ij}) = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ijk}$$

with $\mu_{ij} = E(Y_{ij})$, where g is the link function and p is the number of different fixed effects.

Upon including the random effects, the model becomes:

$$\eta_{ij} = g(\mu_{ij}) = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ijk} + \sum_{l=1}^{c} \alpha_{il} Z_{ijl} \quad (1)$$

where **X** is still assumed to be the matrix for the fixed effects and $Z_{ijl}$ is the $j^{th}$ value for the $l^{th}$ random effect for subject i where $l$ = 1 to c with c being the number of random effects. Also, it is assumed that $\mu_{ij} = E(Y_{ij} \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta})$ and $\text{var}(Y_{ij} \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}) = \phi a_i v(\mu_i)$, where $\phi$ is a dispersion parameter, $v(\cdot)$ is a specified variance function and $a_i$ is a known constant. The random effects $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m)$ are assumed to be independent with mean **0** and $\text{cov}(\boldsymbol{\alpha}_i)$=**D**. It is assumed that the elements of **Y** conditional on $\boldsymbol{\alpha}$ are both independent and drawn from an exponential family distribution. Finally, $\boldsymbol{\alpha}$ is assumed to be distributed $f_{\alpha}(\boldsymbol{\alpha} \mid \mathbf{D})$. Let $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{in_i})^T$. Then, the function becomes:

$$f_{Y_i|\alpha}(\mathbf{Y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) = \exp\left\{ \frac{\mathbf{Y}_i \boldsymbol{\eta}_i - c(\boldsymbol{\eta}_i)}{a(\phi)} + d(\mathbf{Y}_i, \phi) \right\} \quad (2)$$

and the likelihood function is:

$$L(\beta, \phi, \mathbf{D} \mid \mathbf{Y}) = \int \prod_{i=1}^{n} f_{Y_i|\alpha}(\mathbf{Y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) f_{\alpha}(\boldsymbol{\alpha} \mid \mathbf{D}) d\boldsymbol{\alpha} \quad (3)$$

(Breslow &Clayton, 1993; Clayton, 1993; Jiang, 1998; Lin & Breslow, 1996; Lindstrom & Bates, 1990; McCulloch, 1997; Vonesh, 1996).

### Simulation methods

Several methods have been proposed to estimate the solution to the generalized linear mixed model. McCulloch (1997) proposed algorithms for Monte Carlo EM (MCEM) and Monte Carlo Newton-Raphson (MCNR). Lin and Breslow (1996) proposed using a penalized quasi-likelihood approach with bias correction to estimate the model.

The Monte Carlo EM algorithm considers the random effects $\boldsymbol{\alpha}$ to be missing data. Therefore, the complete data would be W=(**Y**, $\boldsymbol{\alpha}$) and the log likelihood for the complete data would be

$$\ell_W = \sum_i \ln f_{Y_i|\alpha}(\mathbf{Y}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) + \ln f_{\alpha}(\boldsymbol{\alpha} \mid \mathbf{D}) \quad (4)$$

Thus, the $\mathbf{Y}_i$'s become independent when the $\boldsymbol{\alpha}$'s are known. Note that $\boldsymbol{\beta}$ and $\phi$ enter into the above equation only in the first term, the maximization with respect to those two terms is similar to a standard GLM computational problem with the $\boldsymbol{\alpha}$'s known. Then maximizing with respect to **D** involves replacing the sufficient statistics with their conditional expected value and then performing maximum likelihood using the distribution of $\boldsymbol{\alpha}$. McCulloch's (1997) algorithm follows:

1. Choose starting values for $\boldsymbol{\beta}^{(0)}$, $\phi^{(0)}$, and $\mathbf{D}^{(0)}$. Set m=0.
2. Calculate (with expectations evaluated under $\boldsymbol{\beta}^{(m)}$, $\phi^{(m)}$, and $\mathbf{D}^{(m)}$):
   a. $\boldsymbol{\beta}^{(m+1)}$ and $\phi^{(m+1)}$ which maximize $E[\ln f_{Y|\alpha}(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) \mid \mathbf{Y}]$
   b. $\mathbf{D}^{(m+1)}$ which maximizes $E[\ln f_{\alpha}(\boldsymbol{\alpha} \mid \mathbf{D}) \mid \mathbf{Y}]$
   c. Set m=m+1
3. If convergence is achieved, declare $\boldsymbol{\beta}^{(m+1)}$, $\phi^{(m+1)}$, and $\mathbf{D}^{(m+1)}$ to be maximum likelihood estimates. Otherwise repeat step two.

Neither expectation in step two can actually be found in closed form. It is, however, possible to produce random draws from the conditional distribution of $\boldsymbol{\alpha}|\mathbf{Y}$ by using the Metropolis algorithm (Vonesh, 1996), which does not require a specification of $f_Y$. Monte Carlo approximations may then be formed in order to estimate the two required expectations. For sufficiently large sample sizes, it was discovered that this method gains likelihood and would converge to a local maximum under appropriate regularity conditions (McCulloch, 1997). Although this holds promise, in variance component problems, such as with GLMMs, the likelihood surfaces are not necessarily unimodal; thus, this method may only converge to a local maximum and never to the global one. A second problem is that it is limited to the binary response with the probit link. Incorporating this Metropolis algorithm into the EM algorithm gives the MCEM algorithm below (McCulloch, 1997):

1. Choose starting values for $\boldsymbol{\beta}^{(0)}$, $\phi^{(0)}$, and $\mathbf{D}^{(0)}$. Set m=0.
2. Generate N values, $\boldsymbol{\alpha}^{(1)}$, $\boldsymbol{\alpha}^{(2)}$, … , $\boldsymbol{\alpha}^{(N)}$ from $f_{\alpha|Y}(\boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{\beta}^{(m)}, \phi^{(m)}, \mathbf{D}^{(m)})$ using the Metropolis algorithm:
   a. Choose $\boldsymbol{\beta}^{(m+1)}$ and $\phi^{(m+1)}$ to maximize a Monte Carlo estimate of $E[\ln f_{Y|\alpha}(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi) \mid \mathbf{Y}]$ that is maximize $\frac{1}{N}\sum_{k=1}^{N} \ln f_{Y|\alpha}(\mathbf{Y} \mid \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}, \phi)$
   b. Choose $\mathbf{D}^{(m+1)}$ to maximize $\frac{1}{N}\sum_{k=1}^{N} \ln f_{\alpha}(\boldsymbol{\alpha}^{(k)} \mid \mathbf{D})$
   c. Set m=m+1
3. If convergence is achieved declare $\boldsymbol{\beta}^{(m+1)}$, $\phi^{(m+1)}$, and $\mathbf{D}^{(m+1)}$ to be maximum likelihood estimates. Otherwise repeat step two.

The next method that McCulloch (1997) used is the Monte Carlo Newton-Raphson method. This method also seemed robust to starting values. Again, since the likelihood surfaces are not unimodal, they are definitely not concave and thus this method may not converge at all, let alone to the global maximum. In practice it was discovered that this method generally got close to the correct answer. The algorithm appears below:

1. Choose starting values for $\boldsymbol{\beta}^{(0)}$, $\phi^{(0)}$, and $\mathbf{D}^{(0)}$. Set m=0.
2. Generate N values, $\boldsymbol{\alpha}^{(1)}$, $\boldsymbol{\alpha}^{(2)}$, … , $\boldsymbol{\alpha}^{(N)}$ from $f_{\alpha|Y}(\boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{\beta}^{(m)}, \phi^{(m)}, \mathbf{D}^{(m)})$ using the Metropolis algorithm and use them to form Monte Carlo estimates of the expectations (denoted as $\hat{E}[\cdot]$):
   a. Calculate

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \hat{E}[\mathbf{X}^{\mathrm{T}}\mathbf{S}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\alpha})\mathbf{X} \mid \mathbf{Y}]^{-1}\mathbf{X}^{\mathrm{T}}([\mathbf{S}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\alpha})\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\Big|_{\theta=\theta^{(m)}} (\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha})) \mid \mathbf{Y}]) \quad (5)$$

   where $\boldsymbol{\mu}_{ij}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = E[\mathbf{Y}_{ij} \mid \boldsymbol{\alpha}_i]$, $\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} = diag\left\{\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\right\}$, and $\mathbf{S}(\boldsymbol{\theta}, \boldsymbol{\alpha})^{-1} = diag\left\{\left(\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\right)^2 \mathrm{var}(Y_{ij} \mid \boldsymbol{\alpha}_i)\right\}$

a. Calculate $\phi^{(m+1)}$ to solve $E\left[\frac{\partial \ln f_{Y|\alpha}(\mathbf{Y} \mid \boldsymbol{\alpha}, \theta)}{\partial \phi} \mid \mathbf{Y}\right] = 0$ or a scoring equation.

c. Choose $\mathbf{D}^{(m+1)}$ to maximize $\frac{1}{N}\sum_{k=1}^{N} \ln f_{\alpha}(\boldsymbol{\alpha}^{(k)} \mid \mathbf{D})$

d. Set m=m+1

3. If convergence is achieved declare $\boldsymbol{\beta}^{(m+1)}$, $\phi^{(m+1)}$, and $\mathbf{D}^{(m+1)}$ to be maximum likelihood estimates. Otherwise repeat step two.

The MCEM and MCNR algorithms are very similar. In fact, the maximization to calculate the fixed effects coefficients in the MCEM algorithm cannot explicitly be carried out for binomial data, and thus, an estimation method is necessary, such as the Newton Raphson Method. Thus, for our purposes, the MCNR is equivalent to the MCEM algorithm.

Breslow and Clayton (1993) proposed performing a method known as penalized quasi-likelihood analysis (PQL) in order to approximate the maximum likelihood estimates. The key feature of this analysis is that it is easy to implement, especially since there exists a SAS macro for this method. The procedure is to repeatedly fit a linear mixed model to a modified dependent variable. They realized that a limitation of the PQL is that when assessing the uncertainty in both random and fixed effects it does not take into account the contribution of the estimated variance components. Lin and Breslow (1996) proposed a four-step procedure of bias correction for the PQL.

Lin and Breslow (1996) provided a four step algorithm to find the bias-corrected penalized quasi-likelihood estimates of the regression coefficients and variance components. They performed simulation studies and found that the bias correction procedure can improve asymptotic performance of the estimates for correlated binary data. They also discovered that this simple correction procedure would effectively reduce the bias of variance components of the PQL estimates and the associated mean square error as long as the sample size is reasonably large.

## Results

The two methods, Monte Carlo Newton-Raphson and penalized quasi-likelihood with bias correction, were used in a simulation study in order to determine which method better estimates the fixed affects as well as the random effects. The MCNR program was written in Matlab. The penalized quasi-likelihood program with bias correction (PQBC) was coded using SAS and the GLIMMIX macro available from SAS's website: http://ftp.sas.com/techsup/download/stat/.

The response vector for each program was generated using a binomial random generator. Binomial probabilities were calculated for each combination of subject, session and concentration.

It was then determined how many simulations of the program should be carried out in order to have results that converge. Thus, each of the programs was run a total of 100 and 1000 times, respectively. Each time, a new response vector was generated. The response vectors were based on the following model, using concentrations from four of the seven subjects,

$$p_{ij} = \frac{e^{-15-3.5*conc_{ij}}}{1+e^{-15-3.5*conc_{ij}}} \qquad (6)$$

where $p_{ij}$ is the probability for the $j^{th}$ concentration of subject i. The model gives the probability to be used for each concentration value. The binomial generator was then used along with the probabilities found in the model to generate fifteen binary responses for each concentration as it occurred. It can be seen, in Table 1, that both programs appear to have converging results with as few as 100 simulations.

Next, it is necessary to test the random effects portion of the programs. For this step, concentrations for four of the seven subjects were used. For each combination of subject, concentration, $\sigma$ level ($\sigma$ = 0.5, 1.5, 2.0, 2.5 and 3.0) and ($\alpha$, $\beta$) pair [values of ($\alpha$, $\beta$) used were as follows: (-10, -2), (-12.5, -2.75), (-15, -3.5),

Table 1. Simulation Size Necessary

| Number of simulations | | MCNR | PQBC |
|---|---|---|---|
| 100 | Intercept | -14.9606 | -15.0603 |
| | Slope | -3.4918 | -3.5031 |
| 1000 | Intercept | -15.0111 | -15.0100 |
| | Slope | -3.5028 | -3.5066 |

(-17.5, -4.25), and (-20, -5)] the following process was used to generate simulation data sets:

Step 1: Generated a random number, $\gamma$, from the $N(0, \sigma^2)$ distribution.

Step 2: Generated a binomial probability using the following model:

$$p = \frac{e^{\alpha+\beta*conc+\gamma}}{1+e^{\alpha+\beta*conc+\gamma}} \qquad (7)$$

Step 3: Used this generated probability to randomly generate data from the binomial distribution with n equal to 15 and the value generated in step 2 for each time the subject/concentration combination occurred. This gave the ability to weight the different concentrations properly for each subject.

This process was repeated 100 times, so that 100 different data sets were generated for each individual combination of subject, concentration, $\sigma$ level and ($\alpha$, $\beta$) pair.

In Table 2, models for ten subjects with a standard deviation of 0.5, 1.5, 2.0, 2.5, and 3.0 are considered. The MCNR program tends to overestimate the slope and intercept, while the PQBC program tends to estimate the slope and intercept accurately. The PQBC program seems to underestimate the standard deviation, yet the MCNR program tends to estimate the standard deviation fairly close to the actual value.

In Table 3, models for twenty simulated subjects with a standard deviation of 0.5, 1.5, 2.0, 2.5, and 3.0 are considered. The MCNR program tends to come close to estimating the slope and the intercept or else slightly overestimate them, while the PQBC program tends to estimate the slope and intercept rather accurately. The PQBC program seems to underestimate the standard deviation only when it is equal to 0.5 and 1.0, otherwise it estimates the standard deviation fairly well. The MCNR program tends to estimate the standard deviation fairly close to the actual value.

Upon considering both of these tables, it is observed that the MCNR program better estimates the standard deviation then the PQBC program does. Both methods do a good job of estimating the slope and intercept; however, the PQBC program cannot accurately estimate the random effect term effectively when the number of subjects is small. It should also be noted that there does exist a procedure in SAS that has recently been developed to fit a general linear mixed effects model. The problem with this procedure is that it currently allows for only one random effect. Therefore, it will not be used here as it has the potential for two random effects, one for subject and one for session.

Based on these findings, it was decided that the MCNR program would be the best program to use to try to fit the actual data since the number of subjects that is present is seven.

Table 2. Simulation Results with Random Effects and 10 subjects

| **sigma=0.5** | | **MCNR** | | | **PQBC** | | |
|---|---|---|---|---|---|---|---|
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -9.9580 | -1.9960 | 0.4441 | -9.9587 | -1.9895 | 0.1402 |
| -12.5 | -2.75 | -12.6137 | -2.7720 | 0.5213 | -12.4233 | -2.7599 | 0.1585 |
| -15 | -3.5 | -14.9111 | -3.4827 | 0.5129 | -14.8682 | -3.5199 | 0.1846 |
| -17.5 | -4.25 | -17.7383 | -4.3082 | 0.5748 | -17.5716 | -4.2465 | 0.1306 |
| -20 | -5 | -20.0249 | -4.9879 | 0.6037 | -19.7835 | -4.9974 | 0.1310 |
| **sigma=1.5** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -9.6906 | -1.9466 | 1.2642 | -10.1027 | -2.0090 | 1.4606 |
| -12.5 | -2.75 | -11.4029 | -2.5562 | 1.2313 | -12.5513 | -2.7654 | 1.3046 |
| -15 | -3.5 | -13.8923 | -3.2693 | 1.2745 | -14.6752 | -3.5230 | 1.6457 |
| -17.5 | -4.25 | -15.9715 | -3.9428 | 1.3471 | -17.3032 | -4.2515 | 0.9298 |
| -20 | -5 | -18.6952 | -4.6878 | 1.3556 | -19.0461 | -4.9920 | 0.6949 |
| **sigma=2.0** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.5403 | -1.7130 | 1.5299 | -10.6636 | -1.9815 | 1.1236 |
| -12.5 | -2.75 | -10.4584 | -2.3086 | 1.4860 | -11.6325 | -2.7387 | 2.7422 |
| -15 | -3.5 | -12.9367 | -3.0179 | 1.6964 | -15.0677 | -3.5473 | 1.9940 |
| -17.5 | -4.25 | -15.5526 | -3.7921 | 1.7620 | -19.0472 | -4.3063 | 1.8099 |
| -20 | -5 | -17.7169 | -4.4776 | 1.7517 | -21.2486 | -5.0573 | 1.8953 |
| **sigma=2.5** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.3160 | -1.7365 | 1.8244 | -8.8011 | -1.9977 | 3.0365 |
| -12.5 | -2.75 | -10.1496 | -2.2878 | 1.9534 | -13.0257 | -2.7455 | 4.0708 |
| -15 | -3.5 | -12.6502 | -3.0247 | 1.9906 | -15.9416 | -3.4950 | 2.6698 |
| -17.5 | -4.25 | -15.6909 | -3.7334 | 2.0197 | -17.1718 | -4.2794 | 2.2737 |
| -20 | -5 | -16.3706 | -4.0743 | 1.9568 | -22.0580 | -5.0519 | 3.5395 |
| **sigma=3.0** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.4546 | -10 | -2 | -8.4546 | -10 | -2 |
| -12.5 | -2.75 | -10.0818 | -12.5 | -2.75 | -10.0818 | -12.5 | -2.75 |
| -15 | -3.5 | -12.3014 | -15 | -3.5 | -12.3014 | -15 | -3.5 |
| -17.5 | -4.25 | -15.6226 | -17.5 | -4.25 | -15.6226 | -17.5 | -4.25 |
| -20 | -5 | -16.2475 | -20 | -5 | -16.2475 | -20 | -5 |

Table 3. Simulation Results with Random Effects and 20 subjects

| **sigma=0.5** | | **MCNR** | | | **PQBC** | | |
|---|---|---|---|---|---|---|---|
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -9.9054 | -1.9816 | 0.4930 | -9.4116 | -1.9873 | 0.2649 |
| -12.5 | -2.75 | -12.4214 | -2.7306 | 0.4904 | -12.5022 | -2.7393 | 0.2649 |
| -15 | -3.5 | -14.9141 | -3.4801 | 0.4938 | -15.0879 | -3.4892 | 0.2626 |
| -17.5 | -4.25 | -17.6045 | -4.2722 | 0.5487 | -17.4891 | -4.2707 | 0.2662 |
| -20 | -5 | -20.2569 | -5.0673 | 0.6018 | -19.9316 | -4.9978 | 0.2381 |
| **sigma=1.5** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -9.3067 | -10 | -2 | -9.3067 | -10 | -2 |
| -12.5 | -2.75 | -11.6219 | -12.5 | -2.75 | -11.6219 | -12.5 | -2.75 |
| -15 | -3.5 | -13.9817 | -15 | -3.5 | -13.9817 | -15 | -3.5 |
| -17.5 | -4.25 | -16.2716 | -17.5 | -4.25 | -16.2716 | -17.5 | -4.25 |
| -20 | -5 | -19.3602 | -20 | -5 | -19.3602 | -20 | -5 |
| **sigma=2.0** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.9972 | -10 | -2 | -8.9972 | -10 | -2 |
| -12.5 | -2.75 | -11.2885 | -12.5 | -2.75 | -11.2885 | -12.5 | -2.75 |
| -15 | -3.5 | -13.4287 | -15 | -3.5 | -13.4287 | -15 | -3.5 |
| -17.5 | -4.25 | -16.0496 | -17.5 | -4.25 | -16.0496 | -17.5 | -4.25 |
| -20 | -5 | -18.2798 | -20 | -5 | -18.2798 | -20 | -5 |
| **sigma=2.5** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.6322 | -1.7685 | 2.1067 | -10.0411 | -1.9679 | 3.0753 |
| -12.5 | -2.75 | -10.8068 | -2.3196 | 2.2214 | -13.0177 | -2.7644 | 3.7780 |
| -15 | -3.5 | -13.0848 | -3.1166 | 2.1031 | -15.2309 | -3.5379 | 3.1254 |
| -17.5 | -4.25 | -15.9064 | -3.6013 | 2.2602 | -17.6589 | -4.2495 | 2.1665 |
| -20 | -5 | -16.8736 | -4.1663 | 2.1241 | -19.0147 | -5.0061 | 1.8221 |
| **sigma=3.0** | | | | | | | |
| True int | True slope | Int | Slope | S.D. | Int | Slope | S.D. |
| -10 | -2 | -8.7646 | -1.7309 | 2.6432 | -9.2498 | -1.9716 | 3.1413 |
| -12.5 | -2.75 | -10.1691 | -2.2667 | 2.4309 | -12.1977 | -2.7603 | 2.8307 |
| -15 | -3.5 | -12.8800 | -3.0994 | 2.5139 | -14.7616 | -3.5233 | 2.9032 |
| -17.5 | -4.25 | -15.8659 | -3.5443 | 2.3783 | -17.6541 | -4.3037 | 3.6437 |
| -20 | -5 | -16.3182 | -4.1412 | 2.4833 | -20.0524 | -5.0549 | 3.5551 |

Final model

Now the generalized linear mixed-effects models will be applied to the actual odor detection data. This begins by performing the MCNR analysis with both session and subject being random variables, while concentration remains the fixed variable. Subjects were chosen as a random effect because our previous analysis found that each subject did yield a different model. One of the main purposes of including random effects is to accommodate different subjects in one model with the random term. Session was chosen as a second possible random effect because it was somewhat significant in our early analysis of the data, and it is random in that the subjects may vary slightly from one session to another.

Five of the subjects had similar coefficients for their individual fixed-effects models,

$$p_{ij} = \frac{e^{\alpha+\beta*conc_{ij}}}{1+e^{\alpha+\beta*conc_{ij}}} \qquad (8)$$

For example, slope estimates for the five subjects were -0.600, -0.695, -0.487, -0.872, and -0.471 while intercept estimates were -2.881, -3.880, -1.468, -4.381, and -2.457. For the remaining two subjects, slope estimates were -3.090 and -2.287, while intercept estimates were -11.577 and -9.225;. It could be speculated that these two groupings indicate that there are two categories of smellers and that it might prove useful to split the group of seven into these two separate groups to lessen the variability of the data for modeling purposes. This began, however, by keeping all seven subjects together and estimating a model. The general form of the model is

$$p_{ij} = \frac{e^{\alpha+\beta*conc_{ij}+\gamma_{subj}+\gamma_{sess}}}{1+e^{\alpha+\beta*conc_{ij}+\gamma_{subj}+\gamma_{sess}}} \qquad (9)$$

where $\gamma_{sess}$ could be zero.

From Table 4, it is evident that the variability was quite large when all seven subjects were together and hence the resulting model had extraordinarily small coefficients. Thus, the data was split into two groups and estimated a separate model for each group; the results appear in Table 4. Therefore, several models based on the split groups of subjects will

Table 4. Final Models

| No. of subjects in model | Both subject and session as random effects | Only subject as random effect | No random effect |
|---|---|---|---|
| 7 | $\alpha$ = -145.8455, $\beta$ = -40.2636, $\gamma_{subject}$=N(0,0.8069), $\gamma_{sess.}$=N(0,0.7281) | $\alpha$ = -368.0081, $\beta$ = -80.3047, $\gamma_{subject}$=N(0,2.7066) | $\alpha$ = -453.2103, $\beta$ = -98.5135 |
| 5 | $\alpha$ = -3.4457, $\beta$ = -0.7252, $\gamma_{subject}$=N(0,1.0605), $\gamma_{sess.}$=N(0,0.1730) | $\alpha$ = -3.5403, $\beta$ = -1.0846, $\gamma_{subject}$=N(0,4.6057) | $\alpha$ = -2.9907, $\beta$ = -0.6833 |
| 2 | $\alpha$ = -3.5596, $\beta$ = -0.6995, $\gamma_{subject}$=N(0,1.1177), $\gamma_{sess.}$=N(0,0.0013) | $\alpha$ = -2.7960, $\beta$ = -0.8711, $\gamma_{subject}$=N(0,2.3808) | $\alpha$ = -3.1377, $\beta$ = -0.6414 |

be considered to see which will yield a better fitting model.

Vonesh, Chinchilli and Pu (1996) observed that for a generalized linear mixed-effects model, a valid measure of the goodness of fit of the model is given by $r_c$. There are several advantages to using $r_c$ over other methods. First, it does not require the specification of a null model. Second, it measures the level of concordance between $y_i$ and $\hat{y}_i$. A higher value of $r_c$ indicates a better fitting model. (Vonesh, Chinchilli & Pu, 1996)

For the three models that involved the 5 subjects, the $r_c$ follows: for the one with two random effects the $r_c$ was 0.418, for the model with only subject as a random effect the $r_c$ was slightly better (higher) with a value of 0.445 and for the model with no random effects the $r_c$ was only 0.342. Thus, for the 5-subject group, the best model is the one that includes only subject as the random effect. Upon looking at the three models that involved the 2 subjects the $r_c$ follows: for the one with two random effects it was 0.371, for the model with only subject as a random effect the $r_c$ was not quite as good with a value of 0.316 and the model with no random effects included yielded an $r_c$ of 0.318. Thus, for the 2 subject group, the best model is the one that includes both subject and session as the random effects.

Therefore, the conclusion is drawn that the data for all seven subjects can be best represented using the following two models. For the group of five subjects the best model is

$$p_{ij} = \frac{e^{-3.5403-1.0846 * conc_{ij} + \gamma_{subject}}}{1 + e^{-3.5403-1.0846 * conc_{ij} + \gamma_{subject}}} \quad (10)$$

where $\gamma$=N(0, 4.6057) and for the group of two subjects the best model is

$$p_{ij} = \frac{e^{-3.5596-0.6995 * conc_{ij} + \gamma_{sub} + \gamma_{sess}}}{1 + e^{-3.5596-0.6995 * conc_{ij} + \gamma_{sub} + \gamma_{sess}}} \quad (11)$$

where $\gamma_{sub}$~N(0, 1.1177) and $\gamma_{sess}$~N(0, 0.0013). Thus, based on this small sample of individuals, it was found that two models will adequately represent the whole sample of seven individuals as opposed to the idea of finding a single model for each subject. It also would make it very difficult to adequately model the population as a whole if there had been individual models for each subject.

## Conclusion

How accurate are people at detecting odors? In general terms, the question could also be stated as sensitivity – what is lowest concentration needed for reliable, if not perfect, detection? From there, there was an attempt to characterize all seven subjects with one generalized linear mixed effects model.

Two methods of fitting the generalized linear mixed effects models were used. Simulations were conducted to discover which method would yield the best results, in terms of stable estimates and a high $r_c$ value, for the data. It was discovered that for this data, the method that would yield the best results was the MCNR method. Once this method was implemented, it was discovered the data was best fit by two models as opposed to just one model. The subjects were split into one group of five and one group of two based on the results discovered in the initial portion of the simulation study. For the group of five, it was necessary to have a random effects term for the subjects and for the group of two, it was necessary to have a random effects term for the subjects and another for the sessions.

Thus, all seven subjects' odor detection ability was able to be modeled for the one chemical tested through the use of two models. This is an improvement over the seven models that were initially investigated. The benefit of the smaller number of models is that it allows one to represent a population's ability to detect odors with just a few models instead of a different model for each individual in the population.

It would be instructive to perform a study with a larger sample in which the same task was asked of participants as in this study, namely: Do you detect an odor or not? An ideal situation would be to have many subjects of each gender and in each age group. This would

allow an expansion of these models to attempt to include a term for gender and also for age. Some researchers have hypothesized that as humans age, they begin to lose their sense of smell (Doty, 1994). Others (Hales, 1999) have wondered if sensitivity varies with gender. By expanding the model, it would begin to answer these questions.

## References


Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9-25.

Clayton, D. (1992). Generalized linear mixed models in biostatistics. *Statistician, 41*, 327-328.

Devanand, DP, Michaels-Marston, KS, Liu, X, Pelton, GH, Padilla, M, Marder, K, Bell, K, Stern, Y, & Mayeux, R (2000). Olfactory deficits in patients with mild cognitive impairment predict Alzheimer's disease at follow-up. *The American Journal of Psychiatry*, *15,* 1399-1405.

Doty RL. (1994). Olfactory dysfunction in the elderly and in Alzheimer's disease. In (Kurihara K, Suzuki N, Ogawa H, eds.) *Olfaction and taste XI: proceedings of the 11th International Symposium on Olfaction and Taste and of the 27th Japanense Symposium on Taste and Smell*. Joint meeting held at Kosei-nenkin Kaikan, Sapporo, Japan, July 12-16, 1993. Tokyo: Springer-Verlag, 597-601.

Environmental Protection Agency (1989). Report to congress on indoor air quality, Vol. ii. Assessment and Control of Indoor Air Pollution

Fisk W. J. & Rosenfeld, A. H. (1997). Estimates of improved productivity and health from better indoor environments *Indoor Air, 7*, 158-172.

Hales, D (1999). *Just Like a Woman: How Gender Science is Revealing What Makes Us Female*. New York, NY: Bantam Books.

Jiang, J. (1998). Consistent Estimators in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *93*, 720-729.

Lin, X & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, *91,* 1007-1016.

Lindstrom, M. J. & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Boimetrics*, *46*, 673-687.

Mann, Arnold. (1998, December 21). This Place Makes me Sick. *Time Magazine*, 152 (25).

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association, 92,* 162-170.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., & Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics, 21*, 1087-1091.

Prah, J. D., S. B. Sears & J. C. Walker (1995) Modern approaches to air-dilution olfactometry. In: *Handbook of Olfaction and Gustation*, R. L. Doty (Ed.) Marcel Dekker, New York, pp. 227-255.

Vonesh, E. F. (1996). A note on the use of laplace's approximation for nonlinear mixed-effects models. *Biometrika*, *83*, 447-452.

Vonesh, E. F., Chinchilli, V.M. & Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics, 52,* 572-587.

Walker, JC, Hall, SB, Walker, DB, Kendal-Reed, MS, Hood, AF & Niu, X-F (2003). Human odor detectability: new methodology used to determine threshold and variation. *Chemical Senses*, *28,* 817-826.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 619-629
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# A Weighted Moving Average Process for Forecasting

Shou Hsing Shih Chris P. Tsokos
University of South Florida

A forecasting model for a nonstationary stochastic realization is proposed based on modifying a given time series into a new k-time moving average time series. The study is based on the autoregressive integrated moving average process along with its analytical constrains. The analytical procedure of the proposed model is given. A stock XYZ selected from the Fortune 500 list of companies and its daily closing price constitute the time series. Both the classical and proposed forecasting models were developed and a comparison of the accuracy of their responses is given.

Key words: ARIMA, moving average, stock, time series analysis

## Introduction

Time series analysis and modeling plays a very important role in forecasting, especially when our initial stochastic realization is nonstationary in nature. Some of the interesting and useful publications related to the subject area are Akaike (1974), Banerjee et al. (1993), Box et al. (1994), Brockwell and Davis (1996), Dickey and Fuller (1979), Dickey et al. (1984), Durbin and Koopman (2001), Gardner et al. (1980), Harvey (1993), Jones (1980), Kwiatkowski et al. (1992), Rogers (1986), Said and Dickey (1984), Sakamoto et al. (1986), Shumway and Stoffer (2006), Tsokos (1973), Wei (2006).

The purpose of this study is to begin with a given time series that characterizes an economic or any other natural phenomenon and as usual, is nonstationary. Box and Jenkins (1994) developed a popular and useful classical procedure to develop forecasting models that have been shown to be effective. In this article, a procedure for developing a forecasting model that is more effective than the classical approach is introduced. For a given stationary or nonstationary time series, $\{x_t\}$, generate a k-day moving average time series, $\{y_t\}$, and the developmental process begins.

Certain basic concepts and analytical methods are reviewed that are essential in structuring the proposed forecasting model. The review is based on the autoregressive integrated moving average processes. The accuracy of the proposed forecasting model is illustrated by selecting from the list of Fortune 500 companies, company XYZ, and considering its daily closing prices for 500 days. The classical time series model for the subject information along with the proposed process was developed. A statistical comparison based on the actual and forecasting residuals is given, both in tabular and graphical form.

Shou Hsing Shih recently received the Ph. D. in Statistics from University of South Florida. Shih's research mainly concentrates on time series forecasting. E-mail address: sshih3@tampabay.rr.com. Chris P. Tsokos is Distinguished University Professor of Mathematics and Statistics at the University of South Florida. He is the author of more than 250 research publications. E-mail address: profcpt@cas.usf.edu

## The Proposed Forecasting Model: k-th Moving Average

It is not appropriate to build a time series model without conforming to certain mathematical constrains, such as stationarity of a given stochastic realization. Almost always, the time series that is given is nonstationary in nature and then, the next step is to reduce it into being stationary. Let $\{x_t\}$ be the original time series. The difference filter is given by

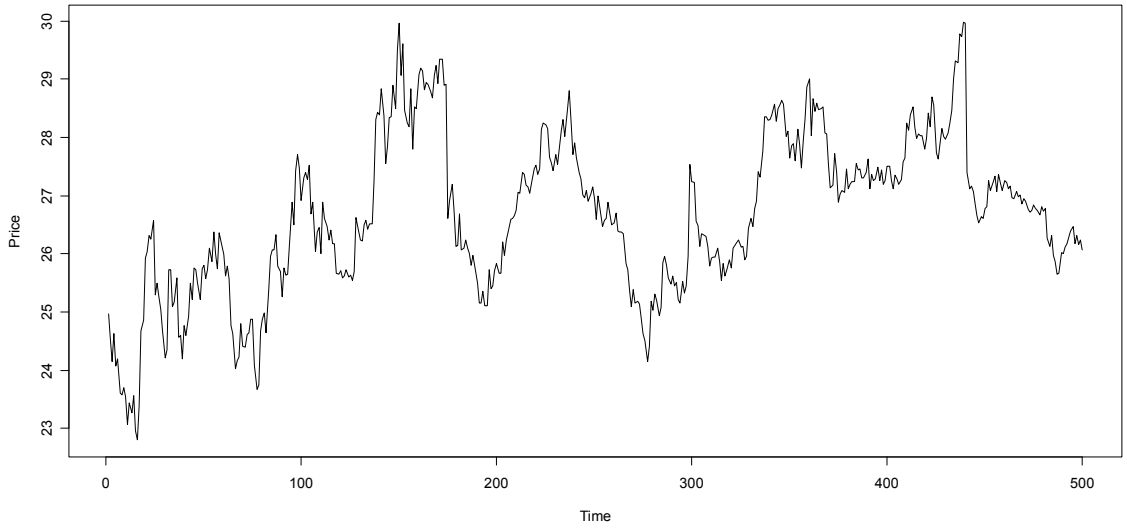



Figure 1. Daily Closing Price for Stock XYZ

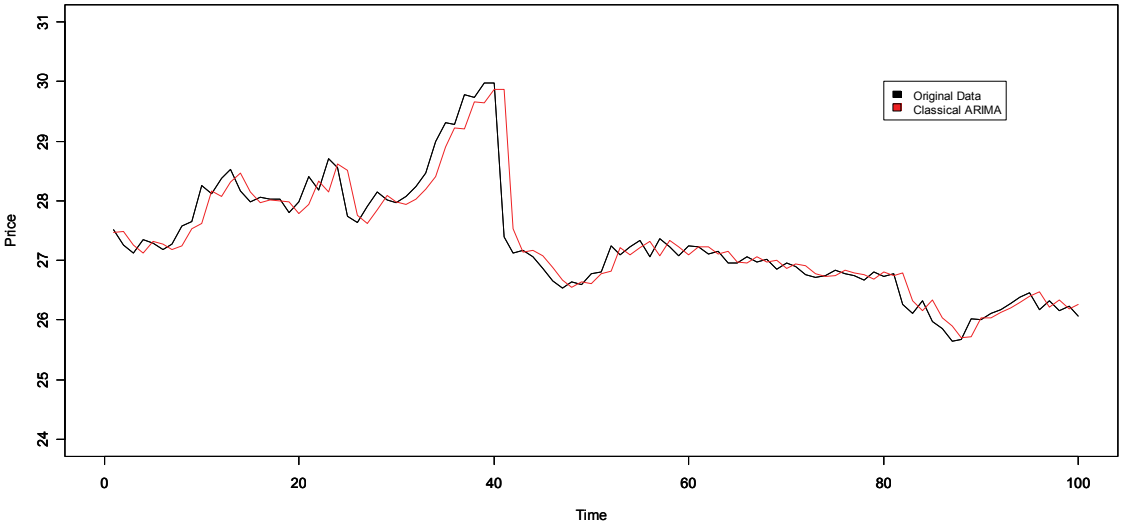



Figure 2. Comparisons on Classical ARIMA Model VS. Original Time Series for the Last 100 Observations

$$(1-B)^d \qquad (1)$$

where $B^j x_t = x_{t-j}$, and d is the degree of differencing of the series.

The primary use for the k-th moving average process is for smoothing a realized time series. It is very useful in discovering a short-term, long-term trends and seasonal components of a given time series. The k-th moving average process of a time series $\{x_t\}$ is defined as follows:

$$y_t = \frac{1}{k}\sum_{j=0}^{k-1} x_{t-k+1+j} \qquad (2)$$

where $t = k, k+1, ..., n$.

As k increases, the number of observations k of $\{y_t\}$ decreases, and $\{y_t\}$ gets closer to the mean of $\{x_t\}$ as k increases. When $k = n$, $\{y_t\}$ reduces to only a single observation, and equals $\mu$, that is

$$y_t = \frac{1}{n}\sum_{j=1}^{n} x_j = \mu \qquad (3)$$

Then, develop the proposed model by transforming the original time series $\{x_t\}$ into $\{y_t\}$ by applying (2). After establishing the new time series, usually nonstationary, begin the process of reducing it into a stationary time series. Kwiatkowski, et al. (1992) introduced the KPSS Test to check the level of stationarity of a time series. Apply the differencing order d to the new time series $\{y_t\}$ for $d = 0,1,2,...$, then verify the stationarity of the series with the KPSS test until the series become stationary. Therefore, one can reduce the nonstationary time series into a stationary one after a proper number of differencing. Then proceed the model building procedure of developing the proposed forecasting model.

After choosing a proper degree of differencing d, proceed with the model building process by assuming different orders for the autoregressive integrated moving average model, ARIMA(p,d,q), also known as Box and Jenkins method, where (p,d,q) represent the order of the autoregressive process, the order of differencing and the order of the moving average process, respectively. The ARIMA(p,d,q) is defined as:

$$\phi_p(B)(1-B)^d y_t = \theta_q(B)\varepsilon_t \qquad (4)$$

where $\{y_t\}$ is the realized time series, $\phi_p$ and $\theta_q$ are the weights or coefficients of the AR and MA that drive the model, respectively, and $\varepsilon_t$ is the random error. Write $\phi_p$ and $\theta_q$ as

$$\phi_p(B) = \qquad (5)$$
$$(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)$$

and

$$\theta_q(B) = \qquad (6)$$
$$(1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q)$$

Sometimes it is difficult to make a decision in selecting the best order of the ARIMA(p,d,q) model when there are several models that all adequately represent a given set of time series. Hence, Akaile's information criterion (AIC) (1974), plays a major role when it comes to model selection. AIC was introduced by Akaike in 1973, and it is defined as:

$$\text{AIC}(M) = \qquad (7)$$
$$\text{-2 ln [maximum likelihood] + 2}M,$$

where M is the number of parameters in the model and the unconditional log-likelihood function suggested by Box, Jenkins, and Reinsel (1994), is given by

$$\ln L(\phi, \mu, \theta, \sigma_\varepsilon^2) = \qquad (8)$$

Table 1. Basic Evaluation Statistics

| $\bar{r}$ | $S_r^2$ | $S_r$ | $S_r/\sqrt{n}$ |
|---|---|---|---|
| 0.02209169 | 0.1445187 | 0.3801562 | 0.0170011 |

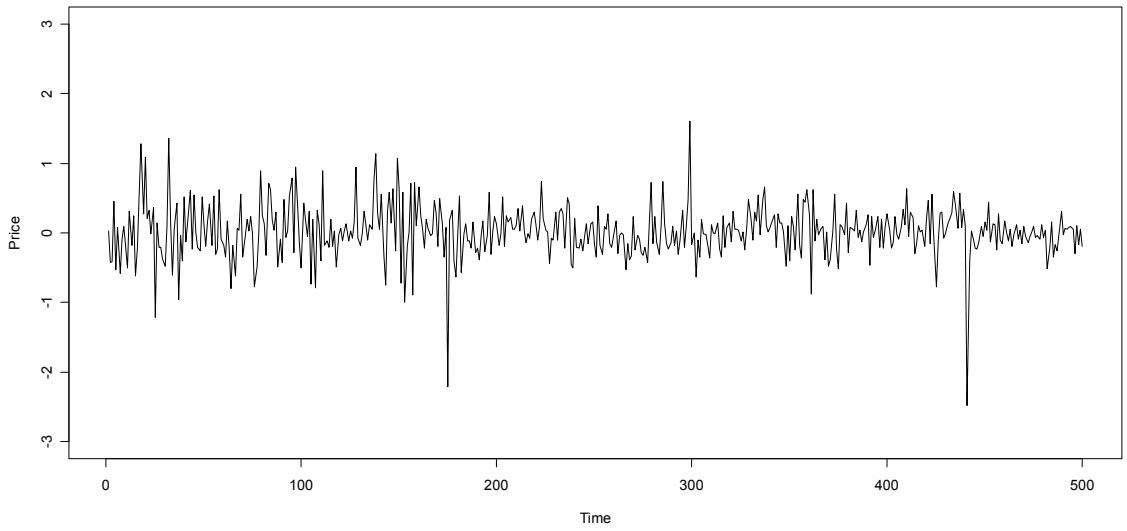



Figure 3. Time Series Plot of the Residuals for Classical Model

$$-\frac{n}{2}\ln 2\pi\sigma_\varepsilon^2 - \frac{S(\phi,\mu,\theta)}{2\sigma_\varepsilon^2}$$

where $S(\phi,\mu,\theta)$ is the unconditional sum of squares function given by

$$S(\phi,\mu,\theta) = \sum_{t=-\infty}^{n}[E(\varepsilon_t|\phi,\mu,\theta,y)]^2 \quad (9)$$

where $E(\varepsilon_t|\phi,\mu,\theta,y)$ is the conditional expectation of $\varepsilon_t$ given $\phi,\mu,\theta,y$.

The quantities $\hat{\phi}$, $\hat{\mu}$, and $\hat{\theta}$ that maximize (8) are called unconditional maximum likelihood estimators. Because $\ln L(\phi,\mu,\theta,\sigma_\varepsilon^2)$ involves the data only through $S(\phi,\mu,\theta)$, these unconditional maximum likelihood estimators are equivalent to the unconditional least squares estimators obtained by minimizing $S(\phi,\mu,\theta)$. In practice, the summation in (9) is approximated by a finite form

$$S(\phi,\mu,\theta) = \sum_{t=M}^{n}[E(\varepsilon_t|\phi,\mu,\theta,y)]^2 \quad (10)$$

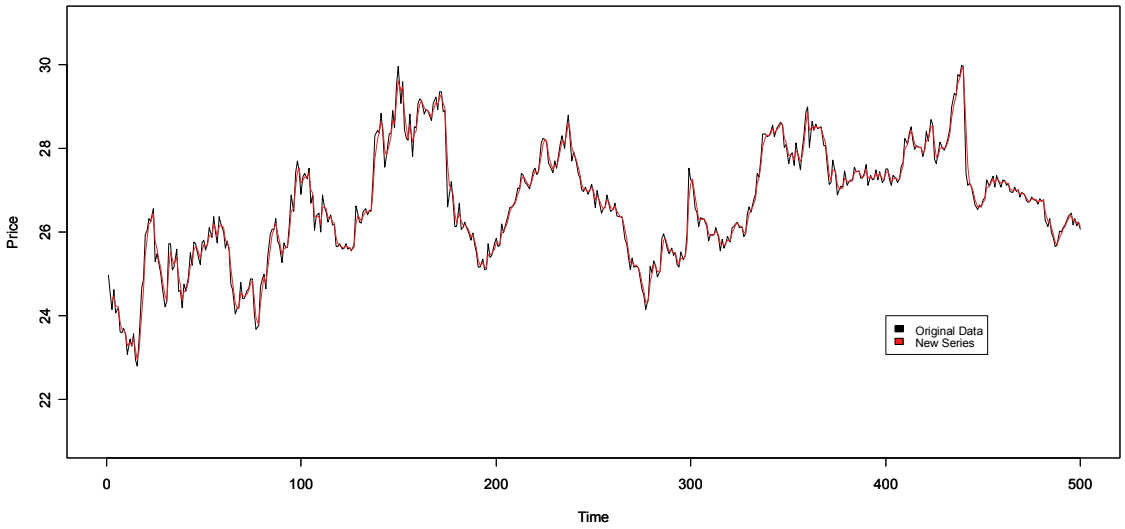



Figure 4. Three Days Moving Average on Daily Closing Price of Stock XYZ Vs. the original time series

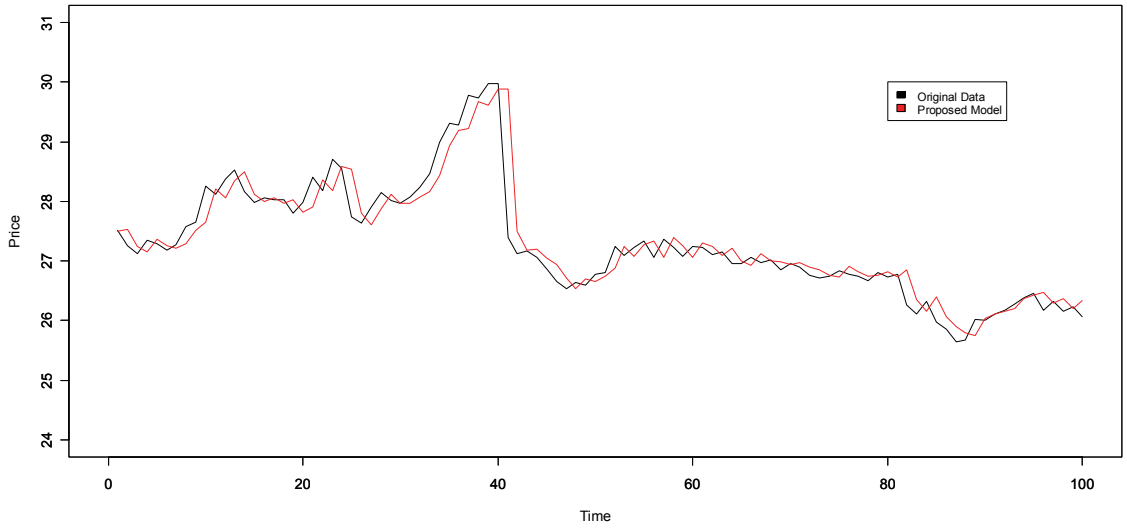



Figure 5. Comparisons on Our Proposed Model VS. Original Time Series for the Last 100 Observations

Table 2. Actual and Predicted Price

| N | Actual Price | Predicted Price | Residuals |
|---|---|---|---|
| 476 | 26.78 | 26.8473 | -0.0673 |
| 477 | 26.75 | 26.7976 | -0.0476 |
| 478 | 26.67 | 26.7673 | -0.0972 |
| 479 | 26.8 | 26.6922 | 0.1078 |
| 480 | 26.73 | 26.8064 | -0.0764 |
| 481 | 26.78 | 26.7490 | 0.0310 |
| 482 | 26.27 | 26.7911 | -0.5211 |
| 483 | 26.12 | 26.3277 | -0.2077 |
| 484 | 26.32 | 26.1631 | 0.1569 |
| 485 | 25.98 | 26.3364 | -0.3564 |
| 486 | 25.86 | 26.0349 | -0.1749 |
| 487 | 25.65 | 25.9068 | -0.2568 |
| 488 | 25.67 | 25.6670 | 0.0031 |
| 489 | 26.02 | 25.7119 | 0.3081 |
| 490 | 26.01 | 26.0335 | -0.0235 |
| 491 | 26.11 | 26.0427 | 0.0674 |
| 492 | 26.18 | 26.1343 | 0.0457 |
| 493 | 26.28 | 26.2032 | 0.0768 |
| 494 | 26.39 | 26.2986 | 0.0914 |
| 495 | 26.46 | 26.4043 | 0.0557 |
| 496 | 26.18 | 26.4743 | -0.2943 |
| 497 | 26.32 | 26.2219 | 0.0981 |
| 498 | 26.16 | 26.3354 | -0.1754 |
| 499 | 26.24 | 26.1953 | 0.0447 |
| 500 | 26.07 | 26.2602 | -0.1902 |

where M is a sufficiently large integer such that the backcast increment $\left|E(\varepsilon_t|\phi,\mu,\theta,y) - E(\varepsilon_{t-1}|\phi,\mu,\theta,y)\right|$ is less than any arbitrary predetermined small $\varepsilon$ value for $t \le -(M+1)$. This expression implies that $E(\varepsilon_t|\phi,\mu,\theta,y) \cong \mu$; hence, $E(\varepsilon_t|\phi,\mu,\theta,y)$ is negligible for $t \le -(M+1)$.

After obtaining the parameter estimates $\hat{\phi}$, $\hat{\mu}$, and $\hat{\theta}$, the estimate $\hat{\sigma}_\varepsilon^2$ of $\sigma_\varepsilon^2$ can then be calculated from

$$\hat{\sigma}_\varepsilon^2 = \frac{S(\hat{\phi},\hat{\mu},\hat{\theta})}{n} \qquad (11)$$

For an ARMA(p,q) model based on n observations, the log-likelihood function is

$$\ln L = -\frac{n}{2}\ln 2\pi\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}S(\phi,\mu,\theta) \qquad (12)$$

Proceed to maximize (12) with respect to the parameters $\phi,\mu,\theta,$ and $\sigma_\varepsilon^2$, from (11),

$$\ln \hat{L} = \qquad (13)$$

•

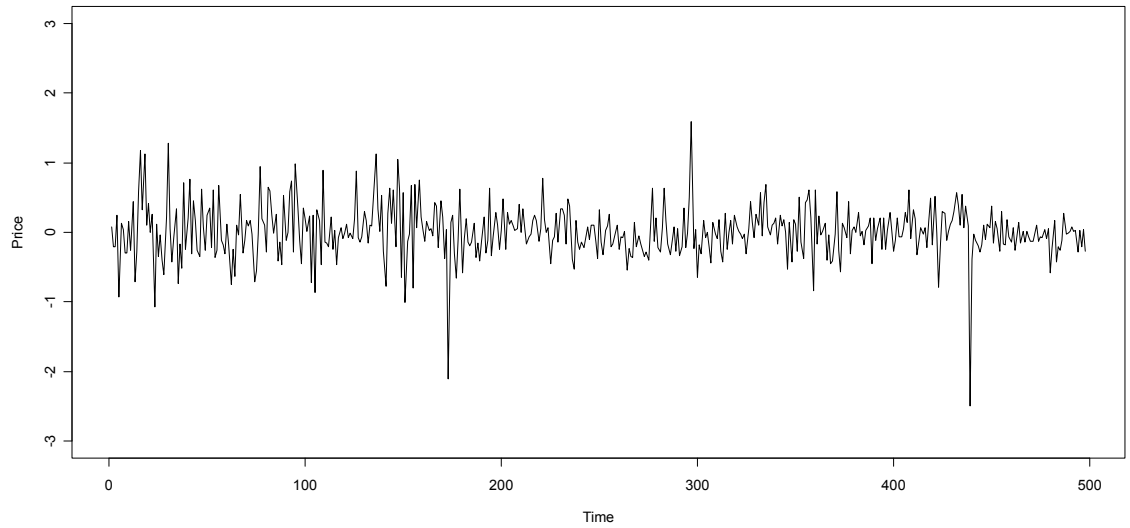



Figure 6. Time Series Plot for Residuals for Our Proposed Model

Table 3. Basic Evaluation Statistics

| $\bar{r}$ | $S_r^2$ | $S_r$ | $S_r/\sqrt{n}$ |
|---|---|---|---|
| 0.01016814 | 0.1437259 | 0.3791119 | 0.01698841 |

$$-\frac{n}{2}\ln\hat{\sigma}_\varepsilon^2-\frac{n}{2}(1+\ln 2\pi)$$

Because the second term in expression (13) is a constant, reduce the AIC to

$$\text{AIC}(M) = n\ln\hat{\sigma}_\varepsilon^2+2M\ . \qquad (14)$$

Thus, we an appropriate time series model is generated and the statistical process with the smallest AIC can be selected. The model identified will possess the smallest average mean square error. The development of the model is summarized as follows.

Transform the original time series $\{x_t\}$ into a new series $\{y_t\}$.

- Check for stationarity of the new time series $\{y_t\}$ by determining the order of differencing d, where $d = 0,1,2,...$ according to KPSS test, until stationarity is achieved.
- Decide the order $m$ of the process, for this case, let $m = 5$ where $p+q=m$.
- After (d, $m$) is selected, list all possible set of (p, q) for $p+q \le m$.
- For each set of (p, q), estimate the parameters of each model, that is, $\phi_1,\phi_2,...,\phi_p,\theta_1,\theta_2,...,\theta_q$
- Compute the AIC for each model, and choose the one with smallest AIC.

Table 4. Actual and Predicted Price

| N | Actual Price | Predicted Price | Residuals |
|---|---|---|---|
| 476 | 26.78 | 26.8931 | -0.1131 |
| 477 | 26.75 | 26.7715 | -0.0215 |
| 478 | 26.67 | 26.7121 | -0.0421 |
| 479 | 26.8 | 26.7239 | 0.0761 |
| 480 | 26.73 | 26.7854 | -0.0554 |
| 481 | 26.78 | 26.6892 | 0.0908 |
| 482 | 26.27 | 26.8292 | -0.5592 |
| 483 | 26.12 | 26.3027 | -0.1827 |
| 484 | 26.32 | 26.0808 | 0.2392 |
| 485 | 25.98 | 26.3603 | -0.3803 |
| 486 | 25.86 | 25.9868 | -0.1268 |
| 487 | 25.65 | 25.8443 | -0.1943 |
| 488 | 25.67 | 25.7115 | -0.0414 |
| 489 | 26.02 | 25.6499 | 0.3701 |
| 490 | 26.01 | 25.9650 | 0.0450 |
| 491 | 26.11 | 26.0526 | 0.0574 |
| 492 | 26.18 | 26.0912 | 0.0888 |
| 493 | 26.28 | 26.1449 | 0.1351 |
| 494 | 26.39 | 26.3090 | 0.0810 |
| 495 | 26.46 | 26.3752 | 0.0848 |
| 496 | 26.18 | 26.4223 | -0.2423 |
| 497 | 26.32 | 26.2461 | 0.0739 |
| 498 | 26.16 | 26.2964 | -0.1364 |
| 499 | 26.24 | 26.1437 | 0.0963 |
| 500 | 26.07 | 26.2678 | -0.1978 |

According to the criterion mentioned above, the ARIMA(p,d,q) model can be obtained that best fit a given time series, where the coefficients are $\phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q$. Using the model that we developed for $\{y_t\}$ and subject to the AIC criteria, we forecast values of $\{y_t\}$ and proceed to apply the back-shift operator to obtain estimates of the original phenomenon $\{x_t\}$, that is,

$$\hat{x}_t = k\hat{y}_t - x_{t-1} - x_{t-2} - ... - x_{t-k+1} \quad (15)$$

The proposed model and the corresponding procedure discussed in this section shall be illustrated with real economic application and the results will be compared with the classical time series model.

The proposed model and the corresponding procedure discussed in this section shall be illustrated with real economic application and the results will be compared with the classical time series model.

Table 5. Basic Comparison on Classical Approach Vs. Our Proposed model

| | $\bar{r}$ | $S_r^2$ | $S_r$ | $S_r/\sqrt{n}$ |
|---|---|---|---|---|
| Classical | 0.02209169 | 0.1445187 | 0.3801562 | 0.0170011 |
| Proposed | 0.01016814 | 0.1437259 | 0.3791119 | 0.01698841 |

First, a time series forecasting model is developed of the given nonstationary data using the ordinary Box and Jenkins methodology. Secondly, the data are modified, Figure 1, to develop the proposed time series forecasting model. A comparison of the two models will be given.

The general theoretical form of the ARIMA(p,d,q) is given by

$$\phi_p(B)(1-B)^d x_t = \theta_q(B)\varepsilon_t \qquad (16)$$

Following the Box and Jenkins' methodology (1994), the classical forecasting model with the best AIC score is the ARIMA(1,1,2). That is, a combination of first order autoregressive (AR) and a second order moving average (MA) with a first difference filter. Write it as

$$(1-.9631B)(1-B)x_t = (1-1.0531B+.0581B^2)\varepsilon_t \qquad (17)$$

After expanding the autoregressive operator and the difference filter,

$$(1-1.9631B+.9631B^2)x_t = (1-1.0531B+.0581B^2)\varepsilon_t \qquad (18)$$

and rewrite the model as

$$x_t = 1.9631x_{t-1} - .9631x_{t-2} + \varepsilon_t - 1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2} \qquad (19)$$

by letting $\varepsilon_t = 0$, there is the one day ahead forecasting time series of the closing price of stock XYZ as

$$\hat{x}_t = 1.9631x_{t-1} - .9631x_{t-2} - 1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2} \qquad (20)$$

Using the above equation, graph the forecasting values obtained by using the classical approach on top of the original time series, as shown by Figure 2.

The basic statistics that reflect the accuracy of model (20) are the mean $\bar{r}$, variance $S_r^2$, standard deviation $S_r$ and standard error $S_r/\sqrt{n}$ of the residuals. Figure 3 gives a plot of the residual and Table 1 gives the basic statistics.

Furthermore, restructure the model (20) with $n = 475$ data points to forecast the last 25 observations only using the previous information. The purpose is to see how accurate our forecast prices are with respect to the actual 25 values that have not been used. Table 2 gives the actual price, predicted price, and residuals between the forecasts and the 25 hidden values. The average of these residuals is $\bar{r} = -0.05608$. Proceed to develop the proposed forecasting model. The original time series of stock XYZ daily closing prices is given by Figure 1. The new time series is being created by $k = 3$ days moving average and the analytical form of $\{y_t\}$ is given by

$$y_t = \frac{x_{t-2} + x_{t-1} + x_t}{3} \tag{21}$$

Figure 4 shows the new time series $\{y_t\}$ along with the original time series $\{x_t\}$, that will be used to develop the proposed forecasting model. The best model that characterizes the behavior of $\{y_t\}$ is ARIMA (2,1,3). That is,

$$(1-.8961B-.0605B^2)(1-B)y_t = \quad (22)$$
$$(1+.0056B-.0056B^2-B^3)\varepsilon_t$$

Expanding the autoregressive operator and the first difference filter, we have

$$(1-1.8961B+.8356B^2+.0605B^3)y_t = \quad (23)$$
$$(1+.0056B-.0056B^2-B^3)\varepsilon_t$$

Thus, write (23) as

$$y_t = \tag{24}$$
$$1.8961y_{t-1}-.8356y_{t-2}-.0605y_{t-3}$$
$$+\ \varepsilon_t+.0056\varepsilon_{t-1}-.0056\varepsilon_{t-2}-\varepsilon_{t-3}$$

The final analytical form of the proposed forecasting model can be written as

$$\hat{y}_t = \tag{25}$$
$$1.8961y_{t-1}-.8356y_{t-2}-.0605y_{t-3}$$
$$+.0056\varepsilon_{t-1}-.0056\varepsilon_{t-2}-\varepsilon_{t-3}$$

Using the above equation, a plot of the developed model (25), showing a one day ahead forecasting along with the new time series, $\{y_t\}$, is displayed by Figure 5.
Note the closeness of the two plots that reflect the quality of the proposed model.

Similar to the classical model approach that we discussed earlier, use the first 475 observations $\{y_1, y_2, ..., y_{475}\}$ to forecast $\hat{y}_{476}$. Then, use the observations $\{y_1, y_2, ..., y_{476}\}$ to forecast $\hat{y}_{477}$, and continue this process until forecasts are obtained for all the observations, that is, $\{\hat{y}_{476}, \hat{y}_{477}, ..., \hat{y}_{500}\}$. From equation (21), the relationship can be seen between the forecasting values of the original series $\{x_t\}$ and the forecasting values of 3 days moving average series $\{y_t\}$, that is,

$$\hat{x}_t = 3\hat{y}_t - x_{t-1} - x_{t-2} \tag{26}$$

Hence, after $\{\hat{y}_{476}, \hat{y}_{477}, ..., \hat{y}_{500}\}$ is estimated, use the above equation, (26), to solve the forecasting values for $\{x_t\}$. Figure 6 is the residual plot generated by the proposed model, and followed by Table 3, that includes the basic evaluation statistics.

Both of the above displayed evaluations reflect on accuracy of the proposed model. The actual daily closing prices of stock XYZ from the 476th day along with the forecasted prices and residuals are given in Table 4. The results given above attest to the good forecasting estimates for the hidden data.

Comparison of the Forecasting Models

In this section, the two developed models are compared. The classical process is given by

$$\hat{x}_t = 1.9631x_{t-1} - .9631x_{t-2} - \tag{27}$$
$$1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2}$$

In the proposed model, the following inversion is used to obtain the estimated daily closing prices of stock XYZ, that is,

$$\hat{y}_t = \tag{28}$$
$$1.8961y_{t-1}-.8356y_{t-2}-.0605y_{t-3}$$
$$+.0056\varepsilon_{t-1}-.0056\varepsilon_{t-2}-\varepsilon_{t-3}$$

in conjunction with

$$\hat{x}_t = 3\hat{y}_t - x_{t-1} - x_{t-2} \qquad (29)$$

Table 5 is a comparison of the basic statistics used to evaluate the two models under investigation. The average mean residuals between the two models shown that the proposed model is overall approximately 54% more effective in estimating one day ahead the closing price of Fortune 500 stock XYZ.

## Conclusion

Based on the average mean residuals the proposed model was significantly more effective in such term of predicting of the closing daily prices of stock XYZ.

## References


Akaike, H. (1974). A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, *AC-19*, 716-723.

Banerjee, A., Dolado, J. J., Galbraith, J. W., & Hendry, D. F. (1993). *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ., 89-99.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994) *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ., 224-247.

Brockwell, P. J., & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*., Springer, New York., Sections 3.3 and 8.3.

Dickey, D. A., & Fuller, W. A. (1979) Distribution and the Estimators for Autoregressive Time Series With a Unit Root., *Journal of the American Statistical Association*, Vol. 74, No. 366, 427-431.

Dickey, D. A., Hasza, D. P., & Fuller, W. A. (1984). Testing for Unit Roots in Seasonal Time Series., *Journal of the American Statistical Association*, Vol. 79, No. 386, 355-367.

Durbin, J., & Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*., Oxford University Press.

Gardner, G., Harvey, A. C., & Phillips, G. D. A. (1980). Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering., *Applied Statistics*, *29*, 311-322.

Harvey, A. C. (1993). *Time Series Models*, 2nd Edition, Harvester Wheatsheaf., sections 3.3 and 4.4.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations., *Technometrics*, *20*, 389-395.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root., *Journal of Econometrics*, *54*, 159-178.

Rogers, A. J. (1986). Modified Lagrange Multiplier Tests for Problems with One-Sided Alternatives, *Journal of Econometrics*, North-Holland., *31*, 341-361.

Said, S. E., & Dickey, D. A. (1984) Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order., *Biometrika*, *71*, 599-607.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*., D. Reidel Publishing Company.

Shumway, R. H., & Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: with R Examples*, 2nd ed., Springer, New York.

Tsokos, C. P. (1973). Forecasting Models from Non-Stationary Time Series-Short Term Predictability of Stocks., *Mathematical Methods in Investment and Finance*., North Holland Publishing Co., 520-63.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed., Pearson Education, Inc.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 630-644
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Longitudinal Evaluation of Estimates in an Establishment Survey After Ration Imputation

Adriana Pérez
University of Louisville

Researchers evaluated a ratio imputation technique used at the US Survey of Graduate Students and Postdoctorates in Science and Engineering, which is an annually conducted cross-sectional establishment survey. Standardized bias was used, mean square error and relative bias to appraise this imputation method on point and variance estimates via simulations.

Key words: Total estimate, variance estimation, establishment data, nonresponse, simulations.

## Introduction

Nonresponse in establishment surveys is an ongoing problem (Kovar & Whitridge, 1995). The problem of nonresponse affects estimates of survey statistics (Little & Rubin DB, 2002; Rubin, 1987; Kovar, et al., 1995; Ruggles & Joint Economic Committee, 2006; Groves, Dillman, Eltinge, & Little, 2002; Groves, et al., 2004). Many imputation methods used in social, demographic and health science settings have been applied within the economic survey framework and very little information is known about the effect of item nonresponse in establishment surveys (Kovar, et al., 1995; Judkins , 2000; West, Butani, & Witt, 1993). There has been a focus on procedures for reducing measurement error, improving sampling strategies (Lee & Croal, 1989), improving estimators (Sirken & Shimizu, 1999), improving response rates (Chun, 1997), response selection, survey coordination, longitudinal analysis(Ruggles, et al., 2006), (Schenker, Treiman, & Weidman, 1988; Heeringa & Lepkowski, 1986), or empirical evaluation of imputation methods (West, et al., 1993; Krenzke, Montaquila, & Mohadjer, 2000; Mueller & Butani, 1995) in establishment surveys.

Many imputation methods are available in the literature. Usually, once a dataset has been imputed, analyses are performed treating the imputed values as observed data. This type of analysis could be misleading because variances and covariances may be underestimated (Kovar, et al., 1995). In this article, the effectiveness of a particular ratio imputation method when applied to an item-nonresponse from an establishment survey including a longitudinal perspective on point and variance estimates is evaluated.

There are a variety of techniques for variance estimation for complex surveys (Wolter, 1985) and few of them incorporate the effect of imputation in their estimation (Shao & Sitter, 1996; Shao & Steel, 1999; Shao, 2002). Most of the time imputation methods in a survey are implemented without theoretical development of the methods (Shao, 2002). Simulation studies make it possible to evaluate and compare estimation techniques in national surveys in any country (U.S.Department of Education.National Center for Education Statistics., 2001). Pseudo-universes from survey

Adriana Pérez is Associate Professor of Biostatistics at the University of Louisville, School of Public Health and Information Sciences, Department of Bioinformatics and Biostatistics. This research was conducted when Dr. Pérez was at the University of Texas Health Science Center at Houston. Her research interests are in statistical methods to handle missing data; statistical methods for epidemiology research; design, conduct and analysis of multi-center clinical trials; sample size estimation and modeling strategies. Email her at adriana.perez@louisville.edu

data can be used instead of national universes (i.e., census data) which are not usually available for simulation studies. Pseudo universes permit a comparison of techniques and sample according to a plan of interest, maintaining the distributions of the variables of interest. Simulations from a pseudo universe can provide estimates of interest and give detailed insight of the estimator performance.

It is the researcher's interest to study the effect on the point and variance estimates of the current imputation plan conducted in the Graduate Students and Postdoctorates in Science and Engineering (GSS)(NSF-NIH, 2005). One of the challenging aspects of any simulation is the creation of an artificial population similar to the one investigated. There are two approaches to create a finite population universe(Katzoff, Jones, & Curtin, 1988; Bernaards, Belin, & Schafer, 2006; Schafer, et al., 1996). One is to create pseudo-random values from an actual multivariate probability model, also known as a hypothetical population. The second is to use an actual large data set to reflect the target population and to define population parameters of interest, also known as a pseudo-universe. Use of a specific probability model is a limitation in the creation of a hypothetical population(Schafer, et al., 1996). Therefore, a pseudo universe was created to impose realistic missing data patterns.

The following describes the generation of the pseudo universe and simulations which allow: (i) appraise the longitudinal missing data patterns in GSS between 1999-2001; (ii) evaluate the effect of current imputation methods in this survey on estimates for different missing data mechanism assumptions in GSS; (iii) assess precision and accuracy measures in the total, and corresponding variance estimates in GSS. In following sections, the GSS survey will be described, the current imputation method, the methodology to evaluate the effect of the current imputation method and the results and conclusions, respectively.

## The Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS)

One of the current surveys conducted at the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) is the NSF-NIH (National Institutes of Health) survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) (NSF-NIH, 2005). This survey (i) measures academic department level information on all U.S. institutions offering graduate programs (masters or PhD degrees) in science, engineering, or health selected field; (ii) provides a description of graduate science and engineering (S&E) student's enrollment in US institutions; and (iii) assesses trends in financial support patterns and shifts in graduate enrollment and postdoctoral appointments.

This cross-sectional establishment survey is conducted annually (NSF-NIH, 2005). Reports from this survey are presented in current year and historical data, setting up a longitudinal structure (National Science Foundation & Division of Science Resources Statistics (SRS), 2005). Total estimates for domains and sub-domains are reported (National Science Foundation & Division of Science Resources Statistics, 2006). Each year, a ratio imputation technique is used to handle item nonresponse based on inflator/deflator factors (NSF-NIH, 2005). For a particular year, these inflator/deflator factors are computed from the current year observed data in combination with previous year observed and imputed data (Morgan M & ORC Macro, 2004). Replacing missing data in the current year with previous year data is an imputation method known in longitudinal human population studies as the last observation carry-forward (LOCF). This imputation method is modified in the GSS by the use of inflator/deflator factors as adjustments when replacing current cycle missing data with adjusted previous cycle data.

Simulations conducted with LOCF, in longitudinal human population studies, indicates that LOCF produces biased estimates for all three types of missing data mechanisms (Missing completely at random (MCAR), Missing at random (MAR) or Missing not at random (MNAR)) and LOCF produces the smallest standard errors that are biased downward (Gadbury, Coffey, & Allison, 2005). For these reasons, evaluation of the current GSS imputation plan is needed.

Imputation at the Graduate Student Survey

The department within an academic institution is the unit of interest of this survey for imputation purposes. This imputation methodology is presented for four variables used in this research only, but can be generalizable to the rest of the variables within this survey.

Creation of inflator/deflator factors

Departments that provided full or partial information about total full-time students, total part-time students, total postdoctorates and total other non-faculty research staff are used for creation of these factors. Specifically, in this study, total full-time students and total part-time students were used. Inflator/deflator factors are computed by highest institutional degree level (doctorate and master's) and by department type (e.g. Biology, Physics, etc.). For a particular variable of interest ($Y_k$), its sum is computed by institutional highest degree level and department type. Then factors are computed by dividing the sum of the variable from the current (t) year by the corresponding sum of the variable from the previous year (t-1). These inflator/deflator factors ($\hat{\psi}_{k_t}$) in mathematical terms are calculated for the k[th] variable and year t.

$$\hat{\psi}_{k_t} = \sum_{j=1}^{r} Y_{jk_t} \Big/ \sum_{j=1}^{r} Y_{jk_{(t-1)}} \qquad (1)$$

$r$ identifies the maximum number of departments in the same institutional degree level and departmental type that provided a variable value $Y_k$ in both years $t$ and $t-1$. Any computed factor less than 0.85 or greater than 1.15 is set to 1 for imputation purposes. In mathematical terms:

$$\hat{\phi}_{k_t} = \begin{cases} \hat{\psi}_{k_t} & \text{if } 0.85 < \hat{\psi}_{k_t} < 1.15 \\ 1 & \text{otherwise} \end{cases} \qquad (2)$$

Using inflator/deflator factors to impute total full- time students all sources of funding and total part-time students of all races

Departments with missing information in total full- time students and/or total part-time students are imputed using equation 3. The imputation value for a particular variable in the current year is obtained by applying $\hat{\phi}_{k_t}$ to previous year information for that variable. This is done at each department institutional level (i.e., MS or PhD) and department type (i.e., Biology, Physics, etc).

$$\hat{Y}_{I(ik_t)} = \hat{\phi}_{k_t} * Y_{ik_{(t-1)}} \qquad (3)$$

$i$ identifies a particular department, $k$ identifies the variable, $t$ identifies the year, $\hat{\phi}_{k_t}$ identifies the inflator/deflator factors, $Y_{ik_{(t-1)}}$ is the $k^{\text{th}}$ variable value for department $i$, in year $t-1$; and $\hat{Y}_{I(ik_t)}$ is the imputed value of the $k^{\text{th}}$ variable for department $i$ at year $t$.

Subsequently, imputed values for total full-time students (from equation 3) are used to impute variables regarding full-time students by: source and mechanism of support. Similarly, imputed values for total part-time students (from Equation 3) are used to impute variables regarding number of part-time students by sex and their distribution by US nationals/permanent residents or foreign students.

Using inflator/deflator factors to impute total part-time students

The imputed value for the total of female part-time students is computed using the same percentage as reported in the previous year on the imputed value from the total part-time students in the current year. Equation 4 shows this in mathematical terms.

$$\hat{Y}_{I(ij_t)} = \hat{Y}_{i_t} \left( Y_{ij_{(t-1)}} \Big/ Y_{i_{(t-1)}} \right) \qquad (4)$$

Where $i$ identifies a particular department, $t$ identifies the year, $j$ identifies women, $\hat{Y}_{i_t}$ represents the observed or imputed value of the total part-time students enrolled for a particular department $i$ at year $t$, $Y_{ij_{(t-1)}}$ represents the observed value of the total part-time women students for year $t-1$ at particular department $i$, $Y_{i_{(t-1)}}$ represents the observed value of the total part-time students for year $t-1$ at a

particular department $i$, and $\hat{Y}_{I(ijt)}$ represents the imputed value of the total part-time women at year $t$ at particular department $i$.

The imputed value for the total of male part-time students is calculated as the difference between the total part-time students in the current year (observed or imputed) and the observed or imputed value for the total of female part-time students.

## Methodology

The purpose is to evaluate the longitudinal effect of imputation on estimates in the GSS, data from the years 1998-2002 (the most recent data through 2005). The GSS survey in 1998 contained 639 variables and 11686 departments and in 2002 contained 639 variables and 12126 departments. Overall, 15379 departments reported any information on the GSS data from the years 1998--2002. The first four variables imputed in this survey were selected for analysis in this research: total full-time graduate students all sources of support, total part-time students of all races, total part-time male students of all races and total part-time female students of all races. To evaluate the effect of the imputation method within this survey a simulation study with a pseudo universe from this survey was conducted.

### Generation of the pseudo universe 1998-2002

A dataset called "Observed 1998-2002" which mainly excluded departments with unit nonresponse between years 1998-2002 was created. If a department reported any missing value for any of the variables of interest in year 1998 and 2002 they were excluded. This is because stable departments were to be used, to exclude new programs (i.e., if a department created a new master or doctoral program in 2002 then previous years would not have reported any information and missing values would appear in the longitudinal structure), and to exclude non current programs. If departments provided information in years 1998 and 2002 this indicated continuity of the master's or PhD degree program at that institution. In summary, one department was excluded because it did not report the type of academic institution (neither school under which this department was associated, nor public or private nor which institutional highest degree is granted). Departments with unit nonresponse were excluded for each year as follows: 3693 within 1998, 936 within 1999, 514 within 2000, 755 within 2001 respectively and 610 within 2002. It was assumed that these departments with unit nonresponse were not stable. Furthermore, the study excluded 328 departments without students enrolled either full-time or part-time in 1998 in any of the four variables of interest, which indicated historically unstable enrollment in that program. This dataset Observed 1998-2002 contained 8542 out of 15379 departments with item nonresponse between the years 1999 and 2001. Using this dataset the researchers generated the longitudinal distributional patterns of missing data in years 1999-2001.

After this, the researchers generated a pseudo universe from this survey by removing any department with missing data in our variables of interest from years 1999-2001. Researchers excluded 685 departments because they did not report full time students for at least one of these years. Forty-five departments that did not report part-time students for at least one of these years were deleted. Furthermore, 127 departments that did not report part time male students for at least one of these years were excluded.

This complete dataset was called and used as Pseudo Universe 1998-2002 and contains 7685 departments with complete information on all these variables. This pseudo universe was used to develop and evaluate the imputation methods used in GSS for the variable totals and their corresponding variability measures. Total estimates coming from this pseudo-universe were treated as parameter values from this pseudo universe. This is notated $\theta_{k_t}$ as the total estimate of the $k^{\text{th}}$-variable of interest for years 1999 to 2001. These parameter values were used for comparison purposes in evaluation the GSS imputation methods.

### Simulation of mechanisms of missingness

Two missingness mechanisms to evaluate the imputation methods at GSS were

Table 1: Actual percentages of missing values in dataset "Observed 1998-2002"

| Year | N | Full time students all sources of support | Part Time students of all races | Part time male students of all races | Part time female students of all races |
|---|---|---|---|---|---|
| 1999 | 11832 | 1.49 | 1.58 | 3.26 | 3.25 |
| 2000 | 11899 | 1.58 | 1.60 | 1.99 | 1.99 |
| 2001 | 11968 | 3.53 | 3.77 | 4.12 | 4.12 |

explored. The first approach was to create an MCAR mechanism.

Actual percentages of missing values were imposed within "Pseudo Universe 1998-2002" on within Pseudo Universe 1998-2002 on each $Y_{k_t}$ independently of any variable in the system. Table 1 illustrates the "Actual" percentage of missing data observed in years 1999—2001 for the four variables of interest in this survey and these percentages were used for creating the MCAR mechanism for evaluation purposes. Our "MCAR dataset" contains these "Actual" percentages imposed randomly as missing. As you may notice these percentages are not high and it will be desired to evaluate the imputation method with this low percentages of missingness.

It was assumed that the occurrence of missing values at the GSS survey is MAR. Under this assumption, the second approach was to impose the Actual percentages of missing values with the same longitudinal distributional patterns of missing data in years 1999-2001 from Observed 1998-2002 within Pseudo Universe 1998-2002. Table 2 shows the observed longitudinal patterns of missing values for these variables, where 0 represents data was missing and 1 represents data was observed.

For the purposes of understanding the effect of the imputation method with increased percentages of missing values, in simulations, the researchers increased these observed longitudinal distributional patterns of missing values from the Observed 1998-2002 in 25%, 50%, 75% and 100% (data not shown) within Pseudo Universe 1998-2002.

Parameter estimation

These datasets, with imposed missing values, can be used to examine many quantities of interest. The total and its corresponding variance estimate were examined for each year. Many other parameters were included in simulations but are not reported for brevity and the research primarily presents the results under the MAR mechanism. Each one of these missingness mechanisms were replicated one thousand times.

Applying the Imputation Method

Inflator/deflator factors for year 1999 were computed using the observed data from the 1998 Pseudo Universe 1998-2002. The ratio imputation methods described in equations 3 and 4 were applied for missing values in year 1999. Then, an imputed and complete 1999 dataset was reached. Similarly, the researchers continued to generate the inflator/deflator factors and to impute missing values in years 2000 and 2001. This procedure produced an Observed and imputed longitudinal 1999-2001 dataset. Cross-sectional 1999-2001 total estimates and their corresponding variances were computed. Estimates after imputation are notated as $\hat{\theta}_{AI_{k_t}}$ for each $k^{\text{th}}$ variable on years 1999--2001.

Evaluation criteria

The performance of the GSS imputation method by the following quantities in years 1999-2001 were evaluated. First, the bias of the total and the variance estimates after imputation of the simulations are described in Equations 5 and 6, respectively.

Table 2. Percentages missing values for each pattern in dataset “Observed 1998-2002”.

| Pattern | Full time students all sources of support | | | Part Time students of all races | | | Part time male students of all races | | | Part time female students of all races | | | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1999 | 2000 | 2001 | 1999 | 2000 | 2001 | 1999 | 2000 | 2001 | 1999 | 2000 | 2001 | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.05 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0.08 |
| 3 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0.01 |
| 4 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0.01 |
| 5 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0.01 |
| 6 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0.89 |
| 7 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0.01 |
| 8 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.06 |
| 9 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.52 |
| 10 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.01 |
| 11 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.01 |
| 12 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.89 |
| 13 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.01 |
| 14 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0.02 |
| 15 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.01 |
| 16 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0.04 |
| 17 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 2.60 |
| 18 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.01 |
| 19 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0.12 |
| 20 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.12 |
| 21 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0.05 |
| 22 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0.01 |
| 23 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0.74 |
| 24 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0.01 |
| 25 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0.21 |
| 26 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0.07 |
| 27 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0.18 |
| 28 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 93.26 |

$$\text{Bias of the total estimate}_{k_t} = E\left(\hat{\theta}_{AI_{k_t}}\right) - \theta_{k_t} \quad (5)$$

$$\text{Bias\_variance\_total}_{k_t} = E\left(\text{Var}(\hat{\theta}_{AI_{k_t}})\right) - \sigma^2_{k_t} \quad (6)$$

where $\sigma^2_{k_t}$ identifies the population variance among the data [$\text{Var}(Y_{k_t})$] within the "Pseudo Universe 1998-2002" and not the variance of the mean estimates. $\text{Var}(\hat{\theta}_{AI_{k_t}})$ identifies the estimated variance after imputation. Second, given that the raw bias can be misleading the standardized bias of the total estimate using equation 7 was computed. A standardized bias of less of 50% in both directions should be considered practically insignificant. Third, the mean square error (MSE) for the total and the variance estimates are described in equations 8 and 9.

$$\text{Standardized Bias}_{k_t} = 100 * \frac{E\left(\hat{\theta}_{AI_{k_t}}\right) - \theta_{k_t}}{SE(\hat{\theta}_{AI_{K_t}})} \quad (7)$$

$$\text{MSE(Total)} = \left(\frac{1}{1000}\right)\sum_{m=1}^{1000}(\hat{\theta}_{AI_{k_t}} - \theta_{k_t})^2 \quad (8)$$

$$\text{MSE(Var)} = \left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\left(\text{Var}(\hat{\theta}_{AI_{k_t}}) - \sigma^2_{k_t}\right)^2 \quad (9)$$

Fourth, the average relative bias of the total and the variance estimates are described in equations 10 and 11. These average relative biases measure the average magnitude of over or under estimation of the imputation method compared with the true value. Finally, the average relative stability of the variance is described in equation 12.

$$\text{Relative Bias of the total} = \left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\frac{(\hat{\theta}_{AI_{k_t}} - \theta_{k_t})}{\theta_{k_t}} \quad (10)$$

$$\text{Relative Bias of the variance} = \left(\frac{1}{10000}\right)\sum_{m=1}^{1000}\left(\frac{\text{Var}(\hat{\theta}_{AI_{k_t}}) - \sigma^2_{k_t}}{\sigma^2_{k_t}}\right) \quad (11)$$

$$\text{Relative Stability} = \frac{\left[\left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\left(\text{Var}(\hat{\theta}_{AI_{k_t}}) - \text{MSE(Var)}\right)^2\right]^{1/2}}{\text{MSE(Var)}} \quad (12)$$

## Results

Table 3 presents the results of the 1000 simulations under MCAR mechanism. The current imputation method underestimates total full-time students and total part-time female students and overestimates part-time students and part-time male students under this mechanism. The underestimation or overestimation of these variables increased yearly from 1999 to 2001. The standardized biases were larger than 50% for many of the variables of interest.

Results of simulations under the MAR mechanism are presented in Tables 4-7. Table 4 shows the results from the evaluation criteria for the imputation method on full-time students all sources of funding. The relative bias of the total estimate of full-time students indicates a 10% underestimation for years 2000 and 2001 with the current amount of missing values. If the amount of missing values increases then this underestimation increased up to 20% for year 2001. It is interesting to note that this imputation method would overestimate the total estimate of full-time students by 40% if the current patterns of missing values were increased by 100% for the year 2000.

Results from the relative bias of the variance of the total estimate of full-time students across the years indicates overestimation between 10% and 30% for year 1999 for increasing percentages of missing values. This overestimation is also observed for year 2001 with a range of 20% to 70%.

Table 3. Results from 1000 replicates under MCAR

| Year | 1999 | 2000 | 2001 |
|---|---|---|---|
| **Bias of the total** | | | |
| Full time students all sources of support | -878 | -665 | -3,629 |
| Part Time students of all races | 1,177 | 2,115 | 2,347 |
| Part time male students of all races | 617 | 1,439 | 2,192 |
| Part time female students of all races | -276 | -478 | -1,250 |
| **Bias of the variance** | | | |
| Full time students all sources of support | -6 | -1 | -78 |
| Part Time students of all races | 6 | 5 | 4 |
| Part time male students of all races | 1 | 4 | 8 |
| Part time female students of all races | 2 | 1 | 16 |
| **MSE of the total** | | | |
| Full time students all sources of support | 818 | 1,372 | 9,166 |
| Part Time students of all races | 98 | 267 | 329 |
| Part time male students of all races | 3 | 34 | 2332 |
| Part time female students of all races | 7 | 22 | 13,175 |
| **MSE of the variance** | | | |
| Full time students all sources of support | 1.11E+06 | 1.11E+06 | 1.45E+07 |
| Part Time students of all races | 1.47E+06 | 4.68E+06 | 5.78E+06 |
| Part time male students of all races | 3.94E+05 | 2.12E+06 | 4.90E+06 |
| Part time female students of all races | 9.09E+04 | 2.67E+05 | 1.72E+06 |
| **Standardized Bias of the variance** | | | |
| Full time students all sources of support | -149.8 | -81.5 | -313.9 |
| Part Time students of all races | 415.5 | 460.8 | 454.0 |
| Part time male students of all races | 526.5 | 653.4 | 734.2 |
| Part time female students of all races | -227.5 | -244.2 | -313.9 |

Table 4. Results from 1000 replicates under MAR for full time students

| Year | Actual% | 25% | 50% | 75% |
|---|---|---|---|---|
| **Bias of the total** | | | | |
| **1999** | -10812 | -15523 | -15156 | -19867 |
| **2000** | -12994 | -22657 | -13833 | -19737 |
| **2001** | -29314 | -46229 | -43331 | -28207 |
| **Bias of the variance** | | | | |
| **1999** | 9.2E+09 | 2.1E+10 | 1.2E+10 | 3.0E+10 |
| **2000** | -2.1E+10 | -3.1E+10 | -2.4E+10 | -2.8E+10 |
| **2001** | 5.4E+10 | 2.7E+10 | 7.2E+10 | 1.0E+11 |
| **Standardized bias** | | | | |
| **1999** | -11.9 | -15.3 | -14.1 | -16.6 |
| **2000** | -10.8 | -16.5 | -9.1 | -12.3 |
| **2001** | -15.3 | -21.1 | -18.9 | -11.3 |
| **MSE of the total** | | | | |
| **1999** | 8.4E+06 | 1.0E+07 | 1.2E+07 | 1.5E+07 |
| **2000** | 1.5E+07 | 1.9E+07 | 2.3E+07 | 2.6E+07 |
| **2001** | 3.8E+07 | 5.0E+07 | 5.4E+07 | 6.3E+07 |
| **MSE of the variance** | | | | |
| **1999** | 5.3E+19 | 8.4E+19 | 7.6E+19 | 1.1E+20 |
| **2000** | 8.8E+19 | 1.2E+20 | 1.4E+20 | 1.5E+20 |
| **2001** | 4.6E+20 | 5.6E+20 | 6.2E+20 | 7.6E+20 |
| **Relative Bias of the total** | | | | |
| **1999** | 0.0 | -0.1 | -0.1 | -0.1 |
| **2000** | -0.1 | -0.1 | -0.1 | -0.1 |
| **2001** | -0.1 | -0.2 | -0.2 | -0.1 |
| **Relative Bias of the Variance** | | | | |
| **1999** | 0.1 | 0.2 | 0.1 | 0.2 |
| **2000** | -0.2 | -0.2 | -0.2 | -0.2 |
| **2001** | 0.3 | 0.2 | 0.5 | 0.7 |
| **Relative Stability of the variance** | | | | |
| **1999** | 1.0 | 0.6 | 0.7 | 0.5 |
| **2000** | 0.6 | 0.5 | 0.4 | 0.4 |
| **2001** | 0.1 | 0.1 | 0.1 | 0.1 |

Table 5. Results from 1000 replicates under MAR for part time students

| Year | Actual% | 25% | 50% | 75% |
|---|---|---|---|---|
| **Bias of the total** | | | | |
| **1999** | 877 | 2366 | 4199 | 5853 |
| **2000** | 17076 | 26118 | 26247 | 37098 |
| **2001** | -9815 | -20641 | -17052 | -27822 |
| **Bias of the variance** | | | | |
| **1999** | -7.1E+09 | -4.7E+09 | -4.8E+09 | 3.2E+09 |
| **2000** | -2.7E+10 | -2.2E+10 | -2.6E+10 | -1.7E+10 |
| **2001** | -1.9E+10 | -3.0E+10 | -1.4E+10 | -1.9E+10 |
| **Standardized bias** | | | | |
| **1999** | 1.3 | 3.0 | 4.8 | 6.2 |
| **2000** | 16.5 | 23.7 | 20.9 | 28.0 |
| **2001** | -5.8 | -11.2 | -8.3 | -12.8 |
| **MSE of the total** | | | | |
| **1999** | 4.5E+06 | 6.3E+06 | 7.7E+06 | 9.0E+06 |
| **2000** | 1.1E+07 | 1.3E+07 | 1.6E+07 | 1.9E+07 |
| **2001** | 2.8E+07 | 3.4E+07 | 4.3E+07 | 4.8E+07 |
| **MSE of the variance** | | | | |
| **1999** | 1.6E+19 | 2.0E+19 | 2.2E+19 | 3.0E+19 |
| **2000** | 5.9E+19 | 6.4E+19 | 8.4E+19 | 7.3E+19 |
| **2001** | 1.8E+20 | 2.3E+20 | 2.9E+20 | 3.2E+20 |
| **Relative Bias of the total** | | | | |
| **1999** | 0.0 | 0.0 | 0.0 | 0.1 |
| **2000** | 0.2 | 0.3 | 0.3 | 0.4 |
| **2001** | -0.1 | -0.2 | -0.2 | -0.3 |
| **Relative Bias of the Variance** | | | | |
| **1999** | 0.1 | 0.2 | 0.1 | 0.2 |
| **2000** | -0.2 | -0.2 | -0.2 | -0.2 |
| **2001** | 0.3 | 0.2 | 0.5 | 0.5 |
| **Relative Stability of the variance** | | | | |
| **1999** | 1.0 | 0.8 | 0.7 | 0.5 |
| **2000** | 0.3 | 0.3 | 0.2 | 0.2 |
| **2001** | 0.1 | 0.1 | 0.1 | 0.1 |

Results from the relative bias of the variance in year 2000 indicate that this imputation method underestimates the variance of the total estimate of full-time students from 20% to 40% depending on the amount of missingness. The MSE of the total and the variance of full-time students using the current imputation method at GSS is large. The MSE of the variance increases for each year of increase and as expected if the percentage of missing values increases then the MSE of the variance will increase. The average relative stability of the variance of the total estimate of full-time students decreases noticeably for each one year increase. This behavior is consistently observed across increasing percentages of missing values.

Table 5 shows the results from the evaluation criteria for the imputation method on part-time students of all races. The relative bias of the total estimate of part-time students indicates a 20% overestimation for year 2000 and a 10% underestimation for year 2001 with the current amount of missing values. If the amount of missing values increases then this

Table 6. Results from 1000 replicates under MAR for part time male students

| Year | Actual% | 25% | 50% | 75% |
|---|---|---|---|---|
| **Bias of the total** | | | | |
| **1999** | 40137 | 51315 | 61580 | 75489 |
| **2000** | 53259 | 66665 | 77239 | 92581 |
| **2001** | 40137 | 51358 | 66098 | 76572 |
| **Bias of the variance** | | | | |
| **1999** | 9.1E+09 | 1.2E+10 | 1.4E+10 | 2.3E+10 |
| **2000** | 9.1E+09 | 1.4E+10 | 1.9E+10 | 2.8E+10 |
| **2001** | 8.5E+09 | 8.2E+09 | 2.0E+10 | 2.6E+10 |
| **Standardized bias** | | | | |
| **1999** | 85.5 | 94.0 | 102.6 | 113.4 |
| **2000** | 75.8 | 91.0 | 92.6 | 107.8 |
| **2001** | 40.0 | 42.8 | 49.0 | 52.4 |
| **MSE of the total** | | | | |
| **1999** | 3.8E+06 | 5.6E+06 | 7.4E+06 | 1.0E+07 |
| **2000** | 7.8E+06 | 9.8E+06 | 1.3E+07 | 1.6E+07 |
| **2001** | 1.4E+07 | 1.7E+07 | 2.3E+07 | 2.7E+07 |
| **MSE of the variance** | | | | |
| **1999** | 3.5E+18 | 5.1E+18 | 5.6E+18 | 7.8E+18 |
| **2000** | 1.8E+19 | 1.7E+19 | 2.4E+19 | 2.0E+19 |
| **2001** | 4.1E+19 | 5.5E+19 | 6.6E+19 | 7.7E+19 |
| **Relative Bias of the total** | | | | |
| **1999** | 0.8 | 1.0 | 1.3 | 1.5 |
| **2000** | 1.1 | 1.4 | 1.6 | 2.0 |
| **2001** | 0.9 | 1.1 | 1.4 | 1.6 |
| **Relative Bias of the Variance** | | | | |
| **1999** | 0.5 | 0.7 | 0.9 | 1.4 |
| **2000** | 0.6 | 0.9 | 1.1 | 1.7 |
| **2001** | 0.5 | 0.5 | 1.2 | 1.6 |
| **Relative Stability of the variance** | | | | |
| **1999** | 1.0 | 0.7 | 0.6 | 0.5 |
| **2000** | 0.2 | 0.2 | 0.1 | 0.2 |
| **2001** | 0.1 | 0.1 | 0.1 | 0.0 |

overestimation increases up to 40% for year 2000 and the underestimation will decrease by at least 20% for year 2001. Results from the relative bias of the variance of the total estimate of part-time students across years indicates increased underestimation for increased year and this behavior seems to follow a U shape for increasing percentages of missing values. Findings about the MSE for the total and variance of full-time students are similar than for part-time students as well as regarding the average relative stability of the variance.

Table 6 shows the results from the evaluation criteria for the imputation method on part-time male students of all races. The relative bias of the total estimate of part-time male students with the current amount of missing values indicates 80%, 110% and 90% overestimation for years 1999, 2000 and 2001, respectively.

Table 7. Results from 1000 replicates under MAR for part time female students

| Year | Actual% | 25% | 50% | 75% |
|---|---|---|---|---|
| **Bias of the total** | | | | |
| **1999** | -39260 | -48949 | -57381 | -69636 |
| **2000** | -36183 | -40547 | -50992 | -55483 |
| **2001** | -829684 | -847999 | -859150 | -880394 |
| **Bias of the variance** | | | | |
| **1999** | -4.8E+09 | -4.8E+09 | -3.9E+09 | -4.9E+09 |
| **2000** | -1.8E+10 | -1.9E+10 | -2.5E+10 | -2.3E+10 |
| **2001** | -2.3E+10 | -2.9E+10 | -3.3E+10 | -4.3E+10 |
| **Standardized bias** | | | | |
| **1999** | -87.2 | -94.6 | -105.9 | -114.1 |
| **2000** | -61.6 | -62.5 | -70.1 | -71.0 |
| **2001** | -924.3 | -842.8 | -768.8 | -744.2 |
| **MSE of the total** | | | | |
| **1999** | 3.6E+06 | 5.1E+06 | 6.2E+06 | 8.6E+06 |
| **2000** | 4.8E+06 | 5.9E+06 | 7.9E+06 | 9.2E+06 |
| **2001** | 7.0E+08 | 7.3E+08 | 7.5E+08 | 7.9E+08 |
| **MSE of the variance** | | | | |
| **1999** | 2.9E+18 | 3.2E+18 | 3.6E+18 | 4.7E+18 |
| **2000** | 4.1E+18 | 5.3E+18 | 7.5E+18 | 8.8E+18 |
| **2001** | 1.1E+19 | 1.4E+19 | 1.9E+19 | 2.1E+19 |
| **Relative Bias of the total** | | | | |
| **1999** | -0.9 | -1.1 | -1.2 | -1.5 |
| **2000** | -0.8 | -0.9 | -1.1 | -1.2 |
| **2001** | -18.1 | -18.5 | -18.8 | -19.2 |
| **Relative Bias of the Variance** | | | | |
| **1999** | -0.3 | -0.3 | -0.2 | -0.3 |
| **2000** | -1.1 | -1.1 | -1.5 | -1.4 |
| **2001** | -1.3 | -1.7 | -1.9 | -2.5 |
| **Relative Stability of the variance** | | | | |
| **1999** | 1.0 | 0.9 | 0.8 | 0.6 |
| **2000** | 0.7 | 0.6 | 0.4 | 0.3 |
| **2001** | 0.3 | 0.2 | 0.2 | 0.1 |

As expected if the amount of missing values increases then this overestimation increases. Results from the relative bias of the variance of the total estimate of part-time male students across years indicates overestimation above 50% and increases for increasing percentages of missing values. Findings about the MSE for the total and variance of full-time students are equal for part-time male students and the average time male students as well as regarding the average relative stability of the variance.

Table 7 shows the results from the evaluation criteria for the imputation method on part-time female students of all races. The relative bias of the total estimate of part-time female students, with the current missing values, indicates a 90%, 80% and 1813% underestimation for 1999, 2000, and 2001.

If the amount of missing values increases then this underestimation increases as well. Results from the relative bias of the variance of the total estimate of part-time female students across years indicates underestimation between 20% and 30% for year 1999 for increasing percentages of missing values. This underestimation is also observed for years 2001 and 2002 with a range from 110% to 270%. Findings about the MSE for the total and variance of full-time students are equal for part-time female students as well as regarding the average relative stability of the variance.

## Conclusion

Overall, the bias and the MSE of the total and the variance estimates are not acceptable under the MCAR mechanism. Our findings under MCAR in this establishment survey are consistent with the literature in human populations where you will expect a higher underestimation or overestimation for increasing percentage of missing values in a variable including the increase as a year passes by.

Overall, the bias of the total estimates for full-time students and part-time students are acceptable under the MAR mechanism. This is because although the estimates across years and for different percentages of increase of current missing values are biased, the standardized biases are less than -50% which means that this bias is practically insignificant. On the contrary, the bias of the total estimates for part-time male and female students are not acceptable under the MAR mechanism using similar criteria of the standardized bias which surpass 50% in either direction for any percentage increase of missing values.

The results of overestimation for 1999 and 2001 using the relative bias of the variance of the total estimate of full-time students and its underestimation in 2000 with this imputation method are in agreement with previous descriptions of variance estimate behaviors after imputation in human population surveys, where imputation methods underestimate or overestimate depending on the variability of the variable. Most of the time it is expected to provide an underestimation of this variance estimate and this is shown in many of the variables chosen for this research.

The MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). Overall, the estimators generated with the current imputation method in GSS do not have good MSE properties because they do not have small combined variance and bias.

The findings regarding the variance estimates using the current imputation methods in this establishment survey for the variables chosen are in agreement with findings with many imputation methods for human population surveys where priority and challenges need to be overcome for improving variance estimates in surveys. The noticeable decrease in the average relative stability of the variance of the total estimates of the variables of interest warrants consideration.

There were many limitations to this study. The chosen pseudo universe represents a best case scenario where departments are fully compliant and provided full information. Furthermore, sampling did not come from this finite population to test the imputation method in full when a sample is selected instead of using the entire population. The entire population was used, which is the best case scenario, being fully efficient in the scenarios regarding the imputation method. It is expected that by selecting different sample sizes will provide worst results than the ones presented here. Also, a good scenario where the current percentages of missing values do not seem very high for each cross-sectional year was used. However, the findings are overwhelming in the large effects that the current GSS imputation method affects the bias of the total estimates of part-time males and females and overall variance estimates. Another limitation is that this study only handles the issue of item-nonresponse when unit non-response was excluded from this research. The results limitation as a best case scenario warrants consideration because worse results would be expected under worse conditions than those presented here.

Currently NSF publishes total estimates from this survey without reporting any variance estimate. Careful attention is needed for those variables where standardized biases are larger

than -50% as well as how to improve the stability of the variance decreasing for increasing percentages of missingness in the cross-sectional and longitudinal setting. Minor discrepancies were observed in the bias and MSE estimates when the unit of analysis is establishments instead of individuals. Further research is needed to identify statistical methods to handle the missing data from this survey and to evaluate this method under a missing not at random mechanism.

## Acknowledgments

This research was supported by the American Statistical Association (ASA) and the Division of Science Resources Statistics (SRS) of the National Science Foundation through the ASA/SRS-NSF Research Program to the first author in 2005. The ASA/SRS-NSF Research Program fosters collaborative and interdisciplinary research efforts. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The use of NSF data does not imply endorsement of the research methods or conclusions contained in this article. The author would like to thank Jeri M. Mulrow for her collaboration during the planning and implementation of this research. It is important to thank Jeri M. Mulrow and Dr. Lemuel A. Moye for helpful comments in a previous version of this manuscript. I also wish to thank Maria L. Fernandez and Gloria L. Sanchez, for data management and macros programming.

## References

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2006). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat.Med, Epub ahead of print*.

Chun YI (1997). Nonresponse follow-up in establishment surveys: A split-half experiment. In V. 2. Alexandria (Ed.), (pp. 988-993). American Statistical Association: Proceedings of the Section on Survey Research Methods.

Gadbury G.L, Coffey C.S, & Allison D.B. (2005). Modern statistical methods for handling repeated measurements in obesity trial data: beyond LOCF. Obesity Reviews 4[3], 175-184.

Groves RM, Dillman DA, Eltinge JL, & Little RJA (2002). *Survey Nonresponse*. New York: John Wiley & Sons.

Groves RM, Fowlwe F.J.Jr, Couper M.P, Lepkowski J.M, Singer E, & Tourangeau R (2004). *Survey Methodology*. (1 ed.) New Jersey: John Wiley & Sons, Inc.

Heeringa SG & Lepkowski JM (1986). Longitudinal imputation for the SIPP. In (pp. 206-210). Proceedings of the Survey Research Methods Section: American Statistical Association.

Judkins DR (2000). Discussion. Session 44: New Developments in Imputation of Business Survey Data. In Alexandria, VA 22314: American Statistical Association.

Katzoff MJ, Jones GK, & Curtin LR (1988). A general system for the empirical evaluation of statsitical methods for data from complex surveys. In (pp. 293-297). American Statistical Association: Proceedings of the Section on Survey Research Methods.

Kovar JG & Whitridge (1995). Imputation of Business Survey Data. In Cox BG, Binder DA, Chinnappa BN, Christianson A, Colledge MJ, & Kott PS (Eds.), *Business Survey Methods* (pp. 403-423). New York: John Wiley & Sons, Inc.

Krenzke T, Montaquila J, & Mohadjer L (2000). Accounting for imputation error variance for establishment surveys: an empirical evaluation. In Alexandria, VA 22314: American Statistical Association.

Lee H & Croal J (1989). A simulation study of various estimators which use auxiliary data in an establishment survey. In (pp. 336-341). Alexandria, VA 22314: American Statistical Association.

Little RJA & Rubin DB (2002). *Statistical analysis with missing data*. (Second ed.) New York: Wiley-Interscience.

Morgan M & ORC Macro. (2004). Memorandum: Imputation Plan for Fall 2002 Graduate Student Survey. Burelli J and NSF/SRS.

Mueller K & Butani S (1995). Nonreponse adjustment in certainty strata for an establishment survey. In (pp. 479-484). Proceedings of the Survey Research Methods Section: American Statistical Association.

National Science Foundation & Division of Science Resources Statistics (2006). Graduate students and Postdoctorates in Science and Engineering: Fall 2003. Detailed Statistical Tables.http://www.nsf.gov/statistics/nsf06307/tables.htm#group1 [On-line].

National Science Foundation & Division of Science Resources Statistics (SRS) (2005). Graduate Students and Postdoctorates in S&E: Fall2003.http://www.nsf.gov/statistics/gradpostdoc/ [On-line].

NSF-NIH (2005). Survey of graduate students and postdoctorates in Science and Engineering. The National Science Foundation and the National Institutes of Health [On-line]. Available:http://www.nsf.gov/statistics/survey.cfm

Rubin DB (1987). *Multiple imputation for nonresponse in surveys*. (vols. John Wiley & Sons, Inc) New York: John Wiley & Sons, Inc.

Ruggles P & Joint Economic Committee (2006). *Longitudinal Analysis of Federal Survey Data* (Rep. No. 112). US Department of Commerce. Bureau of the Census.

Schafer JL, Ezzati-Rice TM, Johnson W, Khare M, Little RJA, & Rubin DB (1996). The NHANES III multiple imputation project. In (pp. 28-37). Alexandria, VA: American Statistical Association.

Schenker N, Treiman DJ, & Weidman L (1988). Multiple imputation of industry and occupation codes for public use files. In (pp. 85-92). Proceedings of the Survey Research Methods Section: American Statistical Association.

Shao J (2002). Replication methods for variance estimation in complex surveys with imputed data. In Robert M Groves, Don A Dillman, John L Eltinge, & Robert JA Little (Eds.), *Survey Nonresponse* (pp. 303-314). New York: John Wiley & Sons.

Shao J & Sitter RR (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association, 91,* 1278-1288.

Shao J & Steel P (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association, 94,* 254-265.

Sirken M & Shimizu I (1999). The Horvitz-Thompson Estimator in population based establishment sample surveys. In (pp. 233-237). Alexandria, VA 22314: American Statistical Association.

U.S.Department of Education.National Center for Education Statistics. (2001). *A study of imputation algorithms* (Rep. No. Working paper No. 2001-17 by Ming-xiu Hu and Sameena Salvucci). Washington, DC: Project Officer, Ralph Lee.

West SA, Butani S, & Witt M (1993). Alternative Imputation Methods for Labor Type Data. In Alexandria, VA 22314: American Statistical Association.

Wolter KM (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 645-648
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# *Brief Reports*
# A Note on Probability Trees

W. J. Hurley
Royal Military College of Canada

Not many introductory probability and statistics textbooks emphasize the use of probability trees to make complex probability calculations. This is puzzling in view of the power that trees bring to organizing such calculations for students. An effective classroom technique is discussed is this note.

## Introduction

Not many introductory probability and statistics textbooks emphasize the use of probability trees to make complex probability calculations, including classics such as Hogg and Craig (1970), Parzen (1960), and Ross (1996). An exception is Aczel (1993). This is puzzling in view of the power that trees bring to organizing such calculations for students.

On the first day of a statistics course (both undergraduate and graduate) I teach, students are given a fairly complex real-world probability problem involving an assessment about whether a particular quality assurance test for an ammunition component is reasonable. After four weeks of lectures on introductory probability theory including probability trees and the binomial distribution, students are asked to revisit the problem for homework, and most are able to make the calculation and are very pleased for being able to do so.

## The Problem

Here is a statement of the problem used in the first lecture, pertaining to defense resource management:

William Hurley is a Professor in the Department of Business Administration at the Royal Military College of Canada. His research interests are military operations research, decision analysis, and game theory. Email: hurley-w@rmc.ca

The primary armament on the Canadian Forces (CF) LAV III (Light Armored Vehicle) is the Bushmaster 242 Cannon. It fires 25mm rounds in three-round bursts at enemy thin-skinned assets. To be able to see where rounds go so that aim can be adjusted, each round comes with tracer. The tracer is the explosive charge that lights up for a brief period of time after the round is fired.

The CF purchases 25mm ammunition in lot sizes of 5,000 and 10,000 rounds. Each lot must be tested to make sure that it satisfies the quality standards specified in the purchase contract. In almost all cases these specifications are governed by operational considerations. If ammunition is not up to specifications, soldiers in an operational environment are put at a higher risk. The specification for the 25mm tracer is that it work 97.5% of the time. That is, the defective rate can be no more than 2.5%. To test whether a lot satisfies this specification, the CF performs the following test.

A random sample of 10 rounds from a lot are fired. If there are 0 or 1 defective, the lot is accepted. If there are 3 or more defectives, it is rejected. If there are 2 defectives, another random sample of 10 rounds is fired. If there are 0 defective in this second sample, the lot is accepted; if there is 1 or more, it is rejected.

> The Weapon Systems Engineer has asked you to determine whether this is a good test. He is worried about accepting a lot when the actual defective rate is higher than 2.5%. When you asked about defective rates, he stated that a 5% defective rate for the tracer was unacceptable and a 10% defective rate was absolutely unacceptable. You are required to assess the CF's chances of accepting a bad lot and report your results to the engineer. Use your intuition to assess whether the chance of accepting a bad lot is high or low.

Students are usually split on whether the probability is high or low and this may be a reflection of the uncertainty they have about the correct answer. Nonetheless, as officers and future officers in the Canadian Forces, they see the value of the problem and want to know how to solve it.

Solution

Over the first month, students are taught how to make probability calculations, including Bayes' Theorem using probability trees. A standard approach is taken, using simple problems such as picking marbles out of urns. With some repetition and homework, most students are able to pick up the mechanics of probability tree calculations very quickly. Once they have the idea with these simple problems, they are given real-world problems, most of which are based on my experience within the Department of National Defense and the Canadian Forces. The problem in the previous section is an example. The problems given for homework are a little different in that students are asked to make some specific calculations. Hence, the closing paragraph as follows:

> The Weapon Systems Engineer has asked you to determine whether this is a good test. He is worried about accepting a lot when the actual defective rate is higher than 2.5%. When you asked about defective rates, he stated that a 5% defective rate for the tracer was unacceptable and a 10% defective rate was absolutely unacceptable. You are required to assess the CF's chances of accepting a bad lot and report your results to the engineer. What is the chance of accepting the lot if the underlying defective rate is 5%? What is the chance of accepting the lot if the underlying defective rate is 10%?

Some students, particularly at the graduate level, will come to my office to see if they have done it properly. Because homework is not graded, it is presumed that they are genuinely interested in the solution strategy.

A tree for this problem is shown in Figure 1 where the base probabilities are shown on each arc. In this diagram, the binomial probability of exactly x successes in 10 trials is represented with *b(x)*. Accepting the lot happens along the top arc of the first stage (0 or 1 defective) and along the combination of the middle arc in the first stage and the top arc in the second stage (2 defective on the first sample and 0 defective on the second). Hence, to get the probability of accepting the lot, $\phi_A$, multiply probabilities along each path and then add the results:

$$\phi_A = b(0) + b(1) + b(2)b(0).$$

Letting $p_D$ be the probability of a defective, we have that

$$b(0) = C(10,0)p_D^0(1-p_D)^{10-0} = (1-p_D)^{10}$$
$$b(1) = C(10,1)p_D^1(1-p_D)^{10-1} = 10p_D(1-p_D)^9$$
$$b(2) = C(10,2)p_D^2(1-p_D)^{10-2} = 45p_D^2(1-p_D)^8$$

and therefore

$$\phi_A = (1-p_D)^{10} + 10p_D(1-p_D)^9 + 45p_D^2(1-p_D)^8(1-p_D)^{10}.$$

Table 1 shows values of $\phi_A$ for $p_D$ = 0.05 and 0.10:

Table 1

| $p_D$ | $\phi_A$ |
|---|---|
| .05 | .959 |
| .10 | .804 |

*Note:* Note that the probabilities of accepting a bad lot are very high. Consequently, the conclusion is that this test is not very good.

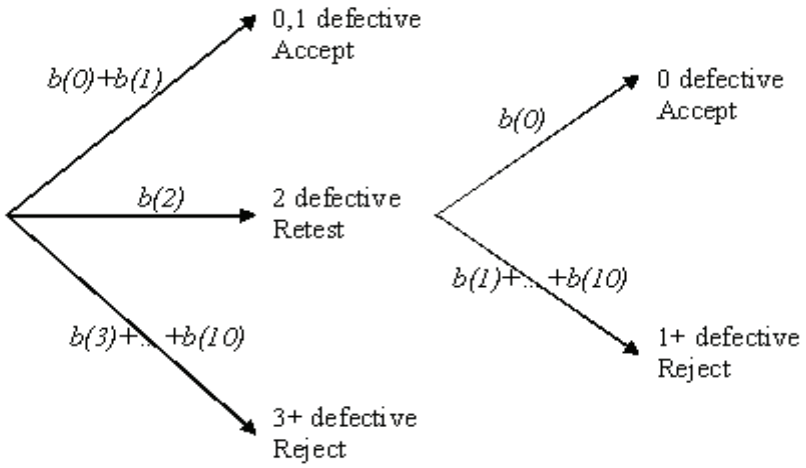



Figure 1. The Probability Tree

There are a number of interesting questions that can be asked at this point. For instance:

1. Regarding operational risk and the maximum 2.5% defective rate, if the underlying defective rate is actually 10%, what is the chance that at least 2 tracers in a burst of three rounds will fire properly (another tree)?
2. How could this test be modified to produce a more favorable result?

To answer this question, students must consider two risks: supplier risk, the risk of rejecting a good lot of ammunition; and soldier risk, the risk of accepting a bad lot of ammunition. Obviously, from the point of view of the Canadian Forces, more weight would be put on soldier risk. This question is usually only pursued with graduate students and by giving specific direction.

Summary

The best feedback I have had on probability trees has come from my graduate students. Most are officers in the Canadian Forces and have an undergraduate engineering degree with at least an introductory course in mathematical statistics. They indicate probability trees are much easier to use that the standard analytic approach. This is particularly true for conditional probability and Bayes' Rule.

The problem presented here is particularly rich. It leads to interesting follow-on questions that can be explored either in the classroom or on assignments.

References

Aczel, Amir, *Complete Business Statistics*, Irwin, Chicago, 1996.

Hogg, R. V. and A. T. Craig, *Introduction to Mathematical Statistics*, Macmillan, New York, 1970.

Parzen, Emanuel, *Modern Probability Theory and Its Application*, Wiley, New York, 1960.

Ross, S. M., *Introduction to Probability Models*, Academic Press, Boston, 1993.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 649-656
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

*Early Scholars*

# Optimum Choice Of Covariates For A Series Of SBIBDS Obtained Through Projective Geometry

Ganesh Dutta
Basanti Devi College
Kolkata, India

Premadhis Das
University of Kalyani
India

Nripes Kumar Mandal
University of Calcutta
India

A block design set up is considered in presence of a number of controllable covariates. The problem is that of choosing the values of the covariates so that for a given block design, it is optimum in the sense of attaining minimum variance for the estimation of each of the covariate parameters. In case of incomplete block designs, the choice of the values of the covariates depends heavily on the allocation of treatments to the plots of blocks; more specifically on the method of construction of the incomplete block design. In this paper the situation where the block design is a member of the complementary series of balanced incomplete block design (BIBD) with parameters $b = v = s^N+s^{N-1}+\ldots+s+1$, $r = k = s^N$, $\lambda=s^N-s^{N-1}$ of symmetric balanced incomplete block design (SBIBD) obtained through projective geometry is considered.

AMS Subject Classification: Primary 62K05; Secondary 62K10.

Key words: block designs, projective geometry, covariates, optimal design, Hadamard matrices.

## Introduction

The following non stochastic controllable covariates model in a block design set up

$$(Y, \mu 1 + X_1\beta + X_2\tau + Z\gamma, \sigma^2 I) \qquad (1)$$

is considered, where $\mu$ is the intercept term, $\sigma^2$ is the common variance of the observations, $\beta$ is the vector of block effects of order $b\times 1$, $\tau$ is the vector of treatment effects of order $v\times 1$, $\gamma$ is the vector of covariate effects of order $c\times 1$ and Y is the uncorrelated observation vector of order $n\times 1$; $X_1$, $X_2$ are the incidence matrices of block effects, treatment effects respectively and Z is a design matrix of covariate effects.

For the covariates, without loss of generality, the (location-scale)- transformed version: $|z_{ij}| \leq 1$ is assumed. It is evident that for orthogonal estimation of treatment and block effect contrasts on one hand and covariate effects on the other, the following condition must be satisfied.

$$Z'X_1 = 0, \qquad Z'X_2 = 0 \qquad (2)$$

For most efficient estimation of each of the regression parameters the following condition must hold (Pukelsheim, 1993)

$$Z'Z = nI_c. \qquad (3)$$

This means that all the elements in each column of Z must be $\pm 1$, and the columns must be mutually orthogonal.

In the block design set up, the optimum

Ganesh Dutta is a Lecturer of Statistics in Basanti Devi College, pursuing the Ph.D. in Statistics from the University of Calcutta, Kolkata, India. Premadhis Das is a Professor in the Department of Statistics, University of Kalyani. Nripes Kumar Mandal is a Professor in the Department of Statistics, University of Calcutta, email: nripes@gmail.com.

properties of randomized block design (RBD) and BIBD with respect to a class of optimality criteria for the estimation of treatment effects are well known (see e.g., Shah & Sinha, 1989). The choice of covariates in a design set up was earlier considered by Troya (1982a, 1982b), Liski et al (2002), Das et al. (2003), Dutta (2004), Rao et al. (2006) and others. Troya (1982a, 1982b) first considered the problem of choice of the levels of the covariates, i.e., Z matrix in a completely randomized design (CRD) model. Das et al. (2003) extended it to the set up of RBD and some series of BIBDs. As mentioned earlier, the choice of covariate values depends heavily on the block design set up as is evidenced from (2). In the case of incomplete block designs, the allocation of treatments to the plots of the blocks depends on the method of construction of designs. Das et al. (2003) considered symmetric balanced incomplete block design (SBIBDs) with parameters b=v, r=k, $\lambda$ constructed through Bose's difference method and some BIBDs with repeated blocks. Dutta (2004) also considered some series of BIBDs obtained through Bose's difference technique together with some arbitrary BIBDs. However, as is well known, there are different methods of construction leading to different series of BIBDs and the choice of the Z matrices also varies from series to series. Here, the problem of choice of Z for the series of complements of SBIBDs, which are obtained through Projective Geometry, is considered. It may be mentioned in this connection that in the series considered in the previous works (Das et al., 2003, Dutta, 2004), the layouts have cyclical pattern which simplified the choice of Z. However, the series of SBIBDs considered here do not have the above cyclical property. Following Das et al. (2003), each column of the Z matrix is transformed to a W-matrix where the element in the $i^{th}$ row and $j^{th}$ column of $W^{(s)}$ is $z_{ij}^{(s)}; z_{ij}^{(s)}$ being the element of Z corresponding to $j^{th}$ treatment in $i^{th}$ block of the design for the $s^{th}$ covariate. Corresponding to the block and treatment classification, conditions (2) and (3) in terms of W-matrices reduce to:

($C_1$) Each W-matrix has all column-sums equal to zero;

($C_2$) Each W-matrix has all row-sums equal to zero;

($C_3$) The grand total of all the entries in the Hadamard product (vide Rao, 1973) of any two distinct W-matrices reduces to zero.

In a BIBD set up with parameters v,b,r,k and $\lambda$, W-matrix of order b×v can be constructed from the incidence matrix of the BIBD by placing judiciously ±1's in the non-zero k-positions in every row and in the non-zero r positions in every column such that each W-matrix satisfies conditions $C_1$, $C_2$ and $C_3$ mentioned above. The paper is organized as follows: In Section 2 an outline of the construction of BIBDs through PG(N,s) and a method of partitioning of the blocks into different sets useful for the choice of W-matrices are given and in Section 3 methods of constructing optimum W-matrices by using sets described in Section 2 have been considered

## BIBDs through Projective Geometry: Partitioning of blocks

With the help of the Galosis field GF(s), a finite projective geometry of N dimensions, to be written as PG(N,s), where $s=p^n$, p is a prime number and n is any positive integer, can be constructed. Any ordered set of (N+1) elements $(x_0,x_1,\ldots,x_N)$ where the $x_i$'s belong to GF(s) and are not simultaneously zero, is called a point of the projective geometry PG(N,s). It is known that the number of points in PG(N,s) is equal to $\phi(N,0,s)=\frac{s^{N+1}-1}{s-1}$ and the number of m-flats is equal to $\phi(N,m,s)$ where

$$\phi(N,m,s)=\frac{(s^{N+1}-1)(s^{N}-1)...(s^{N-m+1}-1)}{(s^{m+1}-1)(s^{m}-1)...(s-1)}$$

By making correspondence between points and m-flats of PG(N,s) with varieties and blocks respectively, a BIBD with parameters v = $\phi(N,0,s)$, b = $\phi(N,m,s)$, r = $\phi(N-1, m-1, s)$, k = $\phi(m,0,s)$, $\lambda$ = $\phi(N-2, m-2, s)$ can be obtained (cf. Bose, 1939). The following series of SBIBDs with m=N-1 has parameters

$b=v=s^N+s^{N-1}+\ldots+s+1$, $r=k=s^{N-1}+\ldots+s+1$,

$$\lambda = s^{N-2}+s^{N-3}+\ldots+s+1. \qquad (4)$$

The complementary SBIBD of (4) has the parameters

$$b = v = s^N+s^{N-1}+\ldots+s+1, \qquad (5)$$
$$r = k = s^N,\ \lambda = s^N-s^{N-1}.$$

It is mentioned above that the choice of the levels of the covariates in BIBD set up depends on the method of its construction and the maximum number of covariates satisfying (2)-(3) varies from series to series.

The blocks of the SBIBD are partitioned into $(s^{N-1} + s^{N-3}+\ldots+s^2+1)$ (=t, say) disjoint sets; each set containing (s+1) blocks such that the portion of the incidence matrix of the complementary design corresponding to each set conforms to that of the incidence matrix of an RBD with suitable parameters. This fact has been used for the choice of the Z matrix.

It is to be noted that the number of (N-1)-flats passing through a particular (N-2)-flat is the number of (N-1)-flats on which a particular (N-2)-flat lies. This number is given by $\phi(1,0,s) = s+1$. Such (s+1), (N-1)-flats passing through a particular (N-2)-flat can be obtained as follows:

Consider an (N-2)-flat of PG(N,s) given by

$$a'x=0, \quad b'x=0 \qquad (6)$$

where, a′ and b′ are two row vectors of a matrix A of order $2\times(N+1)$ with elements from GF(s) such that rank (A)=2.

The (s+1), (N-1)-flats containing the same (N-2)-flat in (6), are given by $(\lambda_1 a'+\lambda_2 b')x=0$; $(\lambda_1,\lambda_2) \neq (0,0)$ and $(\lambda_1,\lambda_2) \equiv \rho(\lambda_1,\lambda_2)$ where, $\rho$ is a non-zero element of GF(s). If N is odd, then the full set of $\phi(N, N-1, s)$, (N-1)-flats can be partitioned into

$$\frac{\phi(N,N-1,s)}{s+1} = \frac{s^{N+1}-1}{(s+1)(s-1)} = (s^{N-1}+s^{N-3}+\ldots+s^2+1)$$

sets each containing (s+1), (N-1)-flats having a common (N-2)-flat. It is clear that the $\frac{\phi(N,N-1,s)}{s+1}$, $(N-1)$-flats passing through a particular (N-2)-flat are disjoint. As the blocks correspond to (N-1)-flats, through one to one correspondence, partition the blocks into $(s^{N-1}+s^{N-3}+\ldots+s^2+1)$ disjoint sets each containing (s+1) blocks. It will be clear from the following two examples covering both the situations where s is prime or prime power.

Example 1: N=3, m=2, s=2. There are 15 blocks which can be partitioned into 5 sets each of size 3 as mentioned above.

$$S_1: \begin{matrix} x_0 = 0 \\ x_1 + x_2 = 0 \\ x_0 + x_1 + x_2 = 0 \end{matrix} \qquad S_2: \begin{matrix} x_1 = 0 \\ x_0 + x_3 = 0 \\ x_0 + x_1 + x_3 = 0 \end{matrix} \qquad S_3: \begin{matrix} x_2 = 0 \\ x_1 + x_3 = 0 \\ x_1 + x_2 + x_3 = 0 \end{matrix}$$

$$S_4: \begin{matrix} x_3 = 0 \\ x_0 + x_2 = 0 \\ x_0 + x_2 + x_3 = 0 \end{matrix} \qquad S_5: \begin{matrix} x_0 + x_1 = 0 \\ x_2 + x_3 = 0 \\ x_0 + x_1 + x_2 + x_3 = 0 \end{matrix}$$

It is to be noted that only two equations in each set $S_i$ are independent and these can conveniently be represented as Ax=0. It is clear that the choice of A matrix in $S_1$ is given by:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

The choice of A matrices for other S's is obvious.

Example 2: N=3, m=2 and $s=2^2$. There are 85 blocks which can be partitioned into 17 sets each of size 5.

Let the elements of $GF(2^2)$ be $\alpha_0=0$, $\alpha_1=1$, $\alpha_2=x$, $\alpha_3=1+x$; x being a primitive element of $GF(2^2)$. Then the 17 sets are:

$$S_1: \begin{matrix} x_0=0 \\ x_1=0 \\ x_0+x_1=0 \\ x_0+\alpha_2x_1=0 \\ x_0+\alpha_3x_1=0 \end{matrix} \qquad S_2: \begin{matrix} x_2=0 \\ x_3=0 \\ x_2+x_3=0 \\ x_2+\alpha_2x_3=0 \\ x_2+\alpha_3x_3=0 \end{matrix} \qquad S_3: \begin{matrix} x_0+x_2=0 \\ x_1+x_3=0 \\ x_0+x_1+x_2+x_3=0 \\ x_0+\alpha_2x_1+x_2+\alpha_2x_3=0 \\ x_0+\alpha_3x_1+x_2+\alpha_3x_3=0 \end{matrix}$$

$$S_4: \begin{matrix} x_0+\alpha_2x_2=0 \\ x_1+\alpha_3x_3=0 \\ x_0+x_1+\alpha_2x_2+\alpha_3x_3=0 \\ x_0+\alpha_2x_1+\alpha_2x_2+x_3=0 \\ x_0+\alpha_3x_1+\alpha_2x_2+\alpha_2x_3=0 \end{matrix} \qquad S_5: \begin{matrix} x_0+\alpha_3x_2=0 \\ x_1+\alpha_2x_3=0 \\ x_0+x_1+\alpha_3x_2+\alpha_2x_3=0 \\ x_0+\alpha_2x_1+\alpha_3x_2+\alpha_3x_3=0 \\ x_0+\alpha_3x_1+\alpha_3x_2+x_3=0 \end{matrix}$$

$$S_6: \begin{matrix} x_0+x_3=0 \\ x_1+x_2+x_3=0 \\ x_0+x_1+x_2=0 \\ x_0+\alpha_2x_1+\alpha_2x_2+\alpha_3x_3=0 \\ x_0+\alpha_3x_1+\alpha_3x_2+\alpha_2x_3=0 \end{matrix} \qquad S_7: \begin{matrix} x_0+\alpha_2x_3=0 \\ x_1+\alpha_3x_2+\alpha_2x_3=0 \\ x_0+x_1+\alpha_3x_2=0 \\ x_0+\alpha_2x_1+x_2+x_3=0 \\ x_0+\alpha_3x_1+\alpha_2x_2+\alpha_3x_3=0 \end{matrix}$$

$$S_8: \begin{matrix} x_0+\alpha_3x_3=0 \\ x_1+\alpha_2x_2+\alpha_3x_3=0 \\ x_0+x_1+\alpha_2x_2=0 \\ x_0+\alpha_2x_1+\alpha_3x_2+\alpha_2x_3=0 \\ x_0+\alpha_3x_1+x_2+x_3=0 \end{matrix} \qquad S_9: \begin{matrix} x_0+x_2+x_3=0 \\ x_1+x_2=0 \\ x_0+x_1+x_3=0 \\ x_0+\alpha_2x_1+\alpha_3x_2+x_3=0 \\ x_0+\alpha_3x_1+\alpha_2x_2+x_3=0 \end{matrix}$$

$$S_{10}: \begin{matrix} x_0+\alpha_2x_2+\alpha_3x_3=0 \\ x_1+\alpha_2x_2=0 \\ x_0+x_1+\alpha_3x_3=0 \\ x_0+\alpha_2x_1+x_2+\alpha_3x_3=0 \\ x_0+\alpha_3x_1+\alpha_3x_2+\alpha_3x_3=0 \end{matrix} \qquad S_{11}: \begin{matrix} x_0+\alpha_3x_2+\alpha_2x_3=0 \\ x_1+\alpha_3x_2=0 \\ x_0+x_1+\alpha_2x_3=0 \\ x_0+\alpha_2x_1+\alpha_2x_2+\alpha_2x_3=0 \\ x_0+\alpha_3x_1+x_2+\alpha_2x_3=0 \end{matrix}$$

.

Example 2 (cont.) N=3, m=2 and $s=2^2$. There are 85 blocks which can be partitioned into 17 sets each of size 5.

$$S_{12}: \begin{matrix} x_0+\alpha_3x_2+\alpha_3x_3=0 \\ x_1+\alpha_2x_2+x_3=0 \\ x_0+x_1+x_2+\alpha_2x_3=0 \\ x_0+\alpha_2x_1+x_3=0 \\ x_0+\alpha_3x_1+\alpha_2x_2=0 \end{matrix} \qquad S_{13}: \begin{matrix} x_0+\alpha_2x_2+\alpha_2x_3=0 \\ x_1+\alpha_3x_2+x_3=0 \\ x_0+x_1+x_2+\alpha_3x_3=0 \\ x_0+\alpha_2x_1+\alpha_3x_2=0 \\ x_0+\alpha_3x_1+x_3=0 \end{matrix}$$

$$S_{14}: \begin{matrix} x_0+x_2+\alpha_3x_3=0 \\ x_1+\alpha_2x_2+\alpha_2x_3=0 \\ x_0+x_1+\alpha_3x_2+x_3=0 \\ x_0+\alpha_2x_1+\alpha_2x_2=0 \\ x_0+\alpha_3x_1+\alpha_2x_3=0 \end{matrix} \qquad S_{15}: \begin{matrix} x_0+x_2+\alpha_2x_3=0 \\ x_1+\alpha_3x_2+\alpha_3x_3=0 \\ x_0+x_1+\alpha_2x_2+x_3=0 \\ x_0+\alpha_2x_1+\alpha_3x_3=0 \\ x_0+\alpha_3x_1+\alpha_3x_2=0 \end{matrix}$$

$$S_{16}: \begin{matrix} x_0+\alpha_2x_2+x_3=0 \\ x_1+x_2+\alpha_2x_3=0 \\ x_0+x_1+\alpha_3x_2+\alpha_3x_3=0 \\ x_0+\alpha_2x_1+\alpha_2x_3=0 \\ x_0+\alpha_3x_1+x_2=0 \end{matrix} \qquad S_{17}: \begin{matrix} x_0+\alpha_3x_2+x_3=0 \\ x_1+x_2+\alpha_3x_3=0 \\ x_0+x_1+\alpha_2x_2+\alpha_2x_3=0 \\ x_0+\alpha_2x_1+x_2=0 \\ x_0+\alpha_3x_1+\alpha_3x_3=0 \end{matrix}$$

where $(x_0, x_1, x_2, x_3)$ is a point of $PG(3,2^2)$. As an illustration, the choice of A matrix corresponding to $S_1$ and $S_4$ are given respectively by

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & \alpha_2 & 0 \\ 0 & 1 & 0 & \alpha_3 \end{pmatrix}.$$

Similarly A matrices for other $S_i$'s can be written.

## Choice of Covariates

From (4), it is seen that any block of the design contains $k = (s^{N-1}+\lambda)$ treatments and any two blocks have exactly $\lambda$ treatments in common. As any two blocks of the set $S_i$ $(i=1(1)t;\ t=(s^{N-1}+s^{N-3}+\ldots+s^2+1))$, have the same $\lambda$ treatments common, without loss of any generality, the portion $N_i$ of the incidence matrix corresponding to the blocks in $S_i$ $(i=1(1)t)$ can be written in the following form (with some rearrangement of blocks if necessary):

$$N_i = \begin{pmatrix} 1'_{s^{N-1}} & 0' & \cdots & 0' & 1'_\lambda \\ 0' & 1'_{s^{N-1}} & \cdots & 0' & 1'_\lambda \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0' & 0' & \cdots & 1'_{s^{N-1}} & 1'_\lambda \end{pmatrix}^{s+1\times v} \qquad (7)$$

The part of the incidence matrix of the design with parameters in (5) corresponding to the part $N_i$ of the design with parameters in (4) is obtained by replacing one's by zero's and zero's by one's in (7) and is given by :

$$N_i^c = \begin{pmatrix} 0'_{s^{N-1}} & 1'_{s^{N-1}} & \cdots & 1'_{s^{N-1}} & 0'_\lambda \\ 1'_{s^{N-1}} & 0'_\lambda & \cdots & 1'_{s^{N-1}} & 0'_\lambda \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1'_{s^{N-1}} & 1'_{s^{N-1}} & \cdots & 0'_{s^{N-1}} & 0'_\lambda \end{pmatrix}^{s+1\times v} \qquad (8)$$

Using the structure (8) above, a method for choosing the values of the covariates optimally for the complementary design with parameters in (5) is developed.

Theorem 1:

If $s=2^p$ where p be any positive integer, $(s^{N-1}-1)(s-1)+(s-1)$, W-matrices for the design with parameters in (5), where N is an odd integer can be constructed.

Proof

Because s is a power of 2, Hadamard matrices of orders $s^{N-1}$ and s exist and can be written as follows:

$$H_{s^{N-1}} = (h_1, \cdots, h_{s^{N-1}-1}, 1)$$
$$H_s = (h_1^*, \cdots, h_{s-1}^*, 1) \qquad (9)$$

where $1'$ is a row vector with all elements equal to one. Again the matrix (8) can be written as

$$N_i^c = (A_{1i}, A_{2i}, \cdots, A_{ji}, \cdots, A_{s+1\,i}, 0_i)$$

where $A_{ji}$ is the $j^{th}$ partitioned matrix in the $j^{th}$ column block of $N_i^c, j = 1(1)(s+1)$. Let the $k^{th}$ non-null row of $A_{ji}$ be replaced by the $k^{th}$ row of $h_m^* h_n'; k = 1(1)s$ and the resultant matrix be denoted by $A_{ji}^*$. The procedure for each $A_{ji}$ is repeated with the same $h_m^* h_n'$. This leads to a matrix $W_{i,m,n}^*$ with elements $\pm 1$ satisfying the properties $C_1$ and $C_2$. Using the same $h_m^*$ and $h_n$ different $W_{i,m,n}^*$'s corresponding to different $N_i^c$'s are obtained. Now for fixed $h_m^*$ and $h_n$

$$W_{m,n}^* = \begin{pmatrix} W_{1,m,n}^* \\ W_{2,m,n}^* \\ \vdots \\ W_{t,m,n}^* \end{pmatrix}$$

satisfies the properties $C_1$ and $C_2$. By varying $h_m$ and $h_n^*$, $(s^{N-1}-1)(s-1)$, $W_{m,n}^* -$ matrices can be constructed. The transformation required to apply on (8) to get back the corresponding portion of the incidence matrix of the design may also be applied on the elements of the above W*- matrix to get the original W-matrix. It is clear that such W-matrices also satisfy all the properties $C_1$, $C_2$ and $C_3$.

Again, note that the number of unit vectors in the rows of $N_i^c$ is s which is the same as that of the elements of $h_m^*$. Let the $q^{th}$ vector $1'_{s^{N-1}}$ be replaced in the first column

block matrix of $N_i^c$ by $+1'_{s^{N-1}}$ or by $-1'_{s^{N-1}}$ according as the $q^{th}$ element of $h_m^*$ is +1 or -1 respectively to get $A_1^{**}$. Now the rows of $A_1^{**}$ are permuted cyclically to get $A_2^{**}, A_3^{**}, \ldots, A_{s+1}^{**}$ and hence a new W-matrix viz. $W_m^{**}$ can be constructed. It is easy to show that these $W_m^{**}$ matrices together with $W_{m,n}^{**}$ satisfy all the conditions $C_1$, $C_2$ and $C_3$. In all, $(s^{N-1}-1)(s-1)+(s-1)$, W-matrices exist. The procedure is illustrated through the following example.

Example: 3

The SBIBD whose blocks are 2-flats of PG (3,2) is considered so that the parameters of the SBIBD are v=b=15, r=k=7, λ=3. Now the complement of this design has parameters v′=b′=15, r′=k′=8, λ′=4.

The sets of blocks of the complementary design of Example 1 where the treatment corresponding to the point $(x_0, x_1, x_2, x_3)$ is indexed by $2^3x_0+2^2x_1+2x_2+x_3$ are:

$S_1$ = [(8,9,10,11,12,13,14,15), (2,3,4,5,10,11,12,13), (2,3,4,5,8,9,14,15)]
$S_2$ = [(4,5,6,7,12,13,14,15), (1,3,5,7,8,10,12,14), (1,3,4,6,8,10,13,15)]
$S_3$ = [(2,3,6,7,10,11, 14,15), (1,3,4,6,9,11,12,14), (1,2,4,7,9,10,12,15)]
$S_4$ = [(1,3,5,7,9,11, 13,15), (2,3,6,7,8,9,12,13), (1,2,5,6,8,11,12,15)]
$S_5$ = [(4,5,6,7,8,9,10,11), (1,2,5,6,9,10,13,14), (1,2,4,7,8,11,13,14)].

The Hadamard matrices of orders 2 and 4 exist and are written as:

$$H_2 = \begin{pmatrix} +1 & +1 \\ -1 & +1 \end{pmatrix} = [h_1, 1]$$

and $$H_4 = \begin{pmatrix} +1 & +1 & +1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 \\ -1 & +1 & -1 & +1 \end{pmatrix} = \left[h_1^*, h_2^*, h_3^*, 1\right]$$

Using $h_1$ and $h_i^*(i = 1(1)3)$ and proceeding as in Theorem 1, three W-matrices can be constructed. The construction of a W-matrix viz. $W_{11}^*$ is illustrated using $h_1$ and $h_1^*$:

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\
0 & +1 & -1 & +1 & -1 & 0 & 0 & 0 & 0 & -1 & +1 & -1 & +1 & 0 & 0 \\
0 & -1 & +1 & -1 & +1 & 0 & 0 & -1 & +1 & 0 & 0 & 0 & 0 & -1 & +1 \\
0 & 0 & 0 & +1 & +1 & -1 & -1 & 0 & 0 & 0 & 0 & +1 & +1 & -1 & -1 \\
+1 & 0 & -1 & 0 & -1 & 0 & +1 & +1 & 0 & -1 & 0 & -1 & 0 & +1 & 0 \\
-1 & 0 & +1 & -1 & 0 & +1 & 0 & -1 & 0 & +1 & 0 & 0 & -1 & 0 & +1 \\
0 & +1 & +1 & 0 & 0 & -1 & -1 & 0 & 0 & +1 & +1 & 0 & 0 & -1 & -1 \\
+1 & 0 & -1 & -1 & 0 & +1 & 0 & 0 & +1 & 0 & -1 & -1 & 0 & +1 & 0 \\
-1 & -1 & 0 & +1 & 0 & 0 & +1 & 0 & -1 & -1 & 0 & +1 & 0 & 0 & +1 \\
+1 & 0 & +1 & 0 & -1 & 0 & -1 & 0 & +1 & 0 & +1 & 0 & -1 & 0 & -1 \\
0 & +1 & -1 & 0 & 0 & -1 & +1 & +1 & -1 & 0 & 0 & -1 & +1 & 0 & 0 \\
-1 & -1 & 0 & 0 & +1 & +1 & 0 & -1 & 0 & 0 & -1 & +1 & 0 & 0 & +1 \\
0 & 0 & 0 & +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 & 0 & 0 & 0 & 0 \\
+1 & -1 & 0 & 0 & -1 & +1 & 0 & 0 & -1 & +1 & 0 & 0 & +1 & -1 & 0 \\
-1 & +1 & 0 & -1 & 0 & 0 & +1 & -1 & 0 & 0 & +1 & 0 & -1 & +1 & 0
\end{pmatrix}$$

Similarly, by taking the combinations $(h_1, h_2^*)$, $(h_1, h_3^*)$ $W_{12}^*$ and $W_{13}^*$ can be constructed. Another matrix $W_1^{**}$ using $h_1$ is given below:

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & +1 & +1 & +1 & +1 & -1 & -1 \\
0 & +1 & +1 & +1 & +1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\
0 & -1 & -1 & -1 & -1 & 0 & 0 & +1 & +1 & 0 & 0 & 0 & 0 & +1 & +1 \\
0 & 0 & 0 & -1 & +1 & -1 & +1 & 0 & 0 & 0 & 0 & +1 & -1 & +1 & -1 \\
+1 & 0 & +1 & 0 & -1 & 0 & -1 & +1 & 0 & +1 & 0 & -1 & 0 & -1 & 0 \\
-1 & 0 & -1 & +1 & 0 & +1 & 0 & -1 & 0 & -1 & 0 & 0 & +1 & 0 & +1 \\
0 & -1 & +1 & 0 & 0 & +1 & -1 & 0 & 0 & -1 & +1 & 0 & 0 & +1 & -1 \\
+1 & 0 & -1 & +1 & 0 & -1 & 0 & 0 & +1 & 0 & -1 & +1 & 0 & -1 & 0 \\
-1 & +1 & 0 & -1 & 0 & 0 & +1 & 0 & -1 & +1 & 0 & -1 & 0 & 0 & +1 \\
-1 & 0 & +1 & 0 & -1 & 0 & +1 & 0 & +1 & 0 & -1 & 0 & +1 & 0 & -1 \\
0 & +1 & -1 & 0 & 0 & +1 & -1 & +1 & -1 & 0 & 0 & +1 & -1 & 0 & 0 \\
+1 & -1 & 0 & 0 & +1 & -1 & 0 & -1 & 0 & 0 & +1 & -1 & 0 & 0 & +1 \\
0 & 0 & 0 & -1 & +1 & +1 & -1 & -1 & +1 & +1 & -1 & 0 & 0 & 0 & 0 \\
+1 & +1 & 0 & 0 & -1 & -1 & 0 & 0 & -1 & -1 & 0 & 0 & +1 & +1 & 0 \\
-1 & -1 & 0 & +1 & 0 & 0 & +1 & +1 & 0 & 0 & +1 & 0 & -1 & -1 & 0
\end{pmatrix}$$

Thus four W–matrices $W_{11}^*$, $W_{12}^*$, $W_{13}^*$ and $W_1^{**}$ are constructed satisfying conditions $C_1$-$C_3$.

## References

Bose, R. C. (1939). On the construction of balanced incomplete block designs. *Annals of Eugenics*, *9*, 353-399.

Das, K., Mandal, N. K., & Sinha, B. K. (2003). Optimal experimental designs with covariates. *Journal of Statistical Planning and Inference*, *115*, 273-285.

Dutta, G. (2004). Optimum choice of covariates in BIBD set up. *Calcutta Statistical Association. Bulletin. 55*, 39-55.

Liski,E.P., Mandal, N. K., Shah, K. R, & Sinha, B. K. (2002). *Topics in optimal design*. Lecture notes in statistics 163, Springer-Verlag, New York.

Troya Lopes, J (1982a). Optimal designs for covariate models. *Journal of Statistical Planning and Inference*, *6*, 373-419.

Troya Lopes, J (1982b). Cyclic designs for a covariate model. *Journal of Statistical Planning and Inference*, *7*, 49-75.

Pukelsheim, F (1993). *Optimal Design of Experiments*. John Wiley & Sons.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. John Wiley & Sons, Inc.

Rao, P.S.S.N.V.P., Rao, S. B., Saha. G. M., &. Sinha, B. K. (2006). Optimal Designs for Covariates' Models and Mixed Orthogonal Arrays. *Discrete Mathematics, 15*, 157-160.

Shah, K. R., & Sinha, B. K. (1989). *Theory of Optimal Designs*. Lecture Notes in Statistics, *54*. Springer Verlag.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 657-666
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# A New Generalization of Negative Polya-Eggenberger Distribution and its Applications

Anwar Hassan
University of Kashmir

Sheikh Bilal Ahmad
Degree College, Baramulla

A new generalization of negative Polya-Eggenberger distribution (GNPED) has been obtained by mixing the negative binomial distribution with generalized beta distribution-II defined by Nadarajah and Kotz (2003). Some special cases and properties of GNPED have been studied. Further, the proposed model has been fitted to two data sets (used by Gupta & Ong, 2004) that provide a satisfactory fit and better alternative as compared to negative binomial and some of its mixture models and extensions. Also, the negative Polya-Eggenberger distribution (NPED), obtained by mixing negative binomial with beta distribution of I-kind, has been fitted to the same data sets for comparison.

Key words: negative binomial distribution, generalized beta distribution-II, generalized negative Polya-Eggenberger distribution (GNPED).

## Introduction

Feller (1943) pointed out there are essentially two kinds of contagious distributions. One type, true contagion, is due to the fact that each favorable event increases (or decreases) the probability of succeeding favorable events. The other type, apparent contagion, is due to an inhomogeneity of the population. Frequently, the data arising in studies of entomology and bacteriology cannot be described by the usual distribution functions but rather by some type of contagious distributions. Some distributions, such as the negative binomial, can apparently be interpreted on the basis of both types of contagion.

A class of contagious distributions is derived from a certain biological model which takes into account the fact that the distribution of larvae over the plots of a field depends upon the fact that the larvae are hatched from egg-masses which appear at random over the field has been derived by Neyman (1939), Evans (1953) and Beal and Rescia (1953). This class of distribution has been successful in accounting for the distribution of some insect populations (ef. Beal-1940). Bliss and Fisher (1953) showed that the negative binomial distribution is useful as a possible underlying distribution for insect populations. Contagious distributions have also been used in the study of accident and medical statistics by Dubourdieu (1939), Greenwood and Yule (1920), Lundberg (1940) and Newbold (1927). Eggenberger and Polya (1923) introduced Polya-Eggenberger distribution (PED) and negative Polya-Eggenberger distribution (NPED) by an urn model and described these as truly contagious distributions.

Anwar Hassan is a Post Graduate, Department. of Statistics. Research interests: Probability and Lagrangian Probability distributions and Statistical Inference. E-mail address: anwar.hassan2007@gmail.com. Sheikh Bilal Ahmad, Department of Statistics, E-mail: sbilal_sbilal@yahoo.com Research interests: Probability distributions and Statistical Inference.

## Negative Polya-Eggenberger Distribution (NPED)

Negative Polya-Eggenberger distribution is related to Polya-Eggenberger distribution in the same way as the negative binomial distribution is related to binomial distribution. It is well known fact that the negative binomial distribution (NBD) has become increasingly popular as a more flexible alternative to the Poisson distribution especially when it is doubtful whether the strict requirements

particularly independence for a Poisson distribution will be statisfied. Negative Polya-Eggenberger distribution is obtained by mixing negative binomial distribution

$$P(X=x)=\binom{n+x-1}{x} p^{x}\ (1-p)^{n}, \quad (1)$$
$$x=0,1,2,....;\ o<p<1$$

With beta distribution of I-Kind

$$P(X=p)=\frac{1}{\beta(\alpha,\gamma)}\quad p^{\alpha-1} \quad (2)$$
$$(1-p)^{\gamma-1},$$
$$0<p<1,$$
$$\alpha,\gamma>0$$

If the parameter p in (1) is not a constant but is varying as beta distribution of I-kind then its probability mass function is (2) and beta mixture of negative binomial distribution is obtained as

$$P(X=x)=$$
$$\binom{n+x-1}{x}\frac{1}{\beta(\alpha,\gamma)}\int_0^1 p^{x+\alpha-1}$$
$$(1-p)^{n+\gamma-1}dp$$
$$x=0,1,2,.....$$
$$=\binom{n+x-1}{x}\frac{\beta(\alpha+x,\gamma+n)}{\beta(\alpha,\gamma)}$$
$$=\binom{n+x-1}{x}\frac{(\alpha+\gamma-1)!}{(\alpha-1)!\,(\gamma-1)!}$$
$$\frac{(\alpha+x-1)!(\gamma+n-1)!}{(\alpha+\gamma+n+x-1)!}$$
$$P(X=x)=$$
$$\binom{n+x-1}{x}$$
$$\frac{\alpha(\alpha+1).....(\alpha+\overline{x-1})\ \gamma(\gamma+1).....(\gamma+\overline{n-1})}{(\alpha+\gamma)(\alpha+\gamma+1).....(\alpha+\gamma+\overline{n+x-1})}$$

Taking $\alpha=a/c$ and $\gamma=b/c$ in the equation above, we get

$$P(X=x)= \quad (3)$$
$$\binom{n+x-1}{x}$$
$$\frac{a(a+c).....(a+\overline{x-1}c)b(b+c).....(b+\overline{n-1}c)}{(a+b)(a+b+c).....(a+b+\overline{n+x-1}c)}$$
$$x=0,1,2,.....$$

Which is negative Polya-Eggenberger distribution with parameters (n, a, b, c).

Generalized Beta Distribution-II

Many generalization of beta distribution of I-kind (2) involving algebraic and exponential function has been proposed in the literature; see chapter 25 in Johnson et al (1995) and Gupta Nadarajah (2004) for detailed accounts. Nadarajah and Kotz (2003) defined a generalization of (2) involving the Gauss hypergeometric function as

$$P(X=p)= \quad (4)$$
$$\frac{b\beta(a,b)}{\beta(a,b+\gamma)}\ p^{a+b-1}\ {}_2F_1[1-\gamma,a;a+b,p],$$

This is known as generalized beta distribution-II. The properties of incomplete beta function and Gauss hypergeometric function can be found in Prudnikov et al (1990, vol.3 sec. 7.3) and Gradshteyn and Ryzhik (2000).

There are various extensions/ modifications of NBD in the literature including Engen's extended NBD (1974, 1978), generalized NBD of Jain and Consul (1971) and weighted NBD; see Johnson et al. (1992) for more details and explanations. A brief list of authors and their work can be seen in Johnson and Kotz (1969), Consul and Famoye (2000), Johnson and Balkrishnan (1995) and Gupta and Nadarajah (2004).

In this article, an attempt has been made to introduce a new generalized contagious distribution, generalized negative Polya-Eggenberger distribution (GNPED), by mixing NBD with generalized beta distribution-II defined by Nadarajah and Kotz's (2003) which is expected to explain data in a better way as compared to distributions obtained by mixing Poison or binomial with other distributions. Further, the proposed model has been fitted to

same data sets previously used by Gupta and Ong (2004) that exhibits a satisfactory fit and better alternative as compared to negative binomial, negative Polya-Eggenberger, Gupta and Ong's (2004) GNBD and Jain and Consul's (1971) GNBD.

The Proposed Model

Let X be a random variable representing the number of independent trails necessary to obtain 'n' occurrences of an event that has a constant probability of occurring at each trail. Then X has a negative binomial distribution with parameters (n, p) and pmf given by (1). But, there are situations in the practical life where probability 'p' of each occurrence of an event is not constant but is following some distribution. In the present case, suppose 'p' is varying as generalized beta distribution-II by Nadarajah and Kotz (2003) with parameters $(a,b,\gamma)$ and pdf given by (4) then the pmf of proposed model is obtained by mixing (1) with (4) as

$$P(X=x)=\binom{x+n-1}{n-1}\frac{b\beta(a,b)}{\beta(a,b+\gamma)}$$
$$\int_0^1 [\,p^{n+a+b-1}(1-p)^x$$
$${}_2F_1[1-\gamma,a;a+b,p]\,dp$$
$$=\binom{x+n-1}{n-1}\frac{b\beta(a,b)}{\beta(a,b+\gamma)}$$
$$\sum_{j=0}^{\infty}\left[\frac{(1-\gamma)^{[j]}a^{[j]}}{(a+b)^{[j]}}\,\frac{1}{j!}\right.$$
$$\left.\int_0^1[\;p^{n+a+b+j-1}(1-p)^x\;]\,dp\;\right]$$

Using Gradibhtyn and Ryzhik's (2000) book, we obtain after few steps

$$P(X=x)= \tag{5}$$
$$\frac{(x+n-1)!}{(n-1)!}\frac{b\beta(a,b)}{\beta(a,b+\gamma)}\frac{(a+b)^{[n]}}{(a+b)^{[n+x+1]}}$$
$$\times\,{}_3F_2[1-\gamma,a,a+b+n;a+b,a+b+n+x+1,1]$$
$$x=0,1,2...$$

This is a probability mass function of the proposed model, generalized negative Polya-Eggenberger distribution (GNPED), which can further be simplified to give

$$P(X=x)= \tag{6}$$
$$\frac{(x+n-1)!}{(n-1)!}\frac{(a+b)^{[\gamma]}(a+b)^{[n]}}{(b+1)^{[\gamma-1]}(a+b)^{[n+x+1]}}$$
$${}_3F_2[1-\gamma,a,a+b+n;a+b,a+b+n+x+1,1]$$
$$x=0,1,.....$$

Some Special Cases of GNPED

Some old and new distributions can be obtained by assigning different values to the parameters of GNPED (6).

For $a+b+\gamma=1$, (6) reduces to

$$P(X=x)= \tag{7}$$
$$\frac{(x+n-1)!}{(n-1)!}\frac{(a+b)^{[\gamma]}(a+b)^{[n]}}{(b+1)^{[\gamma-1]}(a+b)^{[n+x+1]}}$$
$${}_2F_1[a,a+b+n;a+b+n+x+1,1]$$
$$x=0,1,.....$$

The Gauss summation theorem states that

$${}_2F_1[a,b;c,1]=\frac{\Gamma c\;\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \tag{8}$$

Provided $(c-a-b)>0$, $c\neq 0,-1,-2,-3,.....$

Using (8) in (7), we obtain NPED with pmf

$$P(X=x)= \tag{9}$$
$$\binom{x+n-1}{n-1}$$
$$\frac{(1-a)(1-a+1).....(1-a+\overline{x-1})(a+b)(a+b+1)....(a+b+\overline{n-1})}{(1+b)(1+b+1).....(1+b+\overline{n+x-1})}$$

If one puts $(1-a)=\frac{\alpha}{c}$ & $(a+b)=\frac{\beta}{c}$

$\Rightarrow(1+b)=\frac{(\alpha+\beta)}{c}$, then the above equation reduces to NPED in its usual form as

$$P(X=x)=$$

$$\binom{x+n-1}{n-1} \frac{\alpha(\alpha+c).....(\alpha+\overline{x-1}c)\beta(\beta+c).....(\beta+\overline{n-1}c)}{(\alpha+\beta)(\alpha+\beta+c).....(\alpha+\beta+\overline{n+x-1}c)} \quad (10)$$
$$x=0,1,.....$$

For $\gamma=1$, (6) reduces to

$$P(X=x)=\frac{(a+b)}{(a+b+n+x)}\frac{\beta(a+b,x+n)}{\beta(a+b,n)} \quad (11)$$
$$x=0,1,.....$$

Where ${}_3F_2[1-\gamma,a,a+b+n;a+b,$
$a+b+n+x+1,1]=1$ for $\gamma=1$

If in addition, one puts a + b=1, then (11) reduces to factorial distribution

$$P(X=x)=\frac{n}{(n+x)(n+x+1)} \quad (12)$$
$$x=0,1,2,....$$

For b=0, (6) reduces to

$$P(X=x)=\frac{(n+x-1)!}{(n-1)!}\frac{a^{[\gamma]}\ a^{[n]}}{1^{[\gamma-1]}a^{[n+x+1]}}$$
$${}_2F_1[1-\gamma,a+n;a+n+x+1,1]$$

Using Gauss summation theorem (8), we obtain on simplifications

$$P(X=x)=\binom{x+n-1}{n-1}\frac{a^{[n]}\ \gamma^{[x]}}{(a+\gamma)^{[n+x]}}$$
$$x=0,1,.....$$

If one puts $\gamma=\frac{\alpha}{c}$ & $a=\frac{\beta}{c}$ , then above equation reduces to negative Polya-Eggenberger distribution (9).

For a=0, (6) reduces to

$$P(X=x)=\frac{b}{(b+n+x)}\frac{\beta(b,n+x)}{\beta(b,n)} \quad (13)$$
$$x=0,1,2,....$$

Where,

${}_3F_2[1-\gamma,a,a+b+n;a+b,a+b+n+x+1,1]=1$,

for a=0. If in addition one puts b=1, then (13) reduces to (12).

For $(a+b)=1$, (6) reduces to

$$P(X=x)=\frac{n}{(n+x)(n+x+1)}\frac{b!\gamma!}{(b+\gamma-1)!} \quad (14)$$
$${}_3F_2[1-\gamma,a,1+n;1,n+x+2,1]$$

If in addition one puts

i) a=0 , then (14) reduces to

$$P(X=x)=\frac{nb\gamma\beta(b,\gamma)}{(x+x)(x+n+1)}$$

This is a new generalization of factorial distribution (12) as it reduces to (12) for $\gamma=1$ *or* $b=1$

If one replaces x with (x-n), then (6) reduces to

$$P(X=x)=\frac{(x-1)!}{(n-1)!}\frac{b\beta(a,b)}{\beta(a,b+\gamma)}\frac{(a+b)^{[n]}}{(a+b)^{[x+1]}}$$
$${}_3F_2[1-\gamma,a,a+b+n;a+b,a+b+x+1,1]$$

Taking $\gamma=1$ and b=0, this results in a new distribution with pmf

$$P(X=x)=\frac{na}{x}\binom{x}{x-n}\left[\binom{a+x}{x-n}\right]^{-1}$$
$$x=n,n+1,n+2,.....$$

Moment Generating Function of GNPED

Moment generating function of (5) can be obtained as

$$M_X(t) = E(e^{tX})$$
$$= \sum_{x=0}^{\infty} e^{tx} \frac{(x+n-1)!}{(n-1)!} \frac{b\beta(a,b)}{\beta(a,b+\gamma)} \frac{(a+b)^{[n]}}{(a+b)^{[n+x+1]}}$$
$${}_3F_2[1-\gamma, a, a+b+n; a+b, a+b+x+1, 1]$$
$$= \frac{b\beta(a,b)}{\beta(a,b+\gamma)}$$
$$\sum_{j=0}^{\infty} \frac{(1-\gamma)^{[j]}\, a^{[j]}\, (a+b+n)^{[j]}}{(a+b)^{[j]}} \frac{1}{j!}$$
$$\times \sum_{x=0}^{\infty} \frac{n^{[x]}\, 1^{[x]}}{(a+b+n+x+1)^{[j]}(a+b+n)^{[x+1]}}$$
$$\frac{(e^t)^x}{x!}$$

On simplification, this gives moment generating function of GNPED as

$$M_X(t) =$$
$$\frac{b\beta(a,b)}{\beta(a,\gamma+b)} \sum_{j=0}^{\infty} \frac{(1-\gamma)^{[j]}\, a^{[j]} (n+a+b)^{[j]}}{(a+b)^{[j]}(n+a+b)^{[j+1]}} \quad (15)$$
$$\times \frac{{}_2F_1[n, 1; n+a+b+j+1, e^t]}{j!}$$

Remarks:

If one replaces $e^t$ with $(1-t)^{-1}$ in (15), ascending factorial moment generating function is obtained as

$$E(1-t)^{-X} =$$
$$\frac{b\beta(a,b)}{(a+b+n)\beta(a,b+\gamma)}$$
$$\sum_{j=0}^{\infty} \frac{(1-\gamma)^{[j]}\, a^{[j]}\, (a+b+n)^{[j]}}{(a+b)^{[j]}(a+b+n+1)^{[j]}} \frac{1}{j!}$$
$$\times {}_2F_1[n, 1; a+b+n+j+1, (1-t)^{-1}]$$

Similarly, replacing $e^t$ with $(1+t)$ in (15), descending factorial moment generating function of GNPED is obtained as

$$E(1+t)^X =$$
$$\frac{b\beta(a,b)}{(a+b+n)\beta(a,b+\gamma)}$$
$$\sum_{j=0}^{\infty} \frac{(1-\gamma)^{[j]}\, a^{[j]}\, (a+b+n)^{[j]}}{(a+b)^{[j]}(a+b+n+1)^{[j]}} \frac{1}{j!}$$
$$\times {}_2F_1[n, 1; a+b+n+j+1, e^t]$$

Raw Moments of GNPED

The r[th] raw moment of the proposed model (5) can be obtained as

$$E(X^r) = E[E(X^r/p)] \quad (16)$$

Where $E(X^r/p)$ is the conditional rth moment of X for given $p$ and for given $p$, the random variable X has negative binomial distribution (1) with

$$E(X^r/p) = \sum_{x=1}^{r} \binom{x+n-1}{n-1} p^{-x}(1-p)^x \Delta^x 0^r$$

Hence, (16) reduces to

$$E(X^r) = \sum_{x=1}^{r} \binom{x+n-1}{n-1} E[p^{-x}(1-p)^x] \Delta^x 0^r$$

Because $p$ is varying as generalized beta distribution-II (2) with parameters $(a,b,\gamma)$. Therefore,

$$\mu_r' =$$
$$\frac{b\beta(a,b)}{\beta(a,b+\gamma)} \sum_{x=1}^{r} \binom{x+n-1}{n-1} \Delta^x 0^r$$
$$\int_0^1 p^{a+b-x-1}(1-p)^{1+x-1}$$
$${}_2F_1[1-\gamma, a; a+b, p]\, dp$$

This on simplification gives the rth moment of GNPED as

$$\mu_r' = \frac{b\beta(a,b)}{\beta(a,b+\gamma)}$$
$$\sum_{x=1}^{r} \frac{(x+n-1)!(a+b-x-1)!}{(n-1)!(a+b)!} \Delta^x 0^r \quad (17)$$

$$\times {}_3F_2[1-\gamma, a, a+b-x; a+b, a+b+1, 1]$$

Taking r=1, 2, 3, 4 in (17), one gets first four raw moments as

$$\mu_1' = \frac{nb\beta(a,b)}{(a+b)(a+b-1)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1]$$

$$\mu_2' = \frac{nb\beta(a,b)}{(a+b)(a+b-1)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1] + \frac{2n(n+1)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-2; a+b, a+b+1, 1]$$

$$\mu_3' = \frac{nb\beta(a,b)}{(a+b)(a+b-1)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1] + \frac{6n(n+1)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-2; a+b, a+b+1, 1] + \frac{6n(n+1)(n+2)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)(a+b-3)\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-3; a+b, a+b+1, 1]$$

$$\mu_4' = \frac{nb\beta(a,b)}{(a+b)(a+b-1)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1] + \frac{14n(n+1)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-2; a+b, a+b+1, 1] + \frac{36n(n+1)(n+2)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)(a+b-3)\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-3; a+b, a+b+1, 1] + \frac{24n(n+1)(n+2)(n+3)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)(a+b-3)(a+b-4)\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-4; a+b, a+b+1, 1]$$

Descending Factorial Moments of GNPED

The r[th] descending factorial moment of (5) can be obtained as

$$E(X^{(r)}) = E[E(X^{(r)}/p)],$$

Where for given p, the random variable 'X' follows negative binomial distribution with

$$E(X^{(r)}/p) = (n+r-1)^{(r)} p^{-r}(1-p)^r$$

Proceeding in the same way as above, the r[th] descending factorial moment of GNPED is

$$\mu'_{(r)} = \frac{(n+r-1)^{(r)}(a+b-r-1)r!b\beta(a,b)}{(a+b)!\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-r; a+b, a+b+1, 1] \quad (18)$$

Taking r=1, 2, 3, 4 in (18), one gets first four factorial moments as

$$\mu'_{(1)} = \frac{nb\beta(a,b)}{(a+b)(a+b-1)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1]$$

$$\mu'_{(2)} = \frac{2n(n+1)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)\beta(a,b+\gamma)} {}_3F_2[1-\gamma, a, a+b-2; a+b, a+b+1, 1]$$

$$\mu'_{(3)} = \frac{6n(n+1)(n+2)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)(a+b-3)\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-3; a+b, a+b+1, 1]$$

$$\mu'_{(4)} = \frac{24n(n+1)(n+2)(n+3)b\beta(a,b)}{(a+b)(a+b-1)(a+b-2)(a+b-3)(a+b-4)\beta(a,b+\gamma)} \times {}_3F_2[1-\gamma, a, a+b-4; a+b, a+b+1, 1]$$

Central Moments of GNPED

$$\mu_2 = \frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)} {}_3F_2[1-\gamma, a, a+b-1; a+b, a+b+1, 1]$$

$$+\frac{2n(n+1)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)}$$
$${}_3F_2[1-\gamma,a,a+b-2;a+b,a+b+1,1]$$
$$-\left[\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right]^2$$

$$\mu_3=\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}$$
$${}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]$$
$$-3\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^2$$
$$+2\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^3$$
$$+\frac{6n(n+1)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)}$$
$${}_3F_2[1-\gamma,a,a+b-2;a+b,a+b+1,1]$$
$$\times\left\{1-\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}$$
$$+\frac{6(n+1)(n+2)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)(a+b-3)}$$
$$\times{}_3F_2[1-\gamma,a,a+b-3;a+b,a+b+1,1]$$

$$\mu_4=\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}$$
$${}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]$$
$$+\frac{2n(n+1)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)}$$
$${}_3F_2[1-\gamma,a,a+b-2;a+b,a+b+1,1]$$
$$\times\left[\ 7-\frac{12nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\right.$$
$${}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]$$
$$\left.+6\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^2\ \right]$$
$$+\frac{12n(n+1)(n+2)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)(a+b-3)}$$
$$\times{}_3F_2[1-\gamma,a,a+b-3;a+b,a+b+1,1]$$
$$\times\left[\ 3-\frac{2nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\right.$$
$${}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\ \left.\right]$$

$$+\frac{24(n+1)(n+2)(n+3)b\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)(a+b-2)(a+b-3)(a+b-4)}$$
$$\times{}_3F_2[1-\gamma,a,a+b-4;a+b,a+b+1,1]$$
$$-4\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^2$$
$$+6\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^3$$
$$-3\left\{\frac{nb\beta(a,b)}{\beta(a,\gamma+b)(a+b)(a+b-1)}\ {}_3F_2[1-\gamma,a,a+b-1;a+b,a+b+1,1]\right\}^4$$

Goodness of Fit

Gupta and Ong (2004) obtained GNBD by mixing NBD with generalized gamma distribution defined by Amero and Bayrr (1933) and Agarwal and Kalla (1996). The pmf of GNBD is

$$P(X=x)= \tag{19}$$
$$\binom{m+x-1}{x}\left(\frac{\alpha}{1+\alpha}\right)^{m-\lambda}\left(\frac{1}{1+\alpha}\right)^{x}$$
$$\times\frac{\phi(\lambda,\lambda-m+1-x;(\alpha+1)n)}{\phi(\lambda,\lambda-m+1;\alpha n)}$$

Where $(m,\alpha,\lambda,n)$ are the parameters of the distribution and $\phi(\lambda,\lambda-m+1-x;(\alpha+1)n)$ is a confluent hypergeometric function.

Gupta and Ong demonstrated the goodness of fit test for their model (19) with the help of two data sets [Tables (1)-(2)] and observed marked fist than NBD and Jain and Consul's (1971) GNBD.

In this section, the proposed model GNPED has also been fitted to these data sets to show that the proposed model exhibits the best fit as compared to other distributions such as NBD, Jain and Consul's (1971) GNBD and Gupta and Ong's (2004) GNBD. The negative Polya-Eggenberger distribution has also been fitted to these data sets for its comparison with these distributions.

Table 1 Absenteeism among shift-workers in steel industry; data of Arbous and Sichel, 1954

| **Count** | **Observed Frequency** | **EXPECTED FREQUENCY** | | | | |
|---|---|---|---|---|---|---|
| | | **NBD** | **Jain and Consul's (1971) GNBD** | **Ramesh and Ong's (2004) GNBD** | **NPED** | **PROPOSED MODEL GNPED** |
| 0 | 7 | 12.02 | 10.51 | 09.23 | 9.53 | 9.06 |
| 1 | 16 | 16.16 | 17.45 | 16.18 | 15.93 | 16.79 |
| 2 | 23 | 17.77 | 20.38 | 19.86 | 19.06 | 23.62 |
| 3 | 20 | 18.08 | 20.80 | 21.06 | 19.92 | 22.89 |
| 4 | 23 | 17.65 | 19.88 | 20.50 | 19.41 | 21.95 |
| 5 | 24 | 16.80 | 18.34 | 18.78 | 18.17 | 20.67 |
| 6 | 12 | 15.72 | 16.56 | 16.46 | 16.59 | 17.11 |
| 7 | 13 | 14.52 | 14.78 | 14.02 | 14.90 | 15.24 |
| 8 | 09 | 13.28 | 13.08 | 11.79 | 13.25 | 11.04 |
| 9 | 09 | 12.06 | 11.53 | 09.95 | 11.71 | 8.78 |
| 10 | 08 | 10.89 | 10.13 | 08.55 | 10.30 | 8.04 |
| 11 | 10 | 09.78 | 08.89 | 07.54 | 9.04 | 7.21 |
| 12 | 08 | 08.75 | 07.79 | 06.84 | 7.92 | 6.38 |
| 13 | 07 | 07.80 | 16.83 | 06.33 | 6.94 | 5.82 |
| 14 | 02 | 06.93 | 05.99 | 05.94 | 6.08 | 5.24 |
| 15 | 12 | 06.14 | 05.26 | 05.61 | 5.33 | 4.73 |
| 16 | 03 | 05.43 | 04.61 | 05.29 | 4.68 | 4.27 |
| 17 | 05 | 04.79 | 04.05 | 04.97 | 4.12 | 3.96 |
| 18 | 04 | 04.22 | 03.56 | 04.64 | 3.63 | 3.69 |
| 19 | 02 | 03.17 | 03.14 | 04.28 | 3.20 | 3.46 |
| 20 | 02 | 03.23 | 02.76 | 03.92 | 2.83 | 3.27 |
| 21 | 05 | 02.86 | 02.43 | 03.55 | 2.50 | 2.98 |
| 22 | 05 | 02.50 | 02.15 | 03.19 | 2.22 | 2.88 |
| 23 | 02 | 02.91 | 01.90 | 02.84 | 1.97 | 2.67 |
| 24 | 01 | 01.91 | 01.68 | 02.50 | 1.75 | 2.16 |
| 25-48 | 16 | 12.77 | 13.50 | 14.13 | 17.02 | 14.09 |
| **TOTAL** | **248** | **248** | **248** | **248** | **248** | **248** |
| **Estimates** | | p=0.854<br>$n = 1.576$ | $\alpha$ =0.00010775<br>$\beta$ =5978.5288<br>n=29337.0839 | $\lambda$ =0.6226<br>$\alpha$ =0.001<br>m=0.1601<br>n=0.01897 | n=14.962954<br>α=2.492821<br>γ=4.852530 | n=100.09367<br>γ=1.00021<br>a=2.24578<br>b=2.26398 |
| $\chi^2$ | | **14.92** | **27.79** | **8.27** | **10.20108** | **7.621862** |
| **d. f** | | 17 | 17 | 15 | 16 | 15 |

Table 2 Counts of the number of European red mites on apple leaves; data of P.Garman, 1951

| **Count** | **Observed Frequency** | **EXPECTED FREQUENCY** | | | | |
|---|---|---|---|---|---|---|
| | | **NBD** | **Jain and Consul(1971) GNBD** | **Ramesh and Ong's (2004) GNBD** | **NPED** | **PROPOSED MODEL GNPED** |
| 0 | 70 | 69.49 | 69.49 | 70.24 | 69.19 | 70.92 |
| 1 | 38 | 37.60 | 37.60 | 37.05 | 38.27 | 36.87 |
| 2 | 17 | 20.10 | 20.10 | 18.06 | 20.09 | 19.41 |
| 3 | 10 | 10.70 | 10.70 | 11.03 | 10.49 | 10.34 |
| 4 | 09 | 05.69 | 05.69 | 06.89 | 6.51 | 6.57 |
| 5 | 03 | 03.02 | 3.02 | 03.79 | 2.93 | 3.04 |
| 6 | 02 | 01.60 | 1.60 | 01.79 | 1.57 | 1.67 |
| 7 | 01 | 00.85 | 0.85 | 00.74 | 0.86 | 0.93 |
| 8 | 00 | 00.95 | 0.95 | 00.40 | 0.09 | 0.25 |
| **TOTAL** | 150 | 150 | 150 | 150 | 150 | 150 |
| **ML Estimate**<br><br><br><br>$\chi^2$<br>**d.f** | | p=0.5281<br>n=1.0246<br><br><br>**2.484**<br>3 | $\alpha = 0.52810$<br>$\beta = 1.000$<br>$m = 1.0246$<br><br>**2.484**<br>**2** | $\lambda$ =65.170<br>$\alpha$ =1.73908<br>m=66.6914<br>n=0.001<br><br><br>**0.93**<br>**1** | n=22.924537<br>α =1.167381<br>γ=24.297517<br><br>**1.517443**<br>2 | n=37.72660<br>γ=0.99869<br>a=16.91825<br>b=16.91825<br><br><br>**1.257801**<br>1 |

*Note***:** The expected frequencies and the estimates for the parameters of the Jain and Consul's (1971) generalized negative binomial distribution are same at $\beta = 1$ as given by negative binomial distribution shown in column third of the table (2).

The maximum likelihood estimate of the parameters of the proposed model have been obtained and shown at their respective places in the tables. It is mentioned here that due to complicated likelihood function, the ML estimates are determined by the same method as used by Gupta and Ong (2004) i.e. by a direct numerical search for global maximum of the log-likelihood surface. A random start procedure is employed i.e. for a set of random starting points this numerical search is repeated for each starting point in order to verify that the global maximum has been found.

It is evident from the tables 1 and 2 that the chi-square values of the proposed model (GNPED), in all the cases, gives the marked fit as compared to other distributions.

References

Beall, G. (1940).The fit and significance of contagious distributions when applied to observations on larval insects. *Ecology*, *21*, 460-474.

Beall G, & Rescia, R.R. (1953).A generalization of Neyman's contagious distributions. *Biometrics*, *9*, 354-57.

Bliss, C. I.,, Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. *Biometrics*, *9*, 176-200.

Bose, P.K. (1944). On confluent hyper geometric series. *Sankhya*, *6*, 407-412.

Dubourdieu, J. (1939) *Theorie de l'assurance-maladie*, Paris.

Eggenberger, F., & Polya, G. (1923). Uber die Statistik verketteter vorgange. Z.Angew. *Math. Mech.*,*1*, 279-289.

Engen, S. (1974**).** On species frequency model. *Biometrika*, *61*, 263-270

Engen, S. (1978). *Stochastic abundance models*. London: Chapman and Hall.

Evans, D. A. (1953). Experimental evidence concerning contagious distributions in ecology. *Biometrika*, *40*,186-211.

Feller, W. (1943). On a general class of contagious distributions. *Ann. Math. Stat.*, *14*, 389-400.

Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distribution representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Stat. Soc.*, Ser. A 83, 255-279.

Jain G. C., & Consul, P. C. (1971). A Generalized Negative Binomial Distribution. *SIAM. J. Appl. Math*, *21*(.4.) 501-513

Johnson, N. L., & Kotz, S (1969). *Univariate discrete distribution*. NY: John Wiley.

Lundberg, O. (1940). *On random processes and their application to sickness and accident statistics.* Thesis, University of Stockholm.

Nath, P. (1951) Confluent hyper geometric function. *Sankhya*, *11*, 153-166.

Newbold, E. (1927). Practical applications of the statistics of repeated events, particularly to industrial accidents. *J. Roy. Stat. Soc.*, *90*, 487-547.

Neyman, J. (1939). On a new class of contagious distribution applicable in entomology and bacteriology. *Ann. Math. Stat.*, *10*, 35-57.

Slater, L.J (1953).On the evaluation of the confluent hyper geometric function. *Proceedings of the Cambridge Philosophical Society*, *49*, 612-622.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 667
Copyright © 2007 JMASM, Inc.
1538 – 9472/05/$95.00

# *Letters to the Editor*

Lui K. J. (2006). Interval estimation of risk difference in simple compliance randomized trials. *Journal of Modern Applied Statistical Methods,* 5, 395–407.

Ian R. White
MRC Biostatistics Unit

---

Professor Lui (2006) reports a careful comparison of the properties of six possible interval estimators for the causal risk difference among treatment-compliers[1]. He recommends for general use the confidence interval based on a $\tanh^{-1}$ transformation of the causal risk difference, on the grounds that it has at least the nominal coverage and it has the smallest mean length of all the methods.

However, the second of these criteria is not self-evidently the most relevant, and there are other possible criteria which would point to a different choice of interval estimator.

1. Some interval estimators with large mean length are valuable and in common use. An example is the number needed to treat, defined as the inverse of the risk difference. The appropriate confidence interval for the number needed to treat includes the inverse of all values in the confidence interval for the risk difference: in particular, it includes infinity if the confidence interval for the risk difference includes zero[2]. This interval in fact has infinite mean length, but it remains appropriate and widely used, if sometimes misunderstood.
2. More generally, mean confidence interval length is a scale-dependent criterion: when the parameter is transformed to a different scale, confidence intervals retain their coverage properties but not their mean length. Thus mean length on different scales could have been considered.
3. Rather than require coverage to be *at least* the nominal coverage, one could require coverage that is *close to* the nominal coverage. Professor Lui's recommended method has over 98% coverage for nominal 95% confidence intervals in many of the simulation settings.
4. A further criterion in the treatment-compliance setting is that one could require confidence intervals to agree with the intention-to-treat P-value, by excluding zero if and only if the intention-to-treat test is significant. This is an appropriate requirement because the null hypotheses for the intention-to-treat and compliance-adjusted analyses are the same and there is no gain in power from allowing for non-compliance in this setting[3]. Confusion in interpretation could easily arise if adjustment for non-compliance in a particular data set appeared to change a non-significant result into a significant one or vice versa.

The Fieller's theorem confidence interval has properties 3 and 4 above[4]. By its derivation, it agrees exactly with the intention-to-treat P-value computed from an asymptotic test (use of an exact intention-to-treat test would make the equivalence only approximate). Its coverage is therefore close to the nominal, as shown in Professor Lui's simulation study. I therefore believe that the Fieller's theorem confidence interval should also be considered for use in practice, especially when testing the null hypothesis of no intervention effect is important.

## References

Altman DG. Confidence intervals for the number needed to treat. *British Medical Journal* 1998; 317: 1309–1312.

Branson M, Whitehead J. A score test for binary data with patient non-compliance. *Statistics in Medicine* 2003; 22: 3115–3132.

Frost C, Thompson SG. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society (A)* 2000; 163: 173–189.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 668-669
Copyright © 2007 JMASM, Inc.
1538 – 9472/07/$95.00

# Reply

Kung-Jong Lui
San Diego State University

When evaluating the performance of an interval estimator, we generally use the coverage probability to measure the accuracy and the average length to measure the precision (Casella & Berger, 1990). An ideal interval estimator is the one which can consistently cover the underlying true parameter for all parameter values, while its average length is minimal so that one can almost pinpoint the underlying true parameter. In practice, however, such an ideal interval estimator does not exist. Note that an interval estimator, which has a high coverage probability but has a quite wide length, is of little practical value. For example, the interval estimate $(0, \infty)$ has the coverage probability of 100% covering a positive parameter, but is useless due to its length is too wide to be informative. Following the same arguments, we can easily see that the interval estimate [-1, 1] that also has the coverage probability of 100% for the difference between two proportions is also completely useless. Thus, the information on the coverage probability of an interval estimator alone is not sufficient to determine whether it can perform well or not. Given two interval estimators with the same coverage probability, the interval estimator with a shorter average length is obviously preferable to the other with a longer average length. This is because the former can allow us to draw a more precise inference. On the other hand, an interval estimator which has a short average length but has a low coverage probability is also of no practical value. These lead us to consider finding an interval estimator which has the shortest average length among all interval estimators with the coverage probability consistently larger than or equal to the desired confidence level.

Note that obtaining an interval estimate with an infinite length only suggests that the employed interval estimator based on the given data cannot provide us with an accurate estimate of the underlying parameter. This certainly does not imply that the interval estimator with an infinite length is valuable and useful. In fact, there are many problems and concerns by simply inverting the interval estimate for the risk difference to obtain an interval estimate for the number needed to treat (NNT). A systematic list of these concerns and references as well as a simple logic solution to alleviate these concerns can be found elsewhere (Lui, 2004).

It is incorrect and misleading to state that "When the parameter is transformed to a different scale, confidence interval retains their coverage properties, but not their mean length. Thus, mean length on different scales could have been considered". First, this statement about the confidence interval is generally not true unless the transformation is, for example, continuous and monotonic. The mean interval length, just like the standard error, has a unit scale. This certainly does not deter its use once when the parameter of primary interest is selected. The average length for all interval estimators will have the same unit scale as that for the parameter of interest. Thus, there will be no concern that we may compare the average length of different interval estimators at different unit scales. Note also that the relative precision is not invariant with respect to the reciprocal transformation and hence a relatively more precise interval estimate for the risk difference does not necessarily lead to produce a relatively more precise interval estimate for the NNT.

Because the sampling distribution of a statistic on which we are based to derive an interval estimator is not necessarily symmetric, we can obtain an interval estimator with the coverage probability larger than the other one, but the former also has the average length less than the latter. For example, as shown elsewhere (Lui, 2006), we can easily find the situations in which interval estimator (4) using $\tanh^{-1}(x)$ transformation has the largest

coverage probability and the shortest average length among interval estimators considered in the paper. It is senseless to put a penalty on an interval estimator when its coverage probability can be even higher than the desired confidence level without sacrificing its precision. Based on the coverage probability exclusively, we indiscriminately select which interval estimator is the best can be subject to the above concern.

It is certainly desirable that test results between using hypothesis testing and various interval estimators can always be consistent with each other. If readers wish to have this property, test-based confidence intervals will be the choice. However, for given an adequately large sample size, the chance to obtain an inconsistent conclusion between hypothesis testing (in which we generally account for the null conditions when calculating the estimated variance of the test statistic) and interval estimators (in which we calculate the estimated variance of statistic without having the null conditions) should be generally small.

## References

Casella, G. & Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury.

Lui, K.-J. (2004). A simple logical solution to eliminate the limitations of using the number needed to treat. *Evaluation & the Health Professions*, 27, 206-214.

Lui, K.-J. (2006). Interval estimation of risk difference in simple compliance randomized trials. *Journal of Modern Applied Statistical Methods*, 5, 395-407.