5-1-2004

Vol. 3, No. 1 (Full Issue)

JMASM Editors

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

Recommended Citation

Editors, JMASM (2004) "Vol. 3, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 1 , Article 30.

DOI: 10.22237/jmasm/1083369600

Available at: <http://digitalcommons.wayne.edu/jmasm/vol3/iss1/30>

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

Two Years in the Making...

Intel® Visual Fortran 8.0

The next generation of Visual Fortran is here! Intel Visual Fortran 8.0 was developed jointly by Intel and the former DEC/Compaq Fortran engineering team.

Visual Fortran Timeline

- 1997** DEC releases Digital Visual Fortran 5.0
- 1998** Compaq acquires DEC and releases DVF 6.0
- 1999** Compaq ships CVF 6.1
- 2001** Compaq ships CVF 6.6
- 2001** Intel acquires CVF engineering team
- 2003** Intel releases Intel Visual Fortran 8.0

Intel Visual Fortran 8.0

- CVF front-end + Intel back-end
- Better performance
- OpenMP Support
- Real*16

Performance

Outstanding performance on Intel architecture including Intel® Pentium® 4, Intel® Xeon™ and Intel Itanium® 2 processors, as well as support for Hyper-Threading Technology.

Compatibility

- Plugs into Microsoft Visual Studio® .NET
- Microsoft PowerStation4 language and library support
- Strong compatibility with Compaq® Visual Fortran

Support

1 year of free product upgrades and Intel Premier Support

"The Intel Fortran Compiler 7.0 was first-rate, and Intel Visual Fortran 8.0 is even better. Intel has made a giant leap forward in combining the best features of Compaq Visual Fortran and Intel Fortran. This compiler... continues to be a 'must-have' tool for any Twenty-First Century Fortran migration or software development project."

—Dr. Robert R. Trippi
Professor Computational Finance
University of California, San Diego

FREE trials available at:
programmersparadise.com/intel

Programmer's Paradise®

To order or request additional information call:
800-423-9990
Email: intel@programmers.com

Journal Of Modern Applied Statistical Methods

Invited Articles

2 - 12	Rand R. Wilcox, H. J. Keselman	Multivariate Location: Robust Estimators And Inference
13 - 26	James Algina, H. J. Keselman	A Comparison Of Methods for Longitudinal Analysis With Missing Data

Regular Articles

27 - 38	H. J. Keselman, Rand R. Wilcox, James Algina, Katherine Fradette, Abdul R. Othman	A Power Comparison Of Robust Test Statistics Based On Adaptive Estimators
39 - 48	Suzanne Dubnicka	A Rank-based Estimation Procedure For Linear Models With Clustered Data
49 - 53	Yonghong Gao	Depth Based Permutation Test For General Differences In Two Multivariate Populations
54 - 64	Camil Fuchs, Vance W. Berger	Quantifying The Proportion Of Cases Attributable To An Exposure
65 - 71	Todd Headrick	On Polynomial Transformations For Simulating Multivariate Non-normal Distributions
72 - 84	Michael B. C. Khoo	An Alternative Q Chart Incorporating A Robust Estimator Of Scale
85 - 103	Felix Famoye, Carl Lee, Nicholas Eugene	Beta-Normal Distribution: Bimodality Properties And Application
104 - 116	S. James Press	Respondent-Generated Intervals (RGI) For Recall in Sample Surveys
117 - 133	Hani M. Samawi, Laith J. Saeid	Stratified Extreme Ranked Set Sample With Application To Ratio Estimators
134 - 142	Mohammad F. Al-Saleh, Hani M. Samawi	Estimation Using Bivariate Extreme Ranked Set Sampling With Application To The Bivariate Normal Distribution

143 - 148	Omar Eidous, Ayman Baklizi	Kernel-Based Estimation of $P(X < Y)$ With Paired Data
149 - 157	Omar M. Eidous	Some Improvements in Kernel Estimation Using Line Transect Sampling
158 - 164	Mehmet Ali Cengiz	Accurate Binary Decisions For Assessing Coronary Artery Disease
165 - 175	Jeffrey R. Wilson	A Generalized Quasi-Likelihood Model Application To Modeling Poverty Of Asian American Women
176 - 185	Aurelio Tobias, Mark Saez, Manolis Kogevinas	Meta-Analysis Of Results And Individual Patient Data In Epidemiological Studies
186 - 199	David A. Walker	Validation Studies: Matters Of Dimensionality, Accuracy, And Parsimony With Predictive Discriminant Analysis And Factor Analysis
200 - 212	Dongfeng Wu	A Visually Adaptive Bayesian Model In Wavelet Regression

Brief Reports

213 - 214	J. Wanzer Drane, Gregory Thatcher	A Test-Retest Transition Matrix: A Modification of McNemar's Test
215 - 220	Amjad D. Al-Nasser	Estimation Of Multiple Linear Functional Relationships
221 - 226	Shlomo Sawilowsky	Teaching Random Assignment: Do You Believe It Works?

Early Scholars

227 - 233	Jan Perkins, Daniel Wang	A Comparison of Bayesian And Frequentist Statistics As Applied In A Simple Repeated Measures Example
234 - 238	Patric R. Spence	On The Reporting Of Reliability In Content Analysis

JMASM Algorithms and Code

239 - 249	Andrés M. Alonso	JMASM10: A Fortran Routine For Sieve Bootstrap Prediction Intervals
250 - 258	James F. Reed	JMASM11: Comparing Two Small Binomial Proportions

End Matter

259 - 260	Author	Statistical Pronouncements III

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

Three areas are appropriate for *JMASM*: (1) development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) development or study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods. Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

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

Editorial Assistant
Patric R. Spence

Professional Staff
Bruce Fay,
Business Manager
Joe Musial,
Marketing Director

Production Staff
Christina Gase
Jack Sawilowsky

Internet Sponsor
Paula C. Wood,
Dean
College of Education,
Wayne State University

Entire Reproductions and Imaging Solutions
Internet: www.entire-repro.com

248.299.8900 (Phone)
248.299.8916 (Fax)

e-mail:
sales@entire-repro.com

Editorial Board of Journal of Modern Applied Statistical Methods

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

Chris Barker
MEDTAP International
Redwood City, CA

J. Jackson Barnette
Community and Behavioral Health
University of Iowa

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

Ling Chen
Department of Statistics
Florida International University

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

Jai Won Choi
National Center for Health Statistics
Hyattsville, MD

Rahul Dhanda
Forest Pharmaceuticals
New York, NY

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

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

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

Carol J. Etzel
University of Texas M. D.
Anderson Cancer Center

Felix Famoye
Department of Mathematics
Central Michigan University

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

Shiva Gautam
Department of Preventive Medicine
Vanderbilt University

Dominique Haughton
Mathematical Sciences Department
Bentley College

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

Joseph Hilbe
Departments of Statistics/ Sociology
Arizona State University

Peng Huang
Dept. of Biometry & Epidemiology
Medical University of South Carolina

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

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

Harry Khamis
Statistical Consulting Center
Wright State University

Kallappa M. Koti
Food and Drug Administration
Rockville, MD

Tomasz J. Kozubowski
Department of Mathematics
University of Nevada

Kwan R. Lee
GlaxoSmithKline Pharmaceuticals
Collegeville, PA

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

Devan V. Mehrotra
Merck Research Laboratories
Blue Bell, PA

Prem Narain
Freelance Researcher
Farmington Hills, MI

Balgobin Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute

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

Brent Jay Shelton
Department of Biostatistics
University of Alabama at Birmingham

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

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

Joshua M. Tebbs
Department of Statistics
Kansas State University

Dimitrios D. Thomakos
Department of Economics
Florida International University

Justin Tobias
Department of Economics
University of California-Irvine

Jeffrey E. Vaks
Beckman Coulter
Brea, CA

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

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

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

Dongfeng Wu
Dept. of Mathematics & Statistics
Mississippi State University

Chengjie Xiong
Division of Biostatistics
Washington University in St. Louis

Andrei Yakovlev
Biostatistics and Computational Biology
University of Rochester

Heping Zhang
Dept. of Epidemiology & Public Health
Yale University

International
Mohammed Ibrahim Ali Ageel
Department of Mathematics
King Khalid University, Saudi Arabia

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

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

Debasis Kundu
Department of Mathematics
Indian Institute of Technology, India

Christos Koukouvinos
Department of Mathematics
National Technical University, Greece

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

Takis Papaioannou
Statistics and Insurance Science
University of Piraeus, Greece

Mohammad Z. Raqab
Department of Mathematics
University of Jordan, Jordan

Nasrollah Saebi
School of Mathematics
Kingston University, UK

Keming Yu
Statistics
University of Plymouth, UK

INVITED ARTICLES

Multivariate Location: Robust Estimators And Inference

Rand R. Wilcox
Department of
Psychology
University of Southern
California

H. J. Keselman
Department of
Psychology
University of Manitoba

The sample mean can have poor efficiency relative to various alternative estimators under arbitrarily small departures from normality. In the multivariate case, (affine equivariant) estimators have been proposed for dealing with this problem, but a comparison of various estimators by Massé and Plante (2003) indicated that the small-sample efficiency of some recently derived methods is rather poor. This article reports that a skipped mean, where outliers are removed via a projection-type outlier detection method, is found to be more satisfactory. The more obvious method for computing a confidence region based on the skipped estimator (using a slight modification of the method in Liu & Singh, 1997) is found to be unsatisfactory except in the bivariate case, at least when the sample size is small. A much more effective method is to use the Bonferroni inequality in conjunction with a standard percentile bootstrap technique applied to the marginal distributions.

Keywords: Outlier detection; Tukey's halfspace depth, skipped estimators, outlier-projection estimator

---

Introduction

A fundamental problem is estimating a measure of location associated with some multivariate distribution and then computing a confidence region based on the estimator used. Of course, the sample mean performs well under normality based on various well-known criteria. However, from an applied point of view, there are compelling reasons to consider alternative measures of location. One has to do with the effects of outliers on efficiency. Tukey (1960) predicted that outliers are common in applied

work, and modern outlier detection methods indicate that this is indeed the case (outlier detection rules based on the mean and usual covariance matrix, in conjunction with a Mahalanobis distance, are well known to be unsatisfactory; see for example, Rousseeuw & Leroy, 1987).

Moreover, arbitrarily small departures from normality (based on any of several metrics for comparing distributions) can result in outliers commonly appearing in a random sample which in turn can mean poor efficiency when using the sample mean. Another concern is that when sampling from an asymmetric distribution, the population mean can poorly reflect what is typical.

In the univariate case, many alternatives to the sample mean have been proposed (e.g., Andrews et al., 1972). Several maintain relatively high accuracy under normality relative to the sample mean and have high efficiency in situations where the sample mean performs poorly. Simultaneously, inferential (hypothesis testing) methods have been found that perform

---

Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology at the University of Southern California and H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology at the University of Manitoba. Work on this project was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

well over a broad range of situations, including general conditions where methods based on means perform in an unsatisfactory manner (e.g., Wilcox, 1997; 2003).

As for the multivariate case, one could of course simply apply univariate estimators to the marginal distributions, but it is known that usually this does not satisfy a criterion that is frequently imposed. To elaborate, first consider the univariate case, let X be any random variable having distribution F , let $T(X_1, \dots, X_n)$ be some statistic based on the random sample $X_1, \dots, X_n$, and let a and b be any two constants. Then for $T(X_1, \dots, X_n)$ to qualify as a measure of location, a minimum requirement typically imposed is that

$$T[a(X_1, \dots, X_n) + b] = aT(X_1, \dots, X_n) + b$$

(e.g., Staudte & Sheather, 1990). The usual population mean and median satisfy this requirement as do many other robust measures of location. This requirement says, for example, that given a typical measure of temperature in Fahrenheit, if converted to Celsius, the typical measure should be transformed in the obvious way.

Now consider the case where X is any p -variate random variable, A is any nonsingular square matrix, and B is a vector having length p . Then $T(X_1, \dots, X_n)$ is said to be an affine equivariant measure of location if

$$T(X_1 A, \dots, X_n A + B) = T(X_1, \dots, X_n) A + B. \quad (1)$$

So the measure of location is transformed properly under rotations as well as changes in scale and shifts in the possible values of X . There are many robust affine equivariant measures of location in the univariate case, but typically, if they are applied to the marginal distributions in the multivariate case, they are no longer affine equivariant. For example, the marginal medians are not affine equivariant as noted by Donoho and Gasko (1992). So a general goal has been to search for affine equivariant location estimators in the multivariate case that guard against the deleterious effects of outliers.

One of the earliest affine equivariant estimators that guards against outliers was proposed by Rousseeuw and Leroy (1987) and is known as the minimum volume ellipsoid (MVE) estimator. It begins by searching for the ellipsoid containing half of the data that has the smallest volume. If the sample mean is computed based on this half of the data, ignoring the other half, it is evident that it guards against outliers, but efficiency is poor compared to the usual sample mean when sampling from a normal distribution. More recently, Rousseeuw and van Driessen (1999) argued that the MVE estimator be replaced by the minimum covariance determinant (MCD) estimator which searches for a subset of half of the data having the smallest generalized variance. But like the MVE estimator, efficiency is low when sampling from a multivariate normal distribution.

Donoho and Gasko (1992) studied a multivariate location estimator that is based in part on Tukey's notion of halfspace depth. Their approach is of direct interest in this article and details are given later in the paper. But before continuing, a rough outline of their strategy helps. The basic idea is to quantify how deeply each point X_i is nested within the cloud of data, and then eliminate a fixed proportion of those points that are not deeply nested.

That is, use the centrally located data to estimate a measure of location and ignore the data on the edges of the data cloud. In the univariate case, their estimator reduces to a trimmed mean which is known to have many practical advantages. In particular, a 20% trimmed mean (where the largest 20% and the smallest 20% of the observed values are trimmed, and the average of the remaining data is used) maintains reasonably high efficiency under normality (e.g., Rosenberger & Gasko, 1987). This raises the issue of whether a similar amount of trimming performs well when working with multivariate data, and it is found that this is not the case.

Yet another approach was recently proposed by Liu, Parelius and Singh (1999) and represents a generalization of the method studied by Donoho and Gasko (1992). One difference is that Liu et al. consider a wider choice of methods for measuring the depth of a point

within a data cloud. The particular generalization of a univariate trimmed mean described by Liu et al. (1999, pp. 795-796) was considered in this study, but it did not correct the problems with the Donoho and Gasko trimmed mean described later in this article. So for brevity, the complex computational details of their method are not described here.

One more general approach is to first search for outliers using some affine equivariant method, roughly meaning that if the data are transformed as indicated by the left side of equation 1, outliers before transforming the data remain outliers after transformation. Next, eliminate any outliers that are found and simply compute the mean of the remaining data. In the univariate case, this general strategy yields what is known as a skipped estimator, so the term is used here.

The focus here is on one particular outlier detection method for reasons discussed later in the paper. The choice of method is not arbitrary, but it is stressed that alternative outlier detection techniques might be found to have practical value in future studies. It is noted that Massé and Plante (2003) compared the efficiency of several multivariate estimators and found all of the affine equivariant estimators to have relatively poor efficiency under normality. The skipped estimator studied here corrects that problem.

The Estimators Studied

This section provides a formal description of the six estimators considered. Four of the estimators belong to the class of generalized trimmed means studied by Donoho and Gasko (1992); four different amounts of trimming are considered. Results in Massé and Plante (2003) indicate that these estimators can be unsatisfactory, and previous results, based on other distributions and criteria, support their conclusions. The fifth estimator is based on removing outliers with a projection-type method and averaging the values that remain, and the sixth is the usual median of the marginal distributions. Although this last estimator is not affine equivariant, it is included with the goal of adding perspective on the expected accuracy of the other estimators considered.

The Donoho-Gasko Trimmed Mean

The Donoho and Gasko (1992) estimator is based on Tukey's notion of halfspace depth, which represents an approach to generalizing the notion of ranks to multivariate data. An important feature of Tukey's depth is that no assumptions are made about the distribution from which observations are randomly sampled. In particular, it is not assumed that the distribution is elliptical.

A formal definition of Tukey's depth is relegated to an appendix. To provide some intuitive sense of Tukey's strategy we duplicate a description found in Wilcox (in press). Look at Figure 1 which shows a scatterplot of electroencephalographic (EEG) measures taken at two sites in the brain. These data are from Raine, Buchsbaum and LaCasse (1997) where the general goal was to investigate brain abnormalities among murderers.

Consider the left most point indicated by a circle and imagine any line going through this point. Any line forms what are called two halfspaces. The points on or above a line form a closed halfspace, and the same is true for all of the points on or below the line. Because the left point indicated by the circle is located on the edge of the scatterplot, it is evident that a line can be drawn through this point so that it is the only point in one of the closed halfspaces. Now consider the right circle. Because it is more deeply nested within the scatterplot, a relatively large proportion of the scatterplot will be on or above any line drawn through this point, and a relatively large proportion will be on or below the line as well.

For any line L drawn through a point, consider the proportion of points on or above this line, as well as on or below this line, and let P_L be the smaller of these two proportions. Then Tukey's depth is the smallest P_L value among all lines L . For p -variate data (where L becomes a plane), the maximum depth among a scatterplot of points can be as high as $1/2$ or as low as $1/(p+1)$ (Donoho & Gasko, 1992). So for bivariate data ($p=2$), if the depth for every point were computed, it is possible that the

Figure 1. EEG measures used to illustrate Tukey's notion of depth.

largest depth would not exceed $1/3$, but it could be as high as $.5$.

Tukey's notion of depth can be computed exactly in the bivariate case (Rousseeuw & Ruts, 1996). In Figure 1, there are $n = 14$ points, and the halfspace depth for the left circle is $1/14$. For the right circle, the halfspace depth is $5/14$. For p -variate data, $p > 2$, Rousseeuw and Struyf (1998) describe an approximation of Tukey's depth which is used here.

The Donoho-Gasko analog of the γ -trimmed mean, $\hat{\xi}_\gamma$, is the average of all points which are at least γ deep in the sample. That is, points having depth less than γ are trimmed and the mean of the remaining points is computed. For example, suppose $\gamma = .1$ and consider again the data in Figure 1. There are four points that have a depth less than $.1$ so the $.1$ trimmed mean

is the mean after these four points are eliminated. Because the maximum depth is not necessarily $.5$, a generalization of the median, often called Tukey's median, is taken to be the mean of the points having the maximum depth. In Figure 1, the maximum depth is $.357$ which corresponds to only one point: $(.58, -.65)$.

A Skipped Estimator

As previously indicated, a skipped estimator is the sample mean of the data after outliers have been removed. A practical problem is not finding a reasonable outlier detection method for multivariate data, but rather choosing a method from among the many that have been proposed. Rousseeuw, Ruts and Tukey (1999) suggest a method based on the notion of halfspace depth. They focus mainly on the bivariate case, but in principle the method can be used when $p > 2$; also see Liu et al. (1999) as well as Romanazzi (1997).

An approach based on convex hull peeling is discussed by Zani, Riani and Corbellini (1998) but is known to be somewhat less robust than halfspace depth as shown by Donoho and Gasko (1992). Another approach, that has been studied extensively, is related to the strategy behind the MVE and MCD estimators previously described. That is, find the ellipsoid with the smallest volume or smallest covariance determinant that encompasses at least half of the data, and use the corresponding mean and covariance matrix to detect outliers. (See, for example, Davies, 1987; Fung, 1993; Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Rousseeuw & Leroy, 1987; Rousseeuw & van Zomeren, 1990; Rousseeuw & van Driessen, 1999; Tyler, 1991; For additional references, see Peña & Prieto, 2001; cf. Woodruff & Rocke, 1994.)

The main article for detecting outliers based on the minimum volume ellipsoid (MVE) estimator is Rousseeuw and van Zomeren (1990). Rocke and Woodruff (1996) describe a method that uses the MVE and MCD estimators as starting values for computing estimators of location and scatter. Poon, Lew and Poon (2000) suggest a method based in part on a Mahalanobis distance, and yet another approach was recently proposed by Viljoen and Venter (2002).

One more strategy, as suggested by Stahel (1981) and Donoho (1982), is motivated by the fact that each outlier among a multivariate sample must be an extreme point based on some projection of the data. Adopting this view, Peña and Prieto (2001) focus on how far points are from the usual sample mean, and they suggest how to choose interesting projections based on the estimated kurtosis coefficient of the projected observations.

This study uses a projection-type method for detecting outliers for reasons to be described and because software is easily written to perform the calculations. To reduce the number of projections considered, the strategy used by Peña and Prieto (2001) is used where attention is focused on how far a point is from the center of the data. The idea is that by projecting points onto a line that passes through the center of the data, the distances between points on the projected line can be combined

with known properties of univariate outlier detection methods in a manner that are advantageous for the problem at hand.

But rather than use the sample mean, as was done by Peña and Prieto, the Donoho and Gasko (1992) multivariate median estimator $\hat{\xi}_m$, is used instead. Another difference is that n projections are considered. In contrast, with p -variate data, Peña and Prieto search for $2p$ projections instead.

To briefly elaborate, it is noted that the so-called outside rate per observation for an outlier detection method refers to the proportion of points declared outliers based on a sample of size n . When searching for an estimator that performs nearly as well as the sample mean under normality, it seems clear that the outside rate per observation should be reasonably low when sampling is from a multivariate normal distribution. Known results on univariate outlier detection methods, suggest how to control the outside rate per observation when considering projections, so this strategy is used in the current study.

The details of the method used here are as follows. Fix i and for the point X_i , project all n points onto the line connecting $\hat{\xi}_m$ and X_i and let D_j be the distance between $\hat{\xi}_m$ and X_j based on this projection. More formally, let

$$A_i = X_i - \hat{\xi}_m,$$

$$B_j = X_j - \hat{\xi}_m,$$

where both A_i and B_j are column vectors having length p , and let

$$C_j = \frac{A_i' B_j}{B_j' B_j} B_j,$$

$j=1, \dots, n$. Then when projecting the points onto the line between X_i and $\hat{\xi}_m$, the distance of the j th point from $\hat{\xi}_m$ is

$$D_j = \|C_j\|,$$

where $\|C_j\|$ is the Euclidean norm of the vector C_j .

Next, a boxplot rule for detecting outliers is applied to the D_j values, but rather than the standard rule, a modification that has close similarities to one used by Carling (2000) is employed. Let $l = [n/4 + 5/12]$, where $[.]$ is the greatest integer function, and let

$$h = \frac{n}{4} + \frac{5}{12} - l.$$

Let $D_{(1)} \leq \dots \leq D_{(n)}$ be the n distances written in ascending order. The so-called ideal fourths associated with the D_j values are

$$q_1 = (1-h)D_{(j)} + hD_{(j+1)}$$

and

$$q_2 = (1-h)X_{(k)} + hX_{(k-1)}.$$

Then the j th point is declared an outlier if

$$D_j > M_D + \sqrt{\chi_{.95,p}^2} (q_2 - q_1), \quad (2)$$

where M_D is the usual sample median based on the D_j values and $\chi_{.95,p}^2$ is the .95 quantile of a chi-squared distribution with p degrees of freedom (cf. Rousseeuw & van Zomeren, 1999).

The process just described is for a single projection; for fixed i , points are projected onto the line connecting X_i to $\hat{\xi}_m$. Repeating this process for each i , $i = 1, \dots, n$, a point is declared an outlier if for any of these projections, it satisfies equation (2). Removing any outliers found by equation (2), and averaging the values that remain, will be called the OP (outlier-projection) estimator and denoted by $\hat{\xi}_{op}$.

A simple and seemingly desirable modification of the method just described is to replace the interquartile range ($q_2 - q_1$) with the median absolute deviation (MAD) measure of

scale based on the D_j values. So here, MAD is the median of the values

$$|D_1 - M_D|, \dots, |D_n - M_D|.$$

Then the j th point is declared an outlier if

$$D_j > M_D + \sqrt{\chi_{.95,p}^2} \left(\frac{\text{MAD}}{.6745} \right), \quad (3)$$

where the constant .6745 is typically used because under normality, $\text{MAD}/.6745$ estimates the standard deviation. (Equation 3 represents an approximation of the method given by equation 1.3 in Donoho & Gasko, 1992.) One appealing feature of MAD is that it has a higher finite sample breakdown point versus the interquartile range, where the finite sample breakdown point of an estimator refers to the minimum proportion of points that must be altered to make the value of a statistic arbitrarily small or large. MAD has a finite sample breakdown point of approximately .5, while for the interquartile range it is only .25.

In this study, however, the focus is on using a projection-type method in conjunction with the interquartile range, rather than MAD. The reason stems from the outside rate per observation, p_n . As previously suggested, to maintain relatively high accuracy under normality when using a skipped estimator, the outside rate per observation should be reasonably close to zero.

It is common to search for a method with a rate approximately equal to .05; this usually provides good efficiency under normality. A negative feature of equation (3) is that p_n appears to be considerably less stable as a function of n . In the bivariate case, for example, it is approximately .09 with $n = 10$ and drops below .02 as n increases. So the relative accuracy of the corresponding skipped estimator varies with n . For the same situations, p_n based on equation 2 ranged between .043 and .038. So the approach based on equation 3 is not pursued here.

To further elaborate on why the MVE outlier detection method was discarded, it is

noted that under normality, and when all variables are independent, its outside rate per observation is approximately .05, but when there is dependence, the rate can be considerably higher (Wilcox, 2003). The result is that if points declared outliers are removed, and the sample mean of the remaining points is computed, efficiency remains relatively high under independence, but it can be relatively low otherwise, so this approach was abandoned. If the MVE measures of location and scatter are replaced by the MCD estimators previously mentioned, again general situations were found where efficiency is poor under normality, which was not surprising because in these situations the outside rate per observation was even higher than was found for the MVE method.

The outside rate per observation for many outlier detection methods has not been studied and addressing this issue goes beyond the scope of this article. So, of course, some variation of the skipped estimator studied here might give improved results in some sense, but this remains to be determined.

Confidence Region

Given that location is estimated using $\hat{\xi}_{op}$, how should a confidence region for ξ , the parameter estimated by $\hat{\xi}_{op}$, be computed? The initial strategy was to use the bootstrap method in Liu and Singh (1997). A direct application of their method, or some slight variation of it, has been found to perform well for a wide range of problems (Wilcox, 2003). Here, however, this approach was found to be unsatisfactory and was eventually abandoned. A much more satisfactory approach, in simulations, is to proceed as follows.

Let ξ_j represent the j th element of the vector ξ , $j = 1, \dots, p$. Let X_{ij} , $i = 1, \dots, n$; $j = 1, \dots, p$, represent a random sample from some p -variate distribution. Generate a bootstrap sample by resampling with replacement n rows from the n by p matrix corresponding to X_{ij} , and denote this bootstrap sample by X_{ij}^* . Let $\hat{\xi}^* = (\hat{\xi}_1^*, \dots, \hat{\xi}_p^*)$ represent the OP estimate of ξ based on this bootstrap sample.

Repeat this process B times and let $\hat{\xi}_{jb}^*$ be the estimate of ξ_j based on the b th bootstrap sample, $b = 1, \dots, B$. Then from basic principles (e.g., Efron & Tibshirani, 1993), an approximate $1 - \alpha$ confidence interval for ξ_j is given by $(\hat{\xi}_{j(l+1)}^*, \hat{\xi}_{j(u)}^*)$, where for fixed j $\hat{\xi}_{j(1)}^* \leq \dots \leq \hat{\xi}_{j(B)}^*$ are the $\hat{\xi}_{jb}^*$ values written in ascending order, $l = \alpha B / 2$, rounded to the nearest integer, and $u = B - l$. So, to obtain an approximate $1 - \alpha$ confidence region for ξ , a simple strategy is to apply the Bonferroni inequality and compute an α/p confidence interval for ξ_j using the method just described.

Said another way, to test $H_0: \xi_j = \xi_{0j}$, ξ_{0j} given, let p_j^* be the probability that $\hat{\xi}_j^*$ is less than ξ_{0j} . From Liu and Singh (1997), for fixed j , p_j^* has, asymptotically, a uniform distribution. Although p_j^* is not known, it is readily estimated from the data with

$$\hat{p}^* = \frac{A}{B}, \quad (4)$$

where A is the number of bootstrap samples with $\hat{\xi}_{jb}^* < \xi_{0j}$. Then $2\hat{p}_m^*$ is the estimated p-value, where

$$\hat{p}_m^* = \min(\hat{p}^*, 1 - \hat{p}^*). \quad (5)$$

So for fixed j , reject at the α level if $2\hat{p}_m^* \leq \alpha$. To control the familywise error rate (the probability of at least one Type I error) via the Bonferroni inequality when testing all p hypotheses, reject if $2\hat{p}_m^* \leq \alpha/p$.

Methodology

Simulations were used to check both the accuracy of the estimators considered, plus the actual probability coverage when using the method just discussed. Accuracy was measured using the sum of the squared standard errors

associated with the p estimators used to estimate the p parameters. For the OP estimator, again let $\hat{\xi}_j$ be the OP estimate of ξ_j , and let $\bar{X}_j$ be the usual sample mean corresponding to the j th marginal distribution. Then the accuracy of the OP estimator, relative to the sample mean, is measured by R^2 , the sum of the squared standard errors of the estimators associated with $\hat{\xi}_j, j=1, \dots, p$, divided by the sum of the squared standard errors associated with the sample means, $\bar{X}_j$. When dealing with other estimators, the $\hat{\xi}_j$ were replaced with the relevant estimator.

Observations were generated where the marginal distributions have a *g-and-h* distribution (Hoaglin, 1985) which includes normal distributions as a special case. When dealing with accuracy, the focus was on $p=4$ (but when computing a confidence region, both $p=2$ and 4 were considered). More precisely, observations $Z_{ij}=1, \dots, n; j=1, \dots, p$ were initially generated from a multivariate normal distribution having correlation ρ , then the marginal distributions were transformed to

$$X_{ij} = \frac{\exp(g Z_{ij}) - 1}{g} \exp\left(\frac{h Z_{ij}^2}{2}\right)$$

when both g and h were non-zero. When g was zero

$$X_{ij} = Z_{ij} \exp\left(\frac{h Z_{ij}^2}{2}\right),$$

where g and h are parameters that determine the third and fourth moments. Here, $\rho=0$ and .7 are considered. The four (marginal) *g-and-h* distributions used here were the standard normal ($g=h=0$), a symmetric heavy-tailed distribution ($g=0, h=.5$), an asymmetric distribution with relatively light tails ($g=.5, h=0$), and an asymmetric distribution with heavy tails ($g=h=.5$). Also, when dealing with accuracy, simulations were run with $h=1$. This latter case might be viewed as

an extreme departure from normality, but it was considered anyway to see whether any of the estimators performs poorly when sampling from a sufficiently heavy-tailed distribution.

Table 1 shows the theoretical skewness (κ_1) and kurtosis (κ_2) values for each distribution considered. When $g > 0$ and $h > 1/k$, $E(X^k)$ is not defined and the corresponding entry in Table 1 is left blank. Additional properties of the *g-and-h* distribution are summarized by Hoaglin (1985).

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

g	h	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0.0	0.0	0.00	3.0	0.00	3.0
0.0	0.5	0.00	—	0.00	11,896.2
0.5	0.0	1.75	8.9	1.81	9.7
0.5	0.5	—	—	120.10	18,393.6

A possible objection to Table 1 when performing simulations is that the distribution of observations generated on a computer does not always have the theoretical skewness and kurtosis values shown. The reason is that computer observations come from a bounded interval, so the skewness and kurtosis of the distribution will be finite, even when in theory it should be infinite. Accordingly, Table 1 also reports the estimated skewness ($\hat{\kappa}_1$) and kurtosis ($\hat{\kappa}_2$) values based on simulations with 10,000 replications.

Table 2 shows estimates of R based on 5,000 replications, where the first three estimators are the Donoho-Gasko trimmed mean with 10%, 15% and 20% trimming, DGM is the Donoho-Gasko median, OP is the outlier-projection estimator, and M is the usual median. Note that with 20% trimming, accuracy is relatively poor when sampling from a normal distribution ($g=h=0$).

Table 2. Values of R (Accuracy), $n = 40$

g	h	ρ	$\gamma = .10$	$\gamma = .15$	$\gamma = .20$	DGM	OP	M
0.0	0.0	0.0	0.73	0.62	0.50	0.45	0.92	0.81
0.0	0.5	0.0	5.99	5.92	5.40	4.11	6.25	8.48
0.0	1.0	0.0	4660.21	5764.79	5911.29	4643.16	5452.35	10820.14
0.0	0.0	0.7	0.80	0.71	0.61	0.48	0.95	0.44
0.0	0.5	0.7	4.74	4.76	4.50	3.20	4.64	5.44
0.0	1.0	0.7	1082.56	1300.44	1336.63	1005.24	1091.68	1760.98
0.5	0.0	0.0	0.79	0.69	0.54	0.49	0.99	0.99
0.5	0.5	0.0	13.01	12.78	11.82	8.91	14.95	20.66
0.5	1.0	0.0	1908.75	2413.39	2472.07	1852.97	2519.04	4887.50
0.5	0.0	0.7	0.94	0.86	0.69	0.53	1.05	0.99
0.5	0.5	0.7	17.79	18.05	17.22	11.34	17.42	20.66
0.5	1.0	0.7	3005.56	3652.36	3660.06	29996.40	4887.42	4887.40

This is in sharp contrast with the univariate case where a 20% trimmed mean performs reasonably well (e.g., Rosenberger & Gasko, 1983; Wilcox, 1997). Under normality, with $\rho = .7$, the median performs rather poorly, but with sufficiently heavy-tailed distributions, the median performs well. So, if one is willing to sacrifice affine equivariance, applied situations might arise where the usual median has practical advantages. In general, however, the OP estimator seems best for general use. It was found to be the most accurate alternative to the mean under normality, and it remains competitive under fairly extreme kurtosis.

As for probability coverage, when using method OP, Table 3 contains $\hat{\alpha}$, the estimated probability that the confidence region based on the Bonferroni method does not contain the population value when $n = 20$ for $p = 2$ and 4. For this portion of the study, 1,000 replications were used with $B = 1,000$. For asymmetric distributions, the actual value of the parameter was determined by taking the mean of 5,000 estimates based on a sample size of $n = 100$. Bradley (1978) argues that when testing at the .05 level, at a minimum the actual probability of a Type I error should be between .025 and .075. This criterion is satisfied in all cases except when $p = 2$, $(g, h) = (.5, 0)$ and $\rho = 0$, in which case $\hat{\alpha} = .079$. Increasing n to 30, the estimate equals .069.

Table 3: Estimated Type I Error Probabilities Using the OP Estimator, $\alpha = .05, n = 20$.

g	h	ρ	p	$\hat{\alpha}$
0.0	0.0	0.0	2	.071
0.0	0.0	0.7	2	.071
0.0	0.5	0.0	2	.040
0.0	0.5	0.7	2	.040
0.5	0.0	0.0	2	.079
0.5	0.0	0.7	2	.040
0.5	0.5	0.0	2	.056
0.5	0.5	0.7	2	.047
0.0	0.0	0.0	4	.065
0.0	0.0	0.7	4	.069
0.0	0.5	0.0	4	.040
0.0	0.5	0.7	4	.036
0.5	0.0	0.0	4	.063
0.5	0.0	0.7	4	.061
0.5	0.5	0.0	4	.044
0.5	0.5	0.7	4	.040

Conclusion

A criticism of the OP estimator is that it is based on an outlier detection method that has a finite sample breakdown point of at most .25, because when using the interquartile range, the finite sample breakdown point is .25 for any projection. However, this would seem to suffice for many situations, and its efficiency is quite good compared to the mean even when sampling

from a very heavy-tailed distribution. If there are indications that more than 25% of the points are outliers, then one possibility is to use the variation of the OP estimator based on equation (3). The main point is that good efficiency is achieved under normality and a method for computing a confidence region was found that performs reasonably well in simulations.

References

- Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: survey and advances*. Princeton University Press, Princeton: NJ.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144–152.
- Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis*, 33, 249–258.
- Davies, P. L. (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Annals of Statistics*, 15, 1269–1292.
- Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76, 354–362.
- Donoho, D. L., & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, 20, 1803–1827.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- Fung, W. K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, 88, 515–519.
- Golberg, K. M., & Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, 34, 307–320.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes*. (p. 461–515). New York: Wiley.
- Liu, R. Y., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266–277.
- Liu, R. Y., Parelius, J. M., & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, 27, 783–858.
- Masse, J. C., & Plante, J. F. (2003). A Monte Carlo study of the accuracy and robustness of ten bivariate location estimators. *Computational Statistics & Data Analysis*, 42, 1–26.
- Peña, D., & Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43, 286–299.
- Poon, W. Y., Lew, S. F., & Poon, Y. S. (2000). A local influence approach to identifying multiple outliers. *British Journal of Mathematical and Statistical Psychology*, 53, 255–273.
- Raine, A., Buchsbaum, M., & LaCasse, L. (1997). Brain abnormalities in murderers indicated by positron emission tomography. *Biological Psychiatry*, 42, 495–508.
- Rocke, D. M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics*, 24, 1327–1345.
- Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91, 1047–1061.
- Romanazzi, M. (1997). A schematic plot for bivariate data. *Student*, 2, 149–158.
- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Understanding robust and exploratory data analysis*. p. 297–336. New York: Wiley.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York: Wiley.

Rousseeuw, P. J., & Ruts, I. (1996). AS 307: Bivariate location depth. *Applied Statistics*, 45, 516–526.

Rousseeuw, P. J., & Ruts, I., & Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *American Statistician*, 53, 382–387.

Rousseeuw, P. J., & Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8, 193–203.

Rousseeuw, P. J., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association*, 85, 633–639.

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

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Tukey, J. W. (1974). T6: Order statistics. *Mimeographed notes for Statistics 411*, Princeton University.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians*, 2, 523–531.

Tyler, D. E. (1991). Some issues in the robust estimation of multivariate location and scatter. In W. Stahel & S. Weisberg (Eds.), *Directions in robust statistics and diagnostics*, Part II, p. 327–336. New York: Springer-Verlag.

Viljoen, H., & Venter, J. H. (2002). Identifying multivariate discordant observations: a computer-intensive approach. *Computational Statistics & Data Analysis*, 40, 159–172.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (2003). *Applying contemporary statistical methods*. San Diego, CA: Academic Press.

Wilcox, R. R. (in press). Two-sample, bivariate hypothesis testing based on Tukey's depth. *Multivariate Behavioral Research*.

Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics & Data Analysis*, 28, 257–270.

Appendix

Following Liu and Singh (1993), Tukey's depth is defined as follows. Let F be a p -variate distribution. Tukey's depth at the point x is

$$TD(x; F) = \inf_H \{P(H) : H \text{ is a closed space containing } x\}.$$

The sample version is obtained by replacing F with the usual empirical distribution. More precisely, the sample version of $TD(x; F)$ is the smallest proportion of X_i contained in any closed halfplane with boundary line through x . For $p = 1$, $TD(x; F) = \min\{F(x), 1 - F(x^-)\}$.

A Comparison Of Methods For Longitudinal Analysis With Missing Data

James Algina
University of
Florida

H. J. Keselman
University of Manitoba

In a longitudinal two-group randomized trials design, also referred to as randomized parallel-groups design or split-plot repeated measures design, the important hypothesis of interest is whether there are differential rates of change over time, that is, whether there is a group by time interaction. Several analytic methods have been presented in the literature for testing this important hypothesis when data are incomplete. We studied these methods for the case in which the missing data pattern is non-monotone. In agreement with earlier work on monotone missing data patterns, our results on bias, sampling variability, Type I error and power support the use of a procedure due to Overall, Ahn, Shivakumar, and Kalburgi (1999) that can easily be implemented with SAS's PROC MIXED.

Keywords: data, mixed models, split-plot design

Introduction

A randomized parallel-groups design in which participants are randomly assigned to treatments, measured on one pretreatment occasion, and on multiple post treatment occasions, is a common design for investigating treatment effects. One challenge facing researchers who use this design is how to analyze the data when there are missing observations. Little (1995), Overall, Ahn, Shivakumar, and Kalburgi (1999), Wang-Clow, Lange, Laird, and Ware (1995), and Wu

and Bailey (1989) have all suggested procedures for conducting such analyses. Algina and Keselman (2003) compared a number of these methods for designs in which two treatments ($k = 1, 2$) are compared. They concluded that while, in principle, if one has valid information about the type of missing data, the information should be taken into account in selecting a procedure, in practice it may be wise to select a method that performs well over a wide range of methods. Based on their findings, which included empirical estimates of bias, sampling variability, variations of a procedure suggested by Wu and Bailey (1989) might be considered. The principal shortcomings of these three procedures were Type I error rates above the nominal level in some conditions and, for two of the variations, a complicated method of estimation. However, Algina and Keselman acknowledged that their study should be regarded as preliminary in that they studied a limited number of conditions.

James Algina (algina@ufl.edu) is Professor of Educational Psychology at the University of Florida. His research interests are in applied statistics and psychometrics. H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology, University of Manitoba. His research interests are in applied statistics. Work on this project was supported by a grant from the Social Sciences and Humanities Research Council of Canada.

One limiting factor in the Algina-Keselman (2003) study, as well as in Overall et al. (1999), Wang-Clow et al. (1995), and Wu and Bailey (1989), was a monotone pattern for the missing data. That is, once an observation was missing for a participant, no further measurements were available for that participant. Thus, a major purpose of the current investigation was to determine whether the Overall et al. procedure would continue to perform well when the missing data did not occur in a monotone pattern. In addition, the influence of a wider variety of missing data mechanisms than were included by Algina and Keselman and the influence of planned sample size on the methods were investigated. Prior to presentation of the new results, we review missing data mechanisms and the methods we investigated.

Missing Data Mechanisms

Little (1995) reviewed several mechanisms for missing data: missing completely at random (MCAR), covariate dependent (CD), and missing at random (MAR). Following Verbeke and Molenberghs (2000), when the mechanism is not MCAR, CD, or MAR, we refer to it as missing not at random (MNAR). The variables that predict which data are missing determine whether or not the data are MCAR, CD, MAR, or MNAR. In this paper we are concerned with estimation and hypothesis testing when data are missing in a design in which participants in two treatment groups are measured on one pretreatment occasion and several post treatment occasions. In such studies, there are three types of variables that describe the participants.

The first two are the potentially observable variables. These are the measurements on the variable of interest and the covariates. The latter variables include the occasion of measurement, the treatment indicator, and any other variables that are observed prior to the onset of the treatments. The third type comprises the parameters for a subject-specific within-subject model for scores on the repeated measurements. Variables in the third type are latent variables.

When the pattern of missing data at a particular time point is unrelated to the

potentially observable variables and to the latent variables, the data are MCAR. If the pattern of missing data is related only to the covariates the mechanism is CD. It should be noted that some authors (see, for example Diggle & Kenward, 1994) do not distinguish between MCAR and CD missing data mechanisms. If the pattern at a particular time point is related to previous measurements on the variable of interest and the covariates in the model, but not to the actual data values that would have been observed at that time point had there been no missing data, nor to the latent variables, the data are MAR.

Methods of Analysis

In the presentation of the methods we use the following notation: Y_{ijk} , the score for the i th ($i=1, \dots, n_k$) of n_k subjects in the k th ($k=1, 2$) group on the j th ($j=1, \dots, J$) occasion; t_j , an index for the occasion of measurement, and t_{ik} , the index value for the last measurement occasion at which the i th participant in the k th group was observed.

All of the methods, except the endpoint method studied by Overall et al. (1999), assume that if the data were complete they would conform to the following model

$$Y_{ijk} = \beta_{0ik} + \beta_{1ik}t_j + \varepsilon_{ijk} \quad (1)$$

where

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

and, depending on the method for analyzing the data

$$\begin{bmatrix} \beta_{0ik} \\ \beta_{1ik} \end{bmatrix} \sim N(\theta_k, \mathbf{D})$$

or

$$\begin{bmatrix} \beta_{0ik} \\ \beta_{1ik} \end{bmatrix} \sim N(\theta_k, \mathbf{D}_k).$$

The parameters β_{0ik} and β_{1ik} are the subject-specific intercept and slope, respectively, for the

within-subject regression of the dependent variable on time of measurement.

When participants are randomly assigned to groups and it is reasonable to assume that, for each participant, the within-subjects regression is well-described by the simple linear regression model, the test of the treatment effect focuses on the average slope (i.e., the population average) in each treatment. Specifically, to test for a treatment effect one tests whether the average slopes are equal for the treatment groups.

Mixed Model for MAR Data

One method of analysis uses equation (1) as the level-1 model in a multilevel model and the following level-2 models:

$$\beta_{0ik} = \gamma_{00} + \gamma_{01}Z_{ik} + u_{0ik} \quad (2)$$

and

$$\beta_{1ik} = \gamma_{10} + \gamma_{11}Z_{ik} + u_{1ik}, \quad (3)$$

where $Z_{ik} = 1$ if the i th participant is in treatment 2 and 0 otherwise. The estimate of the treatment effect is $\hat{\gamma}_{11}$ and testing $H_0 : \gamma_{11} = 0$ provides a test of the treatment effect. This procedure is known to give correct results provided the data are MCAR, CD, or MAR and, in the case of the latter two mechanisms, provided that the parameters of the missing data mechanism and the parameters of the data model are distinct (Little, 1995). This procedure can be implemented by using the following SAS (SAS, 2000) PROC MIXED code:

```
proc mixed method=ml;
class id group;
model score=time group group*time;
random intercept time/type=un subject=id;
```

The following are definitions of the variables used in this code:

- time—a quantitative index of the time of measurement
- id—a categorical variable identifying the participant
- group—a categorical variable identifying the treatment group

Pattern-Mixture Models (Unweighted Least Squares)

A number of different strategies have been presented over the years to deal with data that are MNAR [see the references provided by Little (1995) and Hedeker & Gibbons (1997)].

Recently, Little provided a general class of models referred to as pattern-mixture models. As Little (1995, p. 1113) noted, “Pattern-mixture models stratify the population by the pattern of dropout, implying a model for the whole population that is a mixture over the patterns.” An advantage of this procedure is that the missing data mechanism is taken into account in the estimation, but a model for the pattern of missing data does not have to be explicitly introduced into the likelihood function.

A pattern-mixture model due to Little (1995), for the design considered in this paper, yields valid estimates of the treatment effect even when the pattern of missing data is related to the covariates and the subject specific slopes and intercepts (a type of MNAR missing data mechanism). The reader should note that Little (1995, p. 1120) indicated that the unweighted least squares (UWLS) estimate of the slope difference is the maximum likelihood (ML) estimator for the pattern-mixture model he used [see Equation (17) in Little] for analysis of longitudinal missing data under normal distribution theory. We implemented the UWLS procedure as follows:

1. Use ordinary least squares (OLS) to estimate the slope for each participant in each treatment group.

2. For each treatment calculate the unweighted average of the subject-specific OLS slopes,

$$\hat{\theta}_{1k} = \frac{\sum_{i=1}^{n_k} \hat{\beta}_{1ik}}{n_k},$$

and calculate the treatment effect as the difference between these two averages.

3. Calculate the sampling variance of the estimated treatment effect by using the (2,2) element of

$$S_{\hat{\theta}_{12}-\hat{\theta}_{11}}^2 = \sum_{k=1}^2 \frac{\sum_{i=1}^{n_k} \hat{\sigma}^2 (\mathbf{X}'_{ik} \mathbf{X}_{ik})^{-1} + \hat{\mathbf{D}}}{n_k^2}$$

where the first column of $\mathbf{X}_{ik}$ is a vector of ones and the second column contains codes for the occasions on which participant i in group k had observed data. Wang-Clow et al. (1995) also used this method, however, they used the method of moments to calculate $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$. We used ML estimation to calculate these quantities. Specifically, we used the PROC MIXED code used to implement the mixed model for MAR data. While these estimates assume that the missing data mechanism is not MNAR, comparison of the variance of $\hat{\theta}_{12} - \hat{\theta}_{11}$, over replications of a condition, to the average value of $S_{\hat{\theta}_{12}-\hat{\theta}_{11}}^2$ suggested that the method provides a consistent estimate of the sampling variance of $\hat{\theta}_{12} - \hat{\theta}_{11}$ for the conditions we studied.

Linear Minimum Variance Unbiased Estimator

Wu and Bailey (1989) presented a method which they called the linear minimum variance unbiased estimator. Later Wang-Clow et al. (1995) referred to the method as the ANCOVA method and we use the latter term in this paper. Wu and Bailey (1989) proposed using the following model within each group

$$\hat{\beta}_{1ik} = \gamma_{10k} + \gamma_{11}t_{ik} + \delta_{ik} \quad (4)$$

where $\hat{\beta}_{1ik}$ is the OLS estimate of the subject-specific slope for participant i in group k . Wu and Bailey propose testing for a treatment effect by calculating an estimate of the expected value of β_{1ik}

$$\hat{E}(\beta_{1ik}) = \hat{\gamma}_{10k} + \hat{\gamma}_{11}\bar{t}_k, \quad (5)$$

where $\bar{t}_k$ is the average in group k of t_{ik} , and comparing the estimates across treatment groups. Noting that the variance of $\hat{\beta}_{1ik}$ varies across treatment groups and the occasions on which the dependent variable was observed for participant i , Wu and Bailey proposed estimating

the γ s by weighted least squares (WLS) with the weight equal to the inverse of the sampling variance of $\hat{\beta}_{1ik}$. The sampling variance is the (2,2) element of $\hat{\sigma}^2 (\mathbf{X}'_{ik} \mathbf{X}_{ik})^{-1} + \hat{\mathbf{D}}_k$. We implemented this WLS procedure. However, whereas Wu and Bailey and Wang-Clow et al. used method of moment estimators of $\hat{\sigma}^2$ and $\hat{\mathbf{D}}_k$, we used ML estimators obtained by using the following code:

```
proc mixed method=ml;
class id group;
model score=time group group*time/solution ;
random intercept time/type=un subject=id
group=group;
```

In the random statement the code group=group specifies that the covariance matrix for the intercept and slope varies across treatment groups.

The procedure described by Wu and Bailey (1989) is fairly complicated because of the necessity of estimating the weights used in the WLS procedure. However, Algina and Keselman (2003) reformulated the Wu and Bailey model as a multilevel model and estimated it by using PROC MIXED, thus eliminating the complication of estimating the weights. Their level-1 model is given by equation (1). The level 2 models are

$$\beta_{0ik} = \gamma_{00} + \gamma_{01}Z_{ik} + \gamma_{02}(t_{ik} - \bar{t}_k) + u_{0ik} \quad (6)$$

and

$$\beta_{1ik} = \gamma_{10} + \gamma_{11}Z_{ik} + \gamma_{12}(t_{ik} - \bar{t}_k) + u_{1ik}. \quad (7)$$

The estimate of the treatment effect is $\hat{\gamma}_{11}$ and testing $H_0: \gamma_{11} = 0$ provides a test of the treatment effect. The approach presented by Wu and Bailey does not include an equation for the intercept. Nevertheless, Algina and Keselman included it because Bryk and Raudenbush (1992) have noted that omitting variables in one level-2 model can impact estimates in a second level-2 model because of the correlated error terms for the level-2 models. The model

represented by equations (1), (6), and (7) can be estimated by using the following PROC MIXED code:

```
proc mixed method=ml;
class id group;
model score=lobsc group time time*lobsc
time*group /solution;
random intercept time/type=un subject=id
group=group;
```

In the preceding code, the variable lobsc is $(t_{ik} - \bar{t}_k)$. The inclusion of lobsc and time*lobsc is intended to improve estimation and testing when the missing data mechanism is MNAR and the missing data pattern is monotone. If the data are MAR, it is known that valid estimates can be obtained with these terms excluded.

Analyses Investigated by Overall et al. (1999)

The simplest method studied by Overall et al. (1999) is an endpoint analysis. This analysis is a two-stage procedure. At stage one a simple change score from baseline to the last available measurement is calculated; at stage two the change scores are the dependent variable in an ANCOVA, using pretest score (Y_1) and time of the last observation as covariates and treatment group as the between-subjects factor.

Overall and his colleagues also used ANCOVA with PROC MIXED to examine the group by time effect (see Overall et al., 1999, pp. 205-209), using Y_1 and t_{ik} as covariates, though their approach differs from the Wu and Bailey (1989) approach. They use the following PROC MIXED code:

```
proc mixed;
class id group;
model score=lobs y1 group time time*group
/solution;
random intercept time/type=un subject=id;
```

There are three major differences between the Overall et al. code and the PROC MIXED code used by Algina and Keselman (2003) to implement the Wu-Bailey procedure. First the time of last observation (lobs) is not centered. Second Y_1 , the pretest score, is included in their

model but not in the Algina-Keselman code. Third, the time by lobs interaction is excluded in their model. The result of this exclusion is that the time code for the last observation on which the participant was observed is excluded from the level-2 model for the slope. Thus, the Overall et al. PROC MIXED ANCOVA is based on the a multilevel model in which the level-1 model is given by equation (1) and the level-2 models are

$$\beta_{0ik} = \gamma_{00} + \gamma_{01}Z_{ik} + \gamma_{02}t_{ik} + \gamma_{03}Y_{1ik} + u_{0ik} \quad (8)$$

and

$$\beta_{1ik} = \gamma_{10} + \gamma_{11}Z_{ik} + \gamma_{12}t_{ik} + u_{1ik} \quad (9)$$

The estimate of the treatment effect is $\hat{\gamma}_{11}$ and testing $H_0: \gamma_{11} = 0$ provides a test of the treatment effect.

Overall et al. (1999) also investigated a two-stage ANCOVA procedure. Like the Wu and Bailey (1989) approach, Overall et al. use OLS in stage 1 to estimate the subject-specific regression coefficients and then these estimates, weighted by lobs, are used in a second stage ANCOVA with Y_1 and t_{ik} used as covariates.

Thus, the previously described analyses can be used to analyze the important group by time interaction effect in longitudinal designs in which data are missing. In this report we assess rates of Type I error and power in testing whether the average slopes are equal for the treatment groups, as well as the bias and variability (i.e., SD) in estimating the average slope difference.

Methodology

Algina and Keselman (2003) investigated three missing data mechanisms (CD, MAR and MNAR), but only considered monotone patterns. In the present investigation, whether or not data are missing for a participant is determined independently for each occasion. Therefore, the pattern of missing data is not monotone. In addition, eight different missing data mechanisms were used:

1. MNAR-Direct Selection (DS) on Y_j . The data point for participant i was missing at occasion j if $Y_{ij} > \delta$. The value of δ was selected so that the probability of missing data at time 3 was 5% for participants in treatment 2. This selection of δ determined the probability of missing data for both groups at time points 3 to 9. Figure 1 shows the probability of missing data at each occasion in treatments 1 and 2 in conditions in which there was a treatment effect. The δ s for the other mechanisms were selected to yield the same probabilities of missing data. In conditions in which there was no treatment effect, the probability of missing data in treatment 1, at a particular occasion, was equal to the probability of missing data that is reported in Figure 1, at that occasion, for treatment 2.

Figure1. Probability of Missing Data by Occasion.

2. CD. The data point for participant i was missing at occasion j if U_{ij} (a uniform random variable) was less than the probability determined for the MNAR condition with direct selection on Y .

3. MAR-DS. The criterion used to determine whether the data point for participant i was missing at occasion j depended on whether the data point for participant i was missing at occasion $j-1$: When the data point for participant i at occasion $j-1$ was not missing, the data point for participant i was missing at

occasion j if $Y_{i(j-1)} > \delta_{j-1}$. When the data point for participant i at occasion $j-1$ was missing, the data point for participant i was missing at occasion j if U_{ij} was less than the probability determined for the MNAR condition with direct selection on Y .

If the first criterion had been used uniformly, the data would have been MNAR because, for a participant with missing data at occasion $j-1$, whether the data were missing at occasion j would depend on the value of a missing score at occasion $j-1$ rather than value of an observed score at occasion $j-1$.

4. MAR-Probabilistic Selection (PS). Again the criterion used to determine whether the data point for participant i was missing at occasion j depended on whether the data point for participant i was missing at occasion $j-1$: When the data point for participant i at occasion $j-1$ was not missing, the data point for participant i was missing at occasion j if $U_{ij} < \phi(\delta_{j-1} + Y_{i(j-1)})$, where $\phi(\bullet)$ is the cumulative normal function. When the data point for participant i at occasion $j-1$ was missing, the data point for participant i was missing at occasion j if U_{ij} was less than the probability determined for the MNAR condition with direct selection on Y .

5. MNAR-DS on Y_{j-1} . The data point for participant i was missing at occasion j if $Y_{i(j-1)} > \delta_{j-1}$. This method employs the first criterion used in the MAR-DS mechanism.

6. MNAR-PS on Y_{j-1} . The data point for participant i was missing at occasion j if $U_{ij} < \phi(\delta_{j-1} + Y_{i(j-1)})$. This method employs the first criterion used in the MAR-PS mechanism.

7. MNAR-PS on Y_j . The data point for participant i was missing at occasion j if $U_{ij} < \phi(\delta_j + Y_{ij})$.

8. MNAR-PS on Slope and Intercept (SI). The data point for participant i was missing at occasion j if $U_{ij} < \phi(\delta_j + .46\beta_{0ik} + .14\beta_{1ik})$.

The U_{ij} for the four probabilistic mechanisms were selected independently for each participant and time point.

It is impossible to know whether or not these eight missing data mechanisms are representative of those found in practice. However, these eight mechanisms represent a wider variety of mechanisms than have been included in previous research.

Seven methods of examining the group by time interaction effect in a randomized parallel groups design were examined; these methods were also examined by Algina and Keselman (2003). Specifically, the methods (with their acronyms) were:

1. Overall et al.'s (1999) two-stage endpoint ANCOVA (OEPAOC),

2. an unweighted least squares (pattern-mixture) analysis (UWLS),

3. the ANCOVA presented by Wang-Clow et al. (1995) (See Section 3.6 in their paper), where the weights for the WLS part of the analysis were obtained from PROC MIXED (WLSAOC),

4. Wu and Bailey's (1989) two-stage ANCOVA implemented in PROC MIXED (WBPMAOC),

5. Overall et al.'s (1999) PROC MIXED analysis that uses Y_1 and t_{ik} as covariates (OPMAOC),

6. Overall et al.'s (1999) two-stage ANCOVA (OTSAOC), and

7. The mixed model analysis, implemented in PROC MIXED, that presumes the data are missing at random (PMMAR).

Theory presented in Little (1995) shows that the UWLS estimator of the treatment effect is consistent when the data are CD or MNAR with missingness (i.e., whether a particular data point is missing) predicted by the slope and intercept. PMMAR is known to yield a consistent estimator when the data are CD or MAR. OEPAOC, WLSAOC, and WBPMAOC were designed to improve performance of the treatment effect estimator when the data are not MCAR or CD, but proofs of consistency have not been presented. Similarly, OPMAOC and OTSAOC were designed to improve control of the Type I error rate and power when the data are not MCAR or CD.

In addition to the eight types of missing data mechanism and the seven tests of the treatment effect, number of planned observations per group ($n_k = 100$ and $n_k = 200$) was also investigated. Overall and his colleagues (see Ahn, Tonidandel & Overall, 2000; Overall et al., 1999; Overall et al., 1996), as well as Algina and Keselman (2003), examined the group by time interaction effect in a parallel-groups design containing a baseline score and eight repeated measurements; thus, for comparative purposes we had nine levels for our number of repeated measurements.

To compare the procedures, we simulated data for a situation in which participants are randomly assigned to treatments. We used the following equation to generate data for the i th participant in group k on the j th occasion:

$$Y_{ijk} = \beta_{0ik} + \beta_{1ik}t_j + \varepsilon_{ijk}. \quad (10)$$

The equation states that the data for the i th person on nine occasions has a linear relationship to the time of measurement. The i subscripts on the intercept (β_{0ik}) and slope (β_{1ik}) indicate that the intercept and slope vary across participants. We assumed

$$\begin{bmatrix} \beta_{0ik} \\ \beta_{1ik} \end{bmatrix} \sim N \left(\begin{bmatrix} \theta_{0k} \\ \theta_{1k} \end{bmatrix}, \mathbf{D} \right).$$

The mean for the intercept was 50 in both groups ($\theta_{01} = \theta_{02}$), implying that both treatment groups had the same population pretest mean. For Type I error data, the mean for the slope was 9.0 in treatments 1 and 2. That is, $\theta_{12} - \theta_{11} = 0$, indicating identical average rates of increase over time, hence a null condition. For our power comparisons, the mean for the slope was 4.5 in treatment 1 and 9.0 in treatment 2. Thus, $\theta_{12} - \theta_{11} = 4.5$. The errors ε_{ijk} were assumed to be uncorrelated for different times of observation. This does not imply that the scores were uncorrelated over time. Allowing the slope and intercept to vary across participants implies that scores were correlated over time. In all

cases the covariance matrix (**D**) for the intercept and slope was

$$\mathbf{D} = \begin{bmatrix} 15.21 & 12.42 \\ 12.42 & 82.81 \end{bmatrix}.$$

The correlation between the slope and intercept was .35, indicating that participants with higher pretest status increased more rapidly. The variance for the residuals, conditional on time was 240. Algina and Keselman (2003) also studied

$$\mathbf{D} = \begin{bmatrix} 15.21 & -12.42 \\ -12.42 & 82.81 \end{bmatrix},$$

but performance of WLSAOC and WBPMAOC was worse when $D_{12} > 0$, and so we have only included $D_{12} > 0$. The variable t_j is an index for observation time and was coded (0, 0.23077, 0.46154, 0.69231, 0.92308, 1.15385, 1.38462, 1.61538, 1.84615). The design of the simulation was based on Wang-Clow et al.'s (1995) study. In their study they had 14 time points, coded from 0 to 3. Our results would also have been obtained if we had coded t_j from 0 to 8 and had multiplied the β_{ik} by 1.84615/8.

Algina and Keselman (2003) also studied experiments with five time points. The performance of WLSAOC and WBPMAOC was worse with nine points and so we have elected to study only nine time points. Without further complications to the methods, the methods can only be applied to participants who have at least two observations. Therefore in our simulated data, every participant had an observation at the pretest and the first follow-up occasion. Each condition was replicated 1000 times. All hypothesis tests were conducted with a nominal alpha of .05.

Results

The slope difference $(\theta_{12} - \theta_{11})$ can be estimated by all procedures except OTSAOC and OEPAOC. For each condition in the study the slope difference was estimated by using each of the remaining five methods. Table 1 contains means and standard deviations of these estimates for the CD and MAR mechanisms. Comparison

of the means to 0 when $\theta_{12} - \theta_{11} = 0$ and to 4.5 when $\theta_{12} - \theta_{11} = 4.5$ provides an indication of bias in the estimates. The standard deviations provide a measure of sampling variability of the estimates. The results indicate that all methods yielded unbiased estimators of the treatment effect when the missing data mechanism was CD and when the missing data mechanism was MAR and $\theta_{12} - \theta_{11} = 0$. However, when the missing data mechanism was MAR and $\theta_{12} - \theta_{11} = 4.5$ only PMMAR and OPMAOC yielded unbiased estimators. For a fixed sample size and a fixed value for the treatment effect there were no notable differences among the methods in the standard deviations of the estimates.

Table 2 contains estimated Type I error rates and power for the CD and MAR mechanisms. For CD data, all procedures had estimated Type I error rates near the nominal value and power differences were small but in favor of OEPAOC (Overall et al.'s, 1999 two-stage end-point procedure). For MAR data, WBPMAOC and WLSAOC had estimated Type I error rates above the nominal level. These two procedures are variations on the method suggested by Wu and Bailey (1989). For MAR data, OEPAOC and OTSAOC tended to have lower power than the other procedures. UWLS, WBPMAOC, and WLSAOC tended to have the best power, but this reflects the positively biased estimator produced by these three procedures. Comparing the two procedures that produced unbiased estimators of the treatment effect, PMMAR tended to have slightly better power than OPMAOC.

Tables 3 and 4 contain means and standard deviations of the estimated treatment effect for conditions in which the missing data mechanism was MNAR. Table 3 contains results for $\theta_{12} - \theta_{11} = 0$ and Table 4 contains results for $\theta_{12} - \theta_{11} = 4.5$. In both tables bold values indicate mean treatment effects that were significantly different from the population treatment effect. In Table 3, there was only one estimated treatment effect that was significantly different from 0 [$t(999) = 1.962$ for WBPMAOC and $n_k = 100$].

Table 1. Means and Empirical Standard Errors of Test Statistics for CD and MAR Conditions

		$\theta_{12} - \theta_{11} = 0$						$\theta_{12} - \theta_{11} = 4.5$					
		CD		MAR-DS		MAR-PS		CD		MAR-DS		MAR-PS	
n_k	Test	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
100	PMMAR	-0.014	1.863	-0.037	1.924	0.056	1.887	4.418	1.846	4.539	1.854	4.569	1.838
	UWLS	-0.013	1.881	-0.034	1.990	0.062	1.940	4.417	1.862	4.858	1.914	4.878	1.894
	OPMAOC	-0.016	1.863	-0.035	1.924	0.056	1.885	4.417	1.846	4.546	1.855	4.576	1.839
	WBPMAOC	-0.013	1.863	-0.046	1.995	0.052	1.948	4.420	1.851	4.911	1.915	4.952	1.881
	WLSAOC	-0.013	1.864	-0.044	1.997	0.055	1.954	4.417	1.852	4.932	1.921	4.973	1.890
200	PMMAR	-0.040	1.349	-0.024	1.296	0.043	1.284	4.501	1.251	4.451	1.310	4.492	1.303
	UWLS	-0.036	1.374	-0.028	1.354	0.049	1.327	4.505	1.259	4.755	1.357	4.793	1.353
	OPMAOC	-0.040	1.350	-0.025	1.296	0.044	1.284	4.501	1.251	4.457	1.310	4.496	.306
	WBPMAOC	-0.039	1.349	-0.026	1.359	0.054	1.330	4.503	1.251	4.828	1.355	4.864	1.347
	WLSAOC	-0.038	1.350	-0.027	1.364	0.054	1.335	4.503	1.250	4.848	1.358	4.884	1.353

Note: PMMAR-Proc Mixed MAR analysis; UWLS-Unweighted least squares analysis which is ML for pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WBPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis. Bold values indicate a statistically significant difference between the mean of $\widehat{\theta_{12} - \theta_{11}}$ and $\theta_{12} - \theta_{11}$.

Table 2. Estimated Type I Error Rates and Power

n_k	Test	CD		MAR-DS		MAR-PS	
		$\hat{\alpha}$	$1 - \hat{\beta}$	$\hat{\alpha}$	$1 - \hat{\beta}$	$\hat{\alpha}$	$1 - \hat{\beta}$
100	PMMAR	0.044	0.670	0.053	0.685	0.048	0.711
	UWLS	0.048	0.661	0.063	0.738	0.057	0.745
	OPMAOC	0.039	0.646	0.044	0.664	0.038	0.687
	WBPMAOC	0.045	0.667	0.084	0.789	0.083	0.799
	WLSAOC	0.044	0.667	0.079	0.787	0.081	0.795
	OEPAOC	0.053	0.694	0.061	0.508	0.047	0.518
	OTSAOC	0.054	0.647	0.059	0.448	0.051	0.468
200	PMMAR	0.054	0.931	0.047	0.920	0.051	0.929
	UWLS	0.052	0.935	0.059	0.935	0.059	0.947
	OPMAOC	0.044	0.923	0.039	0.911	0.042	0.918
	WBPMAOC	0.053	0.930	0.086	0.955	0.076	0.972
	WLSAOC	0.052	0.929	0.082	0.956	0.076	0.971
	OEPAOC	0.044	0.950	0.047	0.797	0.047	0.780
	OTSAOC	0.058	0.919	0.040	0.742	0.043	0.745

Note: PMMAR-Proc Mixed MAR analysis; UWLS-Unweighted least squares analysis which is ML for pattern-mixture models; OPMAOC-Overall et al.'s (1999) Proc Mixed ANCOVA; WBPMAOC- Wu and Bailey's (1989) ANCOVA with PROC Mixed as defined in this paper; WLSAOC- Wang-Clow et al.'s (1995) ANCOVA analysis; WLSAOCMM-Wang-Clow et al.'s ANCOVA using the method of moments for estimation; OEPAOC- Overall et al.'s two-stage endpoint ANCOVA analysis; OTSAOC- Overall et al.'s two-stage ANCOVA.

Table 3 Means and Estimated Standard Errors of Test Statistics: MNAR and $\theta_{12} - \theta_{11} = 0$

n_k	Test	MNAR-DS- Y_{j-1}		MNAR-PS- Y_{j-1}		MNAR-DS- Y_j		MNAR-PS- Y_j		MNAR-PS-SI	
		MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
100	PMMAR	0.068	1.825	0.041	1.802	0.006	1.454	-0.086	1.524	-0.041	1.754
	UWLS	0.077	2.064	0.041	2.090	0.062	1.561	-0.094	1.616	-0.070	3.110
	OPMAOC	0.068	1.827	0.044	1.810	0.008	1.454	-0.086	1.522	-0.047	1.764
	WBPMAOC	0.072	2.039	0.060	2.065	0.012	1.525	-0.098	1.572	-0.071	2.084
	WLSAOC	0.073	2.045	0.064	2.078	0.013	1.519	-0.092	1.568	-0.073	2.019
200	PMMAR	-0.045	1.274	0.044	1.251	0.040	1.077	-0.013	1.048	0.012	1.284
	UWLS	-0.065	1.495	0.045	1.455	0.041	1.193	-0.023	1.146	0.009	2.165
	OPMAOC	-0.045	1.278	0.043	1.258	0.044	1.080	-0.015	1.046	0.019	1.291
	WBPMAOC	-0.050	1.460	0.027	1.432	0.048	1.130	-0.016	1.076	0.062	1.475
	WLSAOC	-0.051	1.468	0.030	1.437	0.049	1.128	-0.016	1.077	0.055	1.447

Note: See note to Table 1.

Table 4. Means and Estimated Standard Errors of Test Statistics: MNAR and $\theta_{12} - \theta_{11} = 4.5$

n_k	Test	MNAR-DS- Y_{j-1}		MNAR-PS- Y_{j-1}		MNAR-DS- Y_j		MNAR-PS- Y_j		MNAR-PS-SI	
		MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
100	PMMAR	4.314	1.875	4.287	1.781	2.937	1.540	2.833	1.493	3.859	1.809
	UWLS	4.990	2.073	4.978	1.998	3.218	1.667	3.141	1.629	4.596	3.048
	OPMAOC	4.364	1.880	4.336	1.785	3.026	1.542	2.921	1.494	4.024	1.819
	WBPMAOC	5.165	2.051	5.132	1.990	3.419	1.589	3.310	1.552	4.992	2.077
	WLSAOC	5.182	2.059	5.149	2.000	3.396	1.585	3.289	1.552	4.845	2.045
200	PMMAR	4.305	1.269	4.294	1.328	2.879	1.007	2.873	1.073	3.815	1.218
	UWLS	4.967	1.417	4.973	1.477	3.168	1.082	3.149	1.140	4.457	2.062
	OPMAOC	4.351	1.272	4.342	1.333	2.970	1.005	2.961	1.067	3.988	1.236
	WBPMAOC	5.140	1.394	5.128	1.455	3.366	1.037	3.340	1.079	4.933	1.458
	WLSAOC	5.158	1.404	5.151	1.468	3.339	1.036	3.319	1.079	4.794	1.425

Note: See note to Table 1.

Table 5. Estimated Type I Error Rates and Power: MNAR Conditions.

n_k	Test	MNAR-DS- Y_{j-1}		MNAR-PS- Y_{j-1}		MNAR-DS- Y_j		MNAR-PS- Y_j		MNAR-PS- SI	
		$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$	$\hat{\alpha}$	$1-\hat{\beta}$
100	PMMAR	0.041	0.661	0.051	0.665	0.058	0.509	0.066	0.477	0.045	0.586
	UWLS	0.058	0.731	0.072	0.738	0.043	0.530	0.050	0.503	0.043	0.362
	OPMAOC	0.034	0.633	0.041	0.641	0.049	0.498	0.060	0.465	0.033	0.580
	WBPMAOC	0.111	0.825	0.116	0.825	0.064	0.632	0.083	0.607	0.092	0.756
	WLSAOC	0.108	0.822	0.117	0.822	0.065	0.625	0.079	0.605	0.066	0.730
	OEPAOC	0.048	0.500	0.047	0.498	0.048	0.484	0.042	0.427	0.037	0.521
	OTSAOC	0.043	0.444	0.051	0.457	0.051	0.418	0.055	0.404	0.048	0.478
200	PMMAR	0.054	0.924	0.048	0.906	0.067	0.762	0.044	0.769	0.039	0.849
	UWLS	0.075	0.948	0.067	0.950	0.053	0.780	0.040	0.775	0.050	0.589
	OPMAOC	0.043	0.912	0.036	0.893	0.050	0.764	0.035	0.775	0.033	0.863
	WBPMAOC	0.115	0.976	0.111	0.978	0.076	0.876	0.056	0.886	0.092	0.957
	WLSAOC	0.111	0.974	0.114	0.977	0.077	0.872	0.059	0.880	0.079	0.948
	OEPAOC	0.042	0.792	0.048	0.797	0.058	0.737	0.051	0.732	0.037	0.836
	OTSAOC	0.051	0.742	0.036	0.741	0.053	0.680	0.046	0.677	0.055	0.749

Note. See note to Table 2. Bold values indicate $\hat{\alpha} > .075$

UWLS, WBPMAOC, and WLSAOC tended to have slightly larger standard deviations than did PMMAR and OPMAOC. In Table 4 all treatment effects were significantly different from 4.5 except for UWLS under the MNAR-PS-SI condition. Again WBPMAOC and WLSAOC tended to have slightly larger standard deviations than did PMMAR and OPMAOC. Except in the MNAR-PS-SI conditions, UWLS tended to have standard deviations similar to those for WBPMAOC and WLSAOC. In the MNAR-PS-SI conditions UWLS had notably larger standard deviations than did the other procedures.

Table 5 contains estimated Type I error rates and power for MNAR missing data mechanisms. With regard to Type I error control we note that, as was true with MAR data, WBPMAOC and WLSAOC did not control the Type I error rate.

Of the methods that control their rates of Type I error, the methods divide into two groups: the more powerful methods (PMMAR, UWLS, and OPMAOC) and the less powerful methods (OEPAOC and OTSAOC). The difference in power between the two groups was quite substantial in most conditions. When missingness was predicted by slopes and

intercepts (MNAR-PS-SI), PMMAR and OPMAOC were more powerful than UWLS. In the other MNAR conditions, UWLS was more powerful than PMMAR or OPMAOC. The power advantage in favor of UWLS was smaller than the power advantage for PMMAR and OPMAOC.

Conclusion

Presented and examined are methods of analysis that, according to the literature, should result in better estimation of unknown parameters and which take MNAR missingness into account in their analyses when data are missing in a parallel groups design. In particular, we investigated methods due to Little (1995), Wu and Bailey (1988, 1989), Wang-Clow et al. (1995) and Overall et al. (1999).

The results, along with those in Algina and Keselman (2003), suggest that whether the pattern of missing data is monotone or not will influence the selection of a method for analyzing the data. Based on bias, control of Type I error rate, and power, Algina and Keselman concluded that Overall et al.'s (1998) mixed model procedure (OPMAOC) is promising when the missing data pattern is monotone.

The present research suggests that OPMAOC works reasonably well when the missing data pattern is not monotone, but that the mixed model for MAR data (PMMAR) and UWLS are very competitive. Comparing OPMAOC and PMMAR, both controlled the Type I error rate in all conditions investigated and power differences were very small. The major difference was that under the MNAR missing data mechanism OPMAOC tended to be slightly less biased than PMMAR was. Comparing OPMAOC and PMMAR, both controlled the Type I error rate. Power differences depended on the missing data mechanism.

When missingness was predicted by the slope and intercept, the missing data mechanism for which UWLS was developed, UWLS was much less powerful than OPMAOC because its standard error was notably larger than the standard error for OPMAOC. In all other conditions, power for the two procedures was

either quite similar or favored UWLS. Bias differences also depended on the missing data mechanism. When missingness was predicted by the slope and intercept, UWLS was unbiased but OPMAOC was not. When the data were MAR, OPMAOC was unbiased but UWLS was not. When missingness on Y_j was predicted by scores on Y_j (MNAR-DS- Y_j and MNAR-PS- Y_j), UWLS was less biased than was OPMAOC. The opposite was true when missingness was predicted by scores on Y_{j-1} (MNAR-DS- Y_{j-1} and MNAR-PS- Y_{j-1}).

Considering the performance of OPMAOC in Algina and Keselman (2003) and in the present study, if a researcher wants to use a single procedure for monotone and non-monotone patterns of missing data, OPMAOC appears promising. Of course, as is true of all empirical studies, the generalizability of the results is limited by the design of the study. The procedures may perform differently if different models for dropping out are adopted.

References

- Algina, J., & Keselman, H. J. (2003). Analyzing effects in longitudinal studies with missing data. *Journal of Modern Applied Statistical Methods*, 2, 50-72.
- Ahn, C., Tonidandel, S., & Overall, J. E. (2000). Issues in use of SAS Proc Mixed to test the significance of treatment effects in controlled clinical trials. *Journal of Biopharmaceutical Statistics*, 10(2), 265-286.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Diggle, P. J., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis, with discussion. *Applied Statistics*, 43 49-73.
- Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64-78.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.

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

Overall, J. E., Ahn, C., Shivakumar, C., & Kalburgi, Y. (1999). Problematic formulations of SAS Proc Mixed models for repeated measurements. *Journal of Biopharmaceutical Statistics*, 9, 189-216.

Overall, J. E., Ghasser, S., & Fiore, J. (1996). Random regression with imputed values for dropouts. *Psychopharmacology Bulletin*, 32, 377-388.

SAS Institute Inc. (2000). *SAS/STAT user's guide, version 8*. Cary, NC: Author.

Wang-Clow, F., Lange, M., Laird, N. M., & Ware, J. H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, 14, 283-297.

Wu, M. C., & Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, 45, 939-955.

Verbeke, G., & Molenberghs. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

A Power Comparison of Robust Test Statistics Based On Adaptive Estimators

H. J. Keselman
University of Manitoba

Rand R. Wilcox
University of Southern California

James Algina
University of Florida

Katherine Fradette
University of Manitoba

Abdul R. Othman
University of Sains Malaysia

Seven test statistics known to be robust to the combined effects of nonnormality and variance heterogeneity were compared for their sensitivity to detect treatment effects in a one-way completely randomized design containing four groups. The six Welch-James-type heteroscedastic tests adopted either symmetric or asymmetric trimmed means, were transformed for skewness, and used a bootstrap method to assess statistical significance. The remaining test, due to Wilcox and Keselman (2003), used a modification of the well-known one-step M-estimator of central tendency rather than trimmed means. The Welch-James-type test is recommended because for nonnormal data likely to be encountered in applied research settings it should be more powerful than the test presented by Wilcox and Keselman. However, the reverse is true for data that are extremely nonnormal.

Key words: Trimmed estimators, symmetric and asymmetric trimming, heteroscedastic test statistic, nonnormality, variance heterogeneity

---

Introduction

Keselman, Wilcox, Othman and Fradette (2002) demonstrated the benefit of testing for symmetry, applying a transformation for skewness, adopting robust estimators and using bootstrapping methodology with a Welch-James-type heteroscedastic statistic in order to obtain a robust test of treatment group equality

when data are nonnormal, heterogeneous and unbalanced in one-way completely randomized designs. In particular, they applied a test for symmetry due to Hogg, Fisher and Randles (1975), modified by Babu, Padmanaban and Puri (1999), in order to determine whether data should be trimmed from each tail of the data distribution (symmetric trimming) per group or whether data should only be trimmed from one-tail of the data distribution (asymmetric trimming) per group prior to applying the Johansen (1980) test for treatment group equality. Furthermore, they investigated the utility of transforming the statistic, to circumvent the biasing effects due to skewness, with methods presented by Johnson (1978) and Hall (1992). Lastly, they assessed statistical significance with and without bootstrapping methodology and concluded that critical values obtained through bootstrapping provided an additional benefit against the deleterious effects of nonnormality and variance heterogeneity.

H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology. Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. James Algina (algina@ufl.edu) is Professor of Educational Psychology. Katherine Fradette (umfradet@cc.umanitoba.ca) is a graduate student in psychology. Her interests are in applied statistical analysis. Abdul Rahman Othman (oarahman@usm.my) is Associate Professor, School of Distance Education. Work on this project was supported by a grant by the Natural Sciences and Engineering Council of Canada.

These authors concluded by recommending that researchers test for treatment group equality by adopting the aforementioned

modifications to the Johansen test with 10% symmetric trimming or 20% asymmetric trimming based on a preliminary test for symmetry. They noted as well that other percentages of symmetric/asymmetric trimming worked quite well with respect to Type I error control (e.g., 15%/30%).

Othman, Keselman, Padmanabhan, Wilcox, and Fradette (2003) compared a number of recently developed adaptive robust methods with respect to their ability to control Type I errors and their sensitivity to detect differences between groups when data were nonnormal, heterogeneous, and the design was unbalanced. In particular, two new approaches to comparing the typical score across treatment groups due to Babu et al. (1999) were compared to two new methods presented by Wilcox and Keselman (2003) and Keselman et al. (2002). The procedures examined exhibited very good Type I error control and the power results clearly favored one of the methods (a method they referred to as MOMT) presented by Wilcox and Keselman; indeed, in the vast majority of the cases investigated, this most favored approach had substantially larger power values compared to the other procedures.

Based on the findings of these two studies an important research question remains. Namely, how does the power of the robust and powerful procedure investigated by Othman et al. (2003) (i.e., MOMT) compare to the sensitivity of the Johansen (1980) Welch-James-(WJ)-type procedure for detecting treatment effects in one-way completely randomized designs? This question is important because other investigators have recommended the WJ test due to its sensitivity to detect effects for other designs [See e.g., Algina & Keselman (1998)] and neither Keselman et al. (2002) or Othman et al. investigated the power of the WJ test.

Test Statistics

The WJ Statistic

Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups

designs can be formulated from a general linear model perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. Their approach is adopted in this article and is presented in abbreviated form.

A general approach for testing hypotheses of mean equality using an approximate degrees of freedom solution is developed using matrix notation. The multivariate perspective is considered first; the univariate model is a special case of the multivariate. Consider the general linear model:

$$Y = X\beta + \xi, \quad (1)$$

where Y is an $N \times p$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, X is an $N \times r$ design matrix consisting entirely of zeros and ones with $\text{rank}(X) = r$, β is an $r \times p$ matrix of nonrandom parameters (i.e., population means), and ξ is an $N \times p$ matrix of random error components. Let Y_j ($j=1, \dots, r$) denote the submatrix of Y containing the scores associated with the n subjects in the j th group (cell) (For the one-way design considered in this paper $n=n_j$). It is typically assumed that the rows of Y are independently and normally distributed, with mean vector β_j and variance-covariance matrix Σ_j [i.e., $N(\beta_j, \Sigma_j)$], where the j th row of β , $\beta_j = [\mu_{j1} \dots \mu_{jp}]$, and $\Sigma_j \neq \Sigma_{j'} (j \neq j')$. Specific formulas for estimating β and Σ_j , as well as an elaboration of Y are given in Lix and Keselman (1995, See their Appendix A).

The general linear hypothesis is

$$H_0 : R\mu = 0, \quad (2)$$

where $R = C \otimes U^T$, C is a $df_C \times r$ matrix which controls contrasts on the independent groups effect(s), with $\text{rank}(C) = df_C \leq r$, and U is a $p \times df_U$ matrix which controls contrasts on the within-subjects effect(s), with $\text{rank}(U) = df_U \leq p$, ' $\otimes$ ' is the Kronecker or direct

product function, and $'^T'$ is the transpose operator. For multivariate independent groups designs, U is an identity matrix of dimension p (i.e., I_p). The R contrast matrix has $df_C \times df_U$ rows and $r \times p$ columns. In Equation 2, $\mu = \text{vec}(\beta^T) = [\beta_1 \dots \beta_r]^T$. In other words, μ is the column vector with $r \times p$ elements obtained by stacking the columns of β^T . The 0 column vector is of order $df_C \times df_U$ [See Lix & Keselman (1995) for illustrative examples].

The generalized test statistic given by Johansen (1980) is

$$T_{WJ} = (R\hat{\mu})^T (R\hat{\Sigma}R^T)^{-1} (R\hat{\mu}) \quad (3)$$

where $\hat{\mu}$ estimates μ , and $\hat{\Sigma} = \text{diag}[\hat{\Sigma}_1/n_1 \dots \hat{\Sigma}_r/n_r]$, a block matrix with diagonal elements $\hat{\Sigma}_j/n_j$.

This statistic, divided by a constant, c (i.e., T_{WJ}/c), approximately follows an F distribution with degrees of freedom $\nu_1 = df_C \times df_U$, and $\nu_2 = \nu_1(\nu_1 + 2)/(3A)$, where $c = \nu_1 + 2A - (6A)/(\nu_1 + 2)$. The formula for the statistic A is provided in Lix and Keselman (1995).

When $p=1$, that is, for a univariate model, the elements of Y are assumed to be independently and normally distributed with mean μ_j and variance σ_j^2 [i.e., $N(\mu_j, \sigma_j^2)$]. To test the general linear hypothesis, C has the same form and function as for the multivariate case, but now $U=1$, $\hat{\mu} = [\hat{\mu}_1 \dots \hat{\mu}_r]^T$ and $\hat{\Sigma} = \text{diag}[\hat{\sigma}_1^2/n_1 \dots \hat{\sigma}_r^2/n_r]$, (See Lix & Keselman's 1995 Appendix A for further details of the univariate model.).

Robust Estimation

In this article robust estimates of central tendency and variability are applied to the T_{WJ} statistic. That is, heteroscedastic ANOVA methods are readily extended to the problem of comparing trimmed means. The goal is to determine whether the effect of a treatment varies across J ($j=1, \dots, J$) groups; that is, to

determine whether a typical score varies across groups. When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the μ_t s. That is, to test the omnibus hypothesis in a one-way completely randomized design, the null hypothesis would be $H_0: \mu_{t1} = \mu_{t2} = \dots = \mu_{tJ}$.

Let $Y_{(1)j} \leq Y_{(2)j} \leq \dots \leq Y_{(n_j)j}$ represent the ordered observations associated with the j th group. Let $g_j = [\gamma n_j]$, where γ represents the proportion of observations that are to be trimmed in each tail of the distribution and $[x]$ is the greatest integer $\leq x$. The effective sample size for the j th group becomes $h_j = n_j - 2g_j$. The j th sample trimmed mean is

$$\hat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)j} \quad (4)$$

Wilcox (1995) suggested that 20% trimming should be used. (See Wilcox, 1995, and the references cited for a justification of the 20% rule.)

The sample Winsorized mean is necessary and is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij} \quad (5)$$

where

$$\begin{aligned} X_{ij} &= Y_{(g_j+1)j} \text{ if } Y_{ij} \leq Y_{(g_j+1)j} \\ &= Y_{ij} \text{ if } Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ &= Y_{(n_j-g_j)j} \text{ if } Y_{ij} \geq Y_{(n_j-g_j)j} \end{aligned}$$

The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of a trimmed mean, is then given by

$$\hat{\sigma}_{wj}^2 = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2 \quad (6)$$

The standard error of the trimmed mean is estimated with $\sqrt{(n_j - 1)\hat{\sigma}_{wj}^2 / [h_j(h_j - 1)]}$.

Under asymmetric trimming, and assuming, without loss of generality, that the distribution is positively skewed so that trimming takes place in the upper tail, the j th sample trimmed mean is

$$\hat{\mu}_{ij} = \frac{1}{h_j} \sum_{i=1}^{n_j - g_j} Y_{(i)j}$$

and the j th sample Winsorized mean is

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij},$$

where

$$\begin{aligned} X_{ij} &= Y_{ij} \text{ if } Y_{ij} \leq Y_{(n_j - g_j)j} \\ &= Y_{(n_j - g_j)j} \text{ if } Y_{ij} \geq Y_{(n_j - g_j)j}. \end{aligned}$$

The sample Winsorized variance is again defined as (given the new definition of $\hat{\mu}_{wj}$)

$$\hat{\sigma}_{wj}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^2$$

and the standard error of the mean again takes its usual form (given the new definition of $\hat{\mu}_{wj}$).

Thus, with robust estimation, the trimmed group means ($\hat{\mu}_{ij,s}$) replace the least squares group means ($\hat{\mu}_{j,s}$), the Winsorized group variances estimators ($\hat{\sigma}_{wj}^2$) replace the least squares variances ($\hat{\sigma}_{j,s}^2$), and h_j replaces n_j and accordingly one computes the robust version of T_{wj} , $T_{wj,t}$ (See Keselman, Wilcox, & Lix, 2003; and Rocke, Downs & Rocke, 1982, for another justification for adopting robust estimates).

Bootstrapping

Now considered is how extensions of the ANOVA method just outlined might be improved. In terms of probability coverage and controlling the probability of a Type I error, extant investigations indicate that the most

successful method, when using a 20% trimmed mean (or some M-estimator), is some type of bootstrap method.

Following Westfall and Young (1993), and as described by Wilcox (1997), let $C_{ij} = Y_{ij} - \hat{\mu}_{ij}$; thus, the C_{ij} values are the empirical distribution of the j th group, centered so that the sample trimmed mean is zero. That is, *the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample*. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value. For each j , obtain a bootstrap sample by randomly sampling with replacement n_j observations from the C_{ij} values, yielding $Y_1^*, \dots, Y_{n_j}^*$. Let $T_{wj,t}^*$ be the value of Johansen's (1980) test based on the bootstrap sample. Now randomly sample (with replacement), B bootstrap samples from the shifted/centered distributions each time calculating the statistic $T_{wj,t}^*$. The B values of $T_{wj,t}^*$ are put in ascending order, that is, $T_{wj,t(1)}^* \leq \dots \leq T_{wj,t(B)}^*$, and an estimate of an appropriate critical value is $T_{wj,t(a)}^*$, where $a = (1 - \alpha)B$, rounded to the nearest integer. One will reject the null hypothesis of location equality (i.e., $H_0: \mu_{t1} = \mu_{t2} = \dots = \mu_{tJ}$) when $T_{wj,t} > T_{wj,t(a)}^*$, where $T_{wj,t}$ is the value of the heteroscedastic statistic based on the original non-bootstrapped data. Keselman et al. (2002) illustrate the use of this procedure for testing both omnibus and sub-effect (linear contrast) hypotheses in completely randomized and correlated groups designs.

Transformations for the Welch-James Statistic

Guo and Luh (2000) and Luh and Guo 1999 found that Johnson's (1978) and Hall's (1992) transformations improved the performance of several heteroscedastic test statistics when they were used with trimmed means, including the WJ statistic, in the presence of heavy-tailed and skewed distributions.

In this study both approaches are compared for removing skewness when applied

to the T_{WJt} statistic. Let $Y_{ij} = (Y_{1j}, Y_{2j}, \dots, Y_{n_{ij}})$ be a random sample from the j th distribution. Let $\hat{\mu}_{ij}$, $\hat{\mu}_{wj}$ and $\hat{\sigma}_{wj}^2$ be, respectively, the trimmed mean, Winsorized mean and Winsorized variance of group j . Define the Winsorized third central moment of group j as

$$\hat{\mu}_{3j} = \frac{1}{n_j} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_{wj})^3.$$

Let

$$\tilde{\sigma}_{wj}^2 = \frac{(n_j - 1)}{h_j - 1} \hat{\sigma}_{wj}^2,$$

$$\tilde{\mu}_{wj} = \frac{n_j}{h_j} \hat{\mu}_{3j},$$

$$q_j = \frac{\tilde{\sigma}_{wj}^2}{h_j},$$

$$w_{ij} = \frac{1}{q_j},$$

$$U_t = \sum_{j=1}^J w_{ij},$$

and

$$\hat{\mu}_t = \frac{1}{U_t} \sum_{j=1}^J w_{ij} \hat{\mu}_{ij}.$$

Luh and Guo (2000) defined a trimmed mean statistic with Johnson's transformation as

$$T_{Johnson_j} = (\hat{\mu}_{ij} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}_{wj}^2 h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}_{wj}^4} (\hat{\mu}_{ij} - \hat{\mu}_t)^2. \quad (7)$$

From Guo and Luh (2000) one can deduce that a trimmed mean statistic with Hall's (1992) transformation would be

$$T_{Hall_j} = (\hat{\mu}_{ij} - \hat{\mu}_t) + \frac{\tilde{\mu}_{wj}}{6\tilde{\sigma}_{wj}^2 h_j} + \frac{\tilde{\mu}_{wj}}{3\tilde{\sigma}_{wj}^4} (\hat{\mu}_{ij} - \hat{\mu}_t)^2 + \frac{\tilde{\mu}_{wj}^2}{27\tilde{\sigma}_{wj}^8} (\hat{\mu}_{ij} - \hat{\mu}_t)^3. \quad (8)$$

Keselman et al. (2002) indicated that sample trimmed means, sample Winsorized variances and trimmed sample sizes can be substituted for the usual sample means, variances and sample sizes in the T_{WJ} statistic. That is,

$$T_{WJ} = \sum_{j=1}^J w_{ij} (\hat{\mu}_{ij} - \hat{\mu}_t)^2,$$

which, when divided by c , is distributed as an F variable with df of $J - 1$ and

$$\nu = (J^2 - 1) \left[3 \sum_{j=1}^J \frac{(1 - w_{ij} / U_t)^2}{h_j - 1} \right]^{-1}$$

where

$$c = (J - 1) \left[1 + \frac{2(J - 2)}{J^2 - 1} \sum_{j=1}^J \frac{(1 - w_{ij} / U_t)^2}{h_j - 1} \right].$$

Now we can define

$$T_{WJ_{Johnson}} = \sum_{j=1}^J w_{ij} (T_{Johnson_j})^2, \quad (9)$$

and

$$T_{WJ_{Hall}} = \sum_{j=1}^J w_{ij} (T_{Hall_j})^2. \quad (10)$$

Then $T_{WJ_{Johnson}}$ and $T_{WJ_{Hall}}$, when divided by c , are also distributed as F variates with no change in degrees of freedom.

A Preliminary Test for Symmetry

A stumbling block to adopting asymmetric versus symmetric trimming has been the inability of researchers to determine when to adopt one form of trimming over the other. Work by Hogg et al. (1975) and Babu et al. (1999), however, may provide a successful solution to this problem. The details of this method are presented in Othman et al. (2002).

The One-Step Modified M-(MOM) Estimator

For J independent groups (this estimator can also be applied to dependent groups) consider the MOM estimator introduced by Wilcox and Keselman (2003). They suggested modifying the well-known one-step M-estimator

$$\frac{1.28(MADN_j)(i_2 - i_1) + \sum_{i=i_1+1}^{n_j-i_2} Y_{(i)j}}{n_j - i_1 - i_2}, \quad (11)$$

by removing $1.28(MADN_j)(i_2 - i_1)$, where $MADN_j = MAD_j / .6745$, MAD_j = the median of the values $|Y_{ij} - \hat{M}_j|, \dots, |Y_{n_{jj}} - \hat{M}_j|$, $\hat{M}_j$ is the median of the j th group, i_1 = the number of observations where $Y_{ij} - \hat{M}_j < 2.24(MADN_j)$ and i_2 = the number of observations where $Y_{ij} - \hat{M}_j > 2.24(MADN_j)$. Thus, the modified M-estimator suggested by Wilcox and Keselman is

$$\hat{\theta}_j = \sum_{i=i_1+1}^{n_j-i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2}. \quad (12)$$

The MOM estimate of location is just the average of the values left after all outliers (if any) are discarded. The constant 2.24 is motivated in part by the goal of having a reasonably small standard error when sampling from a normal distribution. Moreover, detecting outliers with Equation 12 is a special case of a more general outlier detection method derived by Rousseeuw and van Zomeren (1990).

MOMT

MOM estimators, like trimmed means, can be applied to test statistics to investigate the equality of this measure (θ) of the typical score across treatment groups. The null hypothesis is

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_J, \quad (13)$$

where θ_j is the population value of MOM associated with the j th group. Of the two statistics that can be used to test this hypothesis, Othman et al. (in press) found that the one based on the work of Liu and Singh (1997) was most powerful. To obtain the test, let

$$\delta_{jj'} = \theta_j - \theta_{j'} \quad (j < j'). \quad (14)$$

Thus, the $\delta_{jj'}$ s are the all possible pairwise comparisons among the J treatment groups. Now, if all groups have a common measure of location (i.e., $\theta_1 = \theta_2 = \dots = \theta_J$), then $H_0 : \delta_{12} = \delta_{13} = \dots = \delta_{J-1,J} = 0$. A bootstrap method can be used to assess statistical significance. Bootstrap samples are obtained for the Y_{ij} values and one rejects if the zero vector is sufficiently far from the center of the bootstrap estimates of the delta values. Thus, bootstrap samples are obtained from the Y_{ij} values rather than the C_{ij} s. For each bootstrap replication ($B = 599$ is recommended) one computes the robust estimators (i.e., MOM) of location (i.e., $\hat{\theta}_{jb}^*$, $j = 1, \dots, J$; $b = 1, \dots, B$) and the corresponding estimates of $\delta_{jj'b}$ ($\hat{\delta}_{jj'b}^* = \hat{\theta}_{jb}^* - \hat{\theta}_{j'b}^*$). The strategy is to determine how deeply $0 = (0 \ 0 \ \dots \ 0)$ is nested within the bootstrap values $\hat{\delta}_{jj'b}^*$, where 0 is a vector having length $K = J(J-1)/2$. This assessment is made by adopting a modification of Mahalanobis's distance statistic.

For notational convenience, the K differences $\hat{\delta}_{jj'}$ can be rewritten as $\hat{\Delta}_1, \dots, \hat{\Delta}_K$ and their corresponding bootstrap values as $\hat{\Delta}_{kb}^*$ ($k = 1, \dots, K$; $b = 1, \dots, B$). Thus, let

$$\bar{\Delta}_k^* = \frac{1}{B} \sum_{b=1}^B \hat{\Delta}_{kb}^*,$$

and

$$Z_{kb} = \hat{\Delta}_{kb}^* - \bar{\Delta}_k^* + \hat{\Delta}_k.$$

(Note the Z_{kb} s are shifted bootstrap values having mean $\hat{\Delta}_k$.) Now define

$$S_{kk'} = \frac{1}{B-1} \sum_{b=1}^B (Z_{kb} - \bar{Z}_k)(Z_{k'b} - \bar{Z}_{k'}), \quad (15)$$

where

$$\bar{Z}_k = \frac{1}{B} \sum_{b=1}^B Z_{kb}.$$

(Note: The bootstrap population mean of $\bar{\Delta}_k^*$ is known and is equal to $\hat{\Delta}_k$.)

With this procedure, next compute

$$D_b = (\hat{\Delta}_b^* - \hat{\Delta})S^{-1}(\hat{\Delta}_b^* - \hat{\Delta})' , \quad (16)$$

where $\hat{\Delta}_b^* = (\hat{\Delta}_{1b}^*, \dots, \hat{\Delta}_{Kb}^*)$ and $\hat{\Delta} = (\hat{\Delta}_1, \dots, \hat{\Delta}_K)$. Accordingly, D_b measures how closely $\hat{\Delta}_b$ is located to $\hat{\Delta}$. If the null vector (0) is relatively far from $\hat{\Delta}$ one rejects H_0 . Therefore, to assess statistical significance, put the D_b values in ascending order ($D_{(1)} \leq \dots \leq D_{(B)}$) and let $a = (1 - \alpha)B$ (rounded to the nearest integer). Reject H_0 if

$$T \geq D_{(a)} , \quad (17)$$

where

$$T = (0 - \hat{\Delta})S^{-1}(0 - \hat{\Delta})' . \quad (18)$$

It is important to note that $\theta_1 = \theta_2 = \dots = \theta_J$ can be true iff $H_0 : \theta_1 - \theta_2 = \dots = \theta_{J-1} - \theta_J = 0$ (Therefore, it suffices to test that a set of K pairwise differences equal zero.) However, to avoid the problem of arriving at different conclusions (i.e., sensitivity to detect effects) based on how groups are arranged (if all MOMs are unequal), it is recommended that one test the hypothesis that all pairwise differences equal zero.

Methodology

Seven tests for treatment group equality were compared for their sensitivity to detect treatment effects under conditions of nonnormality and variance heterogeneity in an independent groups design with four treatments. The procedures investigated, based on the findings and recommendations of Keselman et al. (2002) and Othman et al. (in press), were:

WJ with preliminary testing for symmetry (Babu et al., 1999)/Symmetric and Asymmetric Trimming:

1.-3. WJJB1020(1530)(2040)-WJ with Johnson's (1978) transformation and bootstrapping. If data are symmetric use 10% (15%) (20%) symmetric trimming, otherwise

use 20% (30%) (40%) one sided trimming.

4.-6. WJHB1020(1530)(2040)-WJ with Hall's (1990) transformation and bootstrapping. If data are symmetric use 10% (15%) (20%) symmetric trimming, otherwise use 20% (30%) (40%) one sided trimming.

7. MOMT.

Four variables were manipulated in the study: (a) sample size, (b) degree of variance heterogeneity, (c) pairing of unequal variances and group sizes, and (d) population distribution.

An unbalanced completely randomized design containing four groups was investigated since previous research has looked at this design (e.g., Keselman et al., 2002; Lix & Keselman, 1998; Othman et al., in press; Wilcox, 1988). The two cases of total sample size and the group sizes were $N = 70$ (10, 15, 20, 25) and $N = 90$ (15, 20, 25, 30). The values of n_j were selected from those used by Lix and Keselman (1998) in their study comparing omnibus tests for treatment group equality; their choice of values was, in part, based on having group sizes that others have found to be generally sufficient to provide reasonably effective Type I error control (e.g., see Wilcox, 1994).

The unequal variances were either in a 36:1:1:1 or 8:1:1:1 ratio. Though a ratio of 36:1:1:1 may seem extreme, ratios similar to this case, and larger, have been reported in the literature. Keselman, et al. (1998) after reviewing articles published in prominent education and psychology journals noted that they found ratios as large as 24:1 and 29:1 in one-way and factorial completely randomized designs. Wilcox (2003) cited data sets where the ratio was 17,977:1!

It is appropriate to compare the test statistics under this condition of variance heterogeneity -- the results under this condition will tell how the tests perform under conditions that either have been reported or may likely be encountered with actual data sets. Furthermore, even assuming that a 36:1:1:1 ratio of variances may be large, it nonetheless seems reasonable to see how well the tests perform under a potentially extreme condition. This will provide researchers with information regarding how well the tests hold up under any degree of heterogeneity they are likely to obtain in their

data, thus providing a generalizable result. Nonetheless, the tests were also compared under a less extreme condition of heterogeneity, i. e., when the variances were in a ratio of 8:1:1:1.

Variances and group sizes were both positively and negatively paired. A positive pairing referred to the case in which the largest n_j was associated with the population having the largest variance; a negative pairing referred to the case in which the largest n_j was associated with the population having the smallest variance. These conditions were chosen since they typically produce conservative and liberal results, respectively.

With respect to the effects of distributional shape on Type I error, we chose to investigate nonnormal distributions in which the data were obtained from a variety of skewed distributions. In addition to generating data from a χ^2_3 distribution, we also used the method described in Hoaglin (1985) to generate distributions with more extreme degrees of skewness and kurtosis. These particular types of nonnormal distributions were selected since educational and psychological research data typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) investigated the effects of eight nonnormal distributions, which were identified by Micceri on the robustness of Student's t test, and they found that only distributions with the most extreme degree of skewness (e.g., $\gamma = 1.64$) affected the Type I error control of the independent sample t statistic. Thus, because the statistics investigated have operating characteristics similar to those reported for the t statistic, it was assumed that this approach to modeling skewed data would adequately reflect conditions in which those statistics might not perform optimally.

For the χ^2_3 distribution, skewness and kurtosis values are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively. The other nonnormal distributions were generated from the g and h distribution (Hoaglin, 1985). Specifically, two g and h distributions were investigated: (a) $g = .5$ and $h = 0$ and (b) $g = .5$ and $h = .5$, where g and h are parameters that determine the moments of a

distribution. To give meaning to these values it should be noted that for the standard normal distribution $g = h = 0$. Thus, when $g = 0$ a distribution is symmetric, and the tails of a distribution will become heavier as h increases in value. Values of skewness and kurtosis corresponding to the investigated values of g and h are (a) $\gamma_1 = 1.75$ and $\gamma_2 = 8.9$, respectively, and (b) $\gamma_1 = \gamma_2 = \text{undefined}$.

These values of skewness and kurtosis for the g and h distributions are theoretical values; Wilcox (1997, p. 73) reported computer generated values, based on 100,000 observations; $\hat{\gamma}_1 = 1.81$ and $\hat{\gamma}_2 = 9.7$ for $g = .5$ and $h = 0$ and $\hat{\gamma}_1 = 120.10$ and $\hat{\gamma}_2 = 18,393.6$ for $g = .5$ and $h = .5$. Thus, the conditions investigated could be described as extreme. They are intended to indicate the operating characteristics of the procedures under substantial departures from homogeneity and normality, with the premise that, if a procedure works under the most extreme of conditions, it is likely to work under most conditions likely to be encountered by researchers.

In terms of the data generation procedure, to obtain pseudo-random normal variates, the SAS generator RANNOR (SAS Institute, 1989) was used. If Z_{ij} is a standard unit normal variate, then $Y_{ij} = \mu_j + \sigma_j \times Z_{ij}$ is a normal variate with mean equal to μ_j and variance equal to σ_j^2 . To generate pseudo-random variates having a χ^2 distribution with three degrees of freedom, three standard normal variates were squared and summed.

To generate data from a g - and h -distribution, standard unit normal variables were converted to random variables via

$$Y_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right),$$

according to the values of g and h selected for investigation. To obtain a distribution with standard deviation σ_j , each Y_{ij} was multiplied by a value of σ_j . It is important to note that this

does not affect the value of the mean when $g = 0$ (see Wilcox, 1994, p. 297). However, when $g > 0$, the population mean for a g - and h -distributed variable is

$$\mu_{gh} = \frac{1}{g(1-h)^{1/2}} (e^{g^2/2(1-h)} - 1)$$

(see Hoaglin, 1985, p. 503). Thus, for those conditions where $g > 0$, μ_{ij} was first subtracted from Y_{ij} before multiplying by σ_j . When working with MOMs, θ_j was first subtracted from each observation (The value of θ_j was obtained from generated data from the respective distributions based on one million observations.). Specifically, for procedures using trimmed means, μ_{ij} was subtracted from the generated variates under every generated distribution. Correspondingly, for the procedure based on MOMs, θ_j was subtracted for all distributions investigated.

The standard deviation of a g - and h -distribution is not equal to one, and thus the values reflect only the amount by which each random variable is multiplied and not the actual values of the standard deviations (see Wilcox, 1994, p. 298). As Wilcox noted, the values for the variances (standard deviations) more aptly reflect the ratio of the variances (standard deviations) between the groups. Five thousand replications of each condition were performed using a .05 statistical significance level. According to Wilcox (1997) and Hall (1986), B was set at 599; that is, their results suggest that it may be advantageous to choose B such that $1 - \alpha$ is a multiple of $(B + 1)^{-1}$.

Lastly, the power of the tests were compared by selected constants to be added to the observations in each group, to avoid ceiling and floor effects; however, values were also selected based on the work of Cohen (1988, pp. 270-272). Specifically, a range for the difference between the groups was selected and then specified this range according to a minimum-, equal-, or maximum-variability difference between the groups. Accordingly, the constants that were added (after centering the data) to the randomly generated data in the four groups were

-1, 0, 0, 1 (minimum variability),
-1, -.5, .5, 1 (equal variability), and
-1, -1, 1, 1 (maximum variability).

Results

Prior to the presentation of power results, the reader should be reminded that the tests examined, very effectively control Type I errors under the conditions studied in this investigation; the Type I error results have been reported in Keselman et al. (2002) and Othman et al. (in press).

The preliminary analysis of the empirical power rates indicated that there were only relatively minor differences between the WJ tests due to type of transformation [i.e., Johnson (1978) or Hall (1992)] for skewness. Accordingly, in Table 1, which contains the empirical power rates, the values tabled for the WJ procedure are based on averaging over the two WJ tests employing the two different transformations for skewness.

Furthermore, no differences existed between the procedures due to sample size and accordingly, the tabled values have been averaged over the two cases of sample size for each test investigated. As well, we note that power rates have been averaged over the type of range investigated (i.e., minimum-, equal- and maximum-variability). Researchers certainly would not be privy to this type of information and thus it seems most reasonable to collapse over this variable.

Based on the values contained in Table 1 we note that: (1) either the WJ1530 and/or the WJ2040 procedure was always at least as powerful as the WH1020 test, (2) the WJ2040 test was at least as powerful as the WJ1530 test for two of the nonnormal distributions investigated (χ_3^2 and $g = .5$ and $h = 0$), while it was marginally less powerful for the remaining nonnormal distribution investigated ($g = .5$ and $h = .5$), and (3) the WJ tests were more powerful than the MOMT test for the χ_3^2 and $g = .5$ and $h = 0$ nonnormal distributions, yet less powerful when the data were $g = .5$ and $h = .5$ distributed.

Table 1. Power Values

Distribution	Max σ^2	Pairing	WJ1020	WJ1530	WJ2040	MOMT	WJ1530- MOMT	WJ2040- MOMT
Chi-Squared	8	Pos	57	60	65	38	22	27
Chi-Squared	36	Pos	52	55	59	34	21	25
Chi-Squared	8	Neg	54	56	61	42	14	19
Chi-Squared	36	Neg	49	50	54	38	12	16
g=.5/h=0	8	Pos	93	94	94	87	07	07
g=.5/h=0	36	Pos	88	90	90	81	09	09
g=.5/h=0	8	Neg	95	95	93	92	03	01
g=.5/h=0	36	Neg	92	92	89	89	03	0
g=.5/h=.5	8	Pos	68	71	69	76	-05	-07
g=.5/h=.5	36	Pos	62	65	64	68	-03	-04
g=.5/h=.5	8	Neg	68	71	68	81	-10	-13
g=.5/h=.5	36	Neg	63	67	65	76	-09	-11
Average			70	72	73	67		

The table also includes values indicating the difference in powers between the WJ1530 and WJ2040 tests and the MOMT test (notated as WJ1530-MOMT and WJ2040-MOMT). These difference scores indicate that power differences favoring the WJ tests were as large as 27 percentage points while those favoring MOMT were at times more powerful by 13 percentage points.

Conclusion

Keselman et al (2002) noted that researchers could achieve robustness to nonnormality and variance heterogeneity by using trimmed means

in a heteroscedastic test statistic [i.e., Johansen (1980)] when data were either trimmed symmetrically or asymmetrically based on a preliminary test for symmetry due to Hogg et al. (1975) and Babu et al. (1999) and when the test was modified by a transformation for skewness due either to Johnson (1978) or Hall (1992) and when statistical significance was assessed through a bootstrap method.

Othman et al. (in press) found that when treatment group equality was assessed with a test statistic suggested by Liu and Singh (1997) comparing across groups a measure of central tendency based on Wilcox and Keselman's (2003) modification of the well-known one-step

M-estimator (i.e., MOM), Type I errors were very effectively controlled under very adverse conditions of nonnormality and variance heterogeneity. Furthermore, and most important to the motivation for the current investigation, they also found that the procedure was substantially more powerful than the other test statistics they investigated.

The purpose of this investigation therefore was to contrast the sensitivity of the test examined by Othman et al. (in press) with the Johansen (1980) Welch-James-type procedure investigated by Keselman et al. (2002) since both methods provide very good Type I error control and good power characteristics have been attributed to the WJ-type test by other researchers (see e.g., Algina & Keselman, 1998), though it has not been compared to the MOMT test nor under conditions examined by Keselman et al. and Othman et al.

For the three nonnormal distributions investigated, it was found that the WJ-type tests were more powerful than the MOMT test when data were moderately to substantially nonnormal (i.e., χ_3^2 and $g = .5$ and $h = 0$ distributed); however, when the data were extremely nonnormal (i.e., $g = .5$ and $h = .5$ distributed), the MOMT test was more powerful than the WJ-type tests. In the former case, the differences favored the WJ-type tests by as much as 27 percentage points while in the latter case MOMT values, at times, exceeded the WJ values by as much as 13 percentage points.

Based on these findings, we recommend, in general, the WJ-type tests that utilize symmetric or asymmetric trimmed means (with the type of trimming based on the Babu et al., 1999, test for symmetry) with a transformation for skewness (due either to Johnson, 1978, or Hall, 1992) and where statistical significance is assessed through the bootstrap method defined in this article (or in Keselman et al., 2002). In particular, the WJ2040 method is recommended. That is, for most nonnormal distributions that researchers are likely to encounter in applied work it is not likely that their data will be as nonnormal as that characterized by the $g = .5$ and $h = .5$ distribution, and thus they are likely to have

greater sensitivity to detect treatment effects with the WJ-type test than with the MOMT test. However, when researchers suspect that their data is extremely nonnormal, in a manner similar to the characteristics of the $g = .5$ and $h = .5$ distribution, then clearly, it will be advantageous to adopt the MOMT test. Numerical results for MOMT can be obtained from Wilcox (2003, pp. 84, 314).

References

- Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and Improved General Approximation tests in the split-plot design. *Journal of Educational and Behavioral Statistics*, 23, 152-169.
- Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*, 413, 321-339.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters*, 49, 1-7.
- Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, 14, 1431-1452.
- Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B*, 54, 221-228.
- 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* (pp. 461-513). New York: Wiley.
- Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association*, 70, 656-661.
- 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.
- Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, 73, 536-544.

- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of Educational Researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, 68(3), 350-386.
- Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A robust approach to hypothesis testing. *Psychophysiology*, 40, 586-596.
- Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, 2, 288-309.
- Liu, R. Y., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92, 266-277.
- Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, 117, 547-560.
- Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, 58, 409-429 (58, 853).
- Luh, W., & Guo, J. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, 52, 303-320.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Othman, A. R., Keselman, H. J., Padmanabhan, A. R., Wilcox, R. R., & Fradette, K. (in press). Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*.
- Othman, A. R., Keselman, H. J., Wilcox, R. R., Padmanabhan, A. R., & Fradette, K. (2002). A description and illustration of a test of symmetry. *Journal of Modern Applied Statistical Methods*, 2, 310-315.
- Rocke, D. M., Downs, G. W., & Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics*, 24(2), 95-101.
- Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers leverage points. *Journal of the American Statistical Association*, 85, 633-639.
- SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6* (1st ed.). Cary, NC: Author.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error probabilities of the t test to departures from population normality. *Psychological Bulletin*, 111, 352-360.
- Welch B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336.
- Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.
- Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology*, 41, 109-117.
- Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, 59, 289-306.
- Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, 65(1), 51-77.
- Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.
- Wilcox, R. R., & Keselman, H. J. (2003). Repeated measures one-way ANOVA based on a modified one-step M-estimator. *British Journal of Mathematical and Statistical Psychology*, 56, 15-25.

A Rank-based Estimation Procedure For Linear Models With Clustered Data

Suzanne R. Dubnicka
Department of Statistics
Kansas State University

A rank method is presented for estimating regression parameters in the linear model when observations are correlated. This correlation is accounted for by including a random effect term in the linear model. A method is proposed that makes few assumptions about the random effect and error distribution. The main goal of this article is to determine the distributions for which this method performs well relative to existing methods.

Key words: R-estimate, random effect, pseudo-sample

Introduction

Consider a situation in which individuals selected for study are not independent of one another. In particular, we consider the situation in which clusters of individuals are observed. These clusters may be families, siblings, littermates, classmates in school, etc. Whatever the origin of the cluster, we consider individuals to be in the same cluster if these individuals are members of a group which, due to this group membership, are more likely to give similar responses than individuals in different groups. Therefore, responses from individuals within a cluster are considered to be correlated while responses from individuals in different clusters are not.

To account for this correlation within clusters, we add a random effect term to the usual linear regression model and consider the following model:

$$Y_i = \alpha \mathbf{1}_{n_i} + X_i \boldsymbol{\beta} + b_i \mathbf{1}_{n_i} + \mathbf{e}_i, \quad i = 1, \dots, m, \quad (1)$$

where Y_i is a $n_i \times 1$ vector of responses for cluster i , X_i is a $n_i \times p$ matrix with j^{th} row, $\mathbf{x}_{ij}^T$, corresponding to the p covariates for observation j in cluster i , is the common unknown intercept, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, b_i is the random effect for cluster i , and $\mathbf{1}_{n_i}$ is a vector of ones of length n_i . We assume that $b_1, \dots, b_m$ are iid continuous random variables, that $e_{11}, \dots, e_{mn_m}$ are iid continuous random variables, and that the b_i and the e_{ij} but these assumptions will depend on the method used for predicting the random effects. These assumptions are discussed in Section 2.1. Thus, there are m clusters with n_i observations within each cluster ($i = 1, \dots, m$), and the total sample size is given by $N = \sum_i n_i$.

Our main interest is to estimate the unknown parameters, and $\boldsymbol{\beta}$. Linear models and generalized linear models with random effects have been studied extensively in a variety of parametric and semiparametric settings in which specific distributions are assumed for the random effects, b_i , and/or the random errors, e_{ij} . For example, Laird and Ware (1982), Ware (1985), Lindstrom and Bates (1988), Schall (1991), Zeger and Rezaul (1991), Waclawiw and Liang (1993), and Chen (2001) all provide methods for fitting such models. In addition, other approaches which also account for correlation within clusters, such as GEE, have

Suzanne R. Dubnicka is an Assistant Professor of Statistics at Kansas State University, 108B Dickens Hall, Manhattan, KS, 66506. Email her at Dubnicka@stat.ksu.edu.

also been developed. For example, see Zeger and Liang (1988) and Lin and Carroll (2001).

In this article, we propose a method for estimating the unknown regression parameters which does not assume a specific distributional form for either b_i or e_{ij} . The proposed method uses rank methods to estimate β and pseudo-samples to predict the random effects b_i . Chen (2001) presents a similar method in which the regression parameters are estimated via rank methods but the random effects are assumed to be normally distributed and are predicted using the best linear unbiased predictor under normality. In using pseudo-samples to estimate the random effects, we do not assume a specific distributional form for these random effects. In addition, unlike Chen, we do not need to estimate the variance of the b_i s or the e_{ij} s with each iteration.

The main purpose of this paper is to evaluate the performance of the proposed method, relative to some existing methods, for a variety of distributions for the random effects and random errors. The more theoretical aspects relating to the proposed method, including asymptotics, are the subject of another paper currently in review (Dubnicka 2004).

The method for estimating β proposed in this paper is an iterative procedure with two major components: the estimation of β given b_i and the prediction of b_i given β . These two components are detailed in Methodology. In Simulations, we evaluate the proposed method and compare it to existing methods via computer simulations. We conclude with a summary of our findings.

Methodology

Consider the model given in (1). We estimate β and b_i using the following iterative steps until the convergence:

1. Estimate β as if the N subjects are independent by solving the usual rank estimating equations given below.
2. Predict the random effects, b_i , using a pseudo-sample approach.
3. Given the estimates of b_i , obtain a rank-based estimate β by solving (13).
4. Repeat steps 2 and 3 until convergence.

Steps 2 and 3 are detailed in the next two sections.

Prediction of the Random Effects

The random effects $b_1, \dots, b_m$ are predicted using pseudo-samples. Since we know only the Y_{ij} , the random effects b_i and the errors e_{ij} are not observable. However, we can use the information in the Y_{ij} to construct a sample of size m that, as $N \rightarrow \infty$, is asymptotically equivalent to the b_i . In particular, we follow the approach of Groggel, Wackerly and Rao (1988) who use pseudo-samples of random effects and random errors to conduct inference on the intraclass correlation in a one-way random effects model. They propose two methods for constructing pseudo-samples: one based on means and another based on medians. We modify their approach in order to predict the random effects in the linear model. The creation of such pseudo-samples requires only a few assumptions regarding the distributions of the random effects, b_i , and the random errors, e_{ij} . The particular assumptions depend on the method used to create the pseudo-samples and are discussed below.

The two methods for creating pseudo-samples proposed by Groggel, Wackerly, and Rao (1988) are the means method and the medians method. With a small adjustment, we can construct a pseudo-sample of the b_i using these methods. Let

$$U_{ij} = Y_{ij} - \mathbf{x}_{ij}^T \beta. \quad (2)$$

Then $U_{ij} = b_i + e_{ij}$ which is in the form of a one-way random effects model considered by Groggel, Wackerly, and Rao (1988).

Pseudo-samples based on means are given by

$$V_{ij} = U_{ij} - \bar{U}_{i\cdot} = e_{ij} - \bar{e}_{i\cdot}. \quad (3)$$

$$W_i = \bar{U}_{i\cdot} - \bar{U}_{..} = b_i + \bar{e}_{i\cdot} - \bar{b} - \bar{e}_{..} \quad (4)$$

where $\bar{U}_{i\cdot} = n_i^{-1} \sum_j U_{ij}$, $\bar{U}_{..} = N^{-1} \sum_i \sum_j U_{ij}$, $\bar{e}_{i\cdot} = n_i^{-1} \sum_j e_{ij}$, $\bar{e}_{..} = N^{-1} \sum_i \sum_j e_{ij}$, and

$\bar{b} = m^{-1} \sum b_i$. If the random effects and the random errors distributions have mean 0, V_{ij} converges in distribution to e_{ij} and W_i converges in distributions to b_i (Dubnicka 2004; Groggel 1983).

Pseudo-samples based on medians are defined in an analogous manner. Let

$$V'_{ij} = U_{ij} - \hat{U}_i = e_{ij} - \hat{e}_i \quad (5)$$

$$W'_i = \hat{U}_i - \hat{U} = b_i - \hat{e}_i - \hat{b} \quad (6)$$

where $\hat{U}_i = \text{med}\{U_{i1}, \dots, U_{in_i}\}$,

$$\hat{U} = \text{med}\{\hat{U}_1, \dots, \hat{U}_m\}, \quad \hat{e}_i = \text{med}\{e_{i1}, \dots, e_{in_i}\}$$

and $\hat{b} = \text{med}\{b_1 + \hat{e}_1, \dots, b_m + \hat{e}_m\}$. If the random effects and the random errors distributions are bounded, V'_{ij} converges in distribution to e_{ij} and V'_{ij} converges in distribution to b_i (Dubnicka 2004; Groggel 1983). Note that for $n_i = 1$ or 2, $V_{ij} = V'_{ij}$.

Therefore, under general conditions, the W_i and W'_i ($i=1, \dots, m$) asymptotically equivalent to the true random effects b_i , $i = 1, \dots, m$. Thus, the W_i and W'_i represent pseudo-samples which predict the random effects b_i , $i = 1, \dots, m$. Throughout the remainder of this paper $\hat{b}_i$ represents the predicted value of b_i based on one of these two methods; that is, $\hat{b}_i = W_i$ or W'_i . Note that in creating pseudo-samples to predict b_i , we can also create pseudo-samples which predict the e_{ij} . These predicted errors are provided by the V_{ij} and the V'_{ij} . However, we do not need these pseudo-samples in our iterative estimation procedure.

Estimation of Regression Parameters

In the proposed iterative procedure, the regression parameters are estimated using rank methods. Rank-based regression requires only very general assumptions on the underlying error distribution. There are several rank-based regression methods from which we can choose.

We present several variations below but focus on the most basic approach.

Consider the linear model

$$Y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i$$

where the e_i are iid random variables. Then the most common rank-based estimate of $\boldsymbol{\beta}$, introduced by Jaeckel (1972), is found by minimizing the dispersion function

$$D^*(\boldsymbol{\beta}) = \sum_{i=1}^n R(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \quad (7)$$

where $R(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})$ is the rank of $Y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ among $Y_1 - \mathbf{x}_1^T \boldsymbol{\beta}, \dots, Y_n - \mathbf{x}_n^T \boldsymbol{\beta}$. Estimates of $\boldsymbol{\beta}$ found by minimizing (7) are called R-estimates.

One generalization of (7) is given by

$$D_a(\boldsymbol{\beta}) = \sum_{i=1}^n a[R(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})] (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \quad (8)$$

where $a(1) \leq a(2) \leq \dots \leq a(n)$ is a set of scores generated by $a(i) = \phi[i/(n+1)]$ for some nondecreasing score function $\phi(u)$ which is defined on $(0,1)$ and satisfies $\int \phi(u) du = 0$ and $\int \phi^2(u) du = 1$. Two commonly used score functions are Wilcoxon scores and sign scores given by $\phi_w(u) = \sqrt{12}(u - 1/2)$ and $\phi_s(u) = \text{sgn}(u - 1/2)$, respectively.

Using Wilcoxon scores produces a dispersion function which is equivalent to (7) and which will produce the usual R-estimate for $\boldsymbol{\beta}$. Sign scores will produce the L_1 estimate of $\boldsymbol{\beta}$. Other score functions which are optimal for specific error distributions have also been proposed. In addition, there are score functions which may be more appropriate for asymmetric errors (Hettmansperger and McKean 1998).

Note that minimizing $D^*(\boldsymbol{\beta})$ is equivalent to minimizing $D(\boldsymbol{\beta}) = \sum_{i < j} [(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - (Y_j - \mathbf{x}_j^T \boldsymbol{\beta})]$. (9)

That is, minimizing $D(\boldsymbol{\beta})$ will also provide the R-estimate for $\boldsymbol{\beta}$. A related approach, introduced by Sievers (1983) and further developed by Naranjo and Hettmansperger (1994), estimates $\boldsymbol{\beta}$ by minimizing

$$D_b(\boldsymbol{\beta}) = \sum_{i < j} \sum \left| (Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) - (Y_j - \mathbf{x}_j^T \boldsymbol{\beta}) \right| \quad (10)$$

where $b_{ij} = b(\mathbf{x}_i, \mathbf{x}_j)$. Properly chosen weights b_{ij} will produce estimates of $\boldsymbol{\beta}$ with a bounded influence function and high breakdown. The bounded-influence estimate, however, tends to be less efficient than the usual R-estimate. Estimates produced by minimizing $D_b(\boldsymbol{\beta})$ are called generalized rank estimates, or GR-estimates.

The proposed iterative procedure can be performed using R-estimation, general score R-estimation, or GR-estimation. In practice, one would carefully evaluate the particular application to determine which is most appropriate. For the remainder of this paper, however, we will use the more common R-estimates of $\boldsymbol{\beta}$.

Return now to our model (1) which includes the random effect. Let $Y_{ij}^* = Y_{ij} - b_i$. Then, given the random effects, we can estimate the regression parameters using the usual rank estimating equations where the Y_{ij} s are replaced by Y_{ij}^* s.

To simplify notation, let $\{Y_1^*, Y_2^*, \dots, Y_N^*\}$ represent $\{Y_{11}^*, \dots, Y_{1n_1}^*, Y_{21}^*, \dots, Y_{2n_2}^*, \dots, Y_{m1}^*, \dots, Y_{mn_m}^*\}$. The vectors of covariates corresponding to these responses can be written in an analogous manner. Then $\hat{\boldsymbol{\beta}}_R$ is the estimators of $\boldsymbol{\beta}$ which minimizes

$$D(\boldsymbol{\beta}|\mathbf{b}) = \sum_{l < k} \sum \left| (Y_l^* - \mathbf{x}_l^T \boldsymbol{\beta}) - (Y_k^* - \mathbf{x}_k^T \boldsymbol{\beta}) \right|. \quad (11)$$

The gradient of $D(\boldsymbol{\beta}|\mathbf{b})$ is given by

$$\begin{aligned} S(\boldsymbol{\beta}|\mathbf{b}) &= -\nabla D(\boldsymbol{\beta}|\mathbf{b}) \\ &= \sum_{l < k} \sum (\mathbf{x}_l - \mathbf{x}_k) \operatorname{sgn} \\ &\quad \left[(Y_l^* - \mathbf{x}_l^T \boldsymbol{\beta}) - (Y_k^* - \mathbf{x}_k^T \boldsymbol{\beta}) \right] \end{aligned} \quad (12)$$

As $D(\boldsymbol{\beta}|\mathbf{b})$ is a piecewise linear, continuous, convex function, minimizing $D(\boldsymbol{\beta}|\mathbf{b})$ is equivalent to solving

$$S(\boldsymbol{\beta}|\mathbf{b}) \doteq \mathbf{0}. \quad (13)$$

Note that it is unlikely $S(\boldsymbol{\beta}|\mathbf{b})$ will equal $\mathbf{0}$ for any value of $\boldsymbol{\beta}$. In the case of one covariate, $S(\boldsymbol{\beta}|\mathbf{b})$ is a nondecreasing step function of $\boldsymbol{\beta}$ which steps down at each sample slope. There may be an interval of solutions $S(\boldsymbol{\beta}|\mathbf{b}) = 0$ or $S(\boldsymbol{\beta}|\mathbf{b})$ may “step across” the horizontal axis. We let $\hat{\boldsymbol{\beta}}_R$ denote this rank estimate of $\boldsymbol{\beta}$ in either case.

Once an estimate for $\boldsymbol{\beta}$ has been obtained, α can be estimated by solving

$$S_1(\alpha|\boldsymbol{\beta}, \mathbf{b}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \operatorname{sgn}(Y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_R - \hat{b}_i) \doteq 0 \quad (14)$$

where $\hat{\boldsymbol{\beta}}_R$ is the estimate of $\boldsymbol{\beta}$ obtained from solving (13) and $\hat{b}_i$ is the predicted value of b_i using one of the pseudo-sample methods of the previous section. The solution to equation (14) is simply the median of the residuals $Y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_R - \hat{b}_i$. That is,

$$\hat{\alpha}_R = \operatorname{median} \{Y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_R - \hat{b}_i\} \quad (15)$$

Simulations

Simulations were conducted to evaluate the performance of our proposed method. These simulations were performed with the intent of answering two questions:

1. How large must m and the n_i be to produce “good” estimates of α and $\boldsymbol{\beta}$?

2. How does this method perform as the random effects distribution and the random error distribution vary?

Recall that the pseudo-samples are only asymptotically equivalent to the true random effects. For small samples, there is some concern that this method will not produce estimates of β and α which are reasonably on target. In this first simulation study, we focus on the cluster sizes and number of clusters rather than the distributions of the random effects and random errors. Therefore, with m clusters of n subjects per cluster, a single covariate $x \sim \text{lognormal}(2, 0.5^2)$, and $(\alpha, \beta) = (2, 2)$, 1000 samples were generated in which the random effects and the random errors were both

normally distributed: $b_i \sim N(0, 0.5^2)$ and $e_{ij} \sim N(0, 0.4^2)$. Note that, for simplicity, we have chosen all of the clusters sizes to be the same ($n_1 = \dots = n_m$).

For comparison, estimates of α and β were also obtained using maximum likelihood (ML) and restricted maximum likelihood (REML) since these methods are included in most existing statistical software.

Results of this first simulation study appear in Table 1. For each method, the means of the 1000 estimates are given with the standard deviations of the 1000 estimates below the estimates in parentheses.

Table 1: Parameter Estimates for Various m and n , $b_i \sim N(0, 0.5^2)$, $e_{ij} \sim N(0, 0.4^2)$

	Mean Method	Median Method	ML	REML
(m, n)	$\hat{\alpha}, \hat{\beta}$	$\hat{\alpha}, \hat{\beta}$	$\hat{\alpha}, \hat{\beta}$	$\hat{\alpha}, \hat{\beta}$
	(StDev)	(StDev)	(StDev)	(StDev)
(5,2)	2.00902, 1.99751 (0.53457, 0.0625)	2.015, 1.99747 (0.54577, 0.06239)	2.00592, 1.99819 (0.48297, 0.05278)	2.003934, 1.99783 (0.47995, 0.05272)
(5,5)	1.99484, 2.0008 (0.29678, 0.02253)	2.00185, 2.00077 (0.31546, 0.02366)	1.99704, 2.00077 (0.28749, 0.02159)	1.99697, 2.0078 (0.28703, 0.02155)
(5,8)	2.00099, 2.00031 (0.27815, 0.01747)	2.00178, 2.00029 (0.28995, 0.01755)	2.00078, 2.00019 (0.27389, 0.01702)	2.00079, 2.00019 (0.27386, 0.01699)
(15,2)	1.98592, 2.00158 (0.26826, 0.0278)	1.98717, 2.00158 (0.28243, 0.0278)	1.98907, 2.00135 (0.24222, 0.0239)	1.98894, 2.00136 (0.24184, 0.02386)
(15,5)	1.99979, 2.00004 (0.17499, 0.01245)	2.00222, 2.00002 (0.18033, 0.01283)	1.99953, 2.00008 (0.16884, 0.01215)	1.99954, 2.00008 (0.16884, 0.01215)
(15,8)	2.00682, 1.99933 (0.159, 0.00919)	2.0056, 1.99942 (0.16353, 0.00925)	2.00641, 1.99942 (0.15728, 0.00901)	2.00643, 1.99942 (0.15724, 0.009)
(30,2)	1.99458, 2.00071 (0.1843, 0.01814)	1.99312, 2.00071 (0.19196, 0.01814)	1.99318, 2.00088 (0.16987, 0.01601)	1.9932, 2.00088 (0.16981, 0.016)
(30,5)	1.99352, 2.00001 (0.11875, 0.00846)	1.99409, 2.00012 (0.12621, 0.00875)	1.99282, 2.0007 (0.1152, 0.00813)	1.99284, 2.00007 (0.11519, 0.00813)
(30,8)	2.00049, 2.00001 (0.11341, 0.00648)	2.0003, 2 (0.11765, 0.00661)	2.00023, 2.00005 (0.11055, 0.00628)	2.00023, 2.00005 (0.1055, 0.00628)

Note that the estimates of α and β obtained from the proposed iterative method using either mean or median pseudo-samples seem to be reasonably unbiased even for small m and n ; see Table 1. As one would expect, when both the random effects and the random errors are normally distributed the standard deviations of the estimates obtained through maximum likelihood and REML are smaller. However, the standard deviations of the estimates obtained through the proposed iterative method are not much larger.

Although the proposed method, using mean or median pseudo-samples, provides estimates which are reasonably on target for small m and n , the procedure failed to converge for some samples regardless of the pseudo-sample method used. Table 2 shows the percentage of times that the mean method and the median method converge for each of the combinations of m and n in the first simulation study. Upon closer investigation, we found that for some of the samples the estimates of α continued to increase (or decrease) as more iterations were completed. For some of the samples, however, the estimates of α seemed to “bounce” between two values. This happened more frequently when both m and n were small

Table 2: Convergence Percentage for Various m and n , $b_i \sim N(0, 0.5^2)$, $e_{ij} \sim N(0, 0.4^2)$

(m, n)	Mean Method	Median Method
(5, 2)	99.1%	99.1%
(5, 5)	96.7%	99.8%
(5, 8)	99.0%	93.0%
(15, 2)	99.7%	99.7%
(15, 5)	98.2%	100%
(15, 8)	99.1%	96.5%
(30, 2)	99.8%	99.8%
(30, 5)	98.8%	100%
(30, 8)	99.7%	98.4%

The remaining simulations were designed to help answer the second question. That is, we wanted to determine the distributions under which the proposed method is superior to the existing methods considered. In these simulations, a variety of distributions for both the random effects and the random error were

used. Table 3 gives the abbreviations for the particular distributions chosen for these simulations.

The simulations conducted are divided into three cases: (1) the random effects distribution is normal and the error distribution varies, (2) the error distribution is normal and the random effects distribution varies, and (3) both distributions are nonnormal but from the same family of distributions. As with the first simulation, 1000 random samples were generated with a single covariate $x \sim \text{lognormal}(2, 0.5^2)$ and $(\alpha, \beta) = (2, 2)$. Furthermore, each sample consists of $m = 50$ clusters of $n = 3$ subjects per cluster. For comparison, estimates of α and β were also obtained using Chen’s method and restricted maximum likelihood (REML). Recall that Chen’s method differs from the proposed method in that the random effect is assumed to be normally distributed. Maximum likelihood estimates were also computed but there were almost identical to the REML estimates.

Table 4 shows the simulation results for normally distributed random effects. Since Chen’s method assumes normality for the b_i but does not assume a specific distribution for the e_{ij} , one would expect Chen’s method to perform better than the proposed methods and REML. To some extent, the simulations support this theory. When the errors follow a contaminated normal or double exponential distribution, the standard deviations of the β estimates using Chen’s method are smaller than those of the other methods. When the errors follow a Cauchy distribution, the standard deviation of the β estimates based on the proposed method with median pseudo-samples is smaller than that of the other approaches. Notice, however, that the standard deviation of the α estimates is smaller for the median method than the other methods. In particular, the estimates of α using the mean method and REML were highly variable in the case of Cauchy errors.

Table 3: Distributions used in Simulations

Abbreviation	Name of Distribution	Description
CN1	Contaminated Normal	$0.9 N(0, 0.4^2) + 0.1 N(0, 1.2^2)$
CN2	Contaminated Normal	$0.9 N(0, 0.3^2) + 0.1 N(0, 0.9^2)$
DE1	Double Exponential	$\frac{1}{2} \lambda \exp(-\lambda x), \lambda = 2.5$
DE2	Double Exponential	$\frac{1}{2} \lambda \exp(-\lambda x), \lambda = 3$
C1	Cauchy	$0.16 \text{ Cauchy}(0.1)$
C2	Cauchy	$0.12 \text{ Cauchy}(0, 1)$
U1	Uniform	$\text{Uniform}(-1.2, 1.2)$
U2	Uniform	$\text{Uniform}(-0.9, 0.9)$

Table 4: Parameter Estimates for $m = 50, n = 3, b_i \sim N(0, 0.3^2)$

	Mean Method	Median Method	Chen	REML
Error Distribution	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)
CN1	1.99406, 2.00087 (0.11152, 0.01085)	1.9978, 2.00054 (0.10595, 0.01008)	1.99548, 2.00074 (0.12488, 0.00974)	1.99809, 2.00047 (0.10788, 0.01052)
DE1	1.99851, 2.00024 (0.11973, 0.01232)	1.99718, 2.00015 (0.11261, 0.01101)	1.99368, 2.00024 (0.12384, 0.01081)	1.99763, 2.00033 (0.11517, 0.01183)
C1	2.12036, 2.00034 (4.5088, 0.01205)	1.99724, 2.00004 (0.09558, 0.00919)	1.99743, 1.99999 (0.10268, 0.00988)	2.1836, 1.99199 (7.35446, 0.68516)
U1	2.00417, 1.99981 (0.16375, 0.01667)	2.00281, 1.99973 (0.15796, 0.0147)	2.0028, 1.99971 (0.18557, 0.01459)	2.0028, 1.99974 (0.13293, 0.01367)

Table 5: Convergence Percentage $m = 50, n = 3, b_i \sim N(0, 0.3^2)$

Error Distribution	Mean Method	Median Method	Chen
CN1	98.5%	100%	96.2%
DE1	98.9%	100%	98.8%
C1	99.2%	100%	66.3%
U1	97.8%	100%	92.1%

The major disadvantage of Chen's method is this situation is that it does not always converge. This is also true, to a lesser extent, for the proposed method with mean pseudo-samples. Table 5 gives the convergence percentage of the three rank-based methods for the simulations in Table 4. Notice that the median method always converged. The mean method and Chen's method converged most of the time when the error distribution was a contaminated normal, double-exponential, or uniform. The mean method also converged most of the time when the error distribution is Cauchy but Chen's method had difficulty converging in this case. Chen (2001) also notes this problem. The main source of the problem is that the Chen's method requires the estimation of the error variance (and the random effect variance) at each iteration, and convergence of the algorithm depends on the convergence of the error variance. In distributions for which the variance is undefined, convergence problems will exist for Chen's method.

Table 6 gives the results for cases in which the errors are normally distributed but the distribution of the random effects is non-normal. In addition, Table 7 gives the convergence percentages of the rank-based methods for these simulations. For the four situations considered, the standard deviations of the REML estimates of β are the smallest. This seems to imply that REML is a relatively efficient method for estimating β even when the random effect distribution is non-normal. Notice that the standard deviations of the median method β estimates are the largest of the four methods but they are not much larger than the REML standard deviations. Also, REML estimates of α also tend to be more precise (smallest standard deviation of the α estimates) except when the random effects distribution is Cauchy.

When the random effects follow a Cauchy distribution the median method provides the most precise estimate of α .

As in the previous simulations, Chen's method did not converge for all samples. In fact, when the error distribution was normal and the random effects distribution was Cauchy, Chen's method only converged half of the time. Notice again that the proposed method with median pseudo-samples always converged, and the proposed method with mean pseudo-samples converged most of the time.

Finally, we consider situations in which neither the error distribution nor the random effects distribution are non-normal. For each situation, the error and random effects distributions are from the same family of distributions. The results appear in Tables 8 and 9. In these situations, there is no clear "winner" with respect to the estimation of β . Under the contaminated normal distributions and double exponential distributions, Chen's β estimates have the smallest standard deviations. Under Cauchy distributions and uniform distributions, REML estimates of β are less variables. However, the proposed method with median pseudo-samples provided the most precise estimates of α under the distributions considered. Again note that the median method converged for all samples while the mean method converged most of the time and Chen's method converged most of the time except under the Cauchy distributions. Under Cauchy distributions, Chen's method only converged half of the time.

Conclusion

The paper introduced a new rank-based method for parameter estimation in linear model with a random effect term. Such a model is useful in accounting for the correlation between subjects that are correlated, as is the case when clusters of subjects are observed. The proposed method uses rank-based regression to estimate the parameters of the linear model and pseudo-samples to predict the random effects. As a result the proposed method requires few assumptions regarding the underlying distributions of the errors and the random effects.

Table 6: Parameter Estimates for $m = 50$, $n = 3$, $e_{ij} \sim N(0, 0.4^2)$

	Mean Method	Median Method	Chen	REML
Error Distribution	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)
<i>CN2</i>	2.00057,2.00014 (0.10673,0.00946)	1.99799,2.00033 (0.11103,0.00984)	2.00262,2.00023 (0.22225,0.00911)	1.99829,2.00035 (0.10245,0.00883)
<i>DE2</i>	1.99463,2.00048 (0.10983,0.00934)	1.99765,2.00025 (0.11096,0.01001)	2.00305,2.00039 (0.28775,0.00898)	1.99458,2.00037 (0.10303,0.00859)
<i>C2</i>	2.0442,2.00018 (4.77186,0.00953)	1.99758,1.99998 (0.10331,0.00993)	1.99871,2.00012 (0.83112,0.00947)	2.04528,1.99997 (4.77105,0.00919)
<i>U2</i>	2.00176,1.99979 (0.10811,0.00911)	1.99978,1.99998 (0.13954,0.01103)	1.98411,1.99978 (0.32091,0.00884)	2.00285,1.99982 (0.10519,0.00871)

Table 7: Convergence Percentage $m = 50$, $n = 3$, $b_i \sim N(0, 0.4^2)$

Error Distribution	Mean Method	Median Method	Chen
<i>CN2</i>	98.3%	100%	89.8%
<i>DE2</i>	98.6%	100%	87.4%
<i>C2</i>	99.3%	100%	47.9%
<i>U2</i>	98.7%	100%	83.8%

Table 8: Parameter Estimates for $m = 50$, $n = 3$

	Mean Method	Median Method	Chen	REML
Distributions	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)	$\hat{\alpha}, \hat{\beta}$ (StDev)
$b \sim CN2, e \sim CN1$	2.00216,2.00001 (0.11928, 0.01108)	2.0031, 1.99981 (0.11469,0.01081)	2.00374, 1.99999 (0.17199,0.01014)	2.00268,1.99983 (0.1196,0.01118)
$b \sim DE2, e \sim DE1$	1.99352,2.00026 (0.13154,0.01244)	1.9962,2.0001 (0.1285,0.01238)	1.99604,2.00036 (0.2035,0.0116)	1.99336,2.0002 (0.13173,0.01256)
$b \sim C2, e \sim C1$	1.07711,1.99969 (27.68038,0.00957)	2.00552,1.9996 (0.09885,0.00989)	1.99952,1.99971 (0.82259,0.00941)	1.07733,1.99966 (27.68138,0.00913)
$b \sim U2, e \sim U1$	2.01073,1.99879 (0.17553,0.01733)	2.00585,1.99932 (0.17511,0.01654)	2.02594,1.99903 (0.33254,0.01616)	2.00635,1.99907 (0.15331,0.01504)

Table 9: Convergence Percentage $m = 50$, $n = 3$, $b_i \sim N(0, 0.4^2)$

Distributions	Mean Method	Median Method	Chen
$b \sim CN2, e \sim CN1$	98.6%	100%	98.0%
$b \sim DE2, e \sim DE1$	98.3%	100%	98.3%
$b \sim C2, e \sim C1$	99.0%	100%	48.5%
$b \sim U2, e \sim U1$	98.4%	100%	96.6%

Results from the simulation studies showed that REML often provided estimates β which were less variable than those of other methods. If the goal of a study is to see how the response changes as the predictors change, then REML might provide the best means for assessing this. However, if the goal is to predict a response for certain values of the predictors, REML may provide inaccurate predictions under some distributions since the REML estimate of α can be highly variable. The three rank-based methods considered (mean pseudo-samples, median pseudo-samples, and Chen) all produce estimates of β with comparable precision to REML. Only the median method seems to provide consistently precise estimates of α under all distributions considered. In general, the proposed method with median pseudo-samples is robust to the underlying distribution of the random effects and errors as it is relatively efficient for all distributions considered. Therefore, if prediction of the goal of study, the proposed method with median pseudo-samples is recommended.

As a final note, it may be possible for the proposed method to perform better than REML with properly chosen scores in (8), but this has not yet been explored.

References

- Chen, C. C. (2001). Rank estimating equations for random effects models. *Statistics and Probability Letters*, 54, 5-12.
- Dubnicka, S. R. (2004). Rank-based Regression with Clustered Data, in review.
- Groggel, D. J. (1983). Asymptotic Nonparametric Confidence Intervals for the Ratio of Scale Parameters in Balanced One-Way Random Effects Models. Ph.D. thesis, University of Florida.
- Groggel, D. J., Wackerly, D., & Rao, P. (1988). Nonparametric estimation in one-way random effects models. *Communications in Statistics - Simulation and Computation* 17, 887-903.
- Hettmansperger, T. P., & McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. Arnold.
- Jaekel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics*, 43, 1449-1458.
- Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 962-974.
- Lin, X., & Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1045-1056.
- Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures. *Journal of the American Statistical Association*, 83, 1014-1022.
- Naranjo, J. D., & Hettmansperger, T. P. (1994). Bounded influence rank regression. *Journal of the Royal Statistical Society, Series B*, 56, 209-220.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719-727.
- Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Communications in Statistics - Theory and Methods*, 12, 1161-1179.
- Waclawiw, M. A., & Liang, K. Y. (1993). Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association*, 88, 171-178.
- Ware, J. H. (1985). Linear models for the analysis of longitudinal data. *The American Statistician*, 39, 95-101.
- Zeger, S. L., & Liang, K. Y. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.
- Zeger, S. L., & Rezaul, K. M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79-86.

Depth Based Permutation Test For General Differences In Two Multivariate Populations

Yonghong Gao

Center for Devices and Radiological Health
Food and Drug Administration

For two p-dimensional data sets, interest exists in testing if they come from the common population distribution. Proposed is a practical, effective and easy to implement procedure for the testing problem. The proposed procedure is a permutation test based on the concept of the depth of one observation relative to some population distribution. The proposed test is demonstrated to be consistent. A small Monte Carlo simulation was conducted to evaluate the power of the proposed test. The proposed test is applied to some numerical examples.

Key words: Depth, Monte Carlo, permutation test, spatial rank

---

Introduction

Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ be independent random samples from continuous p-dimensional populations with cumulative distribution functions $F(x)$ and $G(y)$ respectively. The in question interest is in assessing whether there are any differences whatsoever between the X and Y probability distributions. Thus, the null hypothesis is tested (1.1) against the most general alternative possible (1.2):

$$H_0: F(t) = G(t), \text{ for any } t, \quad (1.1)$$

$$H_1: F(t) \neq G(t), \text{ for at least one } t. \quad (1.2)$$

In the univariate case, a popular statistic is the two-sided two-sample Kolmogorov-Smirnov statistic KS, which is

$$KS = (m n / d) \sup_x \{ |F_m(x) - G_n(x)| \} \quad (1.3)$$

Dr. Yonghong Gao may be reached at the Center for Devices and Radiological Health, Food and Drug Administration, 1350 Piccard Drive, Rockville, MD 20850 U.S.A. Telephone: (301)827-0953, fax: (301)443-8559. Email: yhg@cdrh.fda.gov

where d is the greatest common divisor of m and n, $F_m(x)$ and $G_n(x)$ are the empirical distribution functions for the X and Y samples, respectively. Under the null hypothesis, KS is expected to be small, so the null hypothesis is rejected if $KS > J_\alpha$, where the constant J_α is chosen to make the type I error probability equal to α . When sample sizes are small, values of J_α are given in tables, when sample sizes are large, where $\min\{m, n\} \rightarrow \infty$, Smirnov (1939) derived the asymptotic distribution of the standardized KS and the limiting distribution of KS is quite complex.

Another popular approach to the univariate testing problem is the density-based approach, where the two population density functions are estimated using kernel or spline estimation methods and then the test is defined as the distance (maximum distance or mean distance) between the two estimated density functions. Bowman (1985) uses the L_2 distance and Allen (1997) uses the L_1 distance. Allen (1997) conducts a comprehensive simulation study to compare the power of the KS-test, L_2 distance density test, L_1 density test and t-type permutation test, the simulation results show that there is no uniformly superior test.

In multivariate setting, two special cases of the testing problem (1.1) have been studied by many investigators. The first case (more extensively studied case) is the two-sample location problems:

$$H_0: \mu=0, \text{ where } G(x)=F(x-\mu). \quad (1.4)$$

The Hotelling's T^2 -test is the usual normal theory test for this problem, it is well-known that the Hotelling's T^2 is the best when distribution is multivariate normal. To free the constraint of normality and to gain the benefit of robustness, many sign-based and rank-based nonparameter tests are proposed using the multivariate versions of the Mood median test and Mann-Whitney test, see Marden's (1999) excellent review paper on this topic.

The second case is the testing of homogeneity of covariances problems:

$$H_0: \text{Var}(X) = \text{Var}(y). \quad (1.5)$$

The Box's M-test is the likelihood ratio test for this problem under multivariate normal distributions.

For the general testing problem (1.1), there is not much activity in existing literature. To develop the multivariate analog of Kolmogorov-Smirnov test, the first challenge faced is to define the empirical distribution based on multivariate data, and that challenge has not been met satisfactorily. Marden (1999) notices the association of $F(x)$ and $R(x, F)$ in univariate case: $R(x, F)=2F(x)-1$, where $R(x, F)$ is the rank of x relative to distribution F : $R(x, F)=E(\text{Sign}(x-X))$, with $X \sim F$. Hence Marden (1999) suggests we could use KS_R ,

$$KS_R = \sup_x \{ |R_m(x, F) - R_n(x, G)| \} \quad (1.6)$$

where $R_m(x, F)$ is the multivariate spatial rank of x relative to sample $\{X_i\}$, so far no research activity in investigating the performance of KS_R has been reported yet.

In this article a KS-test is examined from another aspect. The key idea of Kolmogorov-Smirnov's test is to compare the two distribution functions $F(x)$ and $G(x)$. Noticed was that the distribution function $F(x)$ is some sort of measure of the position of x relative to distribution F , for example, if $F(x)$ is close to .5, then x is in the close neighbor of the center of distribution F , if $F(x)$ is close to 0 or 1, then x is on the outskirts of distribution F , which leads to the idea of the depth of one observation relative to a distribution. It is believed that the depth

function $D(x, F)$ of one observation x relative to some distribution F is some continuous function of $F(x)$: $D(x, F)=g(F(x))$. For example, in univariate setting, the rank-based depth $D_r(x, F)$ and the simplex's depth $D_s(x, F)$ are concave functions of $F(x)$:

$$\begin{aligned} D_r(x, F) &= 4 F(x) (1-F(x)), \\ D_s(x, F) &= 2 F(x) (1-F(x)). \end{aligned} \quad (1.7)$$

Unfortunately in higher dimensions there does not exist a similar explicit formula supporting the conjecture that $D(x, F)$ is some continuous function of $F(x)$.

Given the association of $D(x, F)$ and $F(x)$, we use the difference of $D(x, F)$ and $D(x, G)$ to measure the difference of $F(x)$ and $G(x)$. While the depth function and the corresponding empirical version are well defined in multivariate settings.

Methodology

Statistical depth functions have been used to measure the centrality of a multivariate data point with respect to a given data cloud, a center is usually given by a point of maximal depth. This center-outward ordering of the multivariate data provides a foundation for new nonparametric methods in multivariate estimation and inference.

For recent results of different versions of depth function and their applications, see Liu (1990), Liu and Singh (1993), Yeh and Singh (1997) and Zuo, Cui and He (2003). The depth functions usually seen in literature are Tukey's depth proposed by Tukey (1975), simplex depth introduced by Liu (1990), projection depth and Mahalanobis depth. They are all affine invariant and show great potential in multivariate analysis. Mahalanobis's depth is the simplest but least popular one, mainly because it is not robust. Projection depth, Tukey's depth and simplex depth can be quite robust, but the common disadvantage of these three depth functions is that the calculations of these depth functions are quite computationally intensive, especially in high dimensions. Gao (2003) proposes a robust yet easy to calculate depth function based on spatial ranks. In this paper we use this notion of the depth.

For a point x in R^p and a p -variate distribution F , the spatial rank of x relative to F is defined as

$$R(x, F) = E(\text{Sign}(x - Y)), Y \sim F, \quad (2.8)$$

where $\text{Sign}(x)$ is a unit vector in the same direction of x . Then the depth of point x relative to distribution F is

$$D(x, F) = 1 - \|R(x, F)\|^2 \quad (2.9)$$

The sample version of $R(x, F)$ and $D(x, F)$ based on iid sample $X_1, \dots, X_n$ are

$$R_n(x, F) = (\sum \text{Sign}(x - X_i)) / n \quad (2.10)$$

$$D_n(x, F) = 1 - \|R_n(x, F)\|^2 \quad (2.11)$$

Under the null hypothesis (1.1), $D(x, F) = D(x, G)$ for any x , so the proposed test statistic is

$$T(m, n) = \sup_x \{|D_m(x, F) - D_n(x, G)|\} \quad (2.12)$$

and the null hypothesis is rejected when $T(m, n) > t_\alpha$, where t_α is chosen to make the type I error probability equal to α .

Proposition 1

Under the null hypothesis (1.1), when $\min\{m, n\} \rightarrow \infty$, $T(m, n) \rightarrow 0$. The proof of above proposition is based on the following result presented in Gao (2003) about the rank based depth,

$\lim_{n \rightarrow \infty} \sup_x \{|D(x, F) - D_n(x, F)|\} = 0$, for any x, F . Note that test $T(m, n)$ and test KS_R are closely related and produces the following result:

$$T(m, n) \leq 2 KS_R.$$

It is not easy to get the distribution (exact or asymptotic) of $T(m, n)$ under the null hypothesis, bootstrap and permutation resampling methods provide the attractive alternative approaches to determine a critical point for the test. Permutation approach usually shows slightly higher power than the bootstrap approach, hence we use permutation in this paper. The procedure is implemented as the following.

The original two samples are pooled into one large sample $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$. Two resampled data sets are drawn without replacement from the pooled forming the permuted samples $\{X_1^*, \dots, X_m^*\}$ and $\{Y_1^*, \dots, Y_n^*\}$. Each pair of resampled datasets gives a permuted value of the statistic $T^*(m, n)$. We repeat this process B times, yielding B permuted values of $T^*(m, n)$, for a specified level of significance α , the hypothesis (1.1) is rejected if $\#\{T^*(m, n) \geq T(m, n)\} + 1 \leq (B+1)\alpha$.

Example 1: Iris data

The Iris dataset was introduced by R. A. Fisher as an example for discriminate analysis. The data report four characteristics (sepal width, sepal length, petal width and petal length) of three species of Iris flower: Setosa, Versicolor and Virginica. From the scatter plot of the any two variables it can be seen that Setosa is quite different from the other two species. The proposed test is applied, $T(m, n)$, Marden's rank-based test KS_R , Box's M -test T_M and the Hotelling T^2 test on the three pairs of dataset: (Setosa and Versicolor), (Versicolor and Virginica), and (Setosa and Virginica). The values of the test statistics and the p -values (the values within the parenthesis) are shown in table 1. From the table we can see that the three species are all significantly different from each other using any of the three tests.

Table 1. Analysis of Iris Data.

Test	Setosa and Versicolor	Versicolor and Virginica	Setosa and Virginica
$T(m, n)$	.9756 (0)	.9885 (0)	.8843 (0)
KS_R	1.8807 (0)	1.942 (0)	1.372 (0)
T_M	71.302 (0)	116.648 (0)	37.392(0)
T^2	2580.8 (0)	4879.6 (0)	355.4(0)

Example 2: Hotdogs

The Hotdogs (1989) data file contains data on the sodium and calories contained in each of 54 major hot dog brands. The hotdogs are classified by type: beef, poultry, and meat

(mostly pork and beef, but up to 15% poultry meat), the two variables are Sodium (Milligrams of sodium per hot dog) and Calories (Calories per hot dog). Corresponding to three different type of hot dog produces three data sets, the proposed test is used to determine if these three datasets have the same distribution in terms of the two variables being considered. The analysis result is shown in Table 2.

It is shown in Table 2 that the four tests agree on the following conclusions: there is no significant evidence to say that the beef hotdogs and the meat hotdogs are different, but the beef hotdogs and the poultry hotdogs are significantly different. For meat hotdogs and poultry hotdogs, there is some disagreement among the four tests, both depth test and rank test show some but not that strong evidence to say that these two types of hotdogs are different, while Hotelling's T^2 -test and Box's M-test show significant evidence of difference. To explain this disagreement, the data is further analyzed. One outlier is found (with extreme low sodium value) for the Meat-type hotdogs, because of that one observation, the poultry hotdogs look more like part of the meat hotdogs family (the range of meat hotdogs covers the range of poultry hotdogs). The outlier is deleted and compared with the poultry hotdogs again. The result is in Table 2, where MeatN means the new meat hotdogs data set. Then the four test procedures give us the same conclusion that the meat hotdogs and poultry hotdogs are different.

From this example it is seen that the depth-based permutation test is not powerful when the range of one data set covers the range of another data set, and we should always check the data first, clean the data if possible before implementing any formal testing procedure.

Results

Two simulation experiments were conducted studying the empirical power of the proposed test. The first experiment investigates the sensitivity of the test to the mean effect, the second investigates the sensitivity of the tail mass effect (characterized by variance matrix). For comparison purpose we estimate powers of the Hotelling's T^2 -test, Box's M-test T_M and Marden's KS_R test as well in the conducted experiments. For every trial, two samples are generated, one from distribution F and one from G, the hypothesis (1.1) is tested independently using each of the four testing statistics mentioned above. The level of significance is 5%, the bootstrap size B is 199, the sample size is $m=n=30$, and dimension is $p=2$. The trial was repeated 1000 times for each case (corresponding to different pairs of (F, G)), the empirical power (the number of times the null hypothesis was rejected divided by 1000) is recorded for each test and the results are summarized in Table 3 and Table 4.

Let $N_2(\mu, \sigma^2 I_2)$ denote the bivariate normal distribution with mean vector as μ and covariance matrix as $\sigma^2 I_2$. For experiment 1, use $F = N_2((0,0), I_2)$, $G = N_2((a,a), I_2)$, with $a=0, .2, .4, .6$ and $.8$. For experiment 2, use $F = N_2((0,0), I_2)$, $G = N_2((0,0), bI_2)$, with $b=1, 1.2, 1.4, 1.6$ and 1.8 . When the case is the location problem in multivariate normal distribution (corresponding to experiment 1), the Hotelling's T^2 has the highest power as it should be, the permutation test $T(m,n)$ has power as much as 80% of the Hotelling's T^2 test, the Box's M-test has no power in this case since it is location invariant, Marden's KS_R test has some power but lower than $T(m,n)$ test.

Table 2. Analysis of Hotdogs Data.

Test	Beef vs. Meat	Beef vs. Poultry	Meat vs. Poultry	MeatN vs. Poultry
$T(m,n)$	.2208 (.71)	.7712 (.005)	.2183 (.1)	.6301 (0)
KS_R	.6260 (.73)	.9382 (.006)	.8208 (.13)	.8976 (0)
T_M	1.696 (.73)	5.011 (.003)	2.454 (0)	5.411 (0)
T^2	.506 (.78)	119.1 (0)	87.96 (0)	81.82 (0)

Table 3. Simulation Study 1: Study of the sensitivity to the mean effect.

Test	a=0	a=.2	a=.4	a=.6	a=.8
T(m,n)	.047	.148	.287	.509	.781
KS _R	.051	.121	.145	.241	.422
T _M	.052	.048	.049	.049	.051
T ²	.051	.149	.493	.764	.983

Table 4. Simulation Study 2: Study of the sensitivity to the tail mass effect.

Test	b=1	b=1.2	b=1.4	b=1.6	b=1.8
T(m,n)	.047	.089	.101	.149	.356
KS _R	.051	.069	.06	.09	.067
T _M	.052	.099	.11	.198	.511
T ²	.051	.069	.06	.054	.066

For the case of the homogeneity of covariance matrices (corresponding to experiment 2), the Box's M-test has the highest power, the proposed test T(m,n) is the second best the Hotelling's T² test and Marden's KS_R test have no power. From this small simulation study it is determined that the proposed test is competitive at least in those two cases and further research is needed to investigate its properties under other situations.

References

- Anderson, A., Hall, P., & Titterton, D. W. (1994). Two-Sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50, 41-54.
- Allen, D. (1997). Hypothesis testing using an [trial mode]-distance bootstrap. *The American Statistical*, 51, 145-150.
- Bowman, A. W. (1985). A comparative study of some kernel-based nonparametric density estimators. *Journal of Statistical Computation and Simulation*, 21, 313-327.
- Gao, Y. (2003). Data depth based on spatial ranks. *Statistics and Probability Letters*, 65, 217-225.

Liu, R. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18, 405-414.

Liu, R. & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *The Journal of the American Statistical Association*, 88, 252-260.

Marden, J. (1999). Multivariate rank tests. multivariate analysis, design of experiments, and survey sampling. *Marcel Dekker*, 401-432.

Moore, D. S., & Macabe, G. (1989). Introduction to the practice of statistics. original source. *Consumer Reports*, June 1986, pp. 366-367.

Tukey, J. W. (1975). Mathematics and picturing data, proceedings of international congress of mathematics. *Vancouver*, 2, 523-531.

Yeh, A., & Singh, K. (1997). Balanced confidence regions based on Tukey's depth and the bootstrap. *Journal of Royal Statistical Society*, B, 59, 639-652.

Zuo, Y., Cui, H., & He, X. (2003). On the Stahel-Donoho estimator and depth-weighted means of multivariate data, *The Annals of Statistics*, (In Press).

Quantifying The Proportion Of Cases Attributable To An Exposure

Camil Fuchs

Department of Statistics and Operations Research
Tel-Aviv University, Ramat-Aviv, Tel-Aviv Israel

Vance W. Berger

Mathematics and Statistics Department
University of Maryland

The attributable fraction and the average attributable fractions, which are commonly used to assess the relative effect of several exposures to the prevalence of a disease, do not represent the proportion of cases caused by each exposure. Furthermore, the sum of attributable fractions over all exposures generally exceeds not only the attributable fraction for all exposures taken together, but also 100%. Other measures are discussed here, including the directly attributable fraction and the confounding fraction, that may be more suitable in defining the fraction directly attributable to an exposure.

Key words: Attributable fraction, average attributable fraction, directly attributable fraction, confounded fraction, fractional complementary attributable risk, multifactor exposure

---

Introduction

If two identical units are treated differently, and respond differently, then the attribution of the differing responses to the differing treatments follows from the process of elimination, and is unambiguous. The same applies to a situation in which two identical groups are treated differently, even if these groups themselves are heterogeneous. Attribution becomes more of a challenge, however, when the groups differ systematically from each other on many dimensions, or exposures. Various measures of attributable fractions have been proposed in these situations, with many exposures being considered simultaneously; one particularly common one bears the name attributable fraction (AF), and is defined as

$$AF = \{Pr(disease) - Pr(disease/no\ exposure)\} / Pr(disease). \quad (1)$$

Camil Fuchs, Sackler School of Exact Sciences, Tel-Aviv University, Ramat-Aviv, Tel-Aviv, 69978, Israel. E-mail: fuchs@math.tau.ac.il. Vance W. Berger is a Mathematical Statistician at the National Cancer Institute. E-mail: bergerv@mail.nih.gov.

See Deubner et al. (1980); Kelsey et al. (1986); and Last (1983). The AF is generally interpreted as an estimate of either the proportion of the cases attributed to (or caused by) the exposure factor or the proportion of the cases that could be prevented if the exposure factor were eliminated. Its importance has grown lately as a measure for interventions, regulations, and lawsuits concerning the effect of the exposure to various factors. Thus, when the Surgeon General warned that 90% of the lung cancer cases are caused by smoking (Gori, 1989), that figure is based on the AF.

In lawsuits, the AF is used in two main contexts. In individual compensation cases, the court may wish to determine the likelihood that the disease of a particular individual was caused by the exposure at issue. The AF has been interpreted as an estimate of this likelihood (Kleinbaum et al., 1982; Greenland & Robins, 1988). Other cases involve class actions, in which states or HMOs sue manufacturers of a presumably hazardous agent for the medical expenses caused by the exposure factor. The medical expenses claimed to have been caused by the exposure factor are usually computed as the sum of the products of the attributable fractions relevant to the specific diseases and the total medical expenses related to those diseases.

The AF was initially termed the attributable risk (Levin, 1953). Other terms include the etiological factor (Kleinbaum et al.,

1982; Schesselman, 1982), the etiological fraction, and the fraction of etiology (Mietienen, 1974). The term attributable risk (e.g. Benichou, 1991), or its variants such as population attributable risk (e.g. Breslow & Day, 1980; MacMahon & Pugh, 1970) or population attributable risk per cent (Cole & MacMahon, 1971; Hennekens et al, 1987) seem to be used less often.

The AF “does not represent disease risk” (Greenland & Drescher, 1993). That is, the AF does not necessarily reflect the proportion of cases caused by the exposure factor; this has been discussed in the statistical and epidemiological literature (Feinstein 1988, 1995; Ashford, 1992; Gori, 1989). One bias originates from shortcomings inherent to epidemiological studies, which invalidate the collected data as representative of the studied populations.

There are also conceptual problems in the measurement of the effect of the exposure factor in general, and in the measurement of the causal effect in particular. For one thing, the AF lacks the desirable property of additivity; that is, in multifactorial diseases, the sum of AFs of all sources of variation (exposures) will generally exceed not only the AF of all exposures taken together, but also 100%. In fact, “... the total ...attributable to the various causes is not 100% but infinity” (Rothman, 1986), which seems to suggest that “...we could prevent more than 100% of any given disease” (Gori, 1989). Many studies focus on a single exposure factor, so this drawback of the AF is not always evident; nevertheless, it remains relevant.

Eide and Gefeller (1995) and Land and Gefeller (2000) propose other measures for assessing the responsibility of the various factors, specifically average attributable fractions (AAFs) and the multiplicative fractional complementary attributable risks (FCARs), respectively. These measures “divide the indivisible” (Pratt, 1987), as they allocate the overall reduced probability of disease into fractions whose sum equals the total effect of the considered exposures. This is accomplished by averaging over all sequences of exposures similar to the situation in multiple regressions with correlated regressors when considering the relative importance of terms (Kruskal, 1987,

Kruskal & Majors, 1989, Pratt, 1987, Gnizi, 1993).

Although these methods may be appropriate for “solving the problem of shared responsibilities for the prevalence of a disease in the population” (Eide & Gefeller, 1995), additivity is not sufficient to ensure a reasonable measure, and the AAF and the FCAR do not represent adequately the proportion of disease attributable to each exposure separately. The task remains to decompose the attributable fraction for the simultaneous exposure to all exposure factors.

When multiple factors contribute to a disease, the ideal situation of perfect knowledge about the relevant variables and of proper collection of data on those variables at the appropriate levels may be rare. But even in these ideal situations, the AF is not an appropriate measure for the assessment of the proportion of cases that can be attributed to an exposure factor. It is even more certainly not a measure of the proportion of cases caused by the exposure factor.

Proposed here is decomposing the AF for the simultaneous exposure to all factors by using terms that are sequentially conditioned on nested sets of factors. The last term is conditioned on all the previous factors and is called the directly attributable fraction (DAF). The DAF is analogous to the Type III sums of squares (Milliken & Johnson, 1984) in linear model theory, in that the variation attributable to an exposure is limited to the variation that cannot be explained by the totality of all other exposures taken together.

The confounded fraction (CF) is $CF = AF - DAF$; the AF of any exposure may be decomposed into a DAF and a CF. It is argued here that the DAF is a more appropriate measure of the proportion of cases that can be directly attributed to the exposure factor than the AF measure defined in (1) above. The overall effect of the exposure factor on the probability of disease is adequately represented by the pair (DAF, CF).

Methodology

First consider the case in which the risk of disease is potentially affected by a single exposure factor A at L levels, and by M adjusting factors (usually demographic variables such as gender, age, residence, etc.). By convention, the first level of the factor A corresponds to no exposure. Each configuration of a level of exposure and a specific combination of levels of the adjusting variables can be presented as a cell E_{sl} in a two way table, $s=1,2,\dots,S$; $l=1,2,\dots,L$. The rows $r_1, r_2, \dots, r_S$ are the strata constructed from the combinations of levels of the adjusting factors, $S=G_1 \cdot G_2 \cdot \dots \cdot G_M$ and the columns are the levels of A .

The attributable fraction for A , adjusted for the confounded variables, can be written as:

$$AF_A = \{Pr(D) - \sum_s Pr(D/E_{s0}) \cdot Pr(r_s)\} / Pr(D), \quad (2)$$

(Whittemore, 1982).

Furthermore, the contribution of each cell and of each column in the table to AF_A can be computed. Following Eide and Gefeller (1995), define

$$\lambda_{sl} = \{Pr(D/E_{sl}) - Pr(D/E_{s0})\} Pr(E_{sl}) / Pr(D) \text{ and } \lambda_l = \sum_s \lambda_{sl}. \quad (3)$$

Thus, λ_{sl} is the contribution of the configuration E_{sl} to $AF_A = \sum_{sl} \lambda_{sl}$ and λ_l is the contribution of the l -th level of exposure to the risk attributable to A . In particular, if A has only two levels, then the only contribution is due to the second (exposed) level. The extension to the general case of F exposures and M adjusting factors is immediate. The adjusted AF for each factor and for the joint effect of several factors can be computed using the appropriate two-way table representation. The columns of the two-way table are now the combinations of levels for the factors whose joint effect is to be computed. The other exposure factors is added to the set of adjusting variables and set the rows of the table as the combination of levels of the newly defined set of adjusting variables.

In particular, the attributable fractions for the F exposure factors, and especially for the

factor of interest A can be computed. Thus, for the computation of the AF of the first exposure factor, the table has $L_2 \cdot L_3 \cdot \dots \cdot L_F \cdot S$ rows and L_1 columns. The table for the second factor has number $L_1 \cdot L_3 \cdot \dots \cdot L_F \cdot S$ rows and L_2 columns, while the table for the assessment of the joint AF of the first two factors has $L_3 \cdot L_4 \cdot \dots \cdot L_F \cdot S$ rows and $L_1 \cdot L_2$ columns. An important special case assesses the AF for the joint effect of all the exposure factors for which data were collected. The two-way table has S rows and $L_1 \cdot L_2 \cdot \dots \cdot L_F$ columns. The first column represents the category of exposure to none of the risk factors.

Estimation of the various AF's

In a cohort study, let n_{sl} be the number of individuals sampled in the E_{sl} configuration with $n_s = \sum_l n_{sl}$ and $n = \sum_s n_s$. The maximum likelihood estimate (MLE) for the AF under the logistic regression model-adjustment (Miettinen 1974; Walter 1975; Bruzzi et al 1985; Greenland 1987; Benichou & Gail 1990; Greenland & Drescher 1993) is obtained by substituting the proper estimates in equation (2) above. Let $Pr(Y=1|E_{sl})$ be the probability of disease at E_{sl} , where Y is an indicator variable taking the value one if the person is diseased, and zero otherwise. If the vector of carriers $\mathbf{x}$ is extended to include $x_l=1$, then these probabilities are assumed to follow the logistic model:

$$\pi_{sl} = Pr(d=1|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}\boldsymbol{\beta})\}. \quad (4)$$

For the $(s,l)^{\text{th}}$ configuration of covariate levels, let E_{s0} be the configurations of levels that a subject with configuration levels E_{sl} would have if not exposed to the studied factors (e.g. factor A). Furthermore, let p_{sl} be the MLE for π_{sl} and DIS be the proportion of diseased in the sample. The MLE for the AF for the studied factors is given by

$$AF = \{DIS - \sum_s p_{s0} (n_s / n)\} / DIS \quad (5)$$

The weighted-sum adjustment (Walter, 1976; Whittemore, 1982, 1983; Benichou, 1991) is a special case of the logistic regression model-adjustment with the fitted model being the saturated model. In this case the relative

frequencies d_{s0}/n_s substitute the estimated probabilities p_{sl} , where d_{sl} is the observed number of diseased in the configuration E_{sl} . Walter (1980) denotes the weighted-sum adjusted AF by the “proportional effect of A”, and reserves the term “attributable fraction” for the unadjusted measure. In a case-control study with randomly sampled n_D cases (diseased) and n_C controls, the AF can be computed from equation (2) by dividing its numerator and denominator by the probability of disease in the no exposure configuration, i.e.

$$AF_A = 1 - \frac{\sum_s \{RR_{s0} \cdot Pr(r_s)\}}{\sum_{sl} \{RR_{sl} \cdot Pr(E_{sl})\}}. \quad (6)$$

For the estimation of AF_A , under the usual rare disease assumption, the estimates of the proportions of the various exposure configurations are replaced with the appropriate proportions in the sample of controls. The relative ratios RR_{sl} are approximated by the corresponding odds ratios from the sample.

Allocation of the overall effect

Consider two nested sets of exposure variables Q_1 and Q_2 , with the second set being $Q_2 = Q_1 \cup A$, i.e. the second set includes all the variables in Q_1 and the extra factor A . The difference $AF_{Q_1} - AF_{Q_2}$, measures the conditional effect of A , given that all the factors in Q_1 have been removed, i.e. set at the non-exposure level.

In general, for a given ordered set of F exposure variables $A_1, A_2, \dots, A_F$, with sequentially nested sets $Q_1 = A_1, Q_2 = A_1 \cup A_2, \dots, Q_F = \cup_j A_j$, the factors can be remove one at a time to compute the F sequentially attributable fractions (*safs*) $AF_{Q_{(j+1)}} - AF_{Q_j}$ (Eide and Gefeller, 1995, Gefeller & Eide, 1998). The set of exposure factors can be extended to include $Q_0 = \emptyset$ by defining $AF_{Q_0} = 0$. This extension properly defines the AF for the factor A_1 as the difference between $AF_{Q_0} - AF_{Q_1}$.

The j -th difference represents the conditional effect of the variable A positioned in the j -th location in the ordering, given that the previous $j-1$ exposures have been removed. Note that, with the exception of the last exposure, the *saf* for a variable depends on the original ordering.

By considering all $F!$ possible orderings, the *safs* for each variable can be computed, with all the combinations of other exposures being removed prior to its own removal. (Note that since a variable's *saf* depends only on the prior exposures, subsets of its $F!$ *safs* will have equal values.)

Cox (1985), Eide and Gefeller (1995), Gefeller and Eide, (1998) propose to compute the average of all possible *safs* relate to each factor, and suggest that those F average attributable fractions (AAFs) are a reasonable measure of the responsibility of the various factors when it is desired to share the disease load in the population among the analyzed exposures. The AAFs satisfy the important requirement that the AF for the joint effect of all the exposures equals the sum of the allocations (Cox, 1985).

A related approach for allocating the responsibility among several exposure factors has been lately proposed by Land and Gefeller (2000). Using a multiplicative Shapley value, they factorize the $1 - AF_{Q_F}$ into a set of F terms called factorial complementary attributable risks (FCARs) which, under this representation, measure the relative contributions of the exposure to the overall load of disease. Unlike the usual AFs, a small FCAR value represents a large effect of the respective factor.

For each factor, the $Pr(\text{disease/no exposure})$ is now substituted in equation (1) by $Pr(\text{disease}) \cdot \text{FCAR}$. The resulting ratio, called factorial attributable risk, equals $\text{FAR} = 1 - \text{FCAR}$. Those measures of shared responsibility do not possess the property that joint effect of all the exposures equals the sum of the allocations.

Directly attributable and confounded fractions

It has been mentioned before that since the AFs for the various exposure do not sum to the total attributable fractions for those exposures, the AFs cannot be considered as proper measures of the cases attributable to a factor. Furthermore, the same reason precludes the AFs from being a proper measure for apportioning, when some factors have to share together the responsibility (Gefeller & Eide, 1993, Eide & Gefeller, 1995).

The AAFs (and/or the FCARs and FARs) may be reasonable measures for solving

the problem of sharing the responsibility for the prevalence of a disease in a population, but they are not good estimates of the effect of a single specific exposure. The allocation the total attributable effect does, what in the regression models context was called the "division of the indivisible" (Pratt, 1987), with the emphasis on the "indivisible".

To continue the parallel from regression models, note that in those models, the direct effect of a factor is commonly assessed by the extra sums of squares yielded when the factor in question is the last to be included in the model. Similarly, it is suggested that since the estimation of the effect for an exposure factor requires the removal of that factor, its directly attributable effect must be interpreted as the disease reduction when the factor is the last to be removed, and not the first.

Thus, if the attributable responsibility of A is considered to represent the segment of the probability of disease which is not explained by the other exposure factors, a more appropriate measure is obtained by ordering the set of exposures with the factor of interest as A_F , and defining the directly attributable fraction (DAF) as the last sequentially attributable fraction. The use of the last *saf* has been also recently proposed by Wilson et al (1998). They termed that special sequentially attributable fraction, resulting when the factor of interest is the last to be removed, "extra attributable fraction" (see also Eide & Gefeller, 2000). This is indeed appropriate in the estimation of the effect of a factor, derived by methods similar to the extra sum of squares in the linear regression models. Used here is the term directly attributable fraction, in the subject matter context, which assesses the attributability of the various fractions of the total probability of disease, and partitions the fraction in which that factor is involved into a directly attributable and a confounded fraction.

As noted before, the *saf* for the last exposure, does not depend on the original ordering. The calculation of the DAFs does not require the calculation of the intermediary *safs*. The DAF for the factor of interest A is defined as the difference of two well defined *AFs*, i.e.

$$DAF_A = AF_{QF} - AF_{QF \sim A}, \quad (7)$$

where $QF \sim A$ is the set of all the exposure factors, except the factor A . This directly attributable fraction is the conditional attributable fraction for A , after removing the effects of all the other exposure factors.

The difference between the attributable fraction AF_A and the directly or conditional attributable fraction (DAF) as the confounded fraction (CF) of A , i.e. is defined as:

$$CF_A = AF_A - DAF_A \quad (8)$$

The confounded fraction is the segment of the probability of disease which is marginally attributed to A , but which is confounded and could just as well be attributed to the effect of the other exposure factors. The confounded factor can also be written as:

$$CF_A = AF_{QF \sim A} - (AF_{QF} - AF_A). \quad (9)$$

The confounded fraction for A can thus be interpreted as a difference of two *AF* terms related to the *notA* exposure factors. The first is the attributable fraction to all the factors which are *notA*, and the second is the effect of those same factors. after the removal of A (i.e. conditioned on A).

A related measure of conditional exposure effect

The conditional *AF*'s defined above are intuitively appealing since they represent the decomposition of the overall effect of the F exposure factors. An additional measure of the conditional exposure effect (CEE) is suggested here, not as an alternative to those presented above, but rather as yielding complementary information. The overall incidence rate after the removal of $A_1, A_2, \dots, A_t$ is:

$$Pr(D|A_1, A_2, \dots, A_t) = \{1 - Pr(D) \cdot AF_{Q_t}\}. \quad (10)$$

The conditional exposure effect (CEE) can thus be defined as:

$$CEE_{A(t+1)|A_1, A_2, \dots, A_t} = (AF_{Q(t+1)} - AF_{Q_t}) \cdot Pr(D) / Pr(D|A_1, A_2, \dots, A_t) \quad (11)$$

If the correlations between A and the other exposure factors are roughly constant, the

various CEE's which correspond to a specific exposure factor, conditioned on various other effects, can be expected to differ only slightly from each other. Note that as in the case of the directly attributable fraction, the CEE's can be defined for an exposure factor conditioned on any subset of the $F-1$ variables, not only on their union.

$$Pr(D/E_{sl}) - Pr(D/E_{s0}) \} \cdot Pr(E_{sl})$$

Examples

The computations and the interpretation of the statistics presented in the previous sections are illustrated with a hypothetical example originated from Walter (1980, Table3). The data contain three dichotomous exposure factors (A , B , C). Complete information is provided on the proportions in the population for each configuration of levels of the exposure factors (the estimates of the $Pr(E_i)$'s) and with the respective incidence rates (the estimates of the $Pr(D/E_i)$'s). The original $Pr(E)$'s-vector was slightly altered to illustrate the fact that $\sum_j AF_j$, $j=1, \dots, F$ can exceed 100%. All the attributable fractions were computed with weighted-sum adjustments.

Panel (a) of Table 1 presents the data and the sequential vectors of estimated proportions in the populations exposed to each factor, following the various removals of factors. There are three factors which can be removed in stage 1, and the resulting statistics are denoted with the notation of $(*|A)$, $(*|B)$, $(*|C)$ according to the respective removed factor. Similarly, one of the pairs of factors AB or AC or BC , is removed at the end of the second stage.. At stage 3 the remaining factor is removed and the conditional probability of disease is obtained, with all the factors being at the not exposed level.

First note that $Pr(D) = \sum_i Pr(D/E_i) \cdot Pr(E_i) = 0.4\%$, and that when all the three factors are controlled for, $\sum_i Pr(D/E_{i0}) \cdot Pr(E_i) = 0.1\%$, yielding an overall attributable fraction for $A+B+C$ of $AF_{A+B+C} = 75\%$, i.e. the three factors together "can explain" 75% of the overall incident rates. The unconditional individual attributable fractions are 38.1%, 43.1% and 41.3%, respectively, whose sum is 122.5%.

Panel (b) of Table 1 presents all the possible sequences of removal of factors.

Assume that the factor of interest is C . The unconditional AF seems to indicate that exposure to C is responsible for 41.3% of all the disease cases. However, when the effects of variables B and C are controlled for, only 9.4% of the cases can be directly attributable to C , and that the remainder of 31.9% is confounded effect with the other two factors.

Table 1 also presents the conditional exposure effects (CEEs) for all the stages. Unlike the conditional AFs, the CEEs are not necessarily monotonic and they vary less as a function of the removed variables.

Case Control Studies

The calculations are illustrated with the data on the oral cancer distributions among persons at the four configurations of (exposed, not exposed) to the alcohol and tobacco factors. The original data set of Rothman and Keller (1972) and Keller and Terris (1965) contained 598 case-control pairs. The data were further analyzed by Walter (1983).

The data summarized in Table 2 presents as initial data the four odds ratios (used to approximate the relative risks) and the proportions of controls in the four configurations (as estimates for $Pr(E_i)$'s). Panel (a) of Table 2 also presents the $P^{(t)}_g$ -values for $t=1,2..$ Panel (b) presents the attributable fractions for the various levels of conditioning. It can be seen from the table that the two individual AFs for alcohol and tobacco are 66.2% and 72.1%, while the AF for the two factors taken together is 76.2%. Walter (1983) noticed that "very little additional is gained by removing tobacco and alcohol exposure as opposite to preventing exposure to just one of them".

Thus, one can expect that the computed individual AFs decompose into small directly attributable fractions and much larger confounded fractions. The entries in Table 2 confirm this expectation. The DAFs for alcohol and tobacco are 4.2% and 10% as opposed to the initial AFs of 66.2% and 72.1%. The remaining roughly 62% for both alcohol and tobacco are confounded fractions.

Table 1. Computation of the all the possible AFs, DAFs, CFs, and CEEs, for the hypothetical data with dichotomous factors. The P -, I -, AF-, DAF-, CF-, and CEE-values are percentages.

Panel (a)

Design Factors			Initial		Proportions in the population in the 1 st , 2 nd and 3 rd stages						
A	B	C	Pr(D E)	Pr(E)	P ⁽¹⁾ _A	P ⁽¹⁾ _B	P ⁽¹⁾ _C	P ⁽²⁾ _{AB}	P ⁽²⁾ _{AC}	P ⁽²⁾ _{BC}	P ⁽³⁾ _{ABC}
0	0	0	20	0.1	45.0	32.5	32.5	62.5	70	50	100
0	0	1	13	0.2	25.0	17.5	0	37.4	0	0	0
0	1	0	13	0.3	17.5	0	17.5	0	30	0	0
0	1	1	5	0.8	12.5	0	0	0	0	0	0
1	0	0	25	0.2	0	30.0	37.5	0	0	50	0
1	0	1	13	0.5	0	20.0	0	0	0	0	0
1	1	0	5	0.6	0	0	12.5	0	0	0	0
1	1	1	8	1.8	0	0	0	0	0	0	0

Panel (b)

	<i>A</i>	<i>B</i>	<i>C</i>	<i>AB</i>	<i>AC</i>	<i>BC</i>	<i>ABC</i>
AF	38.1	43.1	41.3	65.6	60.0	62.5	75.0
AF _{/A}		27.5	21.9			36.9	
AF _{/B}	22.2		19.4		31.9		
AF _{/C}	18.8	21.3		33.8			
AF _{/AB}			9.4				
AF _{/AC}		15.0					
AF _{/BC}	12.5						
DAF	12.5	15.0	9.4				
CF	25.6	28.1	31.9				
AAF	23.7	27.5	23.8				
FAR	35.8	40.4	40.7				
CEE	38.1	43.1	41.3	65.6	60.0	62.5	75.0
CEE _{/A}		44.4	35.4			59.6	
CEE _{/B}	39.6		34.1		56.0		
CEE _{/C}	31.9	36.2		57.4			
CEE _{/AB}			27.3				
CEE _{/AC}		37.5					
CEE _{/BC}	33.3						

Table 2. Computation of the all the possible AFs, DAFs, CFs, and CEEs, for the case-control oral cancer data with two dichotomous factors.

Panel (a)

<i>Tobacco</i>	<i>Alcohol</i>	<i>Pr(E)-controls</i>	<i>RR(from OR)</i>	$P^{(1)}_{/Alcohol}$	$P^{(1)}_{/Tobacco}$	$P^{(2)}_{/A+T}$
<i>No User</i>	<i>No User</i>	9	1.00	19	24	100
	<i>User</i>	10	1.23	0	76	0
<i>User</i>	<i>No User</i>	15	1.52	81	0	0
	<i>User</i>	66	5.71	0	0	0

Panel (b)

	<i>Alcohol</i>	<i>Tobacco</i>	<i>Both</i>
AF	66.2	72.1	76.2
AF _{/Alcohol}		10.0	
AF _{/Tobacco}	4.2		
DAF	4.2	10.0	
CF	62.0	62.1	
AAF	35.2	41.0	
FAR	46.3	55.7	
CEE	66.2	72.1	76.2
CEE _{/Alcohol}		14.9	
CEE _{/Tobacco}	29.6		

The analysis of the CEEs also reveals an interesting pattern. In the example presented in Table 1, the CEEs were relatively stable as a function of the extra conditioning, and did not differ dramatically from the AFs. On the other hand, in this example, the proportion of the incidence rates explained by the second term (alcohol or tobacco) is very low not only when the denominator is the overall incidence rate, but also when the denominator is the incidence rates remained after the first variable was removed. This is another facet of the highly confounding pattern in this data set.

This simple example also illustrate the contention that the AAFs value may provide an equitable solution for the problem of allocation of shared responsibility but is inappropriate for assessing the attributable fraction for a specific exposure. The corresponding AAFs are 35.1% and 41.0% which sum to the total effect of 76.2%. However, if only one exposure is considered, for example alcohol, its AAF of 35.1% is the average of 66.2% (the original AF) and the value of 4.2% (the DAF, which is the *saf* in the second step). It is very difficult to defend this value with any degree of confidence as representing the percent of cases attributable to alcohol. The same is true for smoking where the AAF of 41.0% is the average of 72.1% and 10%.

Conclusion

In the discussion following the analysis oral cancer presented above, Walter (1980) stated that “although the sum (of the AF’s) exceeds 100%, this does not invalidate the individual (AF’s) estimates; indeed, this phenomenon is more likely as more factors are considered and confounding becomes inevitable. Each measure must be interpreted as the disease reduction if the factor in question were the first to be removed”.

However, when the purpose of the research is the assessment of causation and of attributable responsibility of a specific factor, the fact that the total contribution may exceed 100% does invalidate the AF’s as interpretable measurements.

Assume that while assessing the effect of consumption of alcohol, one controls first for the effect of smoking by assessing the remaining incidence rates after all persons stopped smoking. Following this adjustment, the percent of cases for which the alcohol consumption is still “responsible” is assessed. The computations presented above show that the estimate of the percent of cases for which alcohol is found now responsible is 4%, instead of the initial 66%. The controlling for the tobacco variable didn’t assume any change in the drinking behavior of the population.

Nevertheless, following the control for the smoking behavior, one witnesses a very significant decrease in the percent of cases attributable to alcohol consumption. It is thus clear that a significant proportion of the fraction initially attributed to drinking, can in fact be attributed to the effect of smoking, and vice versa.

The AAFs (and/or the FCARs and FARs) may be reasonable measures for solving the shared responsibility problem, but they are not proper estimates of the effect of a single specific exposure.

In contrast, the DAF has the clear interpretation as the fraction that can be attributed to that factor and which cannot be attributed to any of the other factors on which there are data in the sample. The complementary confounding fraction indicates the portion of the extra cases in which the factor in question may

have been involved, but about which it is impossible to distinguish between its effect and the effects of the other factors.

Finally, note that for all measures of attributable fractions, the assumption that the data include all the relevant variables is cardinal for the validity of the results. As an illustration, constructed in the oral cancer is an artificial latent variable X , and set for the four combinations of X and alcohol (regardless of smoking) the RRs to be 1, 2, 10 and 20. The percents exposed to X in the four combinations of smoking and alcohol were set to be 17%, 6%, 31% and 68%, respectively. The collapsed table over X returns the previous pattern, but when X is considered, the AFs for Alcohol, Tobacco and X are 46%, 0% and 83%, with the AF for Alcohol*Tobacco*X (and also Alcohol*X) explaining 90.5% of the total load, a certainly different picture than in the previous analysis.

References

- Ashford, J. R. (1992). Problems in assessing smoking-related mortality. *Smoking-Related Diseases*, 33, 263-274.
- Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case-control studies: A review. *Statistics in Medicine*, 10, 1753-1773.
- Benichou, J. & Gail, M. H. (1990). Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics*, 46, 991-1003.
- Breslow, N. E., & Day, N. E. (1980). Statistical Methods in Cancer Research Vol 1: The Analysis of Case-control Studies, International Agency for Research on Cancer. *Scientific Publications*, No. 32, Lyon.
- Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A., & Schairer, C. (1985) Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology*, 122, 904-914.
- Cole, P. & Macmahon, B. (1971). Attributable risk percent in case-control studies. *British Journal of Preventive and Social Medicine*, 25, 242-244.

- Deubner, D. C., Wilkinson, W.E., Helms, M. J., Tyroler, H. A., & Hames, C. G. (1980). Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. *American Journal of Epidemiology*, 112, 135-143.
- Eide, E. E., & Gefeller, O. (1995). Sequential and average attributable fractions as aids in the selection of preventive strategies. *Journal of Clinical Epidemiology*, 48, 645-655.
- Eide, G. E., & Gefeller, O. (2000). Re: Attributable risk in practice, (Letter, with replies of Walter, S. D., Wilson, P. D., Loffredo, C.A, Correa-villasenor A. & Ferencz, C.). *American Journal of Epidemiology*, 151, 208-212.
- Feinstein, A. R. (1988) Scientific Standards in Epidemiologic Studies of the Menace of Daily Life. *Science*, 242, 1257-1263.
- Feinstein, A. R. (1995). Biases introduced by confounding and imperfect retrospective and prospective exposure assessments, In: J. D. Graham, (Ed.), *The Role of Epidemiology in Regulatory Risk Assessment*, Elsevier Science, B. V., 29-42.
- Gefeller, O., & Eide, G. E. (1993). Estimation of adjusted attributable risk in case-control studies. Letter. *Statistics in Medicine*, 12, 91-94
- Gefeller, O., Land, M., & Eide, G. E. (1998). Averaging attributable fractions in the multifactorial situation: Assumptions and Interpretation. *Journal of Clinical Epidemiology*, 51, 437-441.
- Gnizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3, 407-420.
- Gori, G. B. (1989). Epidemiology and Concept of Causation in Multifactorial Diseases. *Regulatory Toxicology and Pharmacology*, 9, 263-272.
- Greenland, S. (1987). Variance estimators for attributable fraction estimates, consistent in both large strata and sparse data, *Statistics in Medicine*, 6, 701-708.
- Greenland, S., and Drescher, K. (1993). Maximum Likelihood Estimation of the Attributable Fraction from Logistic Models. *Biometrics*, 49, 865-872.
- Greenland, S., & Robins, J. M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128, 1185-1197.
- Hennekens, C. H., & Buring, J. E. (1987). *Epidemiology in Medicine*. Little, Brown & Co, Boston/Toronto.
- Keller, A. Z., & Terris, M. (1965). The association of alcohol and tobacco with cancer of the mouth and pharynx. *American Journal of Public Health*, 55, 1578-1578.
- Kelsey, J. L., Thompson, W. D., & Evans, A. S. (1986). *Methods in Observational Epidemiology*. New York: Oxford University Press.
- Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, California: Lifetime Learning.
- Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41, 6-10.
- Kruskal, W. & Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician*, 43, 2-6.
- Kuritz, S. J., & Landis, J. R. (1988a). Summary attributable risk estimation from unmatched case-control data. *Statistics in Medicine*, 7, 507-517.
- Kuritz, S. J., & Landis, J. R. (1988b). Attributable risk estimation from matched case-control data. *Biometrics*, 44, 355-367.
- Land, M., & Gefeller O. (2000). A multivariate variant of the Shapley value for factorizing the risk of disease, In: Patrone, F., Garcia-jurado, I. & Tijs, S. (Eds.), *Game theory: Contributions from applied game theory*. Kluwer Academic Publishers.
- Levin, M. L. (1953) The occurrence of lung cancer in man, *Acta Unio Internationalis contra. Cancrum*, 9, 531-541.
- Macmahon, B., & Pugh, T. F. (1970). *Epidemiology; Principles and Methods*. Little, Brown and Co, Boston.
- Miettinen, O. S. (1974). Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology*, 99, 325-332.

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data volume 1, Designed Experiments*. Van Nostrand Reinhold, New York.

Pratt, J. W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained, In: T. Pukkila and S. Puntanen (Eds.), *Proceedings of the Second International Conference in Statistics* (Edited by), 245-260. University of Tampere, Tampere, Finland.

Rothman, K. J. (1986). *Modern epidemiology*. Little, Brown & Co, Boston.

Rothman, K., & Keller, A. (1972). The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *Journal of Chronic Diseases*, 25, 711-716.

Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis*. New York: Oxford University Press.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, 32, 829-849.

Walter, S. D. (1980). Prevention for Multifactorial Diseases'. *American Journal of Epidemiology*, 112, 409-416.

Walter, S. D. (1983). Effects of interaction, confounding and observational error on attributable risk estimation. *American Journal of Epidemiology*, 117, 598-604.

Wilson, P. D., Loffredo, C. A, Correa-villasenor, A., & Ferencz, C. (1998). Attributable fractions for cardiac malformations. *American Journal of Epidemiology*, 148, 414-423.

Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine*, 1, 229-243.

On Polynomial Transformations For Simulating Multivariate Non-normal Distributions

Todd C. Headrick
Southern Illinois University - Carbondale

Procedures are introduced and discussed for increasing the computational and statistical efficiency of polynomial transformations used in Monte Carlo or simulation studies. Comparisons are also made between polynomials of order three and five in terms of (a) computational and statistical efficiency, (b) the skew and kurtosis boundary, and (c) boundaries for Pearson correlations. It is also shown how ranked data can be simulated for specified Spearman correlations and sample sizes. Potential consequences of nonmonotonic transformations on rank correlations are also discussed.

Key words: Correlated data, cumulants, Monte Carlo methods, polynomial transformations, nonnormality

---

Introduction

A common practice used to investigate the relative Type I error and power properties of competing statistical procedures under non-normality is the method of Monte Carlo. For example, consider the following polynomial transformation in general form

$$Y_i = c_0 + \sum_{i=1}^m c_i Z_i^i \quad (1)$$

where $Z_i \sim \text{NID}(0,1)$, and $i \in \mathbb{N} = \{1, 2, \dots, m\}$. Setting $m=3$, Fleishman (1978) derived a system of four equations that would solve for the four coefficients $c_0, \dots, c_3$ in (1) for a specified non-normal distribution. Specifically, these coefficients are determined by simultaneously solving this system of equations for the first four standardized cumulants of a distribution. The coefficients are subsequently entered into (1) to

generate Y_i with the specified cumulants. Equation (1) was extended to $m=5$ by Headrick (2002) for controlling the first six standardized cumulants from a specified probability density function.

The third-order polynomial (Fleishman, 1978) and the fifth-order polynomial (Headrick, 2002) transformations were also extended for the purpose of generating multivariate non-normal distributions (Headrick, 2002, Equation, 26; Headrick & Sawilowsky, 1999, Equation 9; Vale & Maurelli, 1983, Equation 11). These extensions have been demonstrated to be quite useful when there is a need for correlated non-normal data sets in a Monte Carlo study.

Some examples include analysis of covariance (Harwell & Serlin, 1988; Headrick & Sawilowsky, 1999; Headrick & Vineyard, 2001; Klockers & Moses, 2002), hierarchical linear models (Shieh, 2000), regression (Harwell & Serlin, 1989; Headrick & Rotou, 2001; Whittaker, Fauladi, & Williams, 2002) repeated measures (Beasley & Zumbo, 2003; Harwell & Serlin, 1997), and multivariate nonparametric tests (Beasley, 2002; Habib & Harwell, 1989). The multivariate extension of the fifth-order polynomial has also demonstrated to be useful for simulating continuous with ranked or ordinal data structures (Headrick & Beasley, 2003) and for generating systems of correlated non-normal linear statistical equations (Headrick & Beasley, 2004).

Todd C. Headrick is Associate Professor of Statistics. Address: Section on Statistics and Measurement, Department of EPSE, 222-J Wham Building, Mail Code 4618, Southern Illinois University-Carbondale, IL, 62901. His areas of research interest are statistical computing, nonparametric statistics, and optimization. Email address: headrick@siu.edu.

Although the primary advantages of the third and fifth-order polynomials are their ease of execution and computational efficiency, there are limitations to these transformations. More specifically, the primary limitations are (a) the transformations are limited in terms of the possible combinations of skew and kurtosis, (b) the polynomials are not, in general, monotonic transformations and therefore have the potential to produce biased rank correlation coefficients, and (c) distributions with bivariate non-normal structures may have lower and upper boundary points $(-a, a)$ for Pearson correlations (r) such that $r \in [-1 < -a, a < +1]$ and where it is possible, for example, that $|a| < 0.70$. It should be noted that the distribution of Y_1 , in general, is not exact. Headrick (2004) has derived the probability density function and distribution function for Y_1 when the transformation between Y_1 and Z_1 is monotonic.

In view of the above, the purposes of the study are to introduce and discuss methods that minimize the limitations of the polynomial transformations and to develop a procedure for simulating rank correlations. More specifically, the intent is to (a) derive and discuss methods for improving computational and statistical efficiency for a Monte Carlo study, (b) compare and contrast the third and fifth order polynomials in terms of the skew and kurtosis boundary and in terms of boundaries for Pearson correlations, (c) provide a method for simulating Spearman rank correlations with specified samples sizes, and (d) discuss the potential effects of nonmonotonic transformations on rank correlations.

Improving Computational and Statistical Efficiency

Consider (1) with $m = 5$ as

$$Y_1 = c_0 + c_1 Z_1 + c_2 Z_1^2 + c_3 Z_1^3 + c_4 Z_1^4 + c_5 Z_1^5 \quad (2)$$

or

$$Y_1 = c_0 + Z_1 \left(c_1 + Z_1 \left(c_2 + Z_1 \left(c_3 + Z_1 \left(c_4 + c_5 Z_1 \right) \right) \right) \right). \quad (3)$$

If the algorithm used to generate Y_1 is coded in the manner as in (3) instead of (2) then the run time of a Monte Carlo or simulation study can be substantially reduced. To illustrate (briefly), on a Pentium-based PC it took approximately 25 seconds of computer time to draw 100,000 random samples of size $n = 550$ from an approximate exponential distribution using (3). On the other hand, using (2), the sample size had to be reduced to $n = 100$ to obtain the same 100,000 draws within the same 25 second time period. Thus, a considerable gain in computational efficiency can be realized by using (3) in lieu of (2).

Suppose two standardized random variables Y_1 and Y_2 based on (3) are generated. A method that is useful to improve the efficiency of the estimate of $(Y_1 + Y_2)/2$ is by inducing a negative correlation on Y_1 and Y_2 . To demonstrate, if Y_1 and Y_2 were identically distributed, then

$$\text{Var} \left[\frac{Y_1 + Y_2}{2} \right] = \frac{1}{2} + \frac{\text{Corr}[Y_1, Y_2]}{2}. \quad (4)$$

By inspection of (4) it would be advantageous if Y_1 and Y_2 were negatively correlated.

Assume that a monotone relationship between Z_1 and Y_i for $i = 1, 2$ exists. To induce a negative correlation on Y_1 and Y_2 it is only necessary to simultaneously reverse the signs of the coefficients with odd subscripts in Y_2 as

$$Y_1 = f_1(c_0, c_1, c_2, c_3, c_4, c_5, Z_1) \quad (5)$$

$$Y_2 = f_2(c_0, -c_1, c_2, -c_3, c_4, -c_5, Z_1). \quad (6)$$

Because the structure between Y_i and Z_1 is standard bivariate normal, the correlation between Y_1 and Y_2 can be defined as

$$\rho_{Y_1 Y_2} = E[Y_1 Y_2]. \quad (7)$$

Expanding (7) and taking expectations using the moments from the standard normal density yields

$$\rho_{Y_1 Y_2} = c_0^2 - c_1^2 + 2c_0(c_2 + 3c_4) - 6c_1(c_3 + 5c_5) + 3(c_2^2 + 10c_2c_4 - 5(c_3^2 - 7c_4^2 + 14c_3c_5 + 63c_5^2)). \quad (8)$$

Thus, the correlation between Y_1 and Y_2 can be determined by evaluating (8) using specified values for $c_0, \dots, c_5$. For example, evaluating (8) using the coefficients that approximate the exponential density (see Headrick, 2002, Table 1) gives $\rho_{Y_1 Y_2} \cong -0.647$.

The method of inducing a negative correlation between Y_1 and Y_2 is analogous to the method used on distributions generated by the inverse transform method. More specifically, consider generating X_1 and X_2 from the single parameter exponential family with distribution function G and with an inverse distribution function denoted as G^{-1} . Let $X_1 = G^{-1}(V)$ and $X_2 = G^{-1}(1 - V)$ where $V \sim U(0,1)$. Define the parameters for the first and second moments as θ and θ^2 . From the definition of the product moment of correlation exists

$$E[X_1 X_2] = \theta^2 \int_0^1 \ln v \ln(1 - v) dv = \theta^2 (2 - \pi^2/6).$$

As such, the correlation between X_1 and X_2 is

$$\rho_{X_1 X_2} = 1 - \pi^2/6 \cong -0.645. \quad (9)$$

Thus, the approximation given by (8) for the exponential distribution is very close to the exact result given in (9).

Presented in Table 1 below are confidence intervals from a Monte Carlo simulation study that demonstrate the advantage of inducing a negative correlation on Y_1 and Y_2 . By inspection of Table 1 when Y_1 and Y_2 are uncorrelated it takes over 2.5 times the sample size to obtain a confidence interval that has approximately the same width as the data with an induced negative correlation. Thus, whenever possible it is advantageous to induce a negative correlation to improve the computational and statistical efficiency of a Monte Carlo study.

Table 1. Confidence Intervals (CI's) on the estimate of $(Y_1 + Y_2)/2$ with and without a negative correlation induced. Y_1 and Y_2 are approximate exponential distributions with population means of $\gamma_1 = 5$. The CI's are based on 50,000 sample estimates.

Corr[Y_1, Y_2]	Sample Size	95% C.I.
0.000 -0.647	$n = 10$	[4.552, 5.448] [4.715, 5.252]
0.000 -0.647	$n = 26$	[4.726, 5.273] [4.841, 5.158]

Statistical efficiency can also be improved when using the fifth-order polynomial in lieu of the third-order polynomial. For example, consider approximating the uniform distribution. The kurtosis for this distribution is theoretically -1.20 . However, the lower-boundary of kurtosis for the third-order polynomial is -1.15132 (Headrick & Sawilowsky, 2000) whereas the fifth-order polynomial can generate this distribution with the required kurtosis (Headrick, 2002, Table 1). Presented in Table 2 is a comparison between the two polynomials' approximations to the uniform distribution. By inspection of the values of RMSE in Table 2, it is evident that the fifth-order polynomial is superior in its approximation to the standardized cumulants of this distribution.

Lower Boundary Points of Kurtosis

The lower boundary points of kurtosis is another topic of concern because neither the third nor the fifth-order polynomial transformations span the entire skew (γ_3) and kurtosis (γ_4) plane given by the general expression

$$\gamma_4 \geq \gamma_3^2 - 2. \quad (10)$$

Table 2. Estimates of the first six standardized cumulants of the uniform density and the Root Mean Square Errors for the third and fifth-order polynomials. Estimates ($\hat{\gamma}_i$) are based on a sample size of $n=50$ and averaged across 50,000 repetitions. The same random numbers were used in both polynomials.

Standardized Parameters		
Uniform Distribution (γ_i)	$\hat{\gamma}_i$	RMSE
Third-Order Polynomial		
$\gamma_1 = 0.0$	0.000	0.142
$\gamma_2 = 1.0$	1.000	0.132
$\gamma_3 = 0.0$	0.002	0.338
$\gamma_4 = -6/5$	-1.152 ¹	1.673
$\gamma_5 = 0.0$	0.095	15.771
$\gamma_6 = 48/7$	8.711	161.61
Fifth-Order Polynomial		
$\gamma_1 = 0.0$	0.000	0.142
$\gamma_2 = 1.0$	1.000	0.127
$\gamma_3 = 0.0$	0.001	0.278
$\gamma_4 = -6/5$	-1.200	0.354
$\gamma_5 = 0.0$	0.006	0.897
$\gamma_6 = 48/7$	6.841	3.301

¹The lower boundary of kurtosis for the third-order polynomial is -1.15132.

Proof (Eq. 10). For any random variable with finite values of γ_i define

$$\gamma_i = \frac{E[X - E[X]]^i}{\sigma_X^i} = E\left(\frac{X - E[X]}{\sigma_X}\right)^i. \quad (11)$$

Without loss of generality, it can be assumed that the random variable X is standardized such that $E[X]=0$ and $\sigma_X = E[X^2]=1$ in (11). From the covariance (or Schwarz) inequality there is $E[XW]^2 \leq E[X^2]E[W^2]$. If the two random variables in the covariance inequality are X and $X^2 - 1$, then

$$\left(E[X(X^2 - 1)]\right)^2 \leq E[X^2] \times E[(X^2 - 1)^2]$$

$$\left(E[X^3 - X]\right)^2 \leq E[X^4 - 2X + 1]$$

$$\left(E[X^3]\right)^2 \leq E[X^4] - 1$$

$$\gamma_3^2 \leq \gamma_4 - 1, \text{ thus}$$

$$\gamma_4 \geq \gamma_3^2 + 1, \text{ and where}$$

subtracting a constant of 3, such that kurtosis for the normal distribution is zero, gives (10) (It can also be shown that the equality condition in (10) is not possible. However, in the context of this paper, the matter is trivial).

Presented in Table 3 are the lower boundary points of kurtosis for both polynomials. The values of minimum kurtosis (γ_4' , γ_4^*) were obtained by minimizing Equation 14 (Headrick & Sawilowsky, 2000) and Equation 36 (Headrick, 2002) using the command 'NMinimize' (*Mathematica*, Wolfram, 2003, version 5.0). By inspection of Table 3, it is evident that the fifth-order polynomial spans a much larger space in the plane defined by (10) than the third-order polynomial.

Pearson Correlations

As mentioned, the third and fifth-order polynomial transformations are computationally efficient algorithms for generating multivariate non-normal distributions. In general, and in terms of the fifth-order polynomial, the approach taken is to solve the equation given in Headrick (2002, Equation, 26) for pairwise intermediate correlations between k variables.

The intermediate correlations are subsequently assembled into a correlation matrix and factored (e.g., a Cholesky factorization). The components from the factorization are used to generate multivariate standard normal random deviates correlated at an intermediate level. These deviates are then transformed by the polynomials to produce the specified non-normal distributions with the desired intercorrelations.

Table 3. Lower boundaries of kurtosis for the third (γ'_4) and fifth (γ_4^*) order polynomials for a given value of skew (γ_3). The coefficients $c_0, \dots, c_5$ are associated with the fifth-order polynomial.

γ_3	γ'_4	γ_4^*	c_0	c_1	c_2	c_3	c_4	c_5
0.00	-1.151320	-1.385081	0.000000	-1.643734	0.000000	0.320242	0.000000	-0.011361
0.25	-1.045100	-1.296301	-0.160182	-1.597079	0.195003	0.302208	-0.011607	-0.010437
0.50	-0.741671	-1.038260	-0.298119	1.492904	0.036292	-0.266933	-0.021600	0.008682
0.75	-0.252697	-0.614627	-0.419443	1.357093	0.508113	-0.228251	-0.029554	0.006969
1.00	0.424841	-0.020321	-0.529477	1.190353	0.637194	-0.187141	-0.035906	0.005314
1.25	1.297258	0.753833	-0.632000	0.981640	0.754682	-0.141828	-0.040894	0.003602
1.50	2.370670	1.724592	-0.732543	0.690295	0.866255	-0.087835	-0.044570	0.001719
1.75	3.652341	2.757983	-0.503230	0.829259	0.623359	0.006876	-0.040043	-0.002257
2.00	5.151620	3.983870	-0.524421	0.710491	0.645056	0.048321	-0.040213	-0.004000

There are limitations in simulating multivariate distributions using the polynomial transformations. Specifically, the third and fifth-order polynomials may have lower and upper boundary points ($-a$, a) for correlations (r) such that $r \in [-1 < -a, a < +1]$. In the context of the bivariate case, this problem is most pronounced when one distribution is symmetric and the other skewed.

For example, suppose the distributions are approximate chi-square (1df) and normal. The boundaries of correlation for the third-order polynomial are $a = \pm .67481$ whereas the boundaries for the fifth-order polynomial are $a = \pm .82024$. As another example, if the normal distribution is replaced by the coefficients for the uniform distribution, then the boundaries for bivariate correlation are $a = \pm .623033$ and $a = \pm .738553$ for the third and fifth-order polynomials, respectively. Thus, the fifth-order polynomial can be a remedy for cases where it is needed to simulate the often used correlation of $r = .70$ when the distributional conditions make it impossible for the third-order polynomial.

Monotonicity and Spearman Correlations

A monotonic relationship between Y_i and Z_i in (3) is defined as

$$Z_{li} > Z_{lj} \Rightarrow Y_{li} > Y_{lj}, \forall_{i \neq j}. \quad (12)$$

Testing for a monotonic relationship can be accomplished by solving $dY_i/dZ_i = 0$ for Z_i . If only complex solutions of Z_i exist then the transformation between Y_i and Z_i is considered globally monotonic. If real solutions of Z_i exist, then the transformation is considered non-monotonic. For example, all chi-square distributions ($df > 1$) approximated by fifth-order polynomials are globally monotonic transformations. The third-order polynomials, however, are not monotone transformations *for any* approximation of the chi-square family (see Headrick, 2004). The concern for monotonic relationships becomes important when there is a need to simulate ranked data with specified Spearman correlations.

Consider generating Y_1 and Y_2 from equations of the form in (3) with a Pearson correlation $\rho_{Y_1 Y_2}$. Let $R(Y_1)$ and $R(Y_2)$ denote the ranks of Y_1 and Y_2 and $R(Z_1)$ and $R(Z_2)$

denote the ranks of Z_1 and Z_2 . If monotonic relationships hold for both transformations as defined in (12), then $\rho_{R(Y_1)R(Y_2)} = \rho_{R(Z_1)R(Z_2)} = \rho_s$ and where ρ_s denotes the Spearman rank coefficient of correlation.

Because the structure of Z_1 and Z_2 is standard bivariate normal, ranked data can be simulated for specified values of ρ_s and n by making use of the following expression (Moran, 1948)

$$\rho_s = \frac{6}{\pi} \left\{ \frac{n-2}{n+1} \sin^{-1} \left(\frac{\rho_{Z_1 Z_2}}{2} \right) + \frac{1}{n+1} \sin^{-1} (\rho_{Z_1 Z_2}) \right\}. \quad (13)$$

More specifically, to generate $R(Y_1)$ and $R(Y_2)$ with a specified rank correlation ρ_s and sample size, one need only numerically solve (13) for $\rho_{Z_1 Z_2}$ given values of ρ_s and n . For example, suppose it is desired to generate $R(Y_1)$ and $R(Y_2)$ with a Spearman rank correlation of $\rho_s = .70$, $n = 5$, and where the distributions Y_1 and Y_2 are approximate exponential. For this example, it is appropriate to use fifth-order polynomial transformations because (12) holds for this case. Thus, solving (13) for the specified values of ρ_s and n gives an intermediate correlation of $\rho_{Z_1 Z_2} = .811202$.

Conclusion

In terms of the procedure for simulating ranked data with specified Spearman correlations, it should be pointed out that equation (12) is a sufficient condition for monotonicity. However, the procedure will provide adequate simulations of rank data with specified correlations if the polynomial transformations are locally monotonic. More specifically, the simulated rank correlations may be robust to violations of (12) even though real solutions of Z_1 (or Z_2) exist for $dY_1/dZ_1 = 0$ (or $dY_2/dZ_2 = 0$). For example, assume more generally, for two symmetric distributions of the same shape that $Z \pm 3.00$ are real solutions for $dY/dZ = 0$.

These distributions could be considered locally monotonic because the probability associated with drawing such values of $Z: |Z| \geq 3.00$ is only .0027. Because the probability of obtaining such values of Z is very low, the amount of bias introduced into a Monte Carlo or simulation study would be negligible.

To provide an empirical definition of local monotonicity, this author conducted simulations using fifth-order transformations with many different non-normal distributions with nonmonotonic relationships. The simulation results indicated that Spearman correlations were close to what (13) would compute ($\rho_s \pm .025$) if the real solutions of Z for $dY/dZ = 0$ were $|Z| \geq 1.75$.

References

- Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures. *Multivariate Behavioral Research*, 37, 197-226.
- Beasley, T. M. & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics and Data Analysis*, 42, 569-593.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- Habib, A. R., & Harwell, M. R. (1989). An empirical study of the Type I error rate and power of some selected normal theory and nonparametric tests of independence of two sets of variables. *Communications in Statistics: Simulation and Computation*, 18, 793-826.
- Harwell, M. R. & Serlin, R. C. (1988). An experimental study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin*, 104, 268-281.
- Harwell, M. R. & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics*, 14, 351-371.

Harwell, M. R. & Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics: Simulation and Computation*, 26, 605-618.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate non-normal distributions. *Computational Statistics and Data Analysis*, 40, 685-711.

Headrick, T. C. & Beasley, T. M. (2004). A method for simulating correlated non-normal systems of linear statistical equations. *Communications in Statistics: Simulation and Computation*, 33, 19-33.

Headrick, T. C. & Beasley, T. M. (April, 2003). A method for simulating correlated structures of continuous and ranked data. American Educational Research Association: Chicago.

Headrick, T. C. & Rotou (2001). An investigation of the rank transformation in multiple regression. *Computational Statistics and Data Analysis*, 38, 203-215.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated non-normal distributions: extending the Fleishman power method. *Psychometrika*, 64, 25-35.

Headrick, T. C. & Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics: Simulation and Computation*, 29, 1059-1088.

Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, 25, 417-436.

Headrick, T. C. & Vineyard, G. (2001). An empirical investigation of four tests for interaction in the context of factorial analysis of covariance. *Multiple Linear Regression Viewpoints*, 27, 3-15.

Headrick, T. C. (2004). Distribution theory for the power method transformation. The Psychometric Society: Monterey.

Klockars, A. J. & Moses, T. P. (2002). Type I error rates for rank-based tests of homogeneity of regression slopes. *Journal of Modern Applied Statistical Methods*, 1, 452-460.

Moran, P.A.P. (1948). Rank correlation and product-moment correlation. *Biometrika*, 35, 203-206

Shieh, Y., (April, 2000). The effects of distributional characteristics on multi-level modeling parameter estimates and Type I error control of parameter tests under conditions of non-normality. American Educational Research Association: New Orleans.

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

Whittaker, T., Fouladi, R. T., & Williams, N. (2002). Determining predictor importance in multiple regression under varied correlational and distributional conditions. *Journal of Modern Applied Statistical Methods*, 2, 354-366.

Wolfram, S., (2003). *The Mathematica book*, 4th Edition. Wolfram Media-Cambridge University Press, Cambridge.

An Alternative Q Chart Incorporating A Robust Estimator Of Scale

Michael B.C. Khoo
School of Mathematical Sciences
Universiti Sains Malaysia

In overcoming the shortcomings of the classical control charts in a short runs production, Quesenberry (1991 & 1995a – d) proposed Q charts for attributes and variables data. An approach to enhance the performance of a variable Q chart based on individual measurements using a robust estimator of scale is proposed. Monte carlo simulations are conducted to show that the proposed robust Q chart is superior to the present Q chart.

Key words: short runs; Q chart; Q_{MSSD} chart; in-control; out-of-control (o.o.c.)

Introduction

Short runs production or more commonly known as short runs is given more emphasis in manufacturing industries nowadays. The trend which is emphasized now is low volume production. This trend is a result of extra emphasis on just-in-time (JIT) techniques, job shop settings and synchronous manufacturing. Classical SPC charting methods such as $\bar{X}$, R and S charts which assume high volume manufacturing processes require at least 25 or 30 calibration samples of size 4 or 5 each to be available in the estimation of the process parameters before on-line charting begins.

In a short runs production, there is often a paucity of relevant data available for estimating the process parameters and establishing control limits prior to a production run. It is desirable to begin charting at or very near the beginning of the run in these cases.

Michael B. C. Khoo is a lecturer at the University of Science of Malaysia. His research interests are statistical process control and reliability analysis. He is a member of the editorial board of *Quality Engineering*.

In a short runs environment, the process mean, μ , and variance, σ^2 , cannot be known before the production run is begun because they change from run to run. For the individual measurements situation based on variables data, Quesenberry (1991 & 1995a) proposed the following four statistics for cases where μ and σ^2 are known and unknown. The notations in Table 1 are used:

1. Case KK: $\mu = \mu_0$, $\sigma = \sigma_0$, both known

$$Q_r(X_r) = \frac{X_r - \mu_0}{\sigma_0}, r = 1, 2, \dots \quad (1)$$

2. Case UK: μ unknown, $\sigma = \sigma_0$ known

$$Q_r(X_r) = \left(\frac{r-1}{r} \right)^{1/2} \frac{(X_r - \bar{X}_{r-1})}{\sigma_0}, r = 2, 3, \dots \quad (2)$$

3. Case KU: $\mu = \mu_0$ known, σ^2 unknown

For this case, let

$$S_{0,r}^2 = \frac{1}{r} \sum_{j=1}^r (X_j - \mu_0)^2. \text{ Thus,}$$

$$Q_r(X_r) = \Phi^{-1} \left\{ G_{r-1} \left(\frac{X_r - \mu_0}{S_{0,r-1}} \right) \right\}, r = 2, 3, \dots \quad (3)$$

4. Case UU: μ and σ^2 both unknown

$$Q_r(X_r) = \Phi^{-1} \left\{ G_{r-2} \left[\left(\frac{r-1}{r} \right)^{1/2} \left(\frac{X_r - \bar{X}_{r-1}}{S_{r-1}} \right) \right] \right\},$$

$$r = 3, 4, \dots \quad (4)$$

$$\text{where } S_r^2 = \frac{1}{r-1} \sum_{j=1}^r (X_j - \bar{X}_r)^2.$$

Note that in eqs. (1) – (4) above, $\bar{X}_r$ represents the sample mean estimated from the first r

$$\text{observations, i.e., } \bar{X}_r = \frac{\sum_{i=1}^r X_i}{r}.$$

Table 1. Notations for distribution functions.

$\Phi^{-1}(\cdot)$ - The inverse of the standard normal distribution function.
$G_v(\cdot)$ - The student- t distribution function with v degrees of freedom.

Q Chart Incorporating A Robust Estimator Of Scale

The Q chart statistics in eqs. (1) and (2) are based on known variance while that of eqs. (3) and (4) are based on the estimated variance. A simulation study performed by Quesenberry (1995a) shows that the performance of a Q chart for cases KK and UK are superior to that of cases KU and UU. In this paper, a method to improve the performance of a basic Q chart using a robust estimator of scale will be suggested. The robust estimator of scale is based on a modified mean square successive difference (MSSD) approach.

Holmes and Mergen (1993) provide some discussion on this approach. Let the new estimator of the process dispersion be denoted by S_{MSSD} while the new Q statistics be represented by Q_{MSSD} . The standard normal

Q_{MSSD} statistics are shown below for cases KU and UU. All the notations which are used here are similar to that defined in Table 1. Let $X_1, X_2, \dots$ represent measurements made on a sequence of parts as they are produced in time and assume that these values are independently and identically distributed (i.i.d.) with a normal, $N(\mu, \sigma^2)$ process distribution. The two cases are:

1. Case KU: $\mu = \mu_0$ known, σ^2 unknown

For odd numbered observations, i.e., when r is an odd number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD},r-1}} \right] \right\}, r = 3, 5, \dots \quad (5a)$$

$$\text{where } S_{\text{MSSD},r-1}^2 = \frac{2}{r-1} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2.$$

For even numbered observations, i.e., when r is an even number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD},r-2}} \right] \right\}, r = 4, 6, \dots \quad (5b)$$

$$\text{where } S_{\text{MSSD},r-2}^2 = \frac{2}{r-2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2.$$

2. Case UU: μ and σ^2 both unknown

For this case, let $\bar{X}_r = \frac{\sum_{i=1}^r X_i}{r}$. For odd numbered observations, i.e., when r is an odd number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-1}} \right] \right\}, r = 3, 5, \dots \quad (6a)$$

$$\text{where } S_{\text{MSSD},r-1}^2 = \frac{2}{r-1} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2.$$

For even numbered observations, i.e., when r is an even number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-2}} \right] \right\}, r = 4, 6, \dots \quad (6b)$$

$$\text{where } S_{\text{MSSD},r-2}^2 = \frac{2}{r-2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2.$$

Note that the Q_{MSSD} statistics in Eqs. (5a), (5b), (6a) and (6b) are standard normal random variables.

Tests For Shifts In The Process Mean

To enable a comparison in the performances of the proposed Q_{MSSD} chart and the basic Q chart to be made in the next section, the following tests which are used by Quesenberry (1995a) will be considered: Given a sequence of say, Q statistics, $Q_t, Q_{t-1}, \dots$, these tests are defined as follow:

The 1-of-1 test: When Q_t is plotted, the test signals an increase in μ if $Q_t > 3$, and signals a decrease in μ if $Q_t < -3$.

The 9-of-9 test: When Q_t is plotted, the test signals an increase in μ if $Q_t, Q_{t-1}, \dots, Q_{t-8}$ all exceed 0, and a decrease in μ if $Q_t, Q_{t-1}, \dots, Q_{t-8}$ are all less than 0. This test can only be employed if nine consecutive Q statistics are available.

The 3-of-3 test: When Q_t is plotted, the test signals an increase in μ if Q_t, Q_{t-1} and Q_{t-2} all exceed 1, and a decrease in μ if Q_t, Q_{t-1} and

Q_{t-2} are all less than -1 . This test can only be employed if three consecutive Q statistics are available.

The 4-of-5 test: When Q_t is plotted, the test signals an increase in μ if at least four of the five values $Q_t, Q_{t-1}, \dots, Q_{t-4}$ exceed 1, and a decrease in μ if at least four of the five values $Q_t, Q_{t-1}, \dots, Q_{t-4}$ are less than -1 . This test can only be employed if five consecutive Q statistics are available.

The EWMA test: The EWMA statistic Z_t is given by

$$Z_t = \lambda Q_t + (1 - \lambda) Z_{t-1}, \quad t = 1, 2, \dots \quad (7)$$

with $Z_0 = 0$. The $Z_t, t = 1, 2, \dots$, values are plotted on a chart with limits at $\pm K \sqrt{\lambda/(2-\lambda)}$.

The same values of $(\lambda, K) = (0.25, 2.90)$ considered by Quesenberry (1995a) which gives control limits at ± 1.096 are used in the next section. These limits give an in-control ARL of 372.6. If $Z_t > 1.096$, an increase in μ is signaled and if $Z_t < -1.096$, a decrease in μ is signaled.

The CUSUM test: The CUSUM statistics are defined as follow:

$$S_t^+ = \max\{0, S_{t-1}^+ + Q_t - k_s\} \quad (8a)$$

and

$$S_t^- = \min\{0, S_{t-1}^- + Q_t + k_s\} \quad (8b)$$

where $S_0^+ = S_0^- = 0$. An increase in μ is detected if $S_t^+ > h_s$ and a decrease in μ if $S_t^- < -h_s$. Similar to Quesenberry (1995a), the values of $k_s = 0.75$ and $h_s = 3.34$ are used in this study. These values of k_s and h_s give an in-control ARL of 370.5.

Evaluating The Performance Of The Q_{MSSD} Chart

A Monte Carlo simulation study is conducted using SAS version 8 to study the performance of the Q_{MSSD} chart based on cases KU and UU. Let the in-control mean be μ_0 and the o.o.c. mean be $\mu_s = \mu_0 + \delta\sigma$. Here, μ shifts

from μ_0 to μ_s after $r = c$, where $c \in \{5, 20, 100\}$ and $\delta \in \{0, 0.5, 1, 1.5, 2, 3, 4, 5, 6\}$. c observations are generated from a $N(0,1)$ distribution and then 30 additional observations from a $N(\delta,1)$ distribution. The Q_{MSSD} statistics in Eqs. (5a), (5b), (6a) and (6b) are computed as soon as enough values are available to define the particular statistic. This procedure is repeated 5000 times and the proportions of times a signal is observed for the Q_{MSSD} chart from $c + 1$ to $c + 30$, for the first time are recorded and are given in Tables 2, 3 and 4 for $c = 5, 20$ and 100 respectively. Note that this simulation study is conducted under the same condition as that in Quesenberry (1995a) so that a comparison between the performances of the Q_{MSSD} chart and the basic Q chart can be made.

All the six tests discussed in the previous section are used in the simulation study. The results of the six tests for the basic Q chart are obtained from Quesenberry (1995a). Since we are interested to detect positive shifts in the process mean, only the upper sided tests are considered. Here, an o.o.c. is signaled if the chart's statistics plot above the upper control limit of the test.

The results in Tables 2, 3 and 4 show that the proportions of o.o.c. for the 1-of-1 test when $\delta = 0$ for both the Q_{MSSD} and Q charts are about the same, thus the two charts have almost similar Type-I errors. For the 1-of-1 test where δ

> 0 , the Q_{MSSD} chart has higher o.o.c. proportions compared to the Q chart. Therefore, the new approach based on the 1-of-1 test has increased the sensitivity of the basic Q chart while maintaining the same rate of Type-I error. On the whole, the Type-I errors of the 9-of-9, 3-of-3, 4-of-5, EWMA and CUSUM tests for $c = 5$ and 20 are higher for the Q_{MSSD} chart compared to the Q chart. However, for $c = 100$, these tests produce similar Type-I errors for both the Q_{MSSD} and Q charts.

The percentage of an increase in the sensitivity of the Q_{MSSD} chart using these tests in detecting shifts in the mean is greater than its increase in the Type-I error in comparison to the Q chart. For example, in Table 2, using the 3-of-3 test for case UU with $\delta = 0.5$, an increase of about 2.7 fold in the sensitivity of the Q_{MSSD} chart over the Q chart is recorded. For this case, the false alarm rate of the Q_{MSSD} chart increase by only about 1.9 fold in comparison to that of the Q chart. The results in Tables 2, 3 and 4 clearly indicate that the Q_{MSSD} chart is superior to the Q chart.

It should be noted that case UU has lower o.o.c. proportions than case KU for $\delta > 0$ irrespective of the test that is used. The Q_{MSSD} and Q charts become more sensitive to process shifts as the value of c increases.

Table 4. Proportions of signaling an out-of-control for the Q and Q_{MSSD} charts based on $c = 100$.

δ		$c = 100$											
		1-of-1		9-of-9		3-of-3		4-of-5		EWMA		CUSUM	
		Q	Q_{MSSD}	Q	Q_{MSSD}	Q	Q_{MSSD}	Q	Q_{MSSD}	Z	Z_{MSSD}	S^+	S^+_{MSSD}
0.0	KU	0.042	0.040	0.026	0.030	0.101	0.099	0.048	0.052	0.039	0.046	0.041	0.050
	UU	0.042	0.039	0.027	0.033	0.100	0.101	0.051	0.055	0.036	0.053	0.038	0.057
0.5	KU	0.150	0.156	0.276	0.281	0.458	0.457	0.366	0.375	0.486	0.484	0.426	0.430
	UU	0.133	0.153	0.228	0.286	0.394	0.453	0.308	0.379	0.394	0.484	0.337	0.435
1.0	KU	0.352	0.435	0.773	0.780	0.892	0.903	0.866	0.888	0.970	0.973	0.958	0.960
	UU	0.295	0.430	0.661	0.770	0.812	0.888	0.773	0.872	0.910	0.961	0.891	0.948
1.5	KU	0.614	0.787	0.977	0.981	0.994	0.997	0.993	0.996	1.000	1.000	1.000	1.000
	UU	0.532	0.776	0.940	0.982	0.977	0.996	0.976	0.996	0.999	1.000	0.999	1.000
2.0	KU	0.850	0.967	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.771	0.963	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3.0	KU	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	0.988	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4.0	KU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.0	KU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.0	KU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	UU	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

An Example of Application

This example is based on simulated data. Observations are generated using SAS version 8. The first 10 observations are generated from a standard normal, $N(0,1)$ distribution followed by 20 additional observations from a $N(1,1)$ distribution. The first 10 observations represent the in-control (stable) process while the next 20 the out-of-control (o.o.c.) process. The o.o.c. process involves a shift of one standard deviation in the mean. The simulated data and the corresponding computed statistics of $Q_r(X_r)$, Z_r , S_r^+ , $Q_{\text{MSSD},r}(X_r)$, $Z_{\text{MSSD},r}$ and $S_{\text{MSSD},r}^+$ are given in Table 5. The control charts plotted from the Q , EWMA (Z) and CUSUM (S^+) statistics are shown in Figure 1. Figure 2 gives the control charts plotted from the Q_{MSSD} , EWMA (Z_{MSSD}) and CUSUM (S_{MSSD}^+) statistics. All the six tests which are considered in the simulation study will be used

here. Because a positive shift is simulated, only the upper limits of each of the tests are used. The upper limits of 1.096 and 3.34 for the EWMA and CUSUM tests respectively are used. These upper limits correspond to an in-control ARL of approximately 370 for the case of the two-sided charts.

An o.o.c. signal is given by the 3-of-3 test at observation 25 in Figure 1. Note that the other tests fail to detect an o.o.c. in Figure 1. In Figure 2, o.o.c. signals are detected for the first time at observation 17 by the 3-of-3 test and at observation 19 by the 4-of-5 and 9-of-9 tests. Here, the EWMA (Z_{MSSD}) and CUSUM (S_{MSSD}^+) tests issue o.o.c. signals for the first time at observations 25 and 26 respectively. This example shows that the Q_{MSSD} chart is more sensitive to shifts compared to the Q chart proposed by Quesenberry (1991 & 1995a).

Table 5. Simulated data and the computed statistics for 30 observations.

Obs. No., r	X_r	$Q_r(X_r)$	EWMA (Z_r)	CUSUM (S_r^+)	$Q_{\text{MSSD},r}(X_r)$	EWMA ($Z_{\text{MSSD},r}$)	CUSUM ($S_{\text{MSSD},r}^+$)
1	-0.862	—	—	—	—	—	—
2	2.519	—	—	—	—	—	—
3	-1.350	-0.700	-0.175	0	-0.535	-0.134	0
4	-0.332	-0.346	-0.218	0	-0.125	-0.132	0
5	0.228	0.239	-0.104	0	0.105	-0.072	0
6	-1.499	-1.571	-0.470	0	-0.660	-0.219	0
7	0.312	0.515	-0.224	0	0.278	-0.095	0
8	0.384	0.556	-0.029	0	0.280	-0.001	0
9	-0.162	-0.100	-0.047	0	-0.056	-0.015	0
10	-2.233	-2.232	-0.593	0	-1.236	-0.320	0
11	0.972	1.160	-0.155	0.410	0.793	-0.042	0.043
12	2.524	2.209	0.436	1.869	1.588	0.366	0.881
13	0.350	0.234	0.386	1.353	0.437	0.383	0.568
14	0.457	0.311	0.367	0.914	0.509	0.415	0.327
15	1.206	0.906	0.502	1.071	1.068	0.578	0.645
16	1.845	1.354	0.715	1.675	1.466	0.800	1.360
17	2.349	1.608	0.938	2.533	1.863	1.066	2.473
18	0.301	-0.072	0.686	1.711	0.473	0.918	2.196
19	1.317	0.730	0.697	1.691	1.203	0.989	2.649
20	0.148	-0.234	0.464	0.707	0.346	0.828	2.246
21	0.638	0.177	0.392	0.135	0.742	0.807	2.238
22	-1.656	-1.709	-0.133	0	-1.058	0.340	0.429
23	1.640	1.038	0.160	0.288	1.433	0.614	1.112
24	2.245	1.457	0.484	0.996	1.824	0.916	2.186
25	1.871	1.086	0.635	1.332	1.654	1.100	3.090
26	1.390	0.672	0.644	1.254	1.317	1.155	3.657
27	1.690	0.888	0.705	1.392	1.592	1.264	4.498
28	3.085	1.909	1.006	2.551	2.505	1.574	6.253
29	0.717	0.019	0.759	1.820	0.855	1.394	6.358
30	1.278	0.451	0.682	1.521	1.303	1.372	6.912

Figure 1. The charts constructed from the Q , EWMA (Z) and CUSUM (S^+) statistics in Table 5.

Figure 2. The charts constructed from the Q_{MSSD} , EWMA (Z_{MSSD}) and CUSUM (S_{MSSD}^+) statistics in Table 5.

Conclusion

This article demonstrates that the performance of the basic Q chart for cases KU and UU have improved tremendously by incorporating a robust estimator of scale based on a modified mean square successive difference approach. The proofs of how the Q_{MSSD} statistics in eqs. (5a), (5b), (6a) and (6b) which are i.i.d. standard normal random variables are derived will be given in the Appendix.

References

Holmes, D. S., & Mergen, A. E. (1993). Improving the performance of the T^2 control chart. *Quality Engineering*, 5, 619 – 625.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Quesenberry, C. P. (1991). SPC Q charts for start-up processes and short or long runs. *Journal of Quality Technology*, 23, 213 – 224.

Quesenberry, C. P. (1995a). On properties of Q charts for variables. *Journal of Quality Technology*, 27, 184 – 203.

Quesenberry, C. P. (1995b). On properties of binomial Q charts for attributes. *Journal of Quality Technology*, 27, 204 – 213.

Quesenberry, C. P. (1995c). On properties of Poisson Q charts for attributes. *Journal of Quality Technology*, 27, 293 – 303.

Quesenberry, C. P. (1995d). Geometric Q charts for high quality processes. *Journal of Quality Technology*, 27, 304 – 315.

Appendix

Notation used here were defined above. The following theorem (Mood, Graybill & Boes, 1974) is required.

Theorem:

If Z has a standard normal distribution and V has a chi-square distribution with k degrees of freedom, and Z and V are independent, then

$$T = \frac{Z}{\sqrt{V/k}} \quad (9)$$

has a student- t distribution with k degrees of freedom.

Equation 5(a): Case KU

For odd numbered observations, i.e., when r is an odd number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD},r-1}} \right] \right\} \sim N(0,1), \quad r = 3, 5, \dots$$

Proof:

If $X_j, j = 1, 2, \dots$, are i.i.d. $N(\mu_0, \sigma^2)$ variables, then $X_i - X_{i-1} \sim N(0, 2\sigma^2), i = 2, 4, \dots$

and $\frac{1}{\sigma\sqrt{2}}(X_i - X_{i-1}) \sim N(0, 1), i = 2, 4, \dots$ Thus, $\frac{1}{2\sigma^2}(X_i - X_{i-1})^2 \sim \chi_1^2$

and $\frac{1}{2\sigma^2} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2 \sim \chi_{\frac{r-1}{2}}^2$. Let $S_{\text{MSSD}, r-1}^2 = \frac{2}{r-1} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2$

then, $S_{\text{MSSD}, r-1}^2 \sim \frac{4\sigma^2}{r-1} \chi_{\frac{r-1}{2}}^2$. Because $\frac{1}{\sigma}(X_r - \mu_0) \sim N(0, 1)$

then, from Eq. (9) in the above theorem, $\frac{\frac{1}{\sigma}(X_r - \mu_0)}{\sqrt{\left(\frac{r-1}{4\sigma^2}\right) S_{\text{MSSD}, r-1}^2 / \left(\frac{r-1}{2}\right)}} \sim t_{\frac{r-1}{2}},$

i.e., $\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD}, r-1}} \sim t_{\frac{r-1}{2}}$. Thus, $Q_{\text{MSSD}, r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD}, r-1}} \right] \right\} \sim N(0, 1), r = 3, 5, \dots$

Equation 5(b): Case KU

It must be shown that for even numbered observations, i.e., when r is an even number,

$$Q_{\text{MSSD}, r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD}, r-2}} \right] \right\} \sim N(0, 1), r = 4, 6, \dots$$

Proof:

If $X_j, j = 1, 2, \dots$, are i.i.d. $N(\mu_0, \sigma^2)$ variables, then $X_i - X_{i-1} \sim N(0, 2\sigma^2), i = 2, 4, \dots$

and $\frac{1}{\sigma\sqrt{2}}(X_i - X_{i-1}) \sim N(0, 1), i = 2, 4, \dots$ Thus, $\frac{1}{2\sigma^2}(X_i - X_{i-1})^2 \sim \chi_1^2$

and $\frac{1}{2\sigma^2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2 \sim \chi_{\frac{r-2}{2}}^2$. Let $S_{\text{MSSD}, r-2}^2 = \frac{2}{r-2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2$

then, $S_{\text{MSSD}, r-2}^2 \sim \frac{4\sigma^2}{r-2} \chi_{\frac{r-2}{2}}^2$.

Because $\frac{1}{\sigma}(X_r - \mu_0) \sim N(0, 1)$, then, from Eq. (9) in the above theorem, $\frac{\frac{1}{\sigma}(X_r - \mu_0)}{\sqrt{\left(\frac{r-2}{4\sigma^2}\right) S_{\text{MSSD}, r-2}^2 / \left(\frac{r-2}{2}\right)}} \sim$

$t_{\frac{r-2}{2}}$. i.e., $\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD}, r-2}} \sim t_{\frac{r-2}{2}}$. Thus, $Q_{\text{MSSD}, r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{2}(X_r - \mu_0)}{S_{\text{MSSD}, r-2}} \right] \right\} \sim N(0, 1), r = 4, 6, \dots$

Equation 6(a): Case UU

It must be shown that for odd numbered observations, i.e., when r is an odd number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-1}} \right] \right\} \sim N(0,1), r = 3, 5, \dots$$

Proof:

If $X_j, j = 1, 2, \dots$, are i.i.d. $N(\mu_0, \sigma^2)$ variables, then $X_i - X_{i-1} \sim N(0, 2\sigma^2), i = 2, 4, \dots$

and $\frac{1}{\sigma\sqrt{2}}(X_i - X_{i-1}) \sim N(0,1), i = 2, 4, \dots$ Thus, $\frac{1}{2\sigma^2}(X_i - X_{i-1})^2 \sim \chi_1^2$

and $\frac{1}{2\sigma^2} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2 \sim \chi_{\frac{r-1}{2}}^2$.

$$\text{Let } S_{\text{MSSD},r-1}^2 = \frac{2}{r-1} \sum_{i=2,4,6}^{r-1} (X_i - X_{i-1})^2 \text{ then, } S_{\text{MSSD},r-1}^2 \sim \frac{4\sigma^2}{r-1} \chi_{\frac{r-1}{2}}^2.$$

Because $\bar{X}_{r-1} \sim N\left(\mu_0, \frac{\sigma^2}{r-1}\right)$ then, $\frac{1}{\sigma} \sqrt{\frac{r-1}{r}} (X_r - \bar{X}_{r-1}) \sim N(0,1)$

from Eq. (9) in the above theorem, $\frac{\frac{1}{\sigma} \sqrt{\frac{r-1}{r}} (X_r - \bar{X}_{r-1})}{\sqrt{\left(\frac{r-1}{4\sigma^2}\right) S_{\text{MSSD},r-1}^2 / \left(\frac{r-1}{2}\right)}} \sim t_{\frac{r-1}{2}}$. i.e., $\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-1}} \sim t_{\frac{r-1}{2}}$.

$$\text{Thus, } Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-1}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-1}} \right] \right\} \sim N(0,1), r = 3, 5, \dots$$

Equation 6(b): Case UU

It must be shown that for even numbered observations, i.e., when r is an even number,

$$Q_{\text{MSSD},r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD},r-2}} \right] \right\} \sim N(0,1), r = 4, 6, \dots$$

Proof:

If $X_j, j = 1, 2, \dots$, are i.i.d. $N(\mu_0, \sigma^2)$ variables, then $X_i - X_{i-1} \sim N(0, 2\sigma^2), i = 2, 4, \dots$

and $\frac{1}{\sigma\sqrt{2}}(X_i - X_{i-1}) \sim N(0,1), i = 2, 4, \dots$ Thus, $\frac{1}{2\sigma^2}(X_i - X_{i-1})^2 \sim \chi_1^2$

and $\frac{1}{2\sigma^2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2 \sim \chi_{\frac{r-2}{2}}^2$.

$$\text{Let } S_{\text{MSSD},r-2}^2 = \frac{2}{r-2} \sum_{i=2,4,6}^{r-2} (X_i - X_{i-1})^2 \text{ then, } S_{\text{MSSD},r-2}^2 \sim \frac{4\sigma^2}{r-2} \chi_{\frac{r-2}{2}}^2.$$

Because $\bar{X}_{r-1} \sim N\left(\mu_0, \frac{\sigma^2}{r-1}\right)$ then, $\frac{1}{\sigma} \sqrt{\frac{r-1}{r}} (X_r - \bar{X}_{r-1}) \sim N(0,1)$,

from Eq. (9) in the above theorem, $\frac{\frac{1}{\sigma} \sqrt{\frac{r-1}{r}} (X_r - \bar{X}_{r-1})}{\sqrt{\left(\frac{r-2}{4\sigma^2}\right) S_{\text{MSSD}, r-2}^2 / \left(\frac{r-2}{2}\right)}} \sim t_{\frac{r-2}{2}} \text{ i.e.,}$

$\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD}, r-2}} \sim t_{\frac{r-2}{2}} \text{ . Thus, } Q_{\text{MSSD}, r}(X_r) = \Phi^{-1} \left\{ G_{\frac{r-2}{2}} \left[\frac{\sqrt{\frac{2(r-1)}{r}} (X_r - \bar{X}_{r-1})}{S_{\text{MSSD}, r-2}} \right] \right\} \sim N(0,1), r = 4, 6, \dots$

Beta-Normal Distribution: Bimodality Properties and Application

Felix Famoye
Department of Mathematics
Central Michigan University

Carl Lee
Department of Mathematics
Central Michigan University

Nicholas Eugene
Computer Science & Mathematics
Coppin State College

The beta-normal distribution is characterized by four parameters that jointly describe the location, the scale and the shape properties. The beta-normal distribution can be unimodal or bimodal. This paper studies the bimodality properties of the beta-normal distribution. The region of bimodality in the parameter space is obtained. The beta-normal distribution is applied to fit a numerical bimodal data set. The beta-normal fits are compared with the fits of mixture-normal distribution through simulation.

Key words: Bimodal region, percentiles, curve estimation, egg size distribution

Introduction

Bimodal distributions occur in many areas of science. Withington et al. (2000), in their study of cardiopulmonary bypass in infants showed that plasma vecuronium and vecuronium clearance requirements have bimodal distributions. They concluded that their findings on bimodal distributions for plasma vecuronium and vecuronium clearance requirements highlight the need for individual monitoring of neuromuscular blockade. Espinoza et al. (2001) discussed the importance of bimodal distributions in the study of size distribution of metals in aerosols. Bimodal distributions also occur in the study of genetic diversity (Freeland et al., 2000), in the study of agricultural farm size distribution (Wolf & Sumner, 2001), in the study of atmospheric pressure (Zangvil et al., 2001), and in the study of anabolic steroids on animals (Isaacson, 2000).

Let $F(x)$ be the cumulative distribution function (CDF) of a random variable X . The cumulative distribution function for a generalized class of distributions for the random variable X can be defined as the logit of the beta random variable given by

$$G(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^{F(x)} t^{\alpha-1}(1-t)^{\beta-1} dt, \quad 0 < \alpha, \beta < \infty. \quad (1.1)$$

Eugene et al. (2002) considered $F(x)$ as the CDF of the normal distribution with parameters μ and σ . Thus, the random variable X has the beta-normal distribution with probability density function (pdf)

$$g(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \left[\Phi\left(\frac{x-\mu}{\sigma}\right) \right]^{\alpha-1} \left[1 - \Phi\left(\frac{x-\mu}{\sigma}\right) \right]^{\beta-1} \sigma^{-1} \phi\left(\frac{x-\mu}{\sigma}\right) \quad (1.2)$$

where $\phi\left(\frac{x-\mu}{\sigma}\right)$ is the normal pdf and

$\Phi\left(\frac{x-\mu}{\sigma}\right)$ is the normal CDF. We denote the

beta-normal distribution with parameters $\alpha, \beta, \mu,$ and σ as $BN(\alpha, \beta, \mu, \sigma)$.

The distribution in (1.2) may be symmetric, skewed to the left, or skewed to the

Acknowledgement

The first two authors gratefully acknowledge the support received from the President Research Investment Fund at Central Michigan University under the Grant #65139. Address correspondence to the first author at Department of Mathematics, Central Michigan University Mt. Pleasant, Michigan 48859-0001. Email: felix.famoye@cmich.edu. Telephone: (989) 774-5497.

right. The distribution may be unimodal or bimodal. Eugene et al. (2002) discussed the shape properties of the unimodal beta-normal distribution. Furthermore, they considered the estimation of its parameters by the method of maximum likelihood.

In the analysis of bimodal data, a mixture of two normal densities is often used as a model (e.g., Cobb et al., 1983). The mixture of normal distribution is used as a model to analyze bimodal data because the mixture of normal densities can take on bimodal shapes depending on the parameters of the distribution. Eisenberger (1964) showed how the parameters of a mixture of normal distributions determine its shape. When a mixture assumption is not required or justified the beta-normal distribution can serve as a model to analyze data since only one distribution has to be used and one less parameter to estimate.

In the rest of the paper, we provide some bimodality properties of the beta-normal distribution. We obtain the region of bimodality in the parameter space. We also illustrate the application of beta-normal distribution to a numerical data set that exhibits two modes and compare the fit with mixture-normal distribution. A simulation study is conducted to compare the performance between beta-normal and mixture-normal distributions in fitting bimodal data.

Bimodality Properties

In this section, some results on the bimodality properties of beta-normal distribution are obtained.

Fact: A mode of the $BN(\alpha, \beta, \mu, \sigma)$ is any point $x_0 = x_0(\alpha, \beta)$ that satisfies

$$x_0 = \frac{\sigma \phi\left(\frac{x_0 - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)} \{2 - \alpha - \beta\} + \frac{(\alpha - 1) \phi\left(\frac{x_0 - \mu}{\sigma}\right) \sigma}{\Phi\left(\frac{x_0 - \mu}{\sigma}\right) [1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)]} + \mu. \quad (2.1)$$

Proof: Differentiating $BN(\alpha, \beta, \mu, \sigma)$ in (1.2) with respect to x , setting it equal to zero, and solving for x gives the result in (2.1).

Corollary 1: If $\alpha = \beta$ and one mode of $BN(\alpha, \beta, \mu, \sigma)$ is at x_0 , then the other mode is at the point $2\mu - x_0$.

Proof: If $BN(\alpha, \beta, \mu, \sigma)$ is unimodal, then the only mode occurs at the point $x_0 = \mu$. For bimodal case, we need to show that if we replace x_0 with $2\mu - x_0$, then equation (2.1) remains the same. When $\alpha = \beta$, equation (2.1) becomes

$$x_0 = \frac{\sigma \phi\left(\frac{x_0 - \mu}{\sigma}\right)(\alpha - 1)}{\Phi\left(\frac{x_0 - \mu}{\sigma}\right) [1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)]} \left\{1 - 2\Phi\left(\frac{x_0 - \mu}{\sigma}\right)\right\} + \mu. \quad (2.2)$$

If x_0 in (2.2) is replaced with $2\mu - x_0$, we obtain

$$2\mu - x_0 = \frac{\sigma \phi\left(\frac{\mu - x_0}{\sigma}\right)(\alpha - 1)}{\Phi\left(\frac{\mu - x_0}{\sigma}\right) [1 - \Phi\left(\frac{\mu - x_0}{\sigma}\right)]} \left\{1 - 2\Phi\left(\frac{\mu - x_0}{\sigma}\right)\right\} + \mu \quad (2.3)$$

By using $\phi\left(\frac{x_0 - \mu}{\sigma}\right) = \phi\left(\frac{\mu - x_0}{\sigma}\right)$ and $\Phi\left(\frac{x_0 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu - x_0}{\sigma}\right)$ in (2.3) and on simplification, we get the result in (2.2).

Corollary 2: If $BN(\alpha, \beta, \mu, \sigma)$ has a mode at x_0 , then $BN(\beta, \alpha, \mu, \sigma)$ has a mode at $2\mu - x_0$.

Proof: We need to show that if we replace α with β , and $2\mu - x_0$ with x_0 , equation (2.1) remains the same. Equation (2.1) can be written as

$$x_0 = \frac{\sigma \phi\left(\frac{x_0 - \mu}{\sigma}\right)}{\Phi\left(\frac{x_0 - \mu}{\sigma}\right) [1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)]} \left\{(2 - \alpha - \beta) \Phi\left(\frac{x_0 - \mu}{\sigma}\right) + (\alpha - 1)\right\} + \mu. \quad (2.4)$$

If x_0 is replaced with $2\mu - x_0$ and α is replaced with β in (2.4), using $\phi\left(\frac{x_0 - \mu}{\sigma}\right) = \phi\left(\frac{\mu - x_0}{\sigma}\right)$ and $\Phi\left(\frac{x_0 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu - x_0}{\sigma}\right)$, and on simplification, we obtain the result in (2.4).

Corollary 3: The modal point $x_0(\alpha, \beta)$ is an increasing function of α and a decreasing function of β .

Proof: Differentiating the result in (2.1) with respect to α and β gives

$$\frac{\partial x_0(\alpha, \beta)}{\partial \alpha} = \frac{\sigma \phi\left(\frac{x_0 - \mu}{\sigma}\right)}{\Phi\left(\frac{x_0 - \mu}{\sigma}\right)} > 0$$

and

$$\frac{\partial x_0(\alpha, \beta)}{\partial \beta} = \frac{-\sigma \phi\left(\frac{x_0 - \mu}{\sigma}\right)}{1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right)} < 0.$$

Hence $x_0(\alpha, \beta)$ is an increasing function of α and a decreasing function of β .

Eugene et al. (2002) showed that the beta-normal distribution is symmetric about μ when $\alpha = \beta$. From this result and corollary 3, the modal value is greater than μ if $\alpha > \beta$. Also, the modal value is less than μ if $\alpha < \beta$. The beta-normal distribution has a very distinct property in that it can be used to describe both bimodal and unimodal data.

Region of Bimodality

The beta-normal distribution becomes bimodal for certain values of the parameters α and β , and the analytical solution of α and β , where the distribution becomes bimodal, cannot be solved algebraically. A numerical solution is obtained, however, by solving the number of roots of the derivative of $BN(\alpha, \beta, \mu, \sigma)$. Table 1 shows a grid of values where the distribution is bimodal. The “2” in Table 1 indicates that the beta-normal distribution has two turning points which implies that the distribution is bimodal and the “1” indicates that the beta-normal distribution has one turning point which implies that the distribution is unimodal.

Table 1. Number of turning points of $BN(\alpha, \beta, 0, 1)$ for various values of α and β

	Beta	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.02	.21	.22
Alpha																							
.01		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.02		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.03		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.04		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.05		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.06		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.07		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1
.08		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
.09		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
.1		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
.11		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
.12		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1
.13		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
.14		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
.15		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
.16		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1	1
.17		1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1
.18		1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	1	1	1
.19		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	1	1
.2		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1
.21		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
.22		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.23		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.24		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.25		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.26		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: “2” indicates where bimodality occurs and “1” indicates where unimodality occurs

Numerically, the largest value of α or β that gives bimodal property is approximately 0.214. Figure 1 shows a plot of the boundary region of α and β values where $BN(\alpha, \beta, 0, 1)$ is bimodal.

Figure 1. Plot of bimodal region for beta-normal distribution $BN(\alpha, \beta, 0, 1)$

Corollary 4: The bimodal property of $BN(\alpha, \beta, \mu, \sigma)$ is independent of the parameters μ and σ .

Proof: The mode(s) of $BN(\alpha, \beta, \mu, \sigma)$ is at the point $x_0 = x_0(\alpha, \beta)$ given in (2.1). On re-writing (2.1), one obtains (2.4). On taking the μ on the right hand side of (2.4) to the left hand side, dividing through by σ , and replacing $(x_0 - \mu)/\sigma$ by z_0 , one obtains

$$z_0 = \frac{\phi(z_0)}{\Phi(z_0)[1 - \Phi(z_0)]} \{ (2 - \alpha - \beta)\Phi(z_0) + (\alpha - 1) \} \quad (2.5)$$

which is independent of parameters μ and σ .

In corollary 4, we showed that the bimodal property of $BN(\alpha, \beta, \mu, \sigma)$ is robust against the parameters μ and σ . In other words, regardless of the values of μ and σ , the α and β range for the bimodality of $BN(\alpha, \beta, \mu, \sigma)$ remains the same. To get more accurate values

of the pairs of (α, β) values that lie on the boundary of the region where the beta-normal distribution becomes bimodal, regression lines were drawn to estimate each boundary. The regression line that traced the boundaries of Figure 1 was approximated using curve estimation. For the values of α in the interval $[0.01, 0.1943)$, the values of β at the upper boundary in Figure 1 were estimated by $\hat{\beta} = 0.8591\alpha^2 + 0.0453\alpha + 0.1603$. For α in the interval $[0.1943, 0.214]$ we estimated β values by $\hat{\beta} = 4.4113\alpha^2 - 1.1966\alpha + 0.2675$. For the values of α in the interval $[0.16, 0.1785)$, the values of β at the lower boundary were estimated by

$$\hat{\beta} = -116.15\alpha^2 + 45.4657\alpha - 4.2908.$$

For α in the interval $[0.1785, 0.214]$ we obtained the equation

$$\hat{\beta} = -41.972\alpha^2 + 18.9913\alpha - 1.9281$$

to estimate the value of β .

If $BN(\alpha, \beta, \mu, \sigma)$ is unimodal, the distribution is skewed to the right whenever $\alpha > \beta$ and it is skewed to the left whenever $\alpha < \beta$. If $BN(\alpha, \beta, \mu, \sigma)$ is bimodal, the distribution is skewed to the right when $\alpha < \beta$ and it is skewed to the left when $\alpha > \beta$. Thus, the beta-normal distribution provides great flexibility in modeling symmetric, skewed and bimodal distributions.

Percentile of beta-normal distribution

Let $CBN(t)$ denote the cumulative probability of the beta-normal distribution up to a point t , which is given by

$$CBN(t) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{-\infty}^t [\Phi(\frac{x - \mu}{\sigma})]^{\alpha-1} [1 - \Phi(\frac{x - \mu}{\sigma})]^{\beta-1} \sigma^{-1} \phi(\frac{x - \mu}{\sigma}) dx. \quad (2.6)$$

The percentiles in Table 2 are computed by solving (2.6) for t such that $CBN(t)$ takes the values 0.50, 0.75, 0.90, 0.95, and 0.99.

When $\beta = 1$, the result in (2.6) reduces to

$$CBN(t) = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \int_{-\infty}^t [\Phi(\frac{x - \mu}{\sigma})]^{\alpha-1} \sigma^{-1} \phi(\frac{x - \mu}{\sigma}) dx = [\Phi(\frac{t - \mu}{\sigma})]^\alpha.$$

When $\alpha = 1$, (2.6) becomes

$$CBN(t) = \frac{\Gamma(\beta + 1)}{\Gamma(\beta)} \int_{-\infty}^t [1 - \Phi(\frac{x - \mu}{\sigma})]^{\beta-1} \sigma^{-1} \phi(\frac{x - \mu}{\sigma}) dx = 1 - [1 - \Phi(\frac{t - \mu}{\sigma})]^\beta.$$

Notice that if we compare the values of the mean of the unimodal beta-normal distribution with its median in Table 2, the mean of the beta-normal distribution is always greater than its median whenever $\alpha > \beta$. When the distribution is bimodal the mean of the beta-normal distribution is less than its median whenever $\alpha > \beta$. The percentiles in Table 2 are clearly increasing functions of α and decreasing functions of β . A graph of α versus the median (50th percentile) is plotted for $\beta = 0.1, 0.5, 1$, and 10 in Figure 2(a). Similar graphs for the 75th and 90th percentiles show the same pattern in Figure 2(b) and Figure 2(c) respectively.

Table 2. Mean and percentiles of $BN(\alpha, \beta, 0,1)$ for different values of α and β

α	β	Mean	0.50	0.75	0.90	0.95	0.99
0.05	0.05	0.0000	0.0000	4.7784	7.6544	9.2705	12.2382
	0.10	-2.2859	-2.7789	1.6446	4.3972	5.7191	8.0155
	0.50	-4.6527	-4.4917	-2.2674	-0.3400	0.7053	2.2795
	1.00	-5.0344	-4.4630	-2.2795	-1.1671	-0.3625	0.9074
	5.00	-5.5403	-5.1608	-3.3464	-2.1275	-1.5810	-0.9765
	10.00	-5.7016	-5.2981	-3.5470	-2.4100	-1.9186	-1.2382
0.10	0.05	2.2859	2.7789	5.8402	8.3706	9.8727	12.7024
	0.10	0.0000	0.0000	3.1398	5.2216	6.3834	8.5077
	0.50	-2.7633	-2.7086	-0.9199	0.6213	1.4295	2.7266
	1.00	-3.2639	-3.0973	-1.5865	-0.3889	0.2501	1.3069
	5.00	-3.9268	-3.6519	2.4222	-1.5832	-1.1778	-0.5432
	10.00	-4.1331	-3.8377	-2.6793	-1.9194	-1.5638	-1.0233
0.50	0.05	4.6527	4.4917	6.8631	9.1215	10.5189	13.2127
	0.10	2.7633	2.7086	4.5059	6.1677	7.1840	9.1287
	0.50	0.0000	0.0000	1.0518	1.9691	2.5031	3.4843
	1.00	-0.7043	-0.6745	0.1573	0.8779	1.2959	2.0558
	5.00	-1.7558	-1.6779	-1.1274	-0.6831	-0.4351	0.0028
	10.00	-2.0809	-1.9945	-1.5095	-1.1289	-0.9204	-0.5588
1.00	0.05	5.0344	4.7630	7.0477	9.2623	10.6416	13.3109
	0.10	3.2639	3.0973	4.7630	6.3613	7.3520	9.2623
	0.50	0.7043	0.6745	1.5341	2.3263	2.8070	3.7190
	1.00	0.0000	0.0000	0.6745	1.2816	1.6449	2.3264
	5.00	-1.1630	-1.1290	-0.6994	-0.3344	-0.1238	0.2582
	10.00	-1.5388	-1.4988	-1.1290	-0.8215	-0.6468	-0.3344
5.00	0.05	5.5403	5.1608	7.3275	9.4786	10.8308	13.4631
	0.10	3.9268	3.6519	5.1547	6.6640	7.6167	9.4751
	0.50	1.7558	1.6779	2.3038	2.9367	3.3438	4.1515
	1.00	1.1630	1.1290	1.5900	2.0365	2.3187	2.8769
	5.00	0.0000	0.0000	0.2742	0.5216	0.6702	0.9503
	10.00	-0.4556	-0.4517	-0.2233	-0.0201	0.1004	0.3251
10.0	0.05	5.7016	5.2981	7.4264	9.5558	10.8986	13.5179
	0.10	4.1331	3.8377	5.2916	6.7719	7.7117	9.5521
	0.50	2.0809	1.9945	2.5625	3.1526	3.5380	4.3130
	1.00	1.5388	1.4988	1.9055	2.3087	2.5679	3.0889
	5.00	0.4556	0.4517	0.6836	0.8960	1.0248	1.2706
	10.00	0.0000	0.0000	0.1915	0.3640	0.4075	0.6621

Figure 2(a). Plot of 50th percentile versus α for some β values Figure 2(b). Plot of 75th percentile versus α for some β values

Figure 2(c). Plot of 90th Percentile versus α for some β values

The percentiles increase very rapidly when α and β are less than 0.2. This rate of increase is due to the fact that the variation of the beta-normal distribution increases when α or β decreases. When α or β gets closer to 0.2, this variation decreases.

Application to Bimodal Data Egg Size Distribution

Sewell and Young (1997) studied the egg size distributions of echinoderm. In marine invertebrates, a species produces either many small eggs with planktotrophic development or fewer larger eggs with lecithotrophic development, Thorson (1950). The models developed by Vance (1973a, 1973b) viewed planktotrophy and lecithotrophy as extreme forms of larvae development. Subsequent modifications of these models (see references

in Sewell and Young, 1997) predict that eggs of marine invertebrates have bimodal distributions. Christiansen and Fenchel (1979) reported a bimodal distribution of egg sizes within prosobranchs. Emlet et al. (1987) described bimodal distributions in asteroid and echinoid echinoderms.

For echinoids and asteroids (see Tables 2 and 7 of Emlet et al., 1987), the egg diameters for species with planktotrophic larvae have less variation than species with lecithotrophic larvae (see Table 3). Because of this variation, the egg diameters appear to have one mode. However, with logarithmic transformation, the effect of large eggs in lecithotrophic species is reduced and the distribution of eggs becomes bimodal for both echinoids and asteroids. The transformation brings the modes nearer to each other and possibly makes their existence easier to detect.

Table 3. Descriptive statistics for asteroids species data

Types	n	Egg Diameter		Log Egg Diameter	
		Mean	SD	Mean	SD
Planktotrophic	35	153.11	34.26	5.01	0.23
Lecithotrophic	36	828.28	304.20	6.64	0.42
Brooding	17	1496.47	1066.58	7.05	0.77
All Types	88	688.83	705.59	6.07	0.99

Sewell and Young (1997) reported that many of the early studies used data sets that were not appropriate for a valid test of the egg size distribution patterns. They defined three criteria for appropriate data sets. The most widely cited example of bimodality in egg sizes is the data set compiled by Emlet et al. (1987). This data set satisfied the three criteria defined by Sewell and Young.

Sewell and Young (1997) reexamined the asteroid and echinoid egg size data in Emlet et al. (1987) with some additional data from more recent study. The additional data used by Sewell and Young were not available in their published article.

In this article, we have applied the beta-normal distribution to fit the logarithm of the egg diameters of the asteroids data in Emlet et al. (1987). The valid data consists of 88 asteroid species divided into three types consisting of 35 planktotrophic larvae, 36 lecithotrophic larvae, and 17 brooding larvae. These species are from a variety of habitats.

The maximum likelihood estimation method is used for parameter estimation. Eugene et al. (2002) gave the detailed discussion of this estimation technique. The parameter estimates for beta-normal distribution are $\hat{\alpha} = 0.0129$, $\hat{\beta} = 0.0070$, $\hat{\mu} = 5.7466$, and $\hat{\sigma} = 0.0675$. The estimates for α and β fall in the bimodal region in Figure 1. The log-likelihood value is -109.48 . By using the result in (2.1), the two modes for

the beta-normal distribution are at the points (log of egg diameters) 5.16 and 6.55.

A mixture of two normal distributions (Johnson et al. (1994) page 164) with parameters μ_1 , μ_2 , σ_1 , σ_2 , and p is fitted to the asteroids data. The maximum likelihood estimates for the parameters are $\hat{\mu}_1 = 5.0014$, $\hat{\mu}_2 = 6.7462$, $\hat{\sigma}_1 = 0.2232$, $\hat{\sigma}_2 = 0.6056$, and $\hat{p} = 0.3875$. The log-likelihood value for the mixture-normal is -101.31 . A histogram of the data with the beta-normal and mixture-normal distributions superimposed is presented in Figure 3.

The Kolmogorov-Smirnov test (see DeGroot & Schervish, 2002, p. 568) is used to compare the goodness of fit of beta-normal and mixture-normal distributions to the data. In Figure 4, the empirical CDF, the beta-normal CDF, and the mixture-normal CDF for the data are presented. The absolute maximum difference between the empirical cumulative distribution function and the beta-normal cumulative distribution function is $D_n^* = 0.1233$.

This provides a test statistic $\sqrt{n}D_n^* = 1.1570$ with a significance probability of 0.1370. The corresponding results for the mixture-normal distribution are $D_n^* = 0.0654$, $\sqrt{n}D_n^* = 0.6135$ with a p-value of 0.8459. Thus, both the beta-normal and mixture-normal distributions provide an adequate fit to the data. However, the mixture-normal appears to provide a better fit.

Figure 3. Histogram of asteroids data with beta-normal and mixture-normal superimposed

In examining the histogram for the log of egg diameter in Figure 3, both modes appear to have come from two symmetric distributions. This may explain in part why mixture-normal distribution provides a better fit than the beta-normal. Another reason is that mixture-normal has five parameters whereas the beta-normal has four parameters.

Test of Bimodality for the Egg Size Distribution Data

Schilling et al. (2002) derived a condition for the unimodality of mixtures of two normal distributions with unequal variances. If σ_1^2 and σ_2^2 are the variances of two normal

distributions with means μ_1 and μ_2 , the mixture is unimodal for any mixture proportion p if and only if $|\mu_2 - \mu_1| \leq S(r)[\sigma_1 + \sigma_2]$, where $r = \sigma_1^2 / \sigma_2^2$ and

$$S(r) = \left(-2 + 3r + 3r^2 - 2r^3 + 2(1 - r + r^2) \right)^{1/2} / \left(\sqrt{r}(1 + \sqrt{r}) \right).$$

From the fit of mixture-normal to the asteroids data, the parameter estimates gave $|\hat{\mu}_2 - \hat{\mu}_1| = 1.7$ and $S(r)[\hat{\sigma}_1 + \hat{\sigma}_2] = 0.56$. Thus, there is evidence that the parameter values do not lie in the region where the mixture is unimodal for any value of p .

Figure 4. Empirical, beta-normal, and mixture-normal CDF for asteroids data

A Comparison Between Beta-Normal and Mixture-Normal Distributions

A simulation study is conducted to compare the performance between beta-normal and mixture-normal for bimodal data. One hundred simulations, each with sample size $n = 400$, are conducted. In each simulation, data are generated from two Weibull distributions, $W(\lambda=2, \beta=5)$ and $W(\lambda=2, \beta=10)$, where λ and β are the scale and shape parameters respectively. Bimodal data are obtained from mixing the data from the two Weibull distributions in the form

$$[p_1 W(2, 5) + 10] + [(1 - p_1) W(2, 10)]. \quad (5.1)$$

The value 10 that is added to the first quantity in (5.1) is used to adjust the location of the modes. The different mixing proportions p_1 considered in the simulation study are 0.2, 0.3, 0.4, and 0.5.

A variety of other types of mixtures are also considered and the results are similar. Some of the simulations failed due to numerical

difficulty in estimating the beta-normal parameters using S-PLUS on personal computer. The main difficulty is that the optimization algorithm in S-PLUS failed to converge. There is a need for better algorithms to solve this numerical difficulty and this will be taken up in future research.

We wish to compare the mixture-normal (MN) density $f(x)$ and the beta-normal (BN) density $g(x)$. Given these two densities, we test the null hypothesis

$$H_0 : \text{MN and BN are equivalent} \quad \text{against the alternative hypothesis} \quad (5.2)$$

$$H_f : \text{MN is better than BN, or } H_g : \text{BN is better than MN.} \quad (5.3)$$

To test the null hypothesis in (5.2), we use the likelihood ratio test proposed by Vuong (1989).

Vuong's Likelihood Ratio Statistic

The likelihood ratio statistic for testing H_0 in (5.2) is

$$L_* = \sum_{i=1}^n \log \left(\frac{f(x_i)}{g(x_i)} \right). \quad (5.4)$$

Because the mixture-normal and the beta-normal densities are non-nested, the statistic in (5.4) is not chi-square distributed. Vuong (1989) used the Kullback-Liebler Information Criterion to discriminate between two non-nested models and proposed an unadjusted test statistic

$$T_* = \frac{L_*}{\hat{\omega}\sqrt{n}}, \quad (5.5)$$

where

$$\omega^2 = \frac{1}{n} \sum_{i=1}^n \left[\log \left(\frac{f(x_i)}{g(x_i)} \right) \right]^2 - \left[\frac{1}{n} \sum_{i=1}^n \log \left(\frac{f(x_i)}{g(x_i)} \right) \right]^2,$$

is an estimate of the variance of $L_* / \sqrt{n}$.

When comparing the goodness of fit for two non-nested parametric distributions, the number of parameters may not be the same. To account for the different number of parameters, Vuong proposed two adjusted test statistics:

$$T_1 = \frac{L_* - K_1(f, g)}{\hat{\omega}\sqrt{n}}, \text{ and } T_2 = \frac{L_* - K_2(f, g)}{\hat{\omega}\sqrt{n}},$$

where $K_1(f, g) = p - q$ is a correction factor for only the number of parameters and $K_2(f, g) = \ln(n)[(p - q)/2]$ is a correction factor for the number of parameters and the sample size n . In the test statistic, p is the

number of parameters in $f(x)$ and q is the number of parameters in $g(x)$. In this case $p = 5$, $q = 4$, and $n = 400$. We apply both adjusted statistics T_1 and T_2 in our comparison. T_i ($i = 1, 2$) is approximately standard normal distributed under the null hypothesis that the two densities are equivalent (Vuong, 1989).

At significant level α , one compares T_i with $z_{\alpha/2}$. If $T_i < -z_{\alpha/2}$, H_0 is rejected in favor of H_g , BN is better than MN. If $T_i > z_{\alpha/2}$, H_0 is rejected in favor of H_f , MN is better than BN. However, if $|T_i| \leq z_{\alpha/2}$, H_0 is not rejected.

Thus, we do not have sufficient evidence to say that both densities are not equivalent. For each generated data, the test statistics T_1 and T_2 are computed for testing H_0 . From the 100 simulations, we record the number of times the BN density is better than the MN density, the number of times the MN density is better than the BN density and the number of times both densities are equivalent.

Simulation Results and Discussion

Table 4 summarizes the Vuong's T_1 and T_2 goodness of fit statistics from 100 simulated data sets. The comparison is conducted at 10% and 5% level of significance. From Table 4, both beta-normal and mixture-normal distributions fit the mixtures of Weibull distribution data equally well for most cases. In general, beta-normal fits better than the mixture-normal, especially when using the adjusted statistic T_2 .

Figures 5 (A – D) give the histograms and the empirical CDF's of some simulated data sets with $n = 400$ and the corresponding fitted distributions of beta-normal and mixture-normal. The fitted distributions shown in Figure 5 indicate that both BN and MN fit these Weibull mixtures well. The histogram in Figure 5 (A) looks less like a bimodal distribution. The beta-normal distribution fits the data as unimodal distribution, while the mixture-normal distribution fits this data as bimodal distribution.

As mentioned before, we encountered some numerical difficulties when using the S-PLUS optimization routines to estimate the para-

meters of beta-normal distributions. Further research to develop better estimation algorithms will be needed to address this numerical problem in the estimation of BN parameters.

Table 4. Comparison between Beta-Normal and Mixture-Normal densities for Fitting 100 Simulated Mixtures of Weibull Distributions, $[p_1 W(2, 5) + 10] + [(1 - p_1) W(2, 10)]$.

α		$p_1 = 0.2$		$p_1 = 0.3$		$p_1 = 0.4$		$p_1 = 0.5$	
		T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
0.10	BN is better	0	7	9	52	16	55	1	9
	MN is better	6	2	0	0	0	0	1	0
	Both equivalent	94	91	91	48	84	45	98	91
0.05	BN is better	0	4	4	41	9	50	1	6
	MN is better	3	0	0	0	0	0	1	0
	Both equivalent	97	96	96	59	91	50	98	94

Figure 5. Histogram of Weibull mixture data with BN and MN superimposed; Empirical, BN and MN CDF for Weibull Mixture data

(A) Mixing proportion $p_1 = 0.2$

(B) Mixing proportion $p_1 = 0.3$

Histogram of simulated data

(C) Mixing proportion $p_1 =$

Histogram of simulated data

0.4

(D) Mixing proportion $p_1 = 0.5$

Histogram of simulated data

References

- Christiansen, F. B., & Fenchel, T. M. (1979). Evolution of marine invertebrate reproductive patterns. *Theoretical Population Biology*, 16, 267-282.
- Cobb, L., Koppstein, P., & Chen, N. H. (1983). Estimation and moment recursion for multimodal distributions of the exponential family. *Journal of American Statistical Association*, 78, 124-130.
- DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics*, 3rd Ed. New York: Addison Wesley.
- Eisenberger, I. (1964). Genesis of bimodal distributions, *Technometrics*, 6, 357-363.
- Emlet, R. B., McEdward, L. R., & Strathmann, R. R. (1987). Echinoderm larval ecology viewed from the egg. In M. Jangoux & J.M. Lawrence. (Eds.), *Echinoderm studies*, Vol. 2, (55-136). Rotterdam: A. A. Balkema.
- Espinoza, A., Rodriguez, M., de la Rosa, F., & Sanchez, J. (2001). Size of metals in urban aerosol in Serville. *Atmospheric Environment*, 35, 2595-2601.
- Eugene, N., Lee, C., & Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics: Theory and Methods*, 31(4), 497-512.
- Freeland, J. R., Romualdi, C., & Okamura, B. (2000). Gene flow and genetic diversity; a comparison of freshwater bryozoan populations in Europe and North America, *Heredity*, 85, 498-508.
- Isaacson, R. L. (2000). Non linear effects in the retention of an avoidance task induced by anabolic steroids, *European Journal of Pharmacology*, 405, 177-186.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*, 1. 2nd ed. NY: Wiley.
- Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? *The American Statistician*, 56(3), 223-229.
- Sewell, M. A., & Young, C. M. (1997). Are echinoderm egg size distributions bimodal? *Biological Bulletin*, 193, 297-305.
- Thorson, G. (1950). Reproductive and larval ecology of marine bottom invertebrates. *Biological Reviews*, 25, 1-45.
- Vance, R. R. (1973a). On reproductive strategies in marine benthic invertebrates. *The American Naturalist*, 107, 339-352.
- Vance, R. R. (1973b). More on reproductive strategies in marine benthic invertebrates. *The American Naturalist*, 107, 353-361.
- Vuong (1989). Likelihood ratio tests for model selection and non-nested hypothesis, *Econometrica*, 57(2), 307-333.
- Withington, D., Menard, G., Harris, J., Kulkarni, P., Donati, F., & Varin, F. (2000). Vecuronium pharmaco-kinetics and pharmacodynamics during hypothermic cardiopulmonary bypass in infants and children. *Canadian Journal of Anaesthesia*, 47, 1188-1195.
- Wolf, C. A., & Sumner, D. A. (2001). Are farm size distributions bimodal? Evidence from kernel density estimates of dairy farm size distributions, *American Journal of Agricultural Economics*, 83, 77-88.
- Zangvil, A., Portis, D. H., & Lamb, P. J. (2001). Investigation of the large-scale atmospheric moisture field over the Midwestern United States in relation to summer precipitation. Part I: Relationships between moisture budget components on different timescales, *Journal of Climate*, 14, 582-597.

Respondent-Generated Intervals (RGI) For Recall in Sample Surveys

S. James Press
Department of Statistics
University of California, Riverside

Respondents are asked for both a basic response to a recall-type question, their usage quantity, and are asked to provide lower and upper bounds for the (Respondent-Generated) interval in which their true values might possibly lie. A Bayesian hierarchical model for estimating the population mean and its variance is presented.

Key words: Bayes, bounds, bracketing, range, recall, survey

Introduction

Answers to recall-type questions are frequently required for surveys carried out by governmental agencies. While answers to such questions might become available to the agency at considerable expense and expenditure of time and effort through record checks, if the information is available at all, it is sometimes more expedient and efficient to directly question samples of the subpopulations for which the answers are required. Unfortunately, because respondents frequently differ greatly in their abilities to recall the correct answers to such questions, estimates of the population mean often suffer from substantial response bias, resulting in large non-

sampling errors for the population characteristics of interest. A new protocol for asking such recall-type questions in sample surveys is proposed, and an estimation procedure for analyzing the results that can improve upon the accuracy of the usual sample mean is suggested.

This new method is called Respondent-Generated Intervals (RGI). The procedure involves asking respondents not only for a basic answer to a recall-type question (this basic answer is called the “usage quantity”), but also, the respondent is asked for a smallest value his/her true answer could be, and a largest value his/her true answer could be. These values are referred to as the lower and upper bounds provided. It is assumed that the respondent knew the true value at some point but because of imperfect recall, he/she is not certain of the true value, and also, that the respondent is not purposely trying to deceive.

With the RGI protocol it is being implicitly assumed that there is a distinctive recall probability distribution associated with each respondent. To obtain an estimate of the mean usage quantity in a population typically the simple average of the responses from individuals who may have very different abilities to carry out the recall task is formed. But such a simple average may not necessarily account well for typical unevenness in recall ability.

It may be that an improvement upon population estimates can be made by learning more about the different recall abilities, and then taking them into account in the estimation

---

S. James Press, is Distinguished Professor, University of California-Riverside. Email: jpress@ucr.edu. The author thanks Dr. Monroe G. Sirken and Professor Judith M. Tanur for their encouragement, and Professor Judith M. Tanur for her suggestions for improving the presentation. Drs. Arthur Kennickell and Steven Heeringa provided helpful background materials and discussions about related approaches to similar problems. Thanks are also due to Dr. Diane Miller for helpful comments, and for developing the macros for calculating the RGI weighted average estimator and related quantities developed in this article. Norman M. Bradburn and Roger Tourangeau provided helpful discussions.

process. Ideally, the respondents could be asked many additional questions about their recall of their true answers for the recall question. That would permit many fractile points on each of their recall distributions to be assessed. Owing to the respondent burden of a long questionnaire, the sometimes heavy cost limitations of adding questions to a survey, the cost of added interviewer time, etc., there may sometimes be a heavy penalty imposed for each additional question posed in the survey questionnaire. The RGI protocol proposes adding to the usage quantity just two additional bounds questions and thereby obtains three points on each respondent's recall distribution. The interpretation of these three points is discussed in the section on estimation.

It is being proposed that respondents provide bounds on what they believe the true value for recall-type questions could possibly be. While there are other survey procedures that also request that respondents provide bounds-type information under certain circumstances, such procedures are not quantitatively associated with improved estimators, as is the RGI estimator. Usually these other procedures ask respondents to select their responses from several (analyst-generated) pre-assigned intervals (sometimes called "brackets").

Kennickell, (1997) described the 1995 Survey of Consumer Finances (SCF), carried out by the National Opinion Research Center at the University of Chicago, as including opportunities for the respondents who answered either "don't know", or "refuse", to select from 8 pre-assigned ranges, or to provide their own lower and upper bounds ("volunteered ranges"). These respondents were addressing what are traditionally recognized as sensitive questions about their assets. By contrast with the survey approach taken in the current research where the respondent is asked for both a basic response and lower and upper bounds, in the SCF, the respondent is given a choice to either give a basic response, or to select from one of several pre-assigned ranges, or to provide volunteered bounds. The pre-assigned intervals are supplied on "range cards" designed for situations in which the respondent has indicated that he/she does not desire to provide the specific usage quantity requested.

Another related technique that has been proposed is called unfolding brackets (Heeringa, Hill, & Howard, 1995). In this approach, respondents are asked a sequence of binary ("yes"/ "no") types of bracketing questions that successively narrow the range in which the respondent's true value might lie.

Several issues about these bounds-, or range-related techniques are not yet resolved. Which of these approaches, RGI, Range, Unfolding Brackets, or more traditional techniques yields the best results? How do these methods compare to one another under various circumstances? How do these different options affect response rate?

Schwartz and Paulin (2000) carried out a study comparing response rates of different groups of randomly assigned participants who used either range cards, unfolding brackets, or RGI, with respect to income questions. To include RGI in their study, Schwartz and Paulin used an early manuscript version of RGI. Schwartz and Paulin (2000) found that all three approaches studied reduced item non-response in that all three techniques presented a viable method for obtaining some income information from respondents who might otherwise have provided none.

In fact, 30% of the participants in the study selected RGI as their favorite range technique. The participants "claimed that they liked this technique because it allowed them to have control over their disclosures; the RGI intervals they provided tended to be narrower than pre-defined intervals; the RGI intervals did not systematically increase with income levels (as did the other techniques); RGI was the only technique that prompted respondents to provide exact values rather than ranges; and RGI allowed respondents to feel the most confident in the accuracy of the information they were providing."

Conrad and Brown (1994; 1996) and Conrad, Brown and Cashman (1998) studied strategies for estimating behavioral frequency using survey interviews. Conrad and his colleagues suggested that when respondents are faced with a question asking about the frequency of a behavior (the usage quantity), if that behavior is infrequent, respondents attempt to count the instances; if it is frequent, they attempt

to estimate. When the respondents count they tend to underreport, but when they estimate they tend to over-report. This finding may be relevant to RGI reporting.

Methodology

Let y_i, a_i, b_i denote the basic usage quantity response, the lower bound response, and the upper bound response, respectively, of respondent $i, i = 1, \dots, n$. Suppose that the y_i 's are all normally distributed $N(\theta_i, \sigma_i^2)$, that the θ_i 's are exchangeable, and $\theta_i \sim N(\theta_0, \tau^2)$. It is shown in the Appendix, using a hierarchical Bayesian model, that in such a situation, the conditional posterior distribution of the population mean, θ_0 , is given by:

$$(\theta_0 \mid \text{data}, \sigma_i^2, \tau^2) \sim N(\tilde{\theta}, \omega^2), \quad (3.1)$$

where the posterior mean, $\tilde{\theta}$, conditional on the data and (σ_i^2, τ^2) is expressible as a weighted average of the usage quantities and the y_i 's, and the weights are expressible approximately as simple algebraic functions of the interval lengths defined by the bounds. The conditional posterior variance, ω^2 , drives the associated credibility intervals; it is discussed below.

For normally distributed data it is commonly assumed that lower and upper bounds that represent extreme possible values for the respondents can be associated with 3 standard deviations below, and above, the mean, respectively. That interpretation is used to assess values for the σ_i^2 parameters from: $k_1 \sigma_i = b_i - a_i \equiv r_i$, the respondent interval lengths. Analogously, a value for τ^2 is assessed from: $k_2 \tau = \bar{b} - \bar{a} \equiv r_0$, the average respondent interval length. It will generally be assumed that $k_1 = k_2 = k = 6$ (corresponding to 3 standard deviations above and below the mean). The assumption of "3" standard deviations is examined numerically in the examples section, and is applied more generally in the Appendix.

The conditional posterior mean is shown in the Appendix to be given by:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i, \quad (3.2)$$

where the λ_i 's are weights that are given approximately by:

$$\lambda_i \doteq \frac{\left(\frac{1}{r_i^2 + r_0^2} \right)}{\sum_1^n \left(\frac{1}{r_i^2 + r_0^2} \right)}. \quad (3.3)$$

Note the following characteristics of this estimator:

1. The weighted average in Eq (3.3) is simple and quick to calculate, without requiring any computer-intensive sampling techniques. A simple Minitab macro is available for calculating it.

2. It will be seen in the examples section that if the respondents who give short intervals are also the more accurate ones, RGI will tend to give an estimate of the population mean that has smaller bias than that of the sample mean. In the special case in which the interval lengths are all the same, the weighted average reduces to the sample mean, $\bar{y}$, where the weights all equal $(1/n)$. In any case, the lambda weights are all non-negative, and must sum to one.

3. The longer the interval a respondent gives, the less weight is applied to that respondent's usage quantity in the weighted average. The length of respondent i 's interval seems intuitively to be a measure of his/her degree of confidence in the usage quantity he/she gives, so that the shorter the interval, the greater degree of confidence that respondent seems to have in the usage quantity he/she reports. Of course a high degree of confidence does not necessarily imply an answer close to the true value.

4. The lambda weights can be thought of as a probability distribution over the values of the usage quantities in the sample. So λ_i

represents the probability that $y = y_i$ in the posterior mean.

5. From equation (A23) in the Appendix it is seen that the conditional variance of the posterior distribution is given by:

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)} \doteq \frac{1}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}. \quad (3.4)$$

As explained in the discussion just above equation (3.2), it will sometimes be taken to be the case that $k_1 = k_2 = k = 6$ (other values of k are also being studied). So if the precision of a distribution is defined as its reciprocal variance, the quantity $\left\{ \frac{r_i^2 + r_0^2}{36} \right\}$ is the conditional variance in the posterior distribution corresponding to respondent i , and therefore, its reciprocal represents the conditional precision corresponding to respondent i . Summing over all respondent's precisions gives:
total conditional posterior variance

$$= \omega^2 \doteq \frac{1}{\sum_1^n \left(\frac{36}{r_i^2 + r_0^2} \right)}. \quad (3.5)$$

Thus, another interpretation of λ_i is that it is the proportion of the total conditional posterior precision in the data attributable to respondent i .

The variance of the conditional posterior distribution is given in equation. (3.4). The posterior variance is the reciprocal of the posterior total precision. Because the posterior distribution of the population mean, θ_0 , is normal, it is straightforward to find credibility intervals for θ_0 . For example, a 95% credibility interval for θ_0 is given by:

$$(\tilde{\theta} - 1.96\omega, \tilde{\theta} + 1.96\omega). \quad (3.6)$$

That is,

$$P\{(\tilde{\theta} - 1.96\omega \leq \theta_0 \leq \tilde{\theta} + 1.96\omega) | data\} = 95\%. \quad (3.7)$$

More general credibility intervals for other percentiles are given in the appendix. From eqn. (3.1) it is seen that the posterior distribution of the population mean, θ_0 , is normal. It is therefore straightforward to test hypotheses about θ_0 using the Jeffreys procedure for Bayesian hypothesis testing; (Jeffreys, 1961).

The behavior of the RGI Bayesian estimator is illustrated and examined using some numerical examples. It will be seen that for these examples, the way the RGI estimator works is to assign greater weight to the usage quantities of respondents who give relatively short bounding intervals, and less weight to the usage quantities of those who give relatively long intervals. If the respondents who give short intervals are also the more accurate ones, RGI will tend to give an estimate of the population mean that has smaller bias than the sample mean. Also, the credibility intervals will tend to be shorter and closer to the true population values than the associated confidence intervals.

Example 1

Suppose there is a sample survey of size $n = 100$ in which the RGI protocol has been used. Suppose also that the true population mean of interest is to be estimated, and it is given by $\theta_0 = 1000$. In this example the usage quantities and the respondents' bounds, (a_i, b_i) are fixed at $r_i = b_i - a_i$, $i = 1, \dots, n$, arbitrarily, whereas in Example 2 it will be assumed that the data are generated randomly. Define $r_0 = \bar{b} - \bar{a}$. This quantity will be used as an assessment for τ , the common standard deviation of θ_i , the mean for respondent i .

Assume that the first 50 respondents all have excellent memories and are quite accurate in their responses. Suppose the intervals these accurate respondents give are:

$$(a_1, b_1), \dots, (a_{50}, b_{50}) = (975, 975), \dots, (975, 975).$$

That is, they are all not only pretty accurate, but they all believe that they are accurate, so they respond to the bounds questions with degenerate intervals whose lower and upper bounds are the same. Accordingly, these accurate respondents all report intervals of length $r_i = 0$, and usages of equal amounts, $y_i = 975$ (compared with the true value of 1000).

Next suppose that the last 50 respondents all have poor memories and are inaccurate. They report the intervals:

$$(a_{51}, b_{51}), \dots, (a_{100}, b_{100}) \\ = (500, 1500), \dots, (500, 1500),$$

that have lengths of $r_i = 1000$, and they report equal usage quantities of $y_i = 550$. Their true values, θ_i , may all be different from one another, but assume that they all guess 550. It is now found that: $\bar{a} = 737.5$, and $\bar{b} = 1237.5$, so $r_0 = \bar{b} - \bar{a} = 500$.

RGI Bayesian Point Estimate of the Population Mean

The weights are calculated to be given by:

$$\lambda_i = \begin{cases} .0167, & i = 1, \dots, 50 \\ .0033, & i = 51, \dots, 100 \end{cases}$$

It is easy to check that: $\sum_{i=1}^{100} \lambda_i = 1$. It may now

be readily found that the conditional posterior mean RGI estimator of the population mean, θ_0 , is given by:

$$\tilde{\theta} = \sum_{i=1}^{100} \lambda_i y_i = 904.167.$$

The corresponding sample mean is given by:

$\bar{y} = 762.5$. The numerical error (bias) of the posterior mean is given by $1000 - \tilde{\theta} = 1000 - 904.167 = 95.833$. The numerical error (bias) of the sample mean is given by

$1000 - \bar{y} = 1000 - 762.5 = 237.5$. The RGI estimator has reduced the bias error by $237.5 - 95.833 = 141.667$, or about 60%, compared with the standard error of the sample mean.

It is also interesting to compare interval estimates of the population mean by comparing the standard error of $\bar{y}$, with ω , the standard deviation of the posterior distribution of $\tilde{\theta}$. These estimates give rise to the corresponding confidence and credibility intervals for θ_0 , respectively.

From Eq (3.4) it may readily be found that for the data in this example, $\omega = 10.76$. It is also easy to check that for the data, the standard deviation of the data is 213.56. So the standard error for a sample of size 100 is $213.56/10$, or 21.36. Thus, the RGI estimate of standard deviation is less than half that of the sample mean.

Correspondingly, the length of the 95% credibility interval $2(1.96)\omega = 42.18$, while the length of the 95% confidence interval is $2(1.96)(21.36) = 83.74$. The 95% confidence interval is about twice as long as the 95% credibility interval. The 95% credibility interval is given by: (883.081, 925.253). The 95% confidence interval is given by: (720.63, 804.37). Note in this example that:

1. Neither the RGI credibility interval nor the confidence interval covers the true value of 1000 (all usage quantities were biased downward).

2. The confidence and credibility intervals do not even overlap (but the entire credibility interval is closer to the true value).

3. It is expected to find many situations for which the bias error of the RGI estimator is smaller than that of the sample mean; however, the differences may be more, or less, dramatic compared with their values in this example.

Now examine some variations of the conditions in this example to explore the robustness of the RGI estimator with respect to variations in the assumptions.

Variation 1

Suppose that there were only 30 accurate respondents (instead of the 50 assumed in this example), responding in exactly the same

way, and 70 inaccurate respondents (instead of the 50 assumed in the example), the RGI estimate would still have been an improvement in bias error over that of the sample mean, although the improvement in bias error would have been smaller (35.03%).

Variation 2

Now take the example to the extreme by supposing that there were only 1 accurate respondent (instead of the original 50 assumed in the example), responding in exactly the same way, and 99 inaccurate respondents (instead of the 50 assumed in the example), the RGI estimate would still have been an improvement in bias error over that of the sample mean, although the improvement in bias error would have been only 9.5%.

Variation 3

How are the population mean estimates affected by the values selected for k_1 and k_2 ? First recall that as long as k_1 and k_2 are the same, the posterior mean is unaffected by the value of k . However, the posterior variance and the credibility intervals are affected. Continue to take $k_1 = k_2 = k$ but vary the value of k and assume the original split of 50 accurately-reporting respondents and 50 inaccurately-reporting respondents. Table 1 below compares results as a function of the common $k_1 = k_2 = k$ selected.

That is, if a_i denotes the lower bound provided by respondent i , and b_i denotes the same respondent's upper bound, the assessment method being used has been to take

$$\tau = \frac{1}{n} \sum_{i=1}^n (b_i - a_i) = \bar{b} - \bar{a}, \quad \text{where the bar}$$

denotes average. But consider as an alternative the range assessment: $\tau = b_0 - a_0$, where $a_0 \equiv \min_i(a_i)$, $b_0 \equiv \max_i(b_i)$.

Table 1. Effect Of Common Value of “k”

k	$\omega =$ posterior standard deviation	95% credibility interval	Length of credibility interval
4	16.14	(872.54, 935.80)	63.26
5	12.91	(878.86, 929.47)	50.61
6	10.76	(883.08, 925.25)	42.17
7	9.22	(886.09, 922.24)	36.15
8	8.07	(888.35, 919.98)	31.63

Examination of Table 1 suggests that for general purposes, selecting a common k and taking it to be $k = 6$ (bold face) is a reasonable compromise.

Note that the *range of belief* is reflected by the interval $(a_0, b_0) \equiv (500, 1500)$. How will the estimates of the population mean be affected? Results are shown in Table 2 for the two different methods for the 50/50 split of accurate and inaccurate respondents used in the original example.

Table 2 demonstrates that in this example, the “average” procedure used for assessing produces better results than the range procedure: there is less bias, smaller posterior variance, a shorter credibility interval, and a credibility interval that is also closer to the true population mean (the population mean in this example was 1000). It is therefore recommended that τ be assessed by using the average, rather than the range procedure.

Example 2

In this example the usage quantities from appropriate normal distributions are simulated while the respondents' bounds are fixed conveniently. Again assume a survey of 100 respondents and again use $k_1 = k_2 = 6$.

Table 2. Comparing Methods For Assessing τ

	posterior mean	posterior variance	95% credibility interval	length of 95% credibility interval
<i>average</i> assessment procedure for τ	904.167	115.741	(883.081, 925.253)	42.172
<i>range</i> assessment procedure for τ	833.333	370.370	(795.614, 871.053)	75.439

Adopt usage quantities that are generated from distinct normal distributions, the average of whose means is $\theta_0 = 1000$, and whose standard deviations are all 300. Such usage data are included within the framework of the model. The actual usage quantities that were generated are given in Tables 3a and 3b. Assign lower and upper bound intervals of (900, 1100) for the 26 usage quantities (out of 100) between 900 and 1100 (usages that lie close to the true population value), and assign lower and upper bound intervals of (200, 1900) for the other 74 usage quantities (those usages that lie further from the true population value). The lower and upper bounds adopted are given in Tables 3a 3b, as are the values of the calculated lambda da weights (which sum to one).

Bias Reduction

The sample mean for this example is 964.497. The posterior mean or RGI estimator is 973.816. The bias error for the sample mean is 35.503, while that for the RGI estimator is 26.184. The RGI estimator has reduced the bias by 9.319, or 26.2%.

The standard deviation of the usage quantities is 324.1 while the standard error of the sample mean is 32.4. So a 95% confidence interval is (900.993, 1028.001). It has length 127.008.

The standard deviation of the RGI estimator (standard deviation of the posterior distribution of the population mean estimator) is about 30.0, so a 95% credibility interval is (915.023, 1032.61). It has length 117.587.

The result is that both the 95% confidence interval and the 95% credibility intervals cover the true population values, but the credibility interval is shorter.

Conclusion

A new method for asking recall-type questions in sample surveys has been proposed. The method can substantially reduce the non-sampling bias error compared with the error of the sample mean. It is anticipated that over time, even better techniques will be developed to take advantage of this path to improved estimation accuracy. Such techniques will likely prompt respondents who believe they are accurate in their recollection to provide short bounding intervals, and conversely, the techniques will prompt respondents who are uncertain of the quantity to be recalled to give longer bounding intervals.

The RGI technique may also prove to be less threatening to respondents faced with answering sensitive questions. Respondents who might not answer such questions at all, might be willing at least to provide bounds, thereby increasing response rate. Therefore, there also may be a response rate benefit that accrues from the use of the RGI protocol in surveys containing sensitive questions.

The RGI estimator proposed appears to be robust with respect to variations in the distributions of the data and in the assumptions of the model. The data selected for Example 1 didn't follow any familiar distribution. What is important is that one or more accurate respondents also gave short bounding intervals that could be used in the weighted average, independently of any distributional assumptions. This robustness property appears to be very promising for survey applications.

Table 3a. Raw Data and Lambda Weights for Normal Data Example

Number	usage	lo-bound	up-bound	lambda	Number	usage	lo-bound	up-bound	lambda
1	518.67	200	1900	0.007033	26	414.04	200	1900	0.007033
2	1428.18	200	1900	0.007033	27	1282.15	200	1900	0.007033
3	1352.09	200	1900	0.007033	28	1317.67	200	1900	0.007033
4	919.02	900	1100	0.018446	29	466.53	200	1900	0.007033
5	572.47	200	1900	0.007033	30	869.69	200	1900	0.007033
6	822.51	200	1900	0.007033	31	820.92	200	1900	0.007033
7	814.42	200	1900	0.007033	32	1223.39	200	1900	0.007033
8	431.61	200	1900	0.007033	33	1330.02	200	1900	0.007033
9	1099.68	900	1100	0.018446	34	1267.04	200	1900	0.007033
10	1318.16	200	1900	0.007033	35	1123.93	200	1900	0.007033
11	704.25	200	1900	0.007033	36	1155.08	200	1900	0.007033
12	918.67	900	1100	0.018446	37	1206.36	200	1900	0.007033
13	1105.79	200	1900	0.007033	38	1082.61	900	1100	0.018446
14	931.64	900	1100	0.018446	39	997.76	900	1100	0.018446
15	1839.33	200	1900	0.007033	40	1205.7	200	1900	0.007033
16	625.11	200	1900	0.007033	41	675.03	200	1900	0.007033
17	1482.88	200	1900	0.007033	42	1642.04	200	1900	0.007033
18	691.66	200	1900	0.007033	43	909.12	900	1100	0.018446
19	1218.58	200	1900	0.007033	44	834.9	200	1900	0.007033
20	761.49	200	1900	0.007033	45	439.11	200	1900	0.007033
21	1041.2	900	1100	0.018446	46	279.14	200	1900	0.007033
22	283.22	200	1900	0.007033	47	996.48	900	1100	0.018446
23	1276.98	200	1900	0.007033	48	237.63	200	1900	0.007033
24	640.76	200	1900	0.007033	49	1284.94	200	1900	0.007033
25	1442.09	200	1900	0.007033	50	1143.45	200	1900	0.007033

Table 3b. Raw Data and Lambda Weights for Normal Data Example

Number	usage	lo-bound	up-bound	lambda	Number	usage	lo-bound	up-bound	lambda
51	906.23	900	1100	0.018446	76	902.57	900	1100	0.018446
52	903.56	900	1100	0.018446	77	706.01	200	1900	0.007033
53	1223.66	200	1900	0.007033	78	953.28	900	1100	0.018446
54	643.64	200	1900	0.007033	79	909.81	900	1100	0.018446
55	1355.94	200	1900	0.007033	80	1097.39	900	1100	0.018446
56	1270.2	200	1900	0.007033	81	1112.48	200	1900	0.007033
57	744.64	200	1900	0.007033	82	1082.91	900	1100	0.018446
58	1036.39	900	1100	0.018446	83	1013.8	900	1100	0.018446
59	673.09	200	1900	0.007033	84	1084.29	900	1100	0.018446
60	1234.4	200	1900	0.007033	85	750.5	200	1900	0.007033
61	435.02	200	1900	0.007033	86	1306.53	200	1900	0.007033
62	1196.57	200	1900	0.007033	87	875.54	200	1900	0.007033
63	849.43	200	1900	0.007033	88	994	900	1100	0.018446
64	1099.91	900	1100	0.018446	89	605.55	200	1900	0.007033
65	1025.81	900	1100	0.018446	90	982.78	900	1100	0.018446
66	1515.97	200	1900	0.007033	91	828.1	200	1900	0.007033
67	646.71	200	1900	0.007033	92	820.58	200	1900	0.007033
68	1119.5	200	1900	0.007033	93	644.87	200	1900	0.007033
69	460.59	200	1900	0.007033	94	630.19	200	1900	0.007033
70	1046.75	900	1100	0.018446	95	762.45	200	1900	0.007033
71	1378.79	200	1900	0.007033	96	1154.98	200	1900	0.007033
72	1267.88	200	1900	0.007033	97	764.34	200	1900	0.007033
73	799.94	200	1900	0.007033	98	984.5	900	1100	0.018446
74	1796.67	200	1900	0.007033	99	973.23	900	1100	0.018446
75	775.25	200	1900	0.007033	100	657.28	200	1900	0.007033

References

- Conrad, F. G., & Brown, N. R. (1994). Strategies for Estimating Category Frequency: Effects of Abstractness and Distinctiveness, *Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, p. 1345-1350. Erlbaum Associates.
- Conrad, F. G., & Brown, N. R. (1996). Estimating frequency: A multiple strategy approach. In D. J. Hermann, C., McEvoy, C., Hertzog, P., Hertel, & M. K., Johnson (Eds.) *Basic and Applied Memory Research: Practical Applications, Volume 2*. Mahwah, NJ: Lawrence

- Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, 6, 339-366.

- Heeringa, S. G., Hill, D. H., & Howell, D. H. (1995, June). Unfolding brackets for reducing Item non-response in economic surveys. *Health and Retirement Study Working Paper Series*, Paper No. 94-029, Institute for Social Research, University of Michigan.

- Jeffreys, H. (1961). *Theory of Probability*. Third Edition, Oxford: Oxford University Press.

Kennickell, A. B. (1997, January). Using range techniques with CAPI in the 1995 Survey of Consumer Finances. Board Governors of the Federal Reserve System, Washington, D.C. 20551.

Schwartz, L. K., & Paulin, G. D. (2000). Improving Response Rates to Income Questions, *Proceedings of the American Statistical Association*. Section on Survey Research Methods, pp. 965-969.

Appendix

In this Appendix, a hierarchical Bayesian model is developed for estimating the posterior distribution of the population mean for data obtained by using the RGI protocol. Suppose respondent i gives a point response y_i , and bounds (a_i, b_i) , $a_i \leq b_i$, $i = 1, \dots, n$, as his/her answers to a factual recall question. Assume:

$$(y_i | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2). \quad (A1)$$

The normal distribution will often be appropriate in situations for which the usage quantity corresponds to a change in some quantity of interest. Assume the means of the usage quantities are themselves exchangeable, and normally distributed about some unknown population mean of fundamental interest, θ_0 :

$$(\theta_i | \theta_0, \tau^2) \sim N(\theta_0, \tau^2). \quad (A2)$$

Thus, respondent i has a recall distribution whose true value is θ_i (each respondent is attempting to recall a different number of visits to the doctor last year). It is desired to estimate θ_0 . Assume $(\sigma_1^2, \dots, \sigma_n^2, \tau^2)$ are known; they will be assigned later. Denote the column vector of usage quantities by $y = (y_i)$, and the column vector of means by $\theta = (\theta_i)$. Let $\sigma^2 = (\sigma_i^2)$ denote the column vector of data variances. The joint density of the y_i 's is given in summary form

$$p(y | \theta, \sigma^2) \propto \exp \left\{ -\frac{1}{2} \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 \right\}. \quad (A3)$$

The joint density of the θ_i 's is given by:

$$p(\theta | \theta_0, \tau^2) \propto \exp \left\{ -\frac{1}{2} \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2 \right\}. \quad (A4)$$

So the joint density of (y, θ) is given by:

$$p(y, \theta | \theta_0, \tau^2, \sigma^2) = p(y | \theta, \sigma^2) p(\theta | \theta_0, \tau^2)$$

or, multiplying (A3) and (A4), gives:

$$\begin{aligned} p(y, \theta | \theta_0, \tau^2, \sigma^2) &\propto \exp \left\{ -\frac{1}{2} \left[\sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2 \right] \right\} \\ &\propto \exp \left\{ -\frac{A(\theta)}{2} \right\}, \end{aligned} \quad (A5)$$

where:

$$A(\theta) \equiv \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2. \quad (A6)$$

Expand (A6) in terms of the θ_i 's by completing the square. This takes some algebra. Then:

$$A(\theta) = \sum_1^n \left\{ \alpha_i \left[\left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2 + \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2} \right) \right] \right\}, \quad (A7)$$

$$\alpha_i = \frac{1}{\sigma_i^2} + \frac{1}{\tau^2}, \quad \beta_i = \frac{y_i}{\sigma_i^2} + \frac{\theta_0}{\tau^2}, \quad \gamma_i = \frac{\theta_0^2}{\tau^2} + \frac{y_i^2}{\sigma_i^2}. \quad (\text{A8})$$

Now find the marginal density of y by integrating (A5) with respect to θ_0 . Then:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto J(\theta_0) \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\},$$

$$J(\theta_0) \equiv \int \exp \left\{ \left(-\frac{1}{2} \right) \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2 \right\} d\theta_0, \\ \delta_i = \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2} \right). \quad (\text{A9})$$

Rewriting (A9) in vector and matrix form, to simplify the integration, it is found that if

$$f \equiv \left(\frac{\beta_i}{\alpha_i} \right), \quad K^{-1} \equiv \text{diag}(\alpha_1, \dots, \alpha_n),$$

$$(\theta - f)' K^{-1} (\theta - f) = \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2. \quad (\text{A10})$$

Carrying out the (normal) integration gives:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto \frac{1}{|K^{-1}|^{1/2}} \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (\text{A11})$$

Now note that $|K^{-1}| = \prod_1^n \alpha_i = \text{constant}$ and the constant can be absorbed into the proportionality constant, but δ_i depends on θ_0 . So:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (\text{A12})$$

Now apply Bayes' theorem to θ_0 in (A12).

$$p(\theta_0|y, \tau^2, \sigma^2) \propto p(\theta_0) \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}, \quad (\text{A13})$$

where $p(\theta_0)$ denotes a prior density for θ_0 . Prior belief (prior to observing the point and bound estimates of the respondents) is that for the large sample sizes typically associated with sample surveys, the population mean, θ_0 , might lie, with equal probability, anywhere in the interval (a_0, b_0) , where a_0 denotes the smallest lower bound given by any respondent, and b_0 denotes the largest. So adopt a uniform prior distribution on (a_0, b_0) . To be fully confident of covering all possibilities, however, adopt the (improper) prior density on the entire positive real line. Therefore adopt a prior density of the form:

$$p(\theta_0) \propto \text{constant}, \quad (\text{A14})$$

for all θ_0 on the positive half line. (In some survey situations the same survey is carried out repeatedly so that there is strong prior information available for providing a realistic finite range for θ_0 ; in such cases it is possible to improve on the estimator by using a proper prior distribution for θ_0 instead of the one given in eqn. (A14).) Inserting (A14) into (A13), and noting that $p(\theta_0) \propto \text{constant}$, gives:

$$p(\theta_0 | y, \tau^2, \sigma^2) \propto \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (\text{A15})$$

Next substitute for δ_i and complete the square in θ_0 to get the final result:

$$p(\theta_0 | y, \tau^2, \sigma^2) \propto \exp \left\{ \left(-\frac{u}{2} \right) \left(\theta_0 - \frac{v}{u} \right)^2 \right\}, \quad (\text{A16})$$

$$u = \sum_1^n \left(\frac{1}{\tau^2} - \frac{1}{\alpha_i \tau^4} \right), \quad v = \sum_1^n \left(\frac{y_i}{\alpha_i \sigma_i^2 \tau^2} \right). \quad (\text{A17})$$

Thus, the conditional posterior density of θ_0 is seen to be expressible as:

$$(\theta_0 | y, \tau^2, \sigma^2) \sim N(\tilde{\theta}, \omega^2), \quad (\text{A18})$$

$$\text{where: } \tilde{\theta} \equiv \frac{v}{u}, \text{ and } \omega^2 \equiv \frac{1}{u}. \quad (\text{A19})$$

Conditional Posterior Mean Of θ_0 As A Convex Mixture Of Usages

The appropriate measure of location of the posterior distribution in Eq. (A18) to use in any given situation depends upon the loss function that is appropriate. For many cases of interest the quadratic loss function (mean squared error) is appropriate. For such situations, interest centers on the posterior mean (under the normality assumptions in the current model, the conditional posterior distribution of θ_0 is also normal, so the posterior mean, median, and mode are all the same). It can be readily found by simple algebra that if:

$$\lambda_i \equiv \frac{\left(\frac{1}{\sigma_i^2 + \tau^2} \right)}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}, \quad \sum_1^n \lambda_i = 1, \quad (\text{A20})$$

$$\tilde{\theta} = \sum_1^n \lambda_i y_i.$$

Thus, the mean of the conditional posterior density of the population mean is a convex combination of the respondents' point estimates, that is, their usage quantities. It is an unequally weighted average of the usage quantities, as compared with the sample estimator of the population mean, which is an equally weighted estimator, $\bar{y}$. Interpret $(\sigma_i^2 + \tau^2)^{-1}$ as the precision attributable to

respondent i's response, and $\sum_1^n (\sigma_i^2 + \tau^2)^{-1}$ as the total precision attributable to all respondents; then, λ_i is interpretable as the proportion of total precision attributable to respondent i. Thus, the greater his/her precision proportion, the greater the weight that is automatically assigned to respondent i's usage response.

Assessing the Variance Parameters

Take: a) $k_1 \sigma_i = (b_i - a_i)$, for all $i = 1, \dots, n$; for some k_1 , such as $k_1 = 4, 5, 6$. Typically, take $k = 6$ (3 standard deviations on either side of the mean). Define, as above: b) $\bar{a} = \frac{1}{n} \sum_1^n a_i$, and $\bar{b} = \frac{1}{n} \sum_1^n b_i$. Then, take c)

$k_2 \tau = \bar{b} - \bar{a}$ for some pre-assigned k_2 . τ is the same for all respondents. Use an interval of 3 standard deviations on either side of the (normal) mean of the individual recall distribution means for the respondents. It is required to have an assessment that will be reasonable for all respondents. Use the average respondent interval.

Different analysts might interpret the k 's somewhat differently. Using these variance assessments, the weights become approximately:

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}, \quad \sum_1^n \lambda_i = 1, \quad (\text{A21})$$

where: $r_0 \equiv \bar{b} - \bar{a}$. Note that in the special case that $k_1 = k_2$, the k 's cancel out in numerator and denominator, so that the weights do not depend upon the k 's. Then, the weights become:

$$\lambda_i \doteq \frac{\left(\frac{1}{(b_i - a_i)^2 + r_0^2} \right)}{\sum_1^n \left(\frac{1}{(b_i - a_i)^2 + r_0^2} \right)}. \quad (\text{A22})$$

Conditional Posterior Variance Of θ_0

It is straightforward to check that the conditional posterior variance of θ_0 is given by:

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)} \doteq \frac{1}{\sum_1^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{r_0^2}{k_2^2}} \right)}, \quad (\text{A23})$$

the reciprocal of the total precision for all respondents in the sample. For $k_1 = k_2 = k$,

$$\omega^2 \doteq \frac{1}{\sum_1^n \left(\frac{k^2}{(b_i - a_i)^2 + r_0^2} \right)}, \quad (\text{A24})$$

so that in this case, while the conditional posterior mean does not depend upon k , the conditional posterior variance does. So the conditional posterior distribution of the population mean is given by:

$$(\theta_0 | y, \tau^2, \sigma^2) \sim N(\tilde{\theta}, \omega^2), \quad (\text{A25})$$

where $\tilde{\theta}$ and ω^2 are given, respectively, in (A19), (A20), and (A23) or (A24).

Credibility Intervals

Let z_γ denote the $\gamma/2$ -percentile of the standard normal distribution. Then, from (A25), a $(100 - \gamma)\%$ credibility interval for the population mean, θ_0 is given by:

$$(\tilde{\theta} - z_\gamma \omega, \tilde{\theta} + z_\gamma \omega). \quad (\text{A26})$$

That is,

$$P\{\tilde{\theta} - z_\gamma \omega \leq \theta_0 \leq \tilde{\theta} + z_\gamma \omega | y, \tau^2, \sigma^2\} = (100 - \gamma)\%. \quad (\text{A27})$$

Stratified Extreme Ranked Set Sample With Application To Ratio Estimators

Hani M. Samawi

Department of Mathematics & Statistics
Sultan Qaboos University

Laith J. Saeid

Department of Mathematics & Statistics
Sultan Qaboos University

Stratified extreme ranked set sample (SERSS) is introduced. The performance of the combined and separate ratio estimates using SERSS is investigated. Theoretical and simulation study are presented. Results indicate that using SERSS for estimating the ratios is more efficient than using stratified simple random sample (SSRS) and simple random sample (SRS). In some cases it is more efficient than ranked set sample (RSS) and stratified ranked set sample (SRSS), when the underlying distribution is symmetric. An application to real data on the bilirubin level in jaundice babies is introduced to illustrate the method.

Key words: Simple random sample; stratified random sample; ranked set sample; stratified ranked set sample; ratio estimation

---

Introduction

When sampling units in a study can be easily ranked compared to quantification. McIntyre (1952) proposed to use the mean of units based on a ranked set sample (RSS) to estimate the population mean. RSS is conducted by selecting random samples from the target population each of size r . Ranking each element within each set with respect to the random variable of interest. Then an actual measurement is taken of the element with the smallest rank from the first sample.

From the second sample an actual measurement is taken of the element with the second smallest rank, and the procedure is continued until the element with the largest rank is chosen for actual measurement from the r -th sample. Thus we obtain a total of r measured elements; one from each ordered sample of size r and this completed one cycle. The cycle may be repeated m times until $n = rm$ elements have been measured. These n elements form the ranked set sample data.

Samawi et al. (1996) investigated variety of extreme ranked set samples (ERSS) for estimating the population means. Furthermore, Samawi (1996) introduced the principle of stratified ranked set sampling (SRSS); to improve the precision of estimating the population means in case of SSRS.

In many situations the quantity that is to be estimated from a random sample is the ratio of two variables both of which vary from unit to unit. For example, in a household survey, the average number of suits of clothes per adult male is the quantities of interest. Examples of this kind occur frequently when the sampling unit (the household) comprises a group or cluster of elements (adult males) and our interest is in the population mean per element.

Moreover, ratio appears in many other applications, for example, the ratio of loans for building purpose to total loans in a bank or the ratio of acres of wheat to total acres on a farm. Also, this method is to obtain increased precision of estimating the population mean or total by taking advantage of the correlation between an auxiliary variable X and the variable of interest Y .

In the literature, ratio estimators are used in case of SRS as well as in case of SSRS (for an example see Cochran, 1977). Also, SSRS is used in certain types of surveys because it combines the conceptual simplicity of simple

---

Hani M. Samawi is Associate Professor, Yarmouk University. Contact information is P.O. Box 36, Sultan Qaboos University, Al-khod, 123, Sultanate of Oman. Email: hsamawi@squ.edu.om.

random sample with potentially, significant gains in efficiency. It is a convenient technique to use whenever, one wish to ensure that a sample is representative of the population and also to obtain separate estimates for parameters of each sub-domain of the population. There are two methods for estimating ratios that are generally used when the sampling design is stratified random sampling, namely the combined ratio estimate and the separate ratio estimate. Moreover, Samawi and Muttlak (1996) used RSS to estimate the population ratio, and showed that it provided a more efficient estimator compared with using SRS.

Introduce in this article is the idea of stratified extreme ranked set sample (SERSS). Also, the use of the idea of SERSS is proposed to improve the precision of the two methods for estimating the ratio namely the combined ratio estimate and separate ratio estimate. Moreover, studied are the properties of these estimators and comparing them in different situations. Later in the article the principle of SERSS and its properties are introduced. Combined and separate ratio estimators using SERSS are then discussed followed by a simulation study and the results of the simulation including an illustration of the methods using real data about the bilirubin level in jaundice babies.

Methodology

Ranked set sample for bivariate elements

A modification of the above procedure used by Samawi and Muttlak (1996) for the estimation of the ratio. First choose r independent samples each of size r of independent bivariate elements from the target population. Rank each sample with respect to one of the variables Y or X . Suppose that the ranking is done on the variable Y . From the first sample an actual measurement is taken of the element with the smallest rank of Y , together with the value of the variable X associated with the smallest value of Y .

From the second sample an actual measurement is then taken of the element with the second smallest rank of Y , together with the value of the variable X associated with the second smallest value of Y . The procedure is continued until the element with the largest rank

of Y is chosen for measurement from the r -th sample, together with the value of the variable X associated with the largest value of Y . The cycle may be repeated m times until $n = rm$ bivariate elements have been measured. Note that we assume that the ranking of the variable Y will be perfect, while the ranking of the variable X will be with errors in ranking, or at worst of a random order if the correlation between Y and X is close to zero.

Stratified ranked set sample

For the h -th stratum, first choose r_h independent samples each of size r_h of independent elements from the h -th subpopulations, $h = 1, 2, \dots, L$. Rank each sample within each stratum, then use the same sampling scheme described above to obtain L independent RSS samples of sizes $r_1, r_2, \dots, r_L$ respectively. Note that $r_1 + r_2 + \dots + r_L = r$. This complete one cycle of stratified ranked set sample. The cycle may be repeated m times until $n = mr$ elements have been measured (see Samawi, 1996).

The following stricture for the stratified Ranked set sample is used when the ranking on the variable Y in case of bivariate elements: For the k -th cycle, the SRSS is denoted by

$$\begin{aligned} \text{Stratum 1: } & (Y_{1(1)k}, X_{1[1]k}), (Y_{1(2)k}, X_{1[2]k}), \dots, \\ & (Y_{1(r_1)k}, X_{1[r_1]k}) \\ \text{Stratum 2: } & (Y_{2(1)k}, X_{2[1]k}), (Y_{2(2)k}, X_{2[2]k}), \dots, \\ & (Y_{2(r_2)k}, X_{2[r_2]k}) \\ & \vdots \\ \text{Stratum L: } & (Y_{L(1)k}, X_{L[1]k}), (Y_{L(2)k}, X_{L[2]k}), \dots, \\ & (Y_{L(r_L)k}, X_{L[r_L]k}), k = 1, 2, \dots, m \end{aligned}$$

Similarly for the stratified ranked set sample when the ranking on the variable X :

$$\begin{aligned}
\text{Stratum 1: } & (Y_{[1]k}, X_{1(1)k}), (Y_{[2]k}, X_{1(2)k}), \dots, \\
& (Y_{[r_1]k}, X_{1(r_1)k}) \\
\text{Stratum 2: } & (Y_{[1]k}, X_{2(1)k}), (Y_{[2]k}, X_{2(2)k}), \dots, \\
& (Y_{[r_2]k}, X_{2(r_2)k}) \\
& \vdots \\
\text{Stratum L: } & (Y_{[1]k}, X_{L(1)k}), (Y_{[2]k}, X_{L(2)k}), \dots, \\
& (Y_{[r_L]k}, X_{L(r_L)k})
\end{aligned}$$

where $k = 1, 2, \dots, m$.

Extreme Ranked Set Sample

The extreme ranked set sample ERSSs investigated by Samawi et al. (1996). The procedure involves randomly drawing r sets of r units each, from the infinite population for which the mean is to be estimated. It is assumed that the lowest or the largest units of this set can be detected visually or with little cost. For sure, this is a simple and practical process. From the first set of r units the lowest ranked unit is measured. From the second set of r units the largest ranked unit is measured. From the third set of r units the lowest ranked unit is measured, and so on. In this way we obtain the first $(r-1)$ measured units using the first $(r-1)$ sets. The choice of the r -th unit from the r -th (i.e the last) set depends on whether r is even or odd.

a) If r is even the largest ranked unit is measured. ERSSa will denote such a sample.

If r is odd then two options exist:

b) For the measure of the r -th unit we take the average of the measures of the lowest and the largest units in the r -th set. ERSSb will denote such a sample.

c) For the measure of the r -th unit we take the measure of the median. ERSSc will denote such a sample. Note that the choice (c) will be more difficult in application than the choice (a) and (b).

Stratified Extreme Ranked Set Sample

Suppose that the population divided into

L mutually exclusive and exhaustive strata, with subpopulation size $N_1, N_2, \dots, N_L$. Through this article it large subpopulation and symmetric underlying distribution will be assumed. The following notations and results will be introduced for this paper. For all $(i = 1, 2, \dots, r_h)$ and $h = 1, 2, \dots, L$.

$$\text{Let } \mu_{Xh} = E(X_{hij}), \sigma_{Xh}^2 = \text{Var}(X_{hij}),$$

$$\mu_{Xh(i)} = E(X_{hi(i)}), \sigma_{Xh(i)}^2 = \text{Var}(X_{hi(i)}) \text{ and}$$

$$W_h = \frac{N_h}{N} = \frac{n_h}{n} = \frac{r_h}{r} \text{ (proportional allocation).}$$

$$\text{Let } X_{h11}^*, X_{h12}^*, \dots, X_{h1r_h}^* ;$$

$$X_{h21}^*, X_{h22}^*, \dots, X_{h2r_h}^* ; \dots ;$$

$X_{hr_h1}^*, X_{hr_h2}^*, \dots, X_{hr_hr_h}^*$ be r_h independent samples of size r_h , each taken from the h -th stratum ($h = 1, 2, \dots, L$). Assume that each element X_{hij}^* in the sample has the same distribution function $F_h(x)$ with mean μ_{Xh} and variance σ_{Xh}^2 . For simplicity of notation, we will assume that X_{hij} denotes the quantitative measure of the unit X_{hij}^* .

Then, according to our description $X_{h11}, X_{h21}, \dots, X_{hr_h1}$ is the SRS from the h -th stratum. Let $X_{hi(1)}^*, X_{hi(2)}^*, \dots, X_{hi(r_h)}^*$ be the ordered statistics of the i -th sample $X_{hi1}^*, X_{hi2}^*, \dots, X_{hir_h}^*$, ($i = 1, 2, \dots, r_h$), taken from the h -th stratum. If r_h is even then $X_{h1(1)}, X_{h2(r_h)}, X_{h3(1)}, \dots, X_{h\{r_h-1\}(1)}, X_{hr_h(r_h)}$ denotes the ERSSh a for the h -th stratum. If

r_h is odd then

$$X_{h1(1)}, X_{h2(r_h)}, X_{h3(1)}, \dots, X_{h\{r_h-1\}(r_h)}, X_{hr_h\left(\frac{r_h+1}{2}\right)}$$

denotes the ERSSh c for the h -th stratum. Note that this will be repeated for each ($h=1, 2, \dots, L$). The resulting L independent ERSSs from each stratum will be denotes the stratified extreme ranked set sample SERSS. This process can be repeated m independent times.

Estimate of Population Mean Using

To estimate the mean μ using SERSS of size n , assume that there is (a) strata with even set size and $(L-a)$ strata with odd set size. For simplicity of notation, let $m=1$ then $n_h = r_h$ and $n = r$, then the estimate of the mean μ_X using SERSS is given by

$$\bar{X}_{SERSS} = \sum_{h=1}^a W_h \bar{X}_{h(a)} + \sum_{h=a+1}^L W_h \bar{X}_{h(c)}, \text{ where}$$

$$\bar{X}_{h(a)} = \frac{1}{2} (\bar{X}_{h(1)} + \bar{X}_{h(r_h)}),$$

$$\bar{X}_{h(c)} =$$

$$\frac{X_{h1(1)} + X_{h2(r_h)} + X_{h3(1)} + \dots + X_{h\{r_h-1\}(r_h)} + X_{hr_h\left(\frac{r_h+1}{2}\right)}{r_h},$$

$$\bar{X}_{h(1)} = 2 \sum_{i=1}^{\frac{r_h}{2}} \frac{X_{h\{2i-1\}(1)}}{r_h} \quad \text{and}$$

$$\bar{X}_{h(r_h)} = 2 \sum_{i=1}^{\frac{r_h}{2}} \frac{X_{h2i(r_h)}}{r_h}.$$

It can be shown that (Samawi et al., 1996).

$$E(\bar{X}_{h(a)}) = \frac{1}{2} (\mu_{Xh(1)} + \mu_{Xh(r_h)}) \quad \text{and}$$

$$E(\bar{X}_{h(c)}) = \left(\frac{r_h-1}{2r_h} \right) (\mu_{Xh(1)} + \mu_{Xh(r_h)})$$

$$+ \frac{1}{r_h} \mu_{Xh\left(\frac{r_h+1}{2}\right)}$$

Therefore, the mean and variance of $\bar{X}_{SERSS}$ are

$$E(\bar{X}_{SERSS}) = \frac{1}{2} \sum_{h=1}^L W_h (\mu_{Xh(1)} + \mu_{Xh(r_h)}) + \frac{1}{2} \sum_{h=a+1}^L \frac{W_h}{r_h} \left[2\mu_{Xh\left(\frac{r_h+1}{2}\right)} - (\mu_{Xh(1)} + \mu_{Xh(r_h)}) \right],$$

and

$$Var(\bar{X}_{SERSS}) = \frac{1}{2} \sum_{h=1}^L \frac{W_h^2}{r_h} (\sigma_{Xh(1)}^2 + \sigma_{Xh(r_h)}^2) + \frac{1}{2} \sum_{h=a+1}^L \frac{W_h^2}{r_h} \left[2\sigma_{Xh\left(\frac{r_h+1}{2}\right)}^2 - (\sigma_{Xh(1)}^2 + \sigma_{Xh(r_h)}^2) \right].$$

Note that the elements in $\bar{X}_{h(a)}$ are independent and so are the elements in $\bar{X}_{h(c)}$. Furthermore, the elements in $\bar{X}_{h(a)}$ are independent of the elements in $\bar{X}_{h(c)}$ and so are independent of the element in $X_{h\left(\frac{r_h+1}{2}\right)}$. If the underlying distribution for each stratum is symmetric then it can be shown that $E(\tilde{X}_{SERSS}) = \mu_X$ (i.e., an unbiased estimator) and

$$Var(\bar{X}_{SERSS}) = \sum_{h=1}^L \frac{W_h^2}{r_h} \sigma_{Xh(1)}^2 + \sum_{h=a+1}^L \frac{W_h^2}{r_h^2} \left[\sigma_{Xh\left(\frac{r_h+1}{2}\right)}^2 - \sigma_{Xh(1)}^2 \right] \quad (2.1)$$

Note that the estimate of the mean μ using SSRS of size r is given by $\bar{X}_{SSRS} = \sum_{h=1}^L W_h \bar{X}_h$. Also, the mean and variance of $\bar{X}_{SSRS}$ are known to be $E(\bar{X}_{SSRS}) = \mu_X$ (i.e., an unbiased estimator) and

$$Var(\bar{X}_{SSRS}) = \sum_{h=1}^L W_h^2 \frac{\sigma_{Xh}^2}{r_h} \quad (2.2)$$

(see Cochran, 1977).

Theorem: Assume that the underlying distribution for each stratum follows Normal or Logistic distribution.

Then $Var(\bar{X}_{SERSS}) \leq Var(\bar{X}_{SSRS})$.

Proof: Assume large subpopulation sizes $(N_1, N_2, \dots, N_L)$. In case of Normal, or Logistic distribution functions the following are true, $\sigma_{Xh(1)}^2 \geq \sigma_{Xh(2)}^2 \geq \dots \geq \sigma_{Xh(\frac{r_h}{2})}^2$ if r_h is even and $\sigma_{Xh(1)}^2 \geq \sigma_{Xh(2)}^2 \geq \dots \geq \sigma_{Xh(\frac{r_h+1}{2})}^2$ if r_h is odd.

Also note the $\sigma_{Xh(i)}^2 \leq \sigma_{Xh}^2$, $i = 1, 2, \dots, r_h$ (Arnold, 1992.) By comparing (2.1) and (2.2), and since

$$\sigma_{Xh(1)}^2 \leq \sigma_{Xh}^2, \sigma_{Xh(\frac{r_h+1}{2})}^2 \leq \sigma_{Xh(1)}^2 \text{ and}$$

$$\sigma_{Xh(\frac{r_h+1}{2})}^2 - \sigma_{Xh(1)}^2 \leq 0, \text{ therefore}$$

$$\sigma_{Xh(\frac{r_h+1}{2})}^2 - \sigma_{Xh(1)}^2 \leq 0, \text{ therefore}$$

$$\sum_{h=1}^L \frac{W_h^2}{r_h} \sigma_{Xh(1)}^2 + \sum_{h=a+1}^L \frac{W_h^2}{r_h^2} \left[\sigma_{Xh(\frac{r_h+1}{2})}^2 - \sigma_{Xh(1)}^2 \right]$$

$$\leq \sum_{h=1}^L W_h^2 \frac{\sigma_{Xh}^2}{r_h},$$

and hence

$$Var(\bar{X}_{SERSS}) \leq Var(\bar{X}_{SSRS}).$$

Simulation Study

The normal and logistic distribution is used in the simulation. Sample size $r = 10, 20$ and 30 and number of strata $L = 3$ are considered. For each of the possible combination of distribution, sample size and different choice of parameters 2000 data sets were generated. The relative efficiencies of the estimate of the population mean using SERSS with respect to SSRS, SRS, and RSS are obtained.

The values obtained by simulation are given in Table 1. Our Simulation indicates that estimating the population means using SERSS is more efficient than using SSRS or SRS. In some case, when the underling distribution is normal with $(\mu_1 = 1.0, \mu_2 = 3.0, \mu_3 = 5.0, r = 10)$, the simulation indicates that estimating the population mean using SERSS is even more efficient than using RSS, of the same size.

Separate Ratio Estimation using SERSS

In this Section, obtain the separate ratio estimator was obtained using stratified extreme ranked set sample. Also, the asymptotic mean and variance of the estimator were derived. Two cases are considered, the first case if the ranking on variable Y is perfect, while the ranking of the variable X will be with errors in ranking. The second case, when the ranking on variable X is perfect, while the ranking of the variable Y will be with errors in ranking. Also, some comparisons of the two cases are investigated.

Table 1. The relative efficiency of the simulation results.

Distribution function	n	$RE(\bar{X}_{SSRS}, \bar{X}_{SSRSS})$	$RE(\bar{X}_{RSS}, \bar{X}_{SERSS})$	$RE(\bar{X}_{SRS}, \bar{X}_{SERSS})$
Normal				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	1.20	1.43	7.42
$\mu_1 = 1.0, \mu_2 = 3.0, \mu_3 = 5.0$	20	2.34	0.94	8.47
	30	2.80	0.75	9.91
$\sigma_1 = 1.0, \sigma_2 = 1.0, \sigma_3 = 1.0$				
Normal				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	2.07	0.74	3.46
$\mu_1 = 1.0, \mu_2 = 2.0, \mu_3 = 3.0$	20	2.39	0.53	4.04
	30	2.77	0.42	4.70
$\sigma_1 = 1.0, \sigma_2 = 1.0, \sigma_3 = 1.0$				
Normal				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	1.92	1.21	6.17
$\mu_1 = 1.0, \mu_2 = 3.0, \mu_3 = 5.0$	20	2.29	0.80	7.41
	30	2.73	0.67	9.05
$\sigma_1 = 1.0, \sigma_2 = 1.1, \sigma_3 = 1.2$				
Logtic				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	1.72	0.71	3.14
$\mu_1 = 1.0, \mu_2 = 3.0, \mu_3 = 5.0$	20	1.74	0.43	3.20
	30	1.87	0.35	3.56
$\sigma_1 = 1.0, \sigma_2 = 1.0, \sigma_3 = 1.0$				
Logtic				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	1.93	0.51	2.19
$\mu_1 = 1.0, \mu_2 = 2.0, \mu_3 = 3.0$	20	1.78	0.30	1.98
	30	1.89	0.29	2.26
$\sigma_1 = 1.0, \sigma_2 = 1.0, \sigma_3 = 1.0$				
Logtic				
$W_1 = 0.3, W_2 = 0.3, W_3 = 0.4$	10	1.79	0.68	3.26
$\mu_1 = 1.0, \mu_2 = 3.0, \mu_3 = 5.0$	20	1.78	0.40	3.09
	30	1.83	0.34	3.42
$\sigma_1 = 1.0, \sigma_2 = 1.1, \sigma_3 = 1.2$				

Ratio Estimation when ranking on Variable Y .

Assuming that we can only rank on the variable Y so that the ranking of Y will be perfect while the ranking of X will be with error in ranking. If r_h is even then

$$\begin{aligned} & \left(X_{h[1]k}, Y_{h(1)k} \right), \left(X_{h2[r_h]k}, Y_{h2(r_h)k} \right), \\ & \left(X_{h3[1]k}, Y_{h3(1)k} \right), \dots, \left(X_{h[r_h-1][1]k}, Y_{h[r_h-1](1)k} \right), \\ & \left(X_{hr_h[r_h]k}, Y_{hr_h(r_h)k} \right) \end{aligned}$$

denotes the $SERSS1_a$ for the h -th stratum. If r_h is odd then

$$\begin{aligned} & \left(X_{h1[1]k}, Y_{h1(1)k} \right), \left(X_{h2[r_h]k}, Y_{h2(r_h)k} \right), \\ & \left(X_{h3[1]k}, Y_{h3(1)k} \right), \dots, \left(X_{h[r_h-1][r_h]k}, Y_{h[r_h-1](r_h)k} \right), \\ & \left(X_{hr_h\left[\frac{r_h+1}{2}\right]k}, Y_{hr_h\left(\frac{r_h+1}{2}\right)k} \right) \end{aligned}$$

denotes the $SERSS1_c$ for the h -th stratum, $k=1, 2, \dots, m$.

The separate ratio estimate requires knowledge of the stratum totals η_h in order to be used for estimating the population mean or total. Then using the same notation of Section (2.1) of the SERSS when ranking on variable Y , then the ratio can be estimated within each stratum as follows:

$$\hat{R}_{SERSSh1a} = \frac{\bar{Y}_{h(a)}}{\bar{X}_{h[a]}} \quad \text{if } (r_h) \text{ is even}$$

$$\text{and } \hat{R}_{SERSSh1c} = \frac{\bar{Y}_{h(c)}}{\bar{X}_{h[c]}} \quad \text{if } (r_h) \text{ is odd}$$

where

$$\bar{X}_{h[a]} = \frac{1}{2} \left(\bar{X}_{h[1]} + \bar{X}_{h[r_h]} \right)$$

$$\begin{aligned} \bar{X}_{h[1]} &= 2 \sum_{k=1}^m \sum_{i=1}^{\frac{r_h}{2}} \frac{X_{h\{2i-1\}[1]k}}{mr_h}, \\ \bar{X}_{h[r_h]} &= 2 \sum_{k=1}^m \sum_{i=1}^{\frac{r_h}{2}} \frac{X_{h2i[r_h]k}}{mr_h}, \\ \bar{X}_{h[c]} &= \frac{\sum_{k=1}^m \left\{ X_{h1[1]k} + X_{h2[r_h]k} + X_{h3[1]k} + \dots \right. \\ & \quad \left. + X_{h[r_h-1][r_h]k} + X_{hr_h\left[\frac{r_h+1}{2}\right]k} \right\}}{mr_h} \end{aligned}$$

$$\bar{Y}_{h(a)} = \frac{1}{2} (\bar{Y}_{h(1)} + \bar{Y}_{h(r_h)}),$$

$$\bar{Y}_{h(1)} = 2 \sum_{k=1}^m \sum_{i=1}^{\frac{r_h}{2}} \frac{Y_{h\{2i-1\}(1)k}}{mr_h},$$

$$\bar{Y}_{h(r_h)} = 2 \sum_{k=1}^m \sum_{i=1}^{\frac{r_h}{2}} \frac{Y_{h2i(r_h)k}}{mr_h} \quad \text{and } n_h = mr_h,$$

and

$$\bar{Y}_{h(c)} = \frac{\sum_{k=1}^m \left\{ Y_{h1(1)k} + Y_{h2(r_h)k} + Y_{h3(1)k} + \dots + Y_{h[r_h-1](r_h)k} + Y_{hr_h\left(\frac{r_h+1}{2}\right)k} \right\}}{mr_h}$$

Note that the sample sizes are different from one stratum to another. Therefore, assume without loss of generality that the first (a) strata have even set size (r_h), $h = 1, 2, \dots, a$, and the last ($L-a$) strata have odd set size (r_h), $h = a+1, a+2, \dots, L$. This implies that, the separate ratio estimator using stratified extreme ranked set sample when the ranking on variable Y , will be as follows:

$$\begin{aligned} \hat{R}_{SERSS1} &= \\ & \sum_{h=1}^a \frac{\eta_h}{\eta} \frac{\bar{Y}_{h(a)}}{\bar{X}_{h[a]}} + \sum_{h=a+1}^L \frac{\eta_h}{\eta} \frac{\bar{Y}_{h(c)}}{\bar{X}_{h[c]}}, \end{aligned} \quad (3.1)$$

$$\hat{R}_{SERSS1} = \sum_{h=1}^a W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h(a)}}{\bar{X}_{h[a]}} + \sum_{h=a+1}^L W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h(c)}}{\bar{X}_{h[c]}} \quad (3.2)$$

where

$$W_h = \frac{N_h}{N}, \quad \eta_h = N_h \mu_{Xh} \quad \text{and} \quad \eta = N \mu_X \quad (\text{known}).$$

It can be shown using the Taylor series expansion method that

$$E(\hat{R}_{SERSS1}) = \frac{\mu_Y}{\mu_X} + O\left(\min_h (mr_h)^{-1}\right).$$

Also, the approximate variance of $\hat{R}_{SERSS1}$ can be obtain as follows: Since we have independent strata and the assumption of symmetric marginal distribution, then

$$\begin{aligned} Var(\hat{R}_{SERSS1}) &= \\ Var\left(\sum_{h=1}^a W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h(a)}}{\bar{X}_{h[a]}} + \sum_{h=a+1}^L W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h(c)}}{\bar{X}_{h[c]}}\right) &= \\ Var(\hat{R}_{SERSS1}) &= \\ \sum_{h=1}^a W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} V(\hat{R}_{ERSSh1a}) &+ \\ + \sum_{h=a+1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} V(\hat{R}_{ERSSh1c}) & \quad (3.3) \end{aligned}$$

Using similar argument as in Samawi and Muttlak (1996), we have

$$\begin{aligned} Var(\hat{R}_{ERSSh1a}) &= \\ \frac{R_h^2}{r_h m} \left(\frac{\sigma_{Xh[1]}^2}{\mu_{Xh}^2} + \frac{\sigma_{Yh(1)}^2}{\mu_{Yh}^2} - 2 \frac{\sigma_{Xh[1]Yh(1)}}{\mu_{Xh} \mu_{Yh}} \right) & \quad (3.4) \end{aligned}$$

$$- 2 \frac{(r_h - 1) \sigma_{Xh[1]Yh(1)} + \sigma_{Xh\left[\frac{r_h+1}{2}\right]Yh\left(\frac{r_h+1}{2}\right)}}{r_h \mu_{Xh} \mu_{Yh}} \quad (3.5)$$

Therefore, the approximate variance of separate ratio estimator using SERSS (ranking on variable Y) is

$$\begin{aligned} Var(\hat{R}_{ERSSh1c}) &= \\ \frac{R_h^2}{n_h} \left(\frac{(r_h - 1) \sigma_{Xh[1]}^2 + \sigma_{X[1]h\left[\frac{r+1}{2}\right]}^2}{r_h \mu_{Xh}^2} \right. & \\ \left. + \frac{(r_h - 1) \sigma_{Yh(1)}^2 + \sigma_{Yh\left(\frac{r_h+1}{2}\right)}^2}{r_h \mu_{Yh}^2} \right) & \end{aligned}$$

$$\begin{aligned} Var(\hat{R}_{SERSS1}) &= \\ \sum_{h=1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\frac{\sigma_{Xh[1]}^2}{\mu_{Xh}^2} + \frac{\sigma_{Yh(1)}^2}{\mu_{Yh}^2} - 2 \frac{\sigma_{Xh[1]Yh(1)}}{\mu_{Xh} \mu_{Yh}} \right) & \\ + \sum_{h=a+1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\frac{\sigma_{Xh\left[\frac{r_h+1}{2}\right]}^2 - \sigma_{Xh[1]}^2}{r_h \mu_{Xh}^2} \right. & \\ + \frac{\sigma_{Yh\left(\frac{r_h+1}{2}\right)}^2 - \sigma_{Yh(1)}^2}{r_h \mu_{Yh}^2} & \\ - 2 \frac{\sigma_{Xh\left[\frac{r_h+1}{2}\right]Yh\left(\frac{r_h+1}{2}\right)} - \sigma_{Xh[1]Yh(1)}}{r_h \mu_{Xh} \mu_{Yh}} \left. \right) & \quad (3.6) \end{aligned}$$

$$\begin{aligned} \text{where} \quad \frac{\sigma_{Xh[1]}^2}{\mu_{Xh}^2} &= \frac{E(X_{h[1]} - \mu_{Xh[1]})^2}{E(X_h)^2}, \\ \frac{\sigma_{Yh(1)}^2}{\mu_{Yh}^2} &= \frac{E(Y_{h(1)} - \mu_{Yh(1)})^2}{E(Y_h)^2} \end{aligned}$$

$$\frac{\sigma_{Xh\left[\frac{r_h+1}{2}\right]}^2}{\mu_{Xh}^2} = \frac{E\left(X_{h\left[\frac{r_h+1}{2}\right]} - \mu_{Xh\left[\frac{r_h+1}{2}\right]}\right)^2}{E(X_h)^2},$$

$$\frac{\sigma_{Yh\left(\frac{r_h+1}{2}\right)}^2}{\mu_{Yh}^2} = \frac{E\left(Y_{h\left(\frac{r_h+1}{2}\right)} - \mu_{Yh\left(\frac{r_h+1}{2}\right)}\right)^2}{E(Y_h)^2}$$

$$\sigma_{Xh[1]Yh(1)} = E(Y_{h(1)} - \mu_{Yh(1)})(X_{h[1]} - \mu_{Xh[1]})$$

$$\text{and } R_h = \frac{\mu_{Yh}}{\mu_{Xh}}.$$

Ratio Estimation when Ranking on Variable X.

Similarly by changing the notation of perfect ranking ( ), by imperfect ranking [ ], for X and Y. Also, by using the same notation of the SERSS when ranking on variable X, then the separate ratio estimator using SERSS when the ranking on variable X, will be as follows:

$$\hat{R}_{SERSS2} = \sum_{h=1}^a W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h[a]}}{\bar{X}_{h(a)}} + \sum_{h=a+1}^L W_h \frac{\mu_{Xh}}{\mu_X} \frac{\bar{Y}_{h[c]}}{\bar{X}_{h(c)}} \quad (3.7)$$

In the same way as in section (3.1) we get the following results:

$$\begin{aligned} Var(\hat{R}_{SERSS2}) = & \sum_{h=1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\frac{\sigma_{Xh(1)}^2}{\mu_{Xh}^2} + \frac{\sigma_{Yh[1]}^2}{\mu_{Yh}^2} - 2 \frac{\sigma_{Xh(1)Yh[1]}}{\mu_{Xh}\mu_{Yh}} \right) \\ & + \sum_{h=a+1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\frac{\sigma_{Xh\left(\frac{r_h+1}{2}\right)}^2 - \sigma_{Xh(1)}^2}{r_h \mu_{Xh}^2} \right. \\ & \left. + \frac{\sigma_{Yh\left[\frac{r_h+1}{2}\right]}^2 - \sigma_{Yh[1]}^2}{r_h \mu_{Yh}^2} \right. \\ & \left. - 2 \frac{\sigma_{Xh\left(\frac{r_h+1}{2}\right)Yh\left[\frac{r_h+1}{2}\right]} - \sigma_{Xh(1)Yh[1]}}{r_h \mu_{Xh}\mu_{Yh}} \right). \quad (3.8) \end{aligned}$$

Ranking on which variable?

Again, since one can not rank on both variables at the same time and some time it is

easier to rank on one variable than the other, then we need to decide on which variable we should rank. We need to compare the variance of

$\hat{R}_{SERSS1}$ in (3.6) and variance $\hat{R}_{SERSS2}$

Theorem 3.2 : Assume that there are L linear relations between Y_h and X_h , i.e., $|\rho_h| > 0$ and it is easy to rank on variable X. Also assume that the approximation to the variance of the ratio estimators $\hat{R}_{SERSS1}$ and

$\hat{R}_{SERSS2}$ given in equations (3.6) and (3.8)

respectively are valid and the bias of the estimators can be ignored. If underlying distribution are Normal or Logistic distribution, then $Var(\hat{R}_{SERSS2}) \leq Var(\hat{R}_{SERSS1})$.

Proof : To prove the above we consider simple linear regression model between Y_h and X_h , and Y & X each has either Normal or Logistic marginal distribution function.

$$Y_{hi} = \alpha_h + \beta_h X_{hi} + \varepsilon_{hi}, \quad (3.9)$$

$$\mu_{Yh} = \alpha_h + \beta_h \mu_{Xh} \quad (3.10)$$

where α_h and β_h are parameters and ε_{hi} is a random error with $E(\varepsilon_{hi}) = 0$, $Var(\varepsilon_{hi}) = \sigma_h^2$ and $Cov(\varepsilon_{hi}, \varepsilon_{hj}) = 0$ for $i \neq j$, $i = 1, 2, \dots, r_h$, also ε_{hi} and X_{hi} are independent. Let

$$\sigma^2_{Xh[i]} = \begin{cases} \sigma_{Xh(i)}^2 & \text{if the ranking of the i-th order statistic} \\ & \text{in the i-th sample is correct} \\ \sigma_{Xh}^2 & \text{if the ranking of the i-th order statistic} \\ & \text{in the i-th sample is not correct} \\ & \text{i.e., random order} \end{cases}$$

Note that, according to our definition and by the assumption of the underlying distributions

$$\sigma_{Xh(i)}^2 \leq \sigma_{Xh[i]}^2 \quad (\text{Arnold, 1992}).$$

Case 1. If we are ranking on the Y_h variable we get the following model from equation (3.9)

$$Y_{h(i)} = \alpha_h + \beta_h X_{h[i]} + \varepsilon_{h[i]}, \quad (3.11)$$

where $\varepsilon_{h[i]}$ is a random error with $E(\varepsilon_{h[i]}) = 0$, $Var(\varepsilon_{h[i]}) = \sigma_h^2$ and $Cov(\varepsilon_{h[i]}, \varepsilon_{h[j]}) = 0$ for $i \neq j$, $i = 1, 2, \dots, r_h$ also $\varepsilon_{h[i]}$ and $X_{h[i]}$ are independent. The expected value of $Y_{h(i)}$ can be written as

$$\mu_{Yh(i)} = \alpha_h + \beta_h \mu_{Xh[i]}. \quad (3.12)$$

The variance of $Y_{h(i)}$ is

$$\sigma_{Yh(i)}^2 = \beta_h^2 \sigma_{Xh[i]}^2 + \sigma_h^2. \quad (3.13)$$

Now, by subtracting $\mu_{Yh(i)}$ from equation (3.11) and multiply both sides by $(X_{h[i]} - \mu_{Xh[i]})$, and then take the expected value for the both sides we get,

$$\sigma_{Yh(i)Xh[i]} = \beta_h \sigma_{Xh[i]}^2 \quad (3.14)$$

Case 2. If we are ranking on the variable X we get the following model from equation (3.9)

$$Y_{h[i]} = \alpha_h + \beta_h X_{h(i)} + \varepsilon_{h[i]}. \quad (3.15)$$

The expected value of $Y_{h[i]}$ is

$$\mu_{Yh[i]} = \alpha_h + \beta_h \mu_{Xh(i)}. \quad (3.16)$$

Similarly we can show that

$$\sigma_{Yh[i]}^2 = \beta_h^2 \sigma_{Xh(i)}^2 + \sigma_h^2. \quad (3.17)$$

$$\text{and } \sigma_{Yh[i]Xh(i)} = \beta_h \sigma_{Xh(i)}^2 \quad (3.18)$$

Now, equations (3.6) and (3.8) can be written as :

$$\begin{aligned} Var(\hat{R}_{SERSS1}) = & \sum_{h=1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\sigma_{Xh[1]}^2 \frac{(\mu_{Yh} - \beta_h \mu_{Yh})^2}{\mu_{Yh}^2 \mu_{Xh}^2} + \frac{\sigma_h^2}{\mu_{Yh}^2} \right) \\ & + \sum_{h=a+1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{mr_h^2} \left\{ \left[\sigma_{Xh\left[\frac{r_h+1}{2}\right]}^2 - \sigma_{Xh[1]}^2 \right] \right. \\ & \left. \frac{(\mu_{Yh} - \beta_h \mu_{Xh})^2}{\mu_{Yh}^2 \mu_{Xh}^2} \right\} \end{aligned}$$

and

$$\begin{aligned} Var(\hat{R}_{SERSS2}) = & \sum_{h=1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{n_h} \left(\sigma_{Xh(1)}^2 \frac{(\mu_{Yh} - \beta_h \mu_{Yh})^2}{\mu_{Yh}^2 \mu_{Xh}^2} + \frac{\sigma_h^2}{\mu_{Yh}^2} \right) \\ & + \sum_{h=a+1}^L W_h^2 \frac{\mu_{Xh}^2}{\mu_X^2} \frac{R_h^2}{mr_h^2} \left\{ \left[\sigma_{Xh\left(\frac{r_h+1}{2}\right)}^2 - \sigma_{Xh(1)}^2 \right] \right. \\ & \left. \frac{(\mu_{Yh} - \beta_h \mu_{Xh})^2}{\mu_{Yh}^2 \mu_{Xh}^2} \right\} \end{aligned}$$

respectively, therefore

$$Var(\hat{R}_{SERSS2}) \leq Var(\hat{R}_{SERSS1}).$$

Finally in this case it is recommended to rank on variable that will be used in the denominator of the ratio estimator if we wish to estimate the mean or the total of the population using the ratio estimator method when the data is selected according SERSS method.

Combined Ratio Estimation using SERSS

In this Section, combined were ratio estimator using stratified extreme ranked set sample. Two cases were considered, the first is to make the ranking on variable Y perfect, while the ranking on the variable X will be with errors in ranking. The second case, when the ranking on variable X is perfect, while the ranking on the variable Y will be with errors in ranking. Also, the properties of these estimators will be discussed.

Using SERSS as describe in Section 2, when ranking on variable Y . The combined Ratio estimate is defined by:

$$\hat{R}_{SERSS1}(s) = \frac{\bar{Y}_{(SERSS)}}{\bar{X}_{[SERSS]}}, \quad (4.1)$$

where

$$\bar{Y}_{(SERSS)} = \sum_{h=1}^a W_h \bar{Y}_{h(a)} + \sum_{h=a+1}^L W_h \bar{Y}_{h(c)},$$

$$\bar{X}_{[SERSS]} = \sum_{h=1}^a W_h \bar{X}_{h[a]} + \sum_{h=a+1}^L W_h \bar{X}_{h[c]}.$$

Therefore,

$$\hat{R}_{SERSS1} = \frac{\sum_{h=1}^a W_h \bar{Y}_{h(a)} + \sum_{h=a+1}^L W_h \bar{Y}_{h(c)}}{\sum_{h=1}^a W_h \bar{X}_{h[a]} + \sum_{h=a+1}^L W_h \bar{X}_{h[c]}}. \quad (4.2)$$

For fixed r_h , assume that we have finite second moments for X and Y . Since the ratio is a function of the means of X and Y , i.e.,

$$R = \frac{\mu_Y}{\mu_X}, \text{ and hence } R \text{ has at least two}$$

bounded derivations of all types in some neighborhood of (μ_Y, μ_X) provided that $\mu_X \neq 0$. Then, assuming large m , we can use the Multivariate Taylor Series Expansion, to approximate the variance and get the order of the bias of the ratio estimator. Therefore,

$$\text{Var}(\hat{R}_{SERSS1}) = R^2 \sum_{h=1}^a \frac{W_h^2}{mr_h} \left(\frac{\sigma_{Xh[1]}^2}{\mu_X^2} + \frac{\sigma_{Yh(1)}^2}{\mu_Y^2} - 2 \frac{\sigma_{Xh[1]Yh(1)}}{\mu_X \mu_Y} \right)$$

$$+ R^2 \sum_{h=a+1}^L \frac{W_h^2}{mr_h} \left(\frac{(r_h - 1)\sigma_{Xh[1]}^2 + \sigma_{Xh\left[\frac{r_h+1}{2}\right]}^2}{r_h \mu_X^2} + \frac{(r_h - 1)\sigma_{Yh(1)}^2 + \sigma_{Yh\left(\frac{r_h+1}{2}\right)}^2}{r_h \mu_Y^2} \right) +$$

$$- 2 \frac{(r_h - 1)\sigma_{Xh[1]Yh(1)} + \sigma_{Xh\left[\frac{r_h+1}{2}\right]Yh\left(\frac{r_h+1}{2}\right)}}{r_h \mu_X \mu_Y}. \quad (4.3)$$

Ratio Estimation when Ranking on variable X .
Similarly, the estimate is given by:

$$\hat{R}_{SERSS} = \frac{\bar{Y}_{[SERSS]}}{\bar{X}_{(SERSS)}}, \quad (4.4)$$

where

$$\bar{Y}_{(SERSS)} = \sum_{h=1}^a W_h \bar{Y}_{h(a)} + \sum_{h=a+1}^L W_h \bar{Y}_{h(c)}$$

$$\bar{X}_{[SERSS]} = \sum_{h=1}^a W_h \bar{X}_{h[a]} + \sum_{h=a+1}^L W_h \bar{X}_{h[c]}$$

Therefore, in combined case, we get:

$$\hat{R}_{SERSS1} = \frac{\sum_{h=1}^a W_h \bar{Y}_{h[a]} + \sum_{h=a+1}^L W_h \bar{Y}_{h[c]}}{\sum_{h=1}^a W_h \bar{X}_{h(a)} + \sum_{h=a+1}^L W_h \bar{X}_{h(c)}} \quad (4.5)$$

Using the same argument as in section (4.1), $E(\hat{R}_{SERSS2}) \approx R + O(\min_h (mr_h)^{-1})$, and

$$\begin{aligned}
& Var(\hat{R}_{SERSS2}) \\
&= R^2 \sum_{h=1}^a \frac{W_h^2}{mr_h} \left(\frac{\sigma_{Xh(1)}^2}{\mu_X^2} + \frac{\sigma_{Yh[1]}^2}{\mu_Y^2} - 2 \frac{\sigma_{Xh(1)Yh[1]}}{\mu_X \mu_Y} \right) \\
&+ R^2 \sum_{h=a+1}^L \frac{W_h^2}{mr_h} \left(\frac{(r_h-1)\sigma_{Xh(1)}^2 + \sigma_{Xh\left(\frac{r_h+1}{2}\right)}^2}{r_h \mu_h^2} \right. \\
&\quad \left. + \frac{(r_{h-1})\sigma_{Yh[1]}^2 + \sigma_{Yh\left[\frac{r_h+1}{2}\right]}^2}{r_h \mu_Y^2} \right. \\
&\quad \left. - 2 \frac{(r_h-1)\sigma_{Xh[1]Yh(1)} + \sigma_{Xh\left[\frac{r_h+1}{2}\right]Yh\left(\frac{r_h+1}{2}\right)}}{r_h \mu_X \mu_Y} \right). \tag{4.6}
\end{aligned}$$

Ranking on which variable?

Again, since we can not rank on both variables at the same time and some time it is easier to rank on one variable than the other, then we need to decide on which variable we should rank. We need to compare the variance of $\hat{R}_{SERSS1}$ in (4.3) and variance of $\hat{R}_{SERSS2}$ in (4.6).

Theorem 4.2 : Assume that there are L linear relations between Y_h and X_h , i.e., $|\rho_h| > 0$ and it is easy to rank on variable X . Also assume that the approximation to the variance of the ratio estimators $\hat{R}_{SERSS1}$ and $\hat{R}_{SERSS2}$ given in equations (4.8) and (4.11) respectively are valid and the bias of the estimators can be ignored, and if underlying distribution is Normal or Logistic distribution, then $Var(\hat{R}_{SERSS2}) \leq Var(\hat{R}_{SERSS1})$

Proof: The proof is similar to that of Theorem 3.2. Finally in this case it is always recommended to rank on variable that will be used in the denominator of the ratio estimator if we wish to estimate the mean or total of the populaton using the ratio estimator method when the data is selected according SERSS method.

Simulation Study

Computer simulation was conducted to gain insight in the properties of the ratio estimator. Bivariate random observations were generated from a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and correlation coefficient ρ . Also we deviled the data into three strata and in some cases into four strata. The sampling methods described above are used to draw SERSS, SRSS and SSRS with sets of size r . We repeat this process m times to get samples of size $n = rm$. The simulation was performed with $r = 20, 30$, and 40 and with $m = 10$ for the SERSS, SRSS and SSRS data sets. The ratio of the population means was estimated from these samples. Using 2000 replications, estimates of the means, and mean square errors were computed.

The ranking was considered on either variable Y or X i.e., the ranking in one of the two variables would be perfect while the second with errors in ranking. Results of these simulations are summarized by the relative efficiencies of the estimators of the population ratio and by the bias of estimation for different values of the correlation coefficient ρ . Introduced here is only one table for efficiency when ranking on X and one for the bias, since other tables give the same conclusion Results of the simulation is given in Table 2 for the efficiency when ranking on variable X . Table 3 shows the bias of the estimators when ranking on the variable X . The efficiency of the ratio estimator is defined by

$$eff(R_{SSRS}, R_{SERSS}) = \frac{MSE(R_{SSRS})}{MSE(R_{SERSS})}.$$

Results

It is concluded that the highest gain in efficiency is obtained by ranking of the variable X and with large values of negative ρ . For example in Table 2, $eff(R_{SSRS}, R_{SERSS})$ when $(\rho = .90, r = 40 \text{ \& } m = 10)$ is 1.69 while when $eff(R_{SSRS}, R_{SERSS})$ $(\rho = -.90, r = 40 \text{ \& } m = 10)$ is 3.05. Also, our simulation indicates the following:

Table 2. Efficiency when ranking on variable X

 $\mu_{xh} : 2 / 3 / 4$ $\sigma_{xh} : 1 / 1 / 1$ $W_h : .3 / .3 / .4$ $\mu_{yh} : 3 / 4 / 6$ $\sigma_{yh} : 1 / 1 / 1$

R= 1.45		Eff. In Combine		Eff. in Separate	
ρ	r	$\frac{SSRS}{SERSS}$	$\frac{SRSS}{SERSS}$	$\frac{SSRS}{SERSS}$	$\frac{SRSS}{SERSS}$
.99	20	2.1720	0.7562	2.2231	0.7499
	30	2.4724	0.6932	2.5604	0.6853
	40	2.7212	0.6762	2.7328	0.6690
.90	20	1.5512	0.8679	1.5715	0.8590
	30	1.7122	0.8353	1.7312	0.8354
	40	1.6920	0.8103	1.7320	0.8089
.70	20	1.5385	0.9894	1.5480	0.9840
	30	1.5100	0.8221	1.5243	0.8218
	40	1.4787	0.8319	1.4745	0.8280
.50	20	1.5079	0.9055	1.5326	0.9074
	30	1.4670	0.9231	1.4791	0.9270
	40	1.5491	0.8471	1.5630	0.8460
.25	20	1.5039	0.8925	1.5273	0.8921
	30	1.7163	0.8597	1.7399	0.8598
	40	1.6810	0.8471	1.6991	0.8462
-.25	20	1.7585	0.8268	1.7780	0.8235
	30	1.8502	0.8042	1.8654	0.8042
	40	1.9663	0.7738	1.9830	0.7751
-.50	20	1.8968	0.7978	1.9015	0.7978
	30	2.2485	0.7770	2.2664	0.7759
	40	2.3928	0.6844	2.4007	0.6832
-.70	20	2.2086	1.2115	2.2175	1.2149
	30	2.5975	1.3015	2.6208	1.3064
	40	2.4561	1.5529	2.4784	1.5572
-.90	20	2.5302	0.7841	2.5749	0.7788
	30	2.8748	0.7002	2.8924	0.6963
	40	3.0556	0.5546	3.0866	0.5505
-.99	20	2.6394	0.7176	2.7023	0.7219
	30	3.0000	0.6558	3.0266	0.6535
	40	3.1623	0.5275	3.1902	0.5269

1. When ranking on variable X , the efficiency will decrease with decreasing the value of ρ from 0.99 to 0.50, and start to increase as ρ decreases from 0.25 to -0.99.
2. The efficiency will increase when the even sample size increased by increasing the number of elements in each set (r).
3. There will be no change in the efficiency if the sample size increased by increasing the cycle size m .
4. For fix ρ , we noted that in combined case, as r increase the efficiency will increase, for all values of ρ positive or negative except, in some cases when ($r=30$) and ρ positive.
5. Also, for fix ρ , we note that in separate case, as r increase the efficiency will increase, for all values of ρ positive or negative, except in some cases when ($r=30$) and ρ positive.
6. For fix r , we note that in combined case the efficiency will decrease from 0.99 to 0.50, and then after this will increase from 0.25 to -0.99.
7. Also, for fix r and change ρ , we note that in separate case the efficiency will decrease from 0.99 to 0.50, and then after this will increase from 0.25 to -0.99.
8. We note that the efficiency in combined case less than in separate case. That is because the sample size within each stratum is small.
9. The bias will decrease when increasing the number of increases the even sample size r elements in each set.
10. The bias in combined case is less than the corresponding bias in separate case.

Application: Bilirubin Level in Jaundice Babies

Introduced is a real life example about Bilirubin level in jaundice babies who stay in neonatal intensive care. Most of birth surveys on live newborns Birth showed that jaundice is common. Jaundice in new Born can be pathological physiological which start on second day of life and it has relationship with race, method of feeding and Gestational age.

Jaundice is observed during the first week of life, and neonatal jaundice is a common problem. It is possible that the generally accepted levels are too high and may produce some high tone hearing loss.

Most of neonatal jaundice appears on second day of life. Most of normal newborn babies leave the hospital after 24 hours of life. Therefore, the primary concern will be on baby's who staying in neonatal intensive care. Physicians are interesting in the jaundice, according to its important and risk on the hearing, brain and death. We will focus on the weight and bilirubin level in blood (tsb) for the babies. The data were collected on 120 babies, who stay in neonatal intensive care, in four Jordanian hospitals (see Samawi and Al-Sagheer, 2001.) The data were divided into two strata, male stratum of size $N_1=72$ and female stratum of size $N_2=48$.

The following are the exact population values of the data

For Males it was found that:
 $\mu_{X1} = 2.91$, $\sigma_{X1} = 0.75$, $\mu_{Y1} = 11.97$,
 $\sigma_{Y1} = 5.52$ and $\rho_1 = 0.22$.

For Females it was found that:
 $\mu_{X2} = 2.82$, $\sigma_{X2} = 0.64$, $\mu_{Y2} = 9.97$,
 $\sigma_{Y2} = 4.11$ and $\rho_2 = -0.37$.

Also, for the whole data it was found that: $\mu_X = 2.87$, $\sigma_X = 0.71$, $\mu_Y = 11.18$,
 $\sigma_Y = 5.08$ and $\rho = 0.06$ Two strata exist

($L=2$), $m=2$ and $r=10$, which produce $n = r.m$
 $= 20$ $W_1 = \frac{72}{120} = 0.6$, $W_2 = \frac{48}{120} = 0.4$. For

Male : $n_1 = 0.6 \times 20 = 12$; $r_1 = 6$. For Female:
 $n_2 = 0.4 \times 20 = 8$; $r_2 = 4$.

Table 2: Bias of the ratio estimators when ranking on variable $\underline{X}$

$W_h: 0.3/0.3/0.4$		$\mu_{Xh}: 2/3/4$		$\sigma_{Xh}: 1/1/1$			
$R=1.45$		$\mu_{Yh}: 3/4/6$		$\sigma_{Yh}: 1/1/1$			
ρ	r	Combined SRSS	Combined SSRS	Combined SERSS	Separate SRSS	Separate SSRS	Separate SERSS
0.99	20	-0.0002	0.0003	0.0001	-0.0001	0.0009	0.0003
	30	0.0000	-0.0001	0.0000	0.0001	0.0001	0.0000
	40	0.0001	0.0002	0.0000	0.0001	0.0004	0.0000
0.90	20	0.0001	0.0002	0.0003	0.0003	0.0009	0.0006
	30	0.0002	0.0007	0.0000	0.0003	0.0011	0.0001
	40	-0.0002	0.0000	0.0000	-0.0002	0.0003	0.0000
0.70	20	0.0005	0.0003	0.0007	0.0007	0.0013	0.0011
	30	0.0003	0.0000	0.0002	0.0004	0.0006	0.0004
	40	0.0004	0.0000	-0.0001	0.0005	0.0003	0.0000
0.50	20	0.0003	0.0011	-0.0005	0.0006	0.0023	-0.0001
	30	0.0007	0.0003	0.0005	0.0009	0.0011	0.0008
	40	0.0006	0.0003	0.0001	0.0007	0.0009	0.0002
0.25	20	0.0005	0.0010	0.0001	0.0010	0.0025	0.0008
	30	0.0000	-0.0006	0.0002	0.0003	0.0004	0.0005
	40	0.0002	0.0000	-0.0001	0.0003	0.0006	0.0001
-0.25	20	-0.0002	0.0007	0.0003	0.0004	0.0029	0.0012
	30	0.0005	0.0000	0.0001	0.0008	0.0013	0.0005
	40	-0.0005	-0.0003	-0.0003	-0.0003	0.0007	0.0000
-0.50	20	0.0000	-0.0011	0.0003	0.0006	0.0011	0.0013
	30	-0.0002	-0.0002	0.0000	0.000	0.0013	0.0004
	40	-0.0004	-0.0003	-0.0004	-0.0002	0.0009	0.0000
-0.70	20	0.0001	0.0006	-0.0003	0.0009	0.00031	0.0006
	30	-0.0002	0.0000	0.0000	0.0000	0.0016	0.0006
	40	0.0003	0.0007	-0.0005	0.0005	0.0020	-0.0001
-0.90	20	-0.0008	0.0022	-0.0008	0.0000	0.0053	0.0003
	30	0.0002	-0.0004	0.0001	0.0006	0.0014	0.0007
	40	0.0000	0.0021	0.0003	0.0002	0.0036	0.0008
-0.99	20	0.0015	0.0022	0.0003	0.0024	0.0053	0.0016
	30	-0.0006	0.0008	-0.0001	-0.0002	0.0028	0.0004
	40	-0.0002	0.0000	0.0000	0.0000	0.0014	0.0005

Using SERSS & SSRS

We use the method of sampling SERSS and SSRS to get the following samples. Note that the ranking was on variable X (weight). The SERSS which is drawn is in Table 4.

Based on the SERSS it was found that

$$\begin{aligned}\hat{\mu}_{x1} &= 2.83, & \hat{\mu}_{y1} &= 9.78, & \hat{\mu}_{x2} &= 3.237, \\ \hat{\mu}_{y2} &= 11.74, & \hat{\sigma}_{x1}^2 &= 0.286, & \hat{\sigma}_{y1}^2 &= 4.69, \\ \hat{\sigma}_{x2}^2 &= 0.377, & \hat{\sigma}_{y2}^2 &= 12.68.\end{aligned}$$

Also, for the SSRS it was found:

$$\begin{aligned}\tilde{\mu}_{x1} &= 2.58, & \tilde{\mu}_{y1} &= 8.29, & \tilde{\mu}_{x2} &= 2.97, \\ \tilde{\mu}_{y2} &= 11.89, & \tilde{\sigma}_{x1}^2 &= 0.27, & \tilde{\sigma}_{y1}^2 &= 23.81, \\ \tilde{\sigma}_{x2}^2 &= 0.35, & \tilde{\sigma}_{y2}^2 &= 50.41.\end{aligned}$$

$$\text{Now, } \hat{Var}(R_{SERSS2(s)}) = 0.071$$

$$\hat{Var}(R_{SERSS2(c)}) = 0.082$$

$$\hat{Var}(R_{SSRS(s)}) = 0.265$$

$$\hat{Var}(R_{SSRS(c)}) = 0.266$$

note that

$$\hat{Var}(R_{SERSS2(s)}) \leq \hat{Var}(R_{SERSS2(c)}) \text{ and}$$

$$\hat{Var}(R_{SSRS(s)}) \leq \hat{Var}(R_{SSRS(c)})$$

It is clear that this just illustration of the computations only. However, still this conclusion indicates that the results in Sections 3, 4 and 5 are correct.

$$eff(R_{SSRS2(s)}, R_{SERSS2(s)}) = 3.73,$$

$$eff(R_{SSRS(c)}, R_{SERSS2(c)}) = 3.23$$

$$eff(R_{SSRS(c)}, R_{SSRS(s)}) = 1.00$$

$$eff(R_{SERSS2(c)}, R_{SERSS2(s)}) = 1.15.$$

Table 4. The drawn samples using SERSS and SSRS methods.

Cycle Number	Females		Males	
SERSS Sample				
1	tsb	Weight	tbs	Weight
	4.80	4.15	7.06	2.80
	6.90	3.00	5.60	2.75
	7.80	3.15	5.50	3.70
	8.60	3.40	7.53	2.50
			9.50	3.60
2			9.20	1.85
	12.76	2.50	23.41	3.10
	8.82	1.55	10.24	3.50
	13.94	2.85	13.18	4.50
	14.59	2.10	14.00	3.10
			16.20	3.65
1		SSRSS	19.50	3.80
	9.30	2.80	7.70	2.60
	5.50	3.00	6.12	3.20
	7.80	3.15	21.29	4.15
	5.40	2.65	10.94	2.60
			9.50	3.60
2			15.47	2.70
	9.24	2.60	8.71	2.45
	20.41	2.10	7.06	2.80
	13.10	2.85	7.60	2.20
	8.82	1.55	13.60	2.50
			29.24	3.15
		5.50	3.70	

References

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A first course in order statistics*. NY: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling techniques*. 3rd edition. NY: John Wiley & Sons.

McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research* 3, 385-390.

Samawi, H. M. (1996). Stratified ranked set sample. *Pakistan Journal of Statistics*, 12, 1, 9-16.

Samawi, H. M., Ahmed, M. S. & Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal*, 38, 5, 577 – 586.

Samawi, H. M. and Al-Sageer, O. A. M. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal*, 43, 2, 29-45 .

Samawi, H. M., & Muttlak, H. A. (1996). Estimation of ratio using rank set sampling. *Biometrical Journal*, 38, 6, 753-764 .

Takahasi, K. & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the stratified sampling by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.

Estimation Using Bivariate Extreme Ranked Set Sampling With Application To The Bivariate Normal Distribution

Mohammad Fraiwan Al-Saleh
Yarmouk University
Jordan

Hani M. Samawi
Sultan Qaboos University
Sultanate of Oman

In this article, the procedure of bivariate extreme ranked set sampling (BVERSS) is introduced and investigated as a procedure of obtaining more accurate samples for estimating the parameters of bivariate populations. This procedure takes its strength from the advantages of bivariate ranked set sampling (BVRSS) over the usual ranked set sampling in dealing with two characteristics simultaneously, and the advantages of extreme ranked set sampling (ERSS) over usual RSS in reducing the ranking errors and hence in being more applicable. The BVERSS procedure will be applied to the case of the parameters of the bivariate normal distributions. Illustration using real data is also provided.

Key words: Bivariate ranked set sampling; Efficiency; Ranked set sampling; Extreme ranked set sampling; Bivariate extreme ranked set sampling

---

Introduction

Ranked set sampling (RSS) was first suggested by McIntyre (1952) as a method for estimating pasture yields. The supporting mathematical theory was later provided by Takahasi and Wakimoto (1968). The RSS procedure consists of drawing m random samples of size m each from the population of interest, and ranking each of them by judgment with respect to (w.r.t.) the characteristic of interest. Then the i th smallest observation from the i th set is chosen for actual quantification. The RSS consists of these m selected units.

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, Yarmouk University, P.O. Box 36, Sultan Qaboos University, Al-khod, 123, Sultanate of Oman. Email: hsamawi@squ.edu.om.

Although only m units out of m^2 are chosen for quantification, all units contribute information to the m quantified ones. The entire cycle may be repeated, if necessary, r times to produce a RSS sample of size $n = rm$. The mean of the RSS sample, as an unbiased estimator of the population mean (μ), is found to have smaller variance than the mean of a simple random sample (SRS) of the same size.

For recent work, consult Patil et al. (1999), Al-Saleh and Al-Kadiri (2000), Al-Saleh and Samawi (2000), Chen (2000), Samawi (2001), Zheng and Al-Saleh (2002) and Al-Saleh and Al-Omari (2002).

The RSS procedure is rarely applicable with large set size m . Ranking a large set of elements is not possible without committing errors of ranking. Ranking errors can destroy the efficiency gain of using RSS instead of SRS. Extreme Ranked Set Sampling (ERSS), as introduced and investigated by Samawi et al. (1996), is a modified procedure of RSS that consists of choosing for quantification the first and the last (Judgment) ordered statistics. In other words, the ERSS procedure consists of drawing m random samples of size m each from the population.

Then, the smallest observation (identified by judgment) from each of the first $\frac{m}{2}$ sets and the largest observation for each of the last $\frac{m}{2}$ sets are chosen for actual quantification. The ERSS consists of these m selected units, assuming that m is even. It turns out that this procedure, besides being more applicable, can be more efficient than RSS procedure in case of uniform distributions and more efficient than SRS in case of symmetric distributions.

A new RSS plan for multiple characteristics was introduced recently by Al-Saleh and Zheng (2002). For simplicity, they introduced the method for two characteristics and refer to it as *Bivariate Ranked Set Sampling* (BVRSS). It is believed that both characteristics will benefit from this scheme of BVRSS. There are situations, when several attributes are to be studied simultaneously using a single combined study rather than separate studies, one for each characteristic. For example, in situations where quantifications entail destruction of units as in uprooting of plants. Also, analytical procedures such as *spectroscopy* can be used to quantify several contaminants at once (Patil *et al.* 1994); also Mode *et al.* (1999) and Al-Saleh and Zheng (2002) for more applications.

Suppose (X, Y) is a bivariate random vector with pdf $f_{X,Y}(x, y)$. Let θ and μ be the means of X and Y , respectively. To obtain a BVRSS sample follow the five steps described below:

- 1) For a given set size m , a random sample of size m^4 is identified from the population and randomly allocated into m^2 pools of size m each, where each pool is a square matrix with m rows and m columns.
- 2) In the first pool, identify the minimum value by judgment w.r.t. the first characteristic, for each of the m rows.
- 3) For the m minima obtained in Step 2, choose the pair that corresponds to the minimum value of the second characteristic, identified by judgment, for actual quantification.

This pair, which resembles the label $(1, 1)$, is the first element of the BVRSS sample.

- 4) Repeat Steps 2 and 3 for the second pool, but the pair that corresponds to the first minimum value w.r.t. the first characteristic and the second minimum value w.r.t. the second characteristic is chosen for actual quantification.

This pair resembles the label $(1, 2)$.

- 5) The process continues until the label (m, m) is resembled from the m^2 th (last) pool. This process produces a BVRSS sample of size m^2 . If a sample of higher size is required, then the whole process can be repeated r times until the required size $n = 2rm$ is achieved. Note that although m^4 units are identified for the BVRSS sample, only m^2 are chosen for actual quantification. However all m^4 units contribute information to the m^2 quantified units.

In this article, the ERSS is combined with BVRSS to obtain a more applicable procedure namely the *Bivariate Extreme Ranked Set Sampling* (BVERSS). In section 2, the procedure is described and some fundamental properties will be given. Application to bivariate normal distribution is introduced in Section 3. Section 4 provides illustration to the procedure using real data set.

Methodology

Assume that (X, Y) is a bivariate random variable with the joint density function (p.d.f) $f_{X,Y}(x, y)$. To obtain a BVERSS follow the following steps:

- 1) For a given set size m , $4m$ random samples of size m each are drawn from the population.
- 2) For each of the first m samples drawn in (1), the minimum with respect to the X -characteristic is identified by Judgment. Among the m pairs identified in this step, the pair that corresponds to the minimum with respect to the Y -characteristic is identified. This pair is the first element in the BVERSS. This element is chosen for actual quantification.

3) For each of the second m samples drawn in (1), the minimum with respect to the X -characteristic is identified by Judgment. Among the m pairs identified in this step, the pair that corresponds to the maximum with respect to the Y -characteristic is identified. This pair is the second element in the BVERSS. This element is chosen for actual quantification.

4) For each of the third m samples drawn in (1), the maximum with respect to the X -characteristic is identified by Judgment. Among the m pairs identified in this step, the pair that corresponds to the minimum with respect to the Y -characteristic is identified. This pair is the third element in the BVERSS. This element is chosen for actual quantification.

5) For each of the fourth m samples drawn in (1), the maximum with respect to the X -characteristic is identified by Judgment. Among the m pairs identified in this step, the pair that corresponds to the maximum with respect to the Y -characteristic is identified. This pair is the fourth element in the BVERSS. This element is chosen for actual quantification.

The above 5 steps leads to a BVERSS of size 4. The above steps can be repeated, if necessary, r times to obtain a sample of size $n = 4r$.

Denote the elements obtained in the second step by $(X_{(1)j}, Y_{[1]j})$; where for $j=1,2,\dots,m$, $X_{(1)j}$ denotes the minimum of the m elements in the j th set and $Y_{[1]j}$ is the corresponding Y -value, where the squared brackets is used here to denote the induced rank of Y by the actual rank of the X .

let $Y_{[1](1)} = \min_j (Y_{[1]j})$ and let $X_{(1)[1]}$ be the corresponding X -value then $(X_{(1)[1]}, Y_{[1](1)})$ denotes the first element in the BVERSS. The other three elements of the first cycle are defined similarly and will be denoted by

$$\begin{aligned} & (X_{(1)[m]}, Y_{[1](m)}), \\ & (X_{(m)[1]}, Y_{[m](1)}), \\ & (X_{(m)[m]}, Y_{[m](m)}). \end{aligned}$$

Now for the k th cycle, let

$$\{(X_{(1)[1],k}, Y_{[1](1),k}), (X_{(1)[m],k}, Y_{[1](m),k}), \\ (X_{(m)[1],k}, Y_{[m](1),k}), (X_{(m)[m],k}, Y_{[m](m),k})\}$$

be the chosen BVERSS, $k=1,2,\dots,r$. $(X_{(1)[1],k}, Y_{[1](1),k})$ are independent and identically distributed (iid) with common joint density $f_{X_{(1)[1]}, Y_{[1](1)}}$ given by

$$f_{X_{(1)[1]}, Y_{[1](1)}}(x, y) = f_{Y_{[1](1)}}(y) \frac{f_{X_{(1)}}(x) f_{Y|X}(y|x)}{f_{Y_{[1]}}(y)} \quad 2.1$$

where $f_{X_{(1)}}(x)$ is the density of the first order statistics of an iid sample from the marginal density $f_X(x)$, given by

$$f_{X_{(1)}}(x) = m(1 - F_X(x))^{m-1} f_X(x);$$

$$f_{Y_{[1]}}(y) = \int_{-\infty}^{\infty} f_{X_{(1)}}(x) f_{Y|X}(y|x) dx; \quad f_{Y_{[1](1)}}(y)$$

is the density of the first order statistics of an iid sample from $f_{Y_{[1]}}(y)$. Similarly, for the other three quantities the joint densities are respectively given by:

$$f_{X_{(1)[m]}, Y_{[1](m)}}(x, y) = f_{Y_{[1](m)}}(y) \frac{f_{X_{(1)}}(x) f_{Y|X}(y|x)}{f_{Y_{[1]}}(y)} \quad 2.2$$

$$f_{X_{(m)[1]}, Y_{[m](1)}}(x, y) = f_{Y_{[m](1)}}(y) \frac{f_{X_{(m)}}(x) f_{Y|X}(y|x)}{f_{Y_{[m]}}(y)} \quad 2.3$$

$$f_{X_{(m)[m]}, Y_{[m](m)}}(x, y) = f_{Y_{[m](m)}}(y) \frac{f_{X_{(m)}}(x) f_{Y|X}(y|x)}{f_{Y_{[m]}}(y)} \quad 2.4 \text{ (Saleh and Zheng, 2002).}$$

Note that if X and Y are uncorrelated then $X_{(1)[1],k} \underline{\underline{d}} X_{(1)[m]} \underline{\underline{d}} X_{(1)}$ and

$X_{(m)[1],k} \underline{\underline{d}} X_{(m)[m],k} \underline{\underline{d}} X_{(m)}$. Similar statements can be said about the Y 's. Thus, in this case the *BVERSS* is equivalent to an *ERSS* sample of size $4r$ from the X -population, and an *ERSS* sample of size $4r$ from the Y -population. This means that there is no gain of using *BVERSS* instead of using *ERSS*. However, there are situations when several characteristics are to be investigated simultaneously and using a single combined study rather than a separate study for each attribute. (Al-Saleh and Zheng 2002).

On the other extreme, if X and Y are perfectly correlated then $X_{(1)[1],k} \underline{\underline{d}} X_{(1)[1]}$; $X_{(1)[m],k} \underline{\underline{d}} X_{(1)[m]}$, $X_{(m)[1],k} \underline{\underline{d}} X_{(m)[1]}$; and $X_{(m)[m],k} \underline{\underline{d}} X_{(m)[m]}$. Thus, in this case for the first variable, the *BVERSS* is equivalent to two *ERSS* of size $2r$ each one from $f_{X_{(1)}}$ and the other from $f_{X_{(m)}}$. Similar statements can be said about the Y -variable. Therefore, the advantage of *BVERSS* over the (univariate) *ERSS* is obvious.

Let

$$\mu = E(X); \theta = E(Y); \sigma^2 = \text{Var}(X); \tau^2 = \text{Var}(Y) \text{ and } \rho = \text{Corr}(X, Y)$$

Assume that there is a *BVERSS* of size $n=4r$ given by

$$\{(X_{(1)[1],k}, Y_{[1](1),k}), (X_{(1)[m],k}, Y_{[1](m),k}), \\ (X_{(m)[1],k}, Y_{[m](1),k}), (X_{(m)[m],k}, Y_{[m](m),k})\}$$

Let

$$\bar{X}_k = \frac{X_{(1)[1],k} + X_{(1)[m],k} + X_{(m)[1],k} + X_{(m)[m],k}}{4}$$

then $\hat{\mu}_{BVERSS} = \frac{1}{r} \sum_{k=1}^r \bar{X}_k$ 2.5 is an estimator of μ based on the *BVERSS*. Similarly $\hat{\theta}_{BVERSS}$ can be defined as an estimator of θ .

$$\begin{aligned} \text{Now, let } \mu_{(1)[1]} &= E(X_{(1)[1]}); \\ \mu_{(1)[m]} &= E(X_{(1)[m]}); \mu_{(m)[1]} = E(X_{(m)[1]}); \\ \mu_{(m)[m]} &= E(X_{(m)[m]}); \\ \sigma^2_{(1)[1]} &= \text{Var}(X_{(1)[1]}); \sigma^2_{(1)[m]} = \text{Var}(X_{(1)[m]}); \\ \sigma^2_{(m)[1]} &= \text{Var}(X_{(m)[1]}); \\ \sigma^2_{(m)[m]} &= \text{Var}(X_{(m)[m]}). \end{aligned}$$

Then,

$$E(\bar{X}_k) = \frac{\mu_{(1)[1]} + \mu_{(1)[m]} + \mu_{(m)[1]} + \mu_{(m)[m]}}{4}$$

and

$$\text{Var}(\bar{X}_k) = \frac{\sigma^2_{(1)[1]} + \sigma^2_{(1)[m]} + \sigma^2_{(m)[1]} + \sigma^2_{(m)[m]}}{16}$$

Hence,

$$\begin{aligned} E(\hat{\mu}_{BVERSS}) &= \\ E(\bar{X}_k) &= \frac{\mu_{(1)[1]} + \mu_{(1)[m]} + \mu_{(m)[1]} + \mu_{(m)[m]}}{4} \end{aligned} \quad 2.6$$

and

$$\begin{aligned} \text{Var}(\hat{\mu}_{BVERSS}) &= \\ \text{Var}(\bar{X}_k) &= \frac{\sigma^2_{(1)[1]} + \sigma^2_{(1)[m]} + \sigma^2_{(m)[1]} + \sigma^2_{(m)[m]}}{16r} \end{aligned} \quad 2.7$$

Similar formulas can be obtained for $\hat{\theta}_{BVERSS}$.

Note that the performances of $\hat{\mu}_{BVERSS}$ and $\hat{\theta}_{BVERSS}$ depend on the properties of the joint distribution of X and Y . Though not explicitly seen in the above formula, the means and variances of the two estimators depend on the relation between the two variables; Values of $\mu_{(i)[j]}$ and $\sigma^2_{(i)[j]}$ depend on the joint distribution of X and Y .

Now assume that (X, Y) have the joint density $f_{X,Y}(x, y)$ which is symmetric in both variable around (μ, θ) , i.e.

$$f_{X,Y}(x - \mu, y - \theta) = f_{X,Y}(-x + \mu, -y + \theta).$$

Then each of X and Y has a symmetric marginal distribution. As result of that

$X_{(1)} - \mu \underline{\underline{d}} - X_{(m)} + \mu$ and
 $Y_{(1)} - \theta \underline{\underline{d}} - Y_{(m)} + \theta$. The following lemma summarizes other related results.

Lemma (1): Under the above assumptions exist

$$\begin{aligned} \text{i.} \quad & X_{[1]} - \mu \underline{\underline{d}} - X_{[m]} + \mu & \& \\ & Y_{[1]} - \theta \underline{\underline{d}} - Y_{[m]} + \theta \\ \text{ii.} \quad & X_{(1)[1]} - \mu \underline{\underline{d}} - X_{(m)[m]} + \mu & \& \\ & Y_{[1](1)} - \theta \underline{\underline{d}} - Y_{[m](m)} + \theta \\ \text{iii.} \quad & X_{(1)[m]} - \mu \underline{\underline{d}} - X_{(m)(1)} + \mu & \& \\ & Y_{[1](m)} - \theta \underline{\underline{d}} - Y_{[m](1)} + \theta. \end{aligned}$$

Proof: (i) Without loss of generality, assume that $\theta = \mu = 0$. Then exist

$$\begin{aligned} f_{Y_{[1]}}(y) &= \int_{-\infty}^{\infty} f_{X_{(1)}}(x) f_{Y|X}(y|x) dx \\ f_{Y_{[1]}}(-y) &= \int_{-\infty}^{\infty} f_{X_{(1)}}(x) f_{Y|X}(-y|x) dx \\ &= \int_{-\infty}^{\infty} f_{X_{(1)}}(-x) f_{Y|X}(-y|-x) dx \\ &= \int_{-\infty}^{\infty} f_{X_{(m)}}(x) f_{Y|X}(y|x) dx \\ &= f_{Y_{[m]}}(y) \end{aligned}$$

For the other variable, the proof is similar.

$$\begin{aligned} \text{(ii)} \quad & f_{Y_{[1](1)}}(y) = m(1 - F_{Y_{[1]}}(y))^{m-1} f_{Y_{[1]}}(y) \\ & f_{Y_{[1](1)}}(-y) = m(1 - F_{Y_{[1]}}(-y))^{m-1} f_{Y_{[1]}}(-y) \\ & = m(F_{Y_{[m]}}(y))^{m-1} f_{Y_{[m]}}(y) \end{aligned}$$

hence, $Y_{[1](1)} - \theta \underline{\underline{d}} - Y_{[m](m)} + \theta$.

From (2.1), $X_{(1)[1]} - \mu \underline{\underline{d}} - X_{(m)[m]} + \mu$ iff

$Y_{[1](1)} - \theta \underline{\underline{d}} - Y_{[m](m)} + \theta$. (iii) follows similarly.

As a consequence of Lemma (1), the following properties of $\hat{\mu}_{BVERSS}$ and $\hat{\theta}_{BVERSS}$, which can be shown easily.

Lemma (2): Under the above assumptions exist

i. $\hat{\mu}_{BVERSS}$ and $\hat{\theta}_{BVERSS}$ are unbiased estimators of μ and θ ; respectively.

$$\text{ii. } Var(\hat{\mu}_{BVERSS}) = \frac{\sigma^2_{(1)[1]} + \sigma^2_{(1)[m]}}{8r};$$

$$Var(\hat{\theta}_{BVERSS}) = \frac{\tau^2_{[1](1)} + \sigma^2_{[1](m)}}{8r}.$$

Examples: (i) Assume that the marginal distribution of X is uniform on the interval $(0, \delta)$. Then it is straight forward to show that

$$\mu_{(1)} = \frac{\delta}{m+1}; \quad \mu_{(m)} = \frac{m\delta}{m+1};$$

$$\sigma^2_{(1)} = \sigma^2_{(m)} = \frac{m\delta^2}{(m+1)^2(m+2)}.$$
 Thus

the efficiency of $\hat{\mu}_{ERSS}$ with respect to the mean $\bar{X}$ of a simple random sample of equivalent size is

$$eff(\hat{\mu}_{ERSS}; \bar{X}) = \frac{(m+2)(m+1)^2}{12m}.$$

This is the same quantity reported by Samawi et al. (1996). From this formula, $eff(\hat{\mu}_{ERSS}; \bar{X})$ is always larger than 1; its value for $m=2, 4$ and 6 are, respectively, 1.50, 3.13 and 5.44. It can be shown that (with $\delta = 1$) that

$$f_{X_{(1)(1)}}(x) = m^2(1-x)^{m-1}(1-(1-x)^m)^{m-1}$$

$$f_{X_{(1)(m)}}(x) = m^2(1-x)^{m^2-1}$$

$$f_{X_{(m)(1)}}(x) = m^2x^{m-1}(1-x^m)^{m-1}$$

$$f_{X_{(m)(m)}}(x) = m^2x^{m^2-1}$$

Thus, for any given m ; the mean and variance of each of these can be obtained easily. In the

best situation when $\rho = 1$, $eff(\hat{\mu}_{MVERSS}; \bar{X})$ can be obtained for any value of m . The values of this efficiency for $m=2, 4$ and 6 are respectively 2.19; 4.07 and 5.95.

(ii) Assume that the marginal distribution of X is exponential with mean μ . In this case, as shown numerically by Samawi et al. (1996),

$eff(\hat{\mu}_{ERSS}; \bar{X})$ is decreasing in m ; its values

for $m=2,4$ and 6 are respectively 1.33; 1.17 and 0.75. It can be shown that (with $\delta = 1$) that

$$\begin{aligned} f_{X_{(1)(1)}}(x) &= m^2 e^{-m^2 x} \\ f_{X_{(1)(m)}}(x) &= m^2 e^{-mx} (1 - e^{-mx})^{m-1} \\ f_{X_{(m)(1)}}(x) &= m^2 e^{-x} (1 - e^{-x})^{m-1} (1 - (1 - e^{-x})^m)^{m-1} \end{aligned}$$

$$f_{X_{(m)(m)}}(x) = m^2 e^{-x} (1 - e^{-x})^{m^2-1}$$

Thus, for any given m ; the mean and variance of each of these can be obtained easily. In the best situation when $\rho = 1$,

$eff(\hat{\mu}_{BVERSS}; \bar{X})$ can be obtained for any value of m . Using Scientific work place, $eff(\hat{\mu}_{BVERSS}; \bar{X})$ was evaluated for some values of m . For $m=2,4$ and 6, the efficiency found to be respectively, 1.82; 1.36 and 0.78: Thus the estimator here doesn't perform well. Note that the distribution in this case is not symmetric and the estimator is biased. Next presented is a case of bivariate normal distribution.

Assume next that (X, Y) has the bivariate normal density given by

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma\tau\sqrt{1-\rho^2}} \\ &e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{y-\theta}{\tau}\right)^2 + \left(\frac{x-\mu}{\sigma}\right)\left(\frac{y-\theta}{\tau}\right)\right)} \end{aligned}$$

where, $\mu, \theta, \sigma^2, \tau^2$, and ρ are, respectively, the mean of X , the mean of Y , the variance of X , the variance of Y , and the correlation between X and Y . Denote this bivariate normal density by $N_2(\mu, \theta, \sigma^2, \tau^2, \rho)$. Using (2.1) above, the joint density of $X_{(1)[1]}, Y_{[1](1)}$ can be written as

$$f_{X_{(1)[m]}, Y_{[1](m)}}(x, y) = f_{Y_{[1](m)}}(y) \frac{f_{X_{(1)}}(x) f_{Y|X}(y|x)}{f_{Y_{[1]}}(y)} \quad 3.1$$

$$= m^2 (1 - F_{Y_{[1]}}(y))^{m-1} (1 - F_X(x))^{m-1} f_X(x) f_{Y|X}(y|x), \quad 3.2$$

and the joint density of $X_{(1)[m]}, Y_{[1](m)}$ can be written as

$$f_{X_{(m)[1]}, Y_{[1](1)}}(x, y) = f_{Y_{[1](1)}}(y) \frac{f_{X_{(m)}}(x) f_{Y|X}(y|x)}{f_{Y_{[1]}}(y)} \quad 3.3$$

$$= m^2 (F_{Y_{[1]}}(y))^{m-1} (1 - F_X(x))^{m-1} f_X(x) f_{Y|X}(y|x), \quad 3.4$$

For simplicity assume $\mu = \theta = 0$ and $\sigma^2 = \tau^2 = 1$, (easily one can go back to the general case), then it can be shown that

$$f_{Y_{[1]}}(y) = 2\phi(x)\Phi\left(\frac{-\rho x}{\sqrt{2-\rho^2}}\right) = f_{Y_{m[1]}}(-y) \text{ where}$$

ϕ & Φ are, respectively, the density and the cumulative distribution function of the standard normal distribution. Hence,

$$\begin{aligned} f_{X_{(1)[1]}, Y_{[1](1)}}(x, y) &= \\ m^2 (\Phi(-x))^{m-1} \phi_{X,Y}(x, y) \\ \left[\int_{-\infty}^y 2\phi(z)\Phi\left(\frac{-\rho z}{\sqrt{2-\rho^2}}\right) dz \right]^{m-1} \\ f_{X_{(1)[m]}, Y_{[1](m)}}(x, y) &= \\ m^2 (\Phi(-x))^{m-1} \phi_{X,Y}(x, y) \\ \left[1 - \int_{-\infty}^y 2\phi(z)\Phi\left(\frac{-\rho z}{\sqrt{2-\rho^2}}\right) dz \right]^{m-1} \end{aligned}$$

and,

$$\begin{aligned} f_{X_{(1)[1]}}(x) &= \\ m^2 (\Phi(-x))^{m-1} \int_{-\infty}^{\infty} \{ \phi_{X,Y}(x, y) \\ \left[\int_{-\infty}^y 2\phi(z)\Phi\left(\frac{-\rho z}{\sqrt{2-\rho^2}}\right) dz \right]^{m-1} \} dy \end{aligned}$$

$$f_{X_{(1)[m]}}(x) = m^2 (\Phi(-x))^{m-1} \int_{-\infty}^{\infty} \left\{ \phi_{X,Y}(x,y) \left[1 - \int_{-\infty}^y 2\phi(z) \Phi\left(\frac{-\rho z}{\sqrt{2-\rho^2}}\right) dz \right]^{m-1} \right\} dy$$

The mean and variance of $X_{(1)[1]}$ and $X_{(1)[m]}$ can be evaluated numerically and hence the variance of the unbiased estimator $\hat{\mu}_{BVERS}$ can be obtained. Its efficiency with respect to the sample mean of a simple random sample of equivalent sample size can be obtained. Since the efficiency depends on ρ , it will be denoted by $eff(\hat{\mu}_{BVERS}; \bar{X})$.

If $\rho = 0$, then $eff_0(\hat{\mu}_{BVERS}; \bar{X}) = eff(\hat{\mu}_{ERSS}; \bar{X})$. In this case the efficiency was reported by Samawi et al. (1996). For $m=2,4$ and 6, it is, respectively, 1.47; 2.03; 2.39.

If $\rho = 1$, it can be shown that

$$f_{X_{(1)(1)}}(x) = m^2 \Phi^{m^2-1}(-x) \phi(x) \quad 3.9$$

$$f_{X_{(1)(m)}}(x) = m^2 \left[1 - \Phi^m(-x) \right]^{m-1} \Phi^{m-1}(-x) \phi(x). \quad 3.10$$

The variance of $X_{(1)(1)}$ and $X_{(1)(m)}$ were obtained using *Scientific Work Place*. Based on these values $eff(\hat{\mu}_{BVERS}; \bar{X})$ is calculated for some values of m . The results are given in the following table.

Table 1. Efficiency of $eff_1(\hat{\mu}_{BVERS}; \bar{X})$ with respect to $\bar{X}$ for $\rho = 0,1$.

m	$\sigma_{(1)(1)}^2$	$\sigma_{(1)(m)}^2$	$eff_1(\hat{\mu}_{BVERS}; \bar{X})$	$eff_0(\hat{\mu}_{BVERS}; \bar{X}) = eff(\hat{\mu}_{ERSS}; \bar{X})$
2	0.4989	0.4389	2.15	1.47
4	0.2949	0.1996	4.04	2.03
6	0.2344	0.1295	5.05	2.39

Table 2. Efficiency of $\hat{\mu}_{BVERS}$ with respect to $\bar{X}$ based on 5000 simulation.

m	$\rho = 0.10$	$\rho = 0.30$	$\rho = 0.50$	$\rho = 0.70$	$\rho = 0.90$
2	1.49	1.51	1.59	1.74	1.98
4	2.10	2.12	2.33	2.62	3.34
6	2.34	2.54	2.82	3.45	4.42

Table 3. Efficiency of $\hat{\mu}_{BVRSS}$ with respect to $\bar{X}$ based on 5000 simulation.

m	$\rho = 0.00$	$\rho = 0.10$	$\rho = 0.30$	$\rho = 0.50$	$\rho = 0.70$	$\rho = 0.90$
2	1.47	1.49	1.51	1.59	1.74	1.98
4	2.30	2.54	2.60	2.74	3.13	4.17
6	3.05	3.16	3.43	3.67	4.84	6.92

For other values of ρ , the marginal densities of $X_{(1)[1]}$ and $X_{(1)[m]}$ given by (3.7 and 3.8) can't be simplified further. Based on simulation with 5000 replications, the efficiency for some values of ρ is given in Table 2.

Next, to compare *BVERSS* with the usual *BVRSS*, i.e. to find the efficiency of *BVRSS* with respect to *SRS* for estimating μ . This efficiency was calculated for $m=2$ and 3 by Al-Saleh and Zheng (2002). Table 3 contains this efficiency for $m=2,4$ and 6.

It is clear that when $m=2$ then *BVERSS* is the same as *BVRSS*. Table 1 and Table 2 show that the *BVERSS* is substantial more efficient than *SRS* and comparing with Samawi et al. (1996), it is more efficient than *ERSS* and *RSS* in case of bivariate normal distribution. Also, the efficiency of *BVRSS* w.r.t. *SRS* is increasing with increasing the set size m and the correlation coefficient ρ . Although Table 3 shows that *BVRSS* is more efficient than our proposed *BVERSS* assuming no error in ranking, *BVERSS* still more practical than *BVRSS* and less prone to ranking error.

Results

The *BVERSS* estimation procedure is illustrated using a real data set which consists of the height (Y) and the diameter (X) at breast height of 399 trees. See Platt et al. (1988) for a detailed description of the data set. The summary statistics of the original data are reported in Table 4. Note that the correlation coefficient $\rho = 0.908$.

In this article, ranking is performed on the both variables exactly measured. However, in practice ranking is done before any actual quantification. Using a set size $m=4$ and cycle size $r=4$, bivariate *SRS*, *BVRSS* and *BVERSS* of size 16 are drawn. The analysis to the tree data showed that the distributions of X and Y have skewed to the right shape. So to compare between *BVERSS* and *BVRSS* the means for the transformed data by using the natural logarithm were estimated. Table 5 contains all the above proposed estimators using the drawn samples. Also, provided are estimates for the efficiency based on 1000 repeated sampling.

Table 4. Summary statistics of trees data.

Variable	Mean	Variance
Height (Y) in feet	52.36	325.14
Diameter (X) in cm	20.84	310.11

Table 5. Results of the selected samples of transformed trees data.

Variable	Mean	eff (<i>BVRSS</i> ; <i>SRS</i>)	eff (<i>BVERSS</i> ; <i>SRS</i>)
Ln (Height (Y))	3.39	5.02	4.97
Ln (Diameter (X))	2.61	4.82	4.88

Conclusion

From the above results, support exists that *BVERSS* procedure can be, in some situations, much better than the bivariate *SRS*, *ERSS* and *RSS* (using concomitant variable) sampling methods for estimating the distribution means of multiple characteristics. Also, *BVERSS* provides unbiased estimators for distribution means in case of symmetric marginal distributions. Finally, *BVERSS* is more practical than *BVRSS* and less prone to ranking error.

References

- Al-Saleh, M. F., & Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *Australian and New Zealand Journal of Statistics*, 44, 221-232.
- Al-Saleh, M. F., & Al-Omari, A. (2002). Multi-Stage RSS. *Journal of Statistical Planning and Inference*, 102, 273-286.
- Al-Saleh, M. F., & Al-Kadiri, M. (2000). Double ranked set sampling. *Statistics and Probability Letters*, 48, 205-212.
- Al-Saleh, M. F., & Samawi, H. (2000). On the efficiency of Monte Carlo methods using steady state ranked simulated samples. *Communication in Statistics-Simulation and Computation*, 29, 941-954.
- Chen, Z. (2000). The efficiency of ranked set sampling relative to simple random sampling under multi-parameter family. *Statistica Sinica*, 10, 247-263.
- McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385-390.
- Mode, N., Conquest, L., & Marker, D. (1999). Ranked set sampling for ecological research: Accounting for the total costs of sampling. *Environmetrics*, 10, 179-194.
- Patil, G. P., Sinha, A. K. & Taillie, C. (1999). Ranked set sampling: A bibliography. *Environmental Ecology Statistics*, 6, 91-98.
- Platt, W. J., Evans, G. W., Rathbun, S. L. (1988). The population dynamics of a long-lived conifer. *The American Naturalist*, 131, 391-525.
- Samawi, H. M. (2001). On quantiles estimation using ranked samples with some applications. *Journal of the Korean Statistical Society*, 30, 667-678.
- Samawi, H. M., Ahmad, M. S. & Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrics Journal*, 38, 577-586.
- Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.
- Zheng, G., & Al-Saleh, M. F. (2002). Modified maximum likelihood estimators based on ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 54, 641-658.

Kernel-Based Estimation of $P(X < Y)$ With Paired Data

Omar Eidous
Department of Statistics
Irbid – Jordan

Ayman Baklizi
Department of Statistics
Irbid – Jordan

A point estimation of $P(X < Y)$ was considered. A nonparametric estimator for $P(X < Y)$ was developed using the kernel density estimator of the joint distribution of X and Y , may be dependent. The resulting estimator was found to be similar to the estimator based on the sign statistic, however it assigns smooth continuous scores to each pair of the observations rather than the zero or one scores of the sign statistic. The asymptotic equivalence of the sign statistic and the proposed estimator is shown and a simulation study is conducted to investigate the performance of the proposed estimator. Results indicate that the estimator has a good overall performance.

Key words: Kernel density estimation, stress-strength reliability, paired data

Introduction

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sample of n independent pairs of possibly dependent jointly distributed random variables (X, Y) . The aim is to estimate $p = P(X < Y)$ using this information. The problem of estimation the stress-strength reliability arises naturally in the context of mechanical reliability of a system with strength X and stress Y . The system fails any time its strength is exceeded by the stress applied to it. Another interpretation of p is that it measures the effect of the treatment when X is the response for a control group and Y refers to the treatment group. Other applications can be found in Johnson et. al. (1994) and the references therein.

Ayman Baklizi is an Assistant Professor of Applied Statistics. His research interests are in accelerated life tests and censored data. Email: baklizi1@hotmail.com. Omar Eidous is an Assistant Professor of Applied Statistics. His research interests are in line transect sampling and kernel methods. Email: omarm@yu.edu.jo

The sign statistic (Lehmann, 1975) is defined as $S = \sum_{i=1}^n \phi(X_i, Y_i)$ where $\phi(X_i, Y_i) = 1$ if $X_i < Y_i$ and 0 otherwise. It is readily seen that

$$E(S) = E\left(\sum_{i=1}^n \phi(X_i, Y_i)\right) = nP(X < Y),$$

therefore an unbiased estimator for p is given by

$$\hat{p}_1 = \frac{1}{n} E\left(\sum_{j=1}^n \phi(X_j, Y_j)\right).$$

Develop in this article is a new estimator for $P(X < Y)$ using kernel methods (Silverman, 1986). The kernel density estimators are used instead of the true unknown density and the estimator of p is introduced with some of its large sample properties. A simulation study was conducted to evaluate the performance of the proposed estimator and compare it with the estimator $\hat{p}_1$.

Methodology

The Kernel – Based Estimator

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be n independent pairs drawn from the distribution

with joint probability density function $f(x, y)$. The desired parameter p to estimate is

$$p = P(X < Y) = \int_{-\infty}^{\infty} \int_{-\infty}^y f(x, y) dx dy. \quad (1)$$

In this article a nonparametric kernel method is used to estimate p . The kernel estimator of the two dimensional probability density function f at (x, y) is defined as (Scott, 1992).

$$\hat{f}(x, y) = \frac{1}{nh_1 h_2} \sum_{i=1}^n K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right), \quad (2)$$

where h_1 and h_2 are positive numbers control the smoothness of the fitted curve, usually called bandwidths or smoothing parameters. $K(u)$ is a kernel function which is a symmetric probability density. Comprehensive reviews of the kernel method are available in Silverman (1986); Scott (1992); Wand and Jones (1995). The proposed estimator of the parameter p is constituted by substituting formula (2) in (1) as an estimator for $f(x, y)$. The resulting estimator is of the form

$$\hat{p}_2 = \frac{1}{nh_1 h_2} \sum_{i=1}^n \int_{-\infty}^{\infty} \int_{-\infty}^y K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) dx dy. \quad (3)$$

If the two random variables X and Y are defined on the positive real line, transform the positive data by taking logarithms of each observation as suggested by Silverman (1986).

To construct the kernel estimator $\hat{p}_2$ kernel function K and smoothing parameters h_1 and h_2 must be chosen. For example, the widely used criterion is to choose K , h_1 and h_2 that minimize the mean integral square error (*MISE*) of $\hat{f}(x, y)$. As many authors stated, there is very little to choose between the various kernels as they all contribute the similar amount to the *MISE* (See Silverman, 1986 and Wand and Jones, 1995). The based-data formulas to choose h_1 and h_2 are given later in this paper.

Large Sample Properties

Consider the following transformation,

$$u = \frac{x - X_i}{h_1} \quad \text{and} \quad v = \frac{y - Y_i}{h_2} \quad \text{it follows that}$$

$$\hat{p}_2 = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \int_{-\infty}^{(vh_2 + Y_i - X_i)/h_1} K(u) K(v) du dv.$$

Since the kernel function K is a probability density function. When $h_1 \rightarrow 0$ and $h_2 \rightarrow 0$ such that $h_1 = O(h_2)$ as $n \rightarrow \infty$ the summand will be either zero or one depending on whether $X_i > Y_i$ or $X_i < Y_i$. Hence the integral approaches one if $X_i < Y_i$ and zero if $X_i > Y_i$.

Thus the limiting value of $\hat{p}_2$ is

$$\hat{p}_2 = \frac{1}{n} \sum_{i=1}^n \phi(X_i, Y_i) \quad \text{which is the estimator}$$

based on the sign statistic. Consider,

$$\begin{aligned} E(\hat{p}_1) &= \frac{1}{n} \sum_{i=1}^n E(\phi(X_i, Y_i)) \\ &= \frac{1}{n} \sum_{i=1}^n P(X < Y) \\ &= P(X < Y). \end{aligned}$$

Also

$$\begin{aligned} \text{var}(\hat{p}_1) &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(\phi(X_i, Y_i)) \\ &= \frac{1}{n} \text{var}(\phi(X, Y)) \\ &= \frac{1}{n} p(1 - p). \end{aligned}$$

Because $\hat{p}_1$ and $\hat{p}_2$ are asymptotically equivalent, it follows that $\hat{p}_2$ is asymptotically unbiased. Because the variance of $\hat{p}_2$ tends to zero as $n \rightarrow \infty$, the proposed estimator is consistent. The limiting distributions of $\hat{p}_1$ and $\hat{p}_2$ are also the same.

Small Sample Performance of the Estimators

To implement the estimator $\hat{p}_2$ in practice the kernel function K and the smoothing parameters h_1 and h_2 need to be chosen. Consider the following two kernel functions. The standard normal kernel :

$$K_1(u) = (2\pi)^{-1/2} e^{-u^2/2}, \quad -\infty < u < \infty$$

The rectangular kernel :

$$K_2(u) = \begin{cases} 1/2, & -1 < u < 1 \\ 0, & \text{otherwise} \end{cases},$$

Note here that if $u = (x - X_i)/h_1$ then

$$K_1\left(\frac{x - X_i}{h_1}\right) = (2\pi)^{-1/2} e^{-(x - X_i)^2 / 2h_1^2}$$

$-\infty < x < \infty$

and

$$K_2\left(\frac{x - X_i}{h_1}\right) = \begin{cases} 1/2, & -h_1 + X_i < x < h_1 + X_i \\ 0, & \text{otherwise} \end{cases}.$$

By adopting K_1 it is easy to show that (3) becomes

$$\begin{aligned} \hat{p}_2 &= \frac{h_1}{nh_2\sqrt{2\pi}} \sum_{i=1}^n \int_{-\infty}^{\infty} \Phi(t) e^{-(h_1 t + Y_i - X_i)^2 / (2h_2^2)} dt. \end{aligned} \quad (4)$$

where $\Phi(t)$ is the normal distribution function at t . The estimator $\hat{p}_2$ which is given by equation (4) cannot be written in closed form, so derive the estimator $\hat{p}_2$ corresponding to K_2 . If K_2 is adopted in (3) then we need to study six cases to find the double integral arises in (3). Let $a(i) = -h_1 + X_i$; $b(i) = h_1 + X_i$; $c(i) = -h_2 + Y_i$ and $d(i) = h_2 + Y_i$, where $i = 1, 2, \dots, n$. Notice that, $a(i) < b(i)$ and $c(i) < d(i)$. The proposed estimator is given by

$$\hat{p}_2 = \frac{1}{nh_1 h_2} \sum_{i=1}^n Q_i, \text{ where}$$

$$Q_i = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_i}{h_2}\right) dx dy, \text{ then}$$

Case1: If $d(i) < a(i)$ then $Q_i = 0$

Case2: If $b(i) < c(i)$ then $Q_i = h_1 h_2$

Case3: If $c(i) < a(i) < b(i) < d(i)$ then

$$Q_i = \int_{a(i)}^{b(i)} \int_x^{d(i)} \frac{1}{4} dy dx$$

$$= \frac{1}{4} (b(i)d(i) - b(i)^2 / 2 - a(i)d(i) + a(i)^2 / 2)$$

Case4: If $a(i) < c(i) < d(i) < b(i)$ then

$$Q_i = \int_{c(i)}^{d(i)} \int_{a(i)}^y \frac{1}{4} dx dy$$

$$= \frac{1}{4} (d(i)^2 / 2 - a(i)d(i) - c(i)^2 / 2 + a(i)c(i)).$$

Case5: If $c(i) < a(i) < d(i) < b(i)$ then

$$Q_i = \int_{a(i)}^{d(i)} \int_{a(i)}^y \frac{1}{4} dx dy$$

$$= \frac{1}{4} (d(i)^2 / 2 + a(i)^2 / 2 - a(i)d(i)).$$

Case6: If $a(i) < c(i) < b(i) < d(i)$ then

$$Q_i = \int_{a(i)}^{c(i)} \int_{c(i)}^{d(i)} \frac{1}{4} dy dx + \int_{c(i)}^{b(i)} \int_x^{d(i)} \frac{1}{4} dy dx$$

$$= \frac{1}{4} \left(-(c(i)^2 + b(i)^2) / 2 + d(i)(b(i) - a(i)) + a(i)c(i) \right).$$

On the other hand, the simulation results are depended on the following formulas to choose the smoothing parameters h_1 and h_2

which based on minimizing the asymptotic mean integral square error and by assuming the bivariate normal distribution and the rectangular kernel K_2 (Scott, 1992)

$$h_1 = 1.745(1 - \rho^2)^{5/12} (1 + \rho^2 / 2)^{-1/6} \sigma_1 n^{-1/6}$$

$$h_2 = 1.745(1 - \rho^2)^{5/12} (1 + \rho^2 / 2)^{-1/6} \sigma_2 n^{-1/6},$$

where σ_1 , σ_2 and ρ are the standard deviations and the correlation coefficient respectively, they are estimated from the data. The performances of the sign estimator and the proposed estimator were investigated and compared. The criteria of the bias and mean squared error are used. The relative efficiency of the proposed estimator to the sign estimator is calculated as the ratio of mean squared errors.

A simulation study was conducted to investigate the performance of the estimators. The indices of our simulations are: $n = 10, 20, 40$ p : the true value of $p = P(X < Y)$ and is taken to be 0.1, 0.3, ..., 0.9.

The distribution from which the data are generated: two cases were considered;

1) The bivariate normal distribution $(X, Y) \sim BVN(0, \mu, 1, 1, \rho)$ where ρ is taken as -0.8, -0.4, 0, 0.4, 0.8 and μ is chosen such that we get $p = P(X < Y)$ as chosen above.

2) The Gumbel bivariate exponential distribution (Johnson & Kotz, 1970) with probability density function

$$g(x, y) = e^{-(x+y)} (1 + \alpha(2e^{-x} - 1)(2e^{-y} - 1)),$$

$$x > 0, y > 0.$$

The parameter α is chosen such that the correlation (r) between X and Y is -0.4, -0.2, 0, 0.2, 0.4. The variable X is transformed such that we get $p = P(X < Y)$ as chosen above. For each combination of n and p , 1000 samples were generated for (X, Y) . The estimators are calculated and the following quantities are obtained for both estimators:

The bias of the estimators,

$$Bias = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{p}_2^{(i)} - p).$$

Mean squared errors,

$$MSE_j = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{p}_j^{(i)} - p)^2, j = 1, 2.$$

$$Efficiency = \frac{MSE_1}{MSE_2}.$$

The results are presented in Tables 1 - 2.

Conclusion

Because the results for both kernels are similar, only the results for the uniform kernel are given. The results are presented in Table 1. In both cases of bivariate normal parent distribution and the bivariate exponential case, it is clear that the efficiency of the proposed estimator relative to the sign estimator is greater than one in all cases considered. Concerning the bias performance, it appears that the proposed estimator is almost unbiased. Overall it appears that the proposed estimator has a good performance, this performance may be improved when using more sophisticated types of kernels, bandwidth selection rules, and bias corrections.

References

- Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions: Volume 1*. New York: Wiley.
- Lehmann, E. L. (1975) *Nonparametrics: Statistical methods based on ranks*. Holden Day: San Francisco.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.
- Silverman, B. W. (1986). *Density estimation*. London: Chapman & Hall.
- Wand, M., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Table 1: Mean Squared Errors and Efficiencies of the Estimators in the Bivariate Normal Case

n	ρ	p	Var	Bias	MSE	EFF	n	ρ	p	var	Bias	MSE	EFF
10	-0.80	0.10	0.009	0.003	0.009	1.052	40	0.40	0.10	0.005	0.006	0.004	1.128
	-0.80	0.30	0.021	0.001	0.021	1.027		0.40	0.30	0.011	0.003	0.009	1.129
	-0.80	0.50	0.025	-0.011	0.025	1.032		0.40	0.50	0.013	0.005	0.010	1.174
	-0.80	0.70	0.021	0.005	0.020	1.042		0.40	0.70	0.011	0.002	0.009	1.152
	-0.80	0.90	0.009	-0.001	0.009	1.045		0.40	0.90	0.005	0.000	0.004	1.128
	-0.40	0.10	0.009	0.001	0.008	1.132		0.80	0.10	0.005	0.002	0.004	1.114
	-0.40	0.30	0.021	-0.005	0.019	1.083		0.80	0.30	0.011	-0.003	0.009	1.109
	-0.40	0.50	0.025	-0.002	0.023	1.101		0.80	0.50	0.013	-0.004	0.011	1.102
	-0.40	0.70	0.021	0.003	0.019	1.099		0.80	0.70	0.011	0.000	0.009	1.080
	-0.40	0.90	0.009	0.002	0.008	1.120		0.80	0.90	0.005	0.000	0.004	1.175
	0.00	0.10	0.009	-0.002	0.007	1.188		-0.80	0.10	0.003	0.000	0.002	1.037
	0.00	0.30	0.021	0.000	0.019	1.135		-0.80	0.30	0.006	-0.001	0.005	1.023
	0.00	0.50	0.025	-0.002	0.023	1.132		-0.80	0.50	0.007	0.002	0.006	1.032
	0.00	0.70	0.021	0.005	0.018	1.126		-0.80	0.70	0.006	0.003	0.005	1.034
	0.00	0.90	0.009	-0.003	0.008	1.123		-0.80	0.90	0.003	0.000	0.002	1.030
	0.40	0.10	0.009	0.007	0.007	1.184		-0.40	0.10	0.003	0.001	0.002	1.074
	0.40	0.30	0.021	0.003	0.018	1.147		-0.40	0.30	0.006	0.006	0.005	1.079
	0.40	0.50	0.025	-0.002	0.020	1.129		-0.40	0.50	0.007	-0.003	0.006	1.109
	0.40	0.70	0.021	-0.004	0.019	1.178		-0.40	0.70	0.006	0.000	0.005	1.115
	0.40	0.90	0.009	-0.001	0.007	1.134		-0.40	0.90	0.003	-0.001	0.002	1.106
20	0.80	0.10	0.009	0.002	0.007	1.154	40	0.00	0.10	0.003	0.002	0.002	1.112
	0.80	0.30	0.021	0.001	0.019	1.144		0.00	0.30	0.006	0.004	0.005	1.132
	0.80	0.50	0.025	0.002	0.023	1.147		0.00	0.50	0.007	0.000	0.005	1.104
	0.80	0.70	0.021	-0.004	0.019	1.106		0.00	0.70	0.006	0.001	0.004	1.124
	0.80	0.90	0.009	0.006	0.008	1.145		0.00	0.90	0.003	-0.004	0.002	1.164
	-0.80	0.10	0.005	0.000	0.004	1.048		0.40	0.10	0.003	0.000	0.002	1.145
	-0.80	0.30	0.011	0.002	0.011	1.027		0.40	0.30	0.006	-0.002	0.004	1.144
	-0.80	0.50	0.013	0.004	0.012	1.030		0.40	0.50	0.007	-0.001	0.006	1.130
	-0.80	0.70	0.011	0.000	0.011	1.022		0.40	0.70	0.006	0.001	0.005	1.165
	-0.80	0.90	0.005	0.001	0.004	1.049		0.40	0.90	0.003	-0.003	0.002	1.161
	-0.40	0.10	0.005	-0.002	0.004	1.093		0.80	0.10	0.003	-0.001	0.002	1.093
	-0.40	0.30	0.011	0.000	0.009	1.097		0.80	0.30	0.006	0.000	0.005	1.090
	-0.40	0.50	0.013	0.000	0.011	1.092		0.80	0.50	0.007	-0.001	0.006	1.094
	-0.40	0.70	0.011	0.003	0.009	1.113		0.80	0.70	0.006	-0.004	0.005	1.082
	-0.40	0.90	0.005	0.000	0.004	1.111		0.80	0.90	0.003	-0.001	0.002	1.090
	0.00	0.10	0.005	0.001	0.004	1.178							
	0.00	0.30	0.011	0.004	0.009	1.148							
	0.00	0.50	0.013	-0.001	0.011	1.152							
	0.00	0.70	0.011	-0.003	0.009	1.110							
	0.00	0.90	0.005	-0.001	0.004	1.120							

Table 2: Mean Squared Errors and Efficiencies of the Estimators in the Bivariate Exponential Case

(r)	n	p	VAR	Bias	MSE	EFF	(r)	n	P	VAR	Bias	MSE	EFF
-0.4	10	0.1	0.009	-0.019	0.006	1.266	0.2	10	0.1	0.011	0.032	0.009	1.158
-0.4		0.3	0.021	-0.028	0.014	1.303	0.2		0.3	0.021	0.029	0.017	1.268
-0.4		0.5	0.025	0.004	0.019	1.320	0.2		0.5	0.024	-0.007	0.019	1.266
-0.4		0.7	0.021	0.029	0.016	1.266	0.2		0.7	0.022	-0.021	0.018	1.263
-0.4		0.9	0.009	0.017	0.006	1.202	0.2		0.9	0.010	-0.024	0.009	1.150
-0.4	20	0.1	0.005	-0.024	0.003	1.232	0.2	20	0.1	0.005	0.024	0.005	1.153
-0.4		0.3	0.011	-0.035	0.009	1.316	0.2		0.3	0.011	0.029	0.010	1.242
-0.4		0.5	0.013	-0.002	0.009	1.260	0.2		0.5	0.013	-0.003	0.010	1.328
-0.4		0.7	0.011	0.034	0.009	1.334	0.2		0.7	0.011	-0.020	0.009	1.206
-0.4		0.9	0.005	0.021	0.003	1.191	0.2		0.9	0.005	-0.030	0.005	1.119
-0.4	40	0.1	0.003	-0.024	0.002	1.292	0.2	40	0.1	0.003	0.025	0.003	1.056
-0.4		0.3	0.006	-0.036	0.005	1.319	0.2		0.3	0.006	0.024	0.005	1.186
-0.4		0.5	0.007	0.005	0.005	1.298	0.2		0.5	0.007	-0.003	0.005	1.300
-0.4		0.7	0.006	0.036	0.005	1.320	0.2		0.7	0.006	-0.021	0.004	1.183
-0.4		0.9	0.003	0.023	0.002	1.220	0.2		0.9	0.003	-0.025	0.003	1.066
-0.2	10	0.1	0.009	-0.006	0.007	1.190	0.4	10	0.1	0.013	0.046	0.012	1.119
-0.2		0.3	0.021	-0.018	0.016	1.277	0.4		0.3	0.024	0.041	0.019	1.244
-0.2		0.5	0.025	-0.005	0.019	1.326	0.4		0.5	0.025	0.001	0.019	1.294
-0.2		0.7	0.021	0.011	0.016	1.255	0.4		0.7	0.023	-0.037	0.019	1.232
-0.2		0.9	0.009	0.008	0.007	1.200	0.4		0.9	0.013	-0.047	0.011	1.135
-0.2	20	0.1	0.005	-0.010	0.003	1.243	0.4	20	0.1	0.005	0.050	0.008	1.074
-0.2		0.3	0.011	-0.010	0.008	1.275	0.4		0.3	0.011	0.043	0.011	1.161
-0.2		0.5	0.013	0.000	0.009	1.329	0.4		0.5	0.013	-0.001	0.010	1.364
-0.2		0.7	0.011	0.016	0.009	1.303	0.4		0.7	0.011	-0.040	0.010	1.204
-0.2		0.9	0.005	0.009	0.004	1.210	0.4		0.9	0.005	-0.046	0.007	1.098
-0.2	40	0.1	0.003	-0.010	0.002	1.262	0.4	40	0.1	0.003	0.048	0.005	0.986
-0.2		0.3	0.006	-0.013	0.004	1.288	0.4		0.3	0.006	0.034	0.006	1.135
-0.2		0.5	0.007	0.001	0.005	1.292	0.4		0.5	0.007	0.000	0.005	1.302
-0.2		0.7	0.006	0.016	0.004	1.258	0.4		0.7	0.006	-0.039	0.006	1.099
-0.2		0.9	0.003	0.010	0.002	1.241	0.4		0.9	0.003	-0.046	0.004	0.976
0.0	10	0.1	0.005	0.005	0.007	1.232							
0.0		0.3	0.011	0.006	0.017	1.275							
0.0		0.5	0.013	0.001	0.018	1.299							
0.0		0.7	0.011	-0.011	0.017	1.306							
0.0		0.9	0.005	-0.006	0.007	1.192							
0.0	20	0.1	0.005	0.006	0.004	1.225							
0.0		0.3	0.011	0.012	0.009	1.256							
0.0		0.5	0.013	-0.004	0.009	1.271							
0.0		0.7	0.011	-0.007	0.008	1.257							
0.0		0.9	0.005	-0.007	0.004	1.215							
0.0	40	0.1	0.003	0.005	0.002	1.179							
0.0		0.3	0.006	0.006	0.004	1.235							
0.0		0.5	0.007	-0.002	0.005	1.264							
0.0		0.7	0.006	-0.008	0.004	1.278							
0.0		0.9	0.003	-0.005	0.002	1.155							

Some Improvements in Kernel Estimation Using Line Transect Sampling

Omar M. Eidous

Department of Statistics, Faculty of Science
Yarmouk University

Kernel estimation provides a nonparametric estimate of the probability density function from which a set of data is drawn. This article proposes a method to choose a reference density in bandwidth calculation for kernel estimator using line transect sampling. The method based on testing the shoulder condition, if the shoulder condition seems to be valid using as reference the half normal density, while if the shoulder condition does not seem to be valid, we will use exponential reference density. Accordingly, the performances of the resultant estimator are studied under a wide range of underlying models using simulation techniques. The results demonstrate the improvements that can be obtained by applying this technique.

Key words: Line transect method; kernel density estimation; shoulder condition

---

Introduction

Line transect sampling is an important technique to estimate population density D of objects in a given region. In line transect sampling an experimenter moves across the region following a specific line with length L looking to the right and to the left of the line and records the perpendicular distance (X_i) from each detected object to the centerline. Assume that n objects has been sighted and the objects on the transect line are seen with probability one. Burnham and Anderson (1976) introduced the fundamental relation for estimating the density of objects D satisfies the following relationship

$$D = \frac{E(n)f(0)}{2L},$$

where $f(x)$ is the conditional density of the line transect distances, given the object is observed. In order to estimate D , one needs to estimate $f(0)$ which is the crucial problem in line transect estimation. When $f(0)$ estimate by an appropriate estimator $\hat{f}(0)$, D can be estimated by

$$\hat{D} = \frac{n\hat{f}(0)}{2L}.$$

Hence, the key aspects in line transect sampling turns out to be modeling $f(x)$ as well as the estimation of $f(0)$.

Various methods have been proposed to estimate $f(0)$ in literature. A parametric approach assuming that $f(x)$ is a member of a family of proper probability density function of known functional form but depend on an unknown parameter(s) θ , where θ may take a vector value and should be estimated by using the perpendicular distances. Estimate θ by $\hat{\theta}$ will lead to $\hat{f}(0) = f(0, \hat{\theta})$, and there is extensive literature on the use of the maximum likelihood techniques for estimation of $f(0)$. See for example, Burnham and Anderson

---

Dr. Omar Eidous is an Assistant Professor of Applied Statistics. His research interests are in line transect sampling and kernel methods. Email: omarm@yu.edu.jo

(1976); Pollock (1978); Burnham et al. (1980) and Buckland (1985).

To remove the model-dependence of the estimator, nonparametric approaches to estimate $f(0)$ can be implemented. A Fourier series is a nonparametric method has been studied in details by Burnham et al. (1980). Recent works has focused on employing the nonparametric kernel method. Some initial efforts in applying the kernel method to line transect sampling have been made by Buckland (1992); Chen (1996) and Mack and Quang (1998).

It has been widely regarded that the performance of the kernel methods depends largely on the smoothing parameter (bandwidth), and depends very little on the form of the kernel (Silverman, 1986), see also the latest three works mentioned above. In this paper we suggest a new estimator for $f(0)$.

The estimator is developed based on the kernel method itself, while Mach and Quang (1998) recommended using the bandwidth referenced to half normal model; the proposed estimator using the bandwidth reference to half normal or negative exponential models depends on testing the shoulder condition. The bandwidth parameter is selected using the half-normal model as a reference when the shoulder condition is true, that is, $f'(0) = 0$, while the negative exponential model is used when the shoulder condition is not true, that is $f'(0) \neq 0$.

In other words, to apply the proposed estimator we need to test whether the dataset at hand satisfies the shoulder condition or not. The bandwidth parameter is chosen by assuming the half normal as the underlying model if the test is accepted and by assuming the negative exponential model if the test is rejected. This method is studied using the simulation technique and the resultant estimator is compared with Mack and Quang (1998)'s estimator.

Methodology

Let $X_1, X_2, \dots, X_n$ be a random sample of perpendicular distances of size n with unknown probability density function $f(x)$. The kernel

estimator $\hat{f}_k(x)$ of $f(x)$ for $x \geq 0$ is (Chen, 1996)

$$\hat{f}_k(x) = \frac{1}{nh} \sum_{i=1}^n \left[K\left(\frac{x - X_i}{h}\right) + K\left(\frac{x + X_i}{h}\right) \right], \quad x \geq 0, \quad (1)$$

where K is a symmetric kernel function and h is the smoothing parameter usually called bandwidth, where both K and h are under the control of the user. Accordingly, the kernel estimator of $f(0)$ is given by

$$\hat{f}_k(0) = \frac{2}{nh} \sum_{i=1}^n K\left(\frac{X_i}{h}\right). \quad (2)$$

As many authors stated, it is very little to choose between the different kernel functions (See for example Silverman, 1986; Wand & Johns, 1995). The crucial problem in kernel density estimation is to select the bandwidth parameter h . The bandwidth controls the smoothness of the fitted density curve. A larger h gives smoother estimate with smaller variance and larger bias. A smaller h produces a rougher estimate with larger variance and smaller bias. One of the most common methods in nonparametric estimation is to find h that minimizing the asymptotic mean integral square error (AMISE) or to minimize the asymptotic mean square error (AMSE) which compromises between the variance and bias of the estimate. In the remaining of this section we derive the AMSE of $\hat{f}_k(0)$. The expected value of $\hat{f}_k(0)$ - which is given by (2)- is

$$\begin{aligned} E(\hat{f}_k(0)) &= \frac{1}{nh} E\left(K\left(\frac{X_1}{h}\right)\right) \\ &= \frac{1}{nh} \int_0^{\infty} K\left(\frac{x_1}{h}\right) f(x_1) dx_1 \end{aligned}$$

Suppose that the underlying probability density function $f(x)$ has a second-order derivative. Let $u = x_1/h$ and using Taylor's series to expand $f(hu)$ around zero. Then, if $h \rightarrow 0$ as $n \rightarrow \infty$,

$$E(\hat{f}_k(0)) = f(0) + 2hf'(0) \int_0^\infty uK(u)du + h^2 f''(0) \int_0^\infty u^2 K(u)du + O(h^3)$$

Suppose that the shoulder condition is true (i.e. $f'(0) = 0$), then the bias of $\hat{f}_k(0)$ is

$$\text{Bias}(\hat{f}_k(0)) = h^2 f''(0) \int_0^\infty u^2 K(u)du + O(h^3),$$

this indicates that the asymptotic bias of kernel estimator is of order $O(h^2)$ under assumption that the shoulder condition holds. If h is related to n in such a way that $h \rightarrow 0$ and $nh \rightarrow \infty$ as $n \rightarrow \infty$, then the variance of $\hat{f}_k(0)$ is

$$\begin{aligned} \text{Var}(\hat{f}_k(0)) &= \frac{4}{nh^2} \text{Var}\left(K\left(\frac{X_1}{h}\right)\right) \\ &= \frac{4f(0)}{nh} \int_0^\infty K^2(u)du + O(n^{-1}). \end{aligned}$$

It is obvious that as $nh \rightarrow \infty$, a $O(nh)^{-1}$ variance is achieved. Accordingly, the AMSE of $\hat{f}_k(0)$ is given by

$$\begin{aligned} \text{AMSE}(\hat{f}_k(0)) &= h^4 f''^2(0) \left(\int_0^\infty u^2 K(u)du \right)^2 \\ &+ \frac{4f(0)}{nh} \int_0^\infty K^2(u)du \end{aligned} \quad (3)$$

where the first term in the right hand side of (3) is the squared bias and the second term is the variance.

Kernel and Bandwidth Selections

Consider the $\text{AMSE}(\hat{f}_k(0))$ - that is given by (3) - as a function of h (say $g(h)$), then differentiate $g(h)$ with respect to h and equating to zero, we get

$$h = \left\{ \frac{f(0) \int_0^\infty K^2(u)du}{\left(f''(0) \int_0^\infty u^2 K(u)du \right)^2} \right\}^{1/5} n^{-1/5}. \quad (4)$$

If the kernel function is chosen as the standard normal function, that is $K_1(u) = \phi(u)$, it is at once apparent that Equation (4) further simplifies to

$$h_1 = 0.892 \left\{ \frac{f(0)}{f''^2(0)} \right\}^{1/5} n^{-1/5}. \quad (5)$$

The value of h_1 can be substituted back into (3) to give as the minimum achievable AMSE for $\hat{f}_k(0)$ given by

$$0.7684 [f''(0)]^{2/5} [f(0)]^{4/5} n^{-4/5}. \quad (6)$$

On the other hand, if the kernel function is chosen as the rectangular function, that is $K_2(u) = 1 - |u|$, if $|u| < 1$ and zero otherwise, (Silverman, 1986) then Equation (4) simplifies to

$$h_2 = 2.169 \left\{ \frac{f(0)}{f''^2(0)} \right\}^{1/5} n^{-1/5} \quad (7)$$

Correspondingly, if the value of h_2 is substituted back into (3) then the minimum achievable AMSE for $\hat{f}_k(0)$ is given by

$$0.7908 [f''(0)]^{2/5} [f(0)]^{4/5} n^{-4/5}. \quad (8)$$

Comparing (6) and (8), the two quantities has the same convergence rates as $n \rightarrow \infty$. If $n < \infty$ then (6) is slightly smaller than (8). In other words, the efficiency that can be obtained when K_2 is used instead of K_1 is less than 3% in the basis of the AMSE.

This conclusion supports the well known result that says, there is a very little to choose between the different kernel functions as

they all contribute the similar amounts to the *AMSE*. Actually, among the different five kernels which are given in Silverman (1986, pp. 43) if the Epanchnikov kernel is used instead of the standard normal kernel then we obtain the maximum efficiency which is less than 4.1%.

Silverman (1986) presented a table contains the efficiencies of different kernel functions with respect to Epanchnikov kernel. His comparative study is achieved on the basis of the *AMISE* instead of the *AMSE* (that we adopted here) and with data support defined on the entire of the real line, while in this study the data support is defined on the positive half of the real line. Accordingly, in the rest of this paper our derivations and computations are based on the standard normal kernel function (K_1). This kernel is differentiable and has all-order derivatives that are required.

The value of h_1 is based on the parameter $f(0)$ that we aim to estimate it. Buckland (1992) and Mack and Quang (1998) overcame this problem by assuming the half normal model as the underlying model of the data and their formula based on minimizing the *AMISE* of the kernel estimator is given by

$$h = 1.06 \sigma n^{-1/5} \quad (9)$$

where σ is estimated practically by its maximum likelihood estimator $\hat{\sigma} = \sqrt{\sum_{i=1}^n x_i^2 / n}$.

Buckland (1992) used Equation (9) for the deer data and reported very similar results to those obtained by Hermite polynomial method. Mack and Quang (1998) recommended the above formula canceling the constant term, that is, $h = \sigma n^{-1/5}$ which is slightly different from Equation (9). Chen (1996) stated that Equation (9) performs quite well when the underlying distribution is close to the half normal distribution, while when the true $f(x)$ is not close to the half normal, the result can be misleading. He suggested an alternative method called "Least Square Cross-Validation Method" (*LSCVM*).

The primary simulation results indicated that the advantage of using the *LSCVM* over

using Formula (9) is not significant despite the computer-intensive procedures that need to apply it. By interesting the last three works mentioned above and the work introduced by Zhang (2001) which was concerned the testing of the shoulder condition, we suggest to use two reference models to choose the bandwidth h . One of these two models is the negative exponential model which does not satisfy the shoulder condition at the origin, and the other is the half normal model which satisfies the shoulder condition at the origin. The criterion to choose between them is by testing the shoulder condition as illustrates in the following section.

Testing the Shoulder Condition

A motivation to assume the half normal or the negative exponential as the underlying model to apply Formula (5) is that, the first model has a shoulder at the origin, while the second one does not. In other words, we expect the reference model that should be used to choose h is the half normal model when the data have a shoulder at the origin, whereas the negative exponential should be used when the data do not have a shoulder at the origin. Accordingly, assume that we are not sure whether the data have the shoulder at the origin or not, in this case and before we decide which model should be used we need to test the shoulder condition.

Zhang (2001) proposed a procedure for testing the shoulder condition of a model based on line transect sampling. Assume that a random sample $x_1, x_2, \dots, x_n$ of perpendicular distances is drawn from a distribution with probability density function $f(x)$. Consider the test $H_0 : f'(0) = 0$ vs $H_1 : f'(0) \neq 0$, according to Zhang (2001), we reject H_0 for

$$\text{large value of } Z = \frac{\sqrt{\sum_{i=1}^n x_i^2}}{\sum_{i=1}^n x_i}. \quad \text{Zhang}$$

constructed a table of critical values of the sampling distribution for Z with respect to different sample sizes by Monte Carlo simulation. For example, at level of significant $\alpha = 0.5$ we reject H_0 in favor of H_1 if

$Z > q$, where $q = 0.1880, 0.1308, 0.0914$ for $n = 50, 100, 200$ respectively.

Accordingly, to choose the bandwidth h we consider the following two steps

1) If H_0 is not rejected, use the half-normal model as a reference model and then compute the value of h by using the following formula

$$h = 0.933 \hat{\sigma} n^{-1/5} \quad (10)$$

2) If H_0 is rejected, use the negative exponential model as a reference model and then compute the value of h by using

$$h = 0.892 \hat{\lambda} n^{-1/5} \quad (11)$$

where $\hat{\sigma}$ is as defined in Section (3) and $\hat{\lambda}$ is the maximum likelihood estimator for the scale parameter λ in the case of the negative exponential density, which is given by

$$\hat{\lambda} = \sum_{i=1}^n x_i / n.$$

Results

To assess the practical impact of our technique, we undertook some numerical investigations in which we compared our proposed estimator with an ordinary estimator given by Mack and Quang (1998). In this numerical study we considered several parent densities. These densities are those considered by Barabesi (2001) which are commonly used in line transect studies. The exponential power family (Pollock, 1978)

$$f(x) = \frac{1}{\Gamma(1+1/\beta)} e^{-x^\beta}, \quad x \geq 0, \quad \beta \geq 1,$$

The hazard-rate family (Hayes and Buckland, 1983)

$$f(x) = \frac{1}{\Gamma(1-1/\beta)} \left(1 - e^{-x^{-\beta}} \right), \quad x \geq 0, \quad \beta > 1$$

and the beta model (Eberhardt, 1968)

$$f(x) = (1 + \beta)(1 - x)^\beta, \quad x \geq 0, \quad \beta \geq 0.$$

In our simulation design, these three families were truncated at some distance w . Four models were selected from the exponential power family with parameter values $\beta = 1.0, 1.5, 2.0, 2.5$ and corresponding truncation points given by $w = 5.0, 3.0, 2.5, 2.0$ (Figure 1a). Four models were selected from the hazard-rate family with parameter values $\beta = 1.5, 2.0, 2.5, 3.0$ and corresponding truncation points given by $w = 20, 12, 8, 6$ (Figure 1b).

Moreover, four models were selected from beta model with parameter values $\beta = 1.5, 2.0, 2.5, 3.0$ and $w = 1$ for all cases (Figure 1c). The considered models cover a wide range of perpendicular distance probability density functions which vary near zero from spike to flat. It should be remarked that the truncated exponential power model with $\beta = 1$ and the beta model do not satisfy the shoulder condition. This choice was made in order to assess the robustness of the considered estimators with respect to the shoulder condition.

For each model and for sample sizes $n = 50, 100, 200$ one thousand samples of distances were randomly drawn. For each model and for each sample size, Table (1) reports the simulated value of the relative bias (RB):

$$RB_i = \frac{E(\hat{f}(0)) - f(0)}{f(0)}, \quad i = 1, 2;$$

the relative mean error (RME)

$$RME_i = \frac{\sqrt{MSE(\hat{f}(0))}}{f(0)}, \quad i = 1, 2,$$

for each considered estimator, and the efficiency (EFF) of the proposed estimator with respect to Mack and Quang (1998)'s estimator,

$$EFF = \frac{MSE_2}{MSE_1},$$

Table (1). Relative Biases (RB) and Relative Mean Error (RME) for the proposed estimator and for the kernel estimator using bandwidth rule based on half normal model.

n	β	w	$RB1$	$RME1$	$RB2$	$RME2$	EFF
Exponential Power Model							
50	1.0	5.0	-0.245	0.283	-0.333	0.347	1.226
100			-0.217	0.238	-0.305	0.313	1.312
200			-0.191	0.207	-0.276	0.282	1.362
50	1.5	3.0	-0.134	0.201	-0.168	0.205	1.021
100			-0.101	0.159	-0.136	0.164	1.032
200			-0.079	0.127	-0.114	0.135	1.065
50	2.0	2.5	-0.067	0.160	-0.084	0.150	0.942
100			-0.059	0.124	-0.071	0.119	0.961
200			-0.044	0.099	-0.053	0.096	0.965
50	2.5	2.0	-0.047	0.144	-0.055	0.137	0.949
100			-0.022	0.119	-0.029	0.112	0.944
200			-0.023	0.091	-0.027	0.088	0.965
Hazard Rate Model							
50	1.5	20.0	-0.174	0.236	-0.402	0.417	1.765
100			-0.118	0.167	-0.354	0.363	2.174
200			-0.072	0.114	-0.306	0.311	2.730
50	2.0	12.0	-0.063	0.166	-0.247	0.276	1.658
100			-0.034	0.119	-0.206	0.225	1.890
200			-0.012	0.086	-0.159	0.172	2.007
50	2.5	8.0	-0.016	0.161	-0.119	0.173	1.077
118			-0.001	0.121	-0.083	0.124	1.032
119			0.009	0.094	-0.053	0.087	0.934
50	3.0	6.0	0.001	0.156	-0.049	0.132	0.845
100			0.000	0.118	-0.034	0.095	0.807
200			0.012	0.099	-0.011	0.073	0.737
Beta Model							
50	1.5	1.0	-0.167	0.218	-0.183	0.219	1.005
100			-0.150	0.182	-0.163	0.186	1.023
200			-0.128	0.149	-0.139	0.155	1.039
50	2.0	1.0	-0.186	0.235	-0.208	0.239	1.018
100			-0.158	0.193	-0.177	0.199	1.030
200			-0.140	0.169	-0.159	0.176	1.037
50	2.5	1.0	-0.205	0.247	-0.231	0.256	1.033
100			-0.170	0.208	-0.198	0.218	1.045
200			-0.152	0.178	-0.179	0.191	1.073
50	3.0	1.0	-0.201	0.249	-0.237	0.261	1.048
100			-0.176	0.213	-0.212	0.228	1.074
200			-0.149	0.178	-0.187	0.199	1.118

where $RB1$, $RME1$ are the RB and RME of the proposed estimator $\hat{f}_k(0)$ and $RB2$, $RME2$ are the RB and RME for the Mack and Quang's estimator $\hat{f}_M(0)$ say.

Depending on the simulation results given in Table (1), several conclusions can be drawn from examining the results in regard to model robustness (RB) and (RME). The estimators $\hat{f}_M(0)$ is with large $|RB2|$ for the exponential power model with $\beta = 1$ and for the hazard rate model with $\beta = 1.5, 2.0$, the maximum $|RB2|$ value turns out to be 0.402 for the hazard rate model with $\beta = 1.5$ (). For the exponential power and the hazard rate models, the $|RB2|$ for $\hat{f}_M(0)$ increases as the shape parameter β decreases, while it decreases as β increases for the beta model. On the other hand, the $RME2$ ranges in $[0.132, 0.417]$ if $n = 50$, in $[0.095, 0.363]$ if $n = 100$ and in $[0.073, 0.311]$ if $n = 200$.

As to our proposed estimator $\hat{f}_k(0)$, it generally produces rather small $|RB1|$ s, the maximum $|RB1|$ value turns out to be 0.283 for the exponential power model with $\beta = 1.0$. Comparing the $|RB2|$ s of $\hat{f}_M(0)$ with that of $\hat{f}_h(0)$, the simulation results demonstrated clearly that the $|RB1|$ s of $\hat{f}_k(0)$ are smaller than the corresponding $|RB2|$ s of $\hat{f}_M(0)$, especially for the exponential power model with $\beta = 1.0, 1.5$ (in which the shapes are spike and has a moderate shoulder respectively); the hazard rate model with $\beta = 1.5, 2.0$ (in which, the two shapes have the shoulder at $x = 0$ but when $\beta = 2.0$ the curve drops sharply –but less than that of $\beta = 1.5$ – when we move far from $x = 0$) and for the beta model with different values of β (the shapes do not have the shoulder at $x = 0$). The different shapes of these models are depicted in Figure 1.

On the other hand, the $RME1$ ranges in $[0.144, 0.283]$ if $n = 50$, in $[0.118, 0.238]$ if $n = 100$ and in $[0.086, 0.207]$ if $n = 200$. Regarding the stability of the accurate of the two estimators, the performance of $\hat{f}_k(0)$ is more stable and hence its performance is better than that of $\hat{f}_M(0)$. The efficiency (EFF) values in Table 1 show that, for some of the models investigated, a considerable gain in the accuracy of the proposed estimator is achieved. The efficiency values increase as the sample size n increases for the exponential power model with $\beta = 1.0, 1.5, 2.0$; the hazard rate model with $\beta = 1.5, 2.0$ and for the beta model with different values of β , in the cases where the proposed estimator performs better than $\hat{f}_M(0)$. In the other cases where the shoulder condition is –in some sense– large the efficiency is less than one which indicates the performance of $\hat{f}_M(0)$ is better than $\hat{f}_k(0)$ but the efficiency remains acceptable in these cases.

Numerical Example

We apply the proposed estimator to the classical wooden stakes data set, given in Burnham *et. al.* (1980, p:61). The data are collected from line transect survey to estimate the density of stakes in a given area. The stakes data are the perpendicular distances (in meters) of detected a stake to the transect line, in which 150 stakes were placed at random in an area of 1000 meters long. Out of 150 stakes, 68 stakes are detected using line transect technique. The true form of $f(x)$ is unknown, but the true value of $f(0)$ is known which equals $f(0) = 0.110294$, thus the actual density of stakes was 37.5 stakes/ha. Calculation shows that $Z = 0.1624$, the empirical critical value for $\alpha = 0.05$ and $n = 68$ is 0.1605 (Zhang, 2001), so the shoulder condition is rejected. In this case, the formula $h = 0.892 \hat{\lambda} n^{-1/5}$ should be used, where the computed value of $\hat{\lambda} = 6.115$, so $h = 2.346$. In turn, the resulting

Figure (1). (a) Exponential power model for $\beta = 1, 1.5, 2, 2.5$. (b) Hazard-rate model for $\beta = 1.5, 2, 2.5, 3$ and (c) Beta model for $\beta = 1.5, 2, 2.5, 3$.

estimate is $\hat{f}_k(0) = 0.10463$ and $\hat{D} = 35.6$ stakes/ha. By adopting the Mack and Quang's estimator, $h = 0.933 \hat{\sigma} n^{-1/5}$, computation gives $\hat{\sigma} = 8.19$ and $h = 3.522$. In turn the resulting estimate is $\hat{f}_M(0) = 0.10005$ and $\hat{D} = 34.01$ stakes/ha. Burnham *et. al.* (1980) analyze the same data by using a cosine series estimator, and they obtain an estimate for $f(0)$ given by 0.1148 with corresponding density estimate $\hat{D} = 39.00$ stakes/ha. It should be remarked that the cosine series estimator employs an exact value for the maximum perpendicular distance (take to be 20 meters for this example), that is, more information is used in this case.

References

- Barabesi, L. (2001). Local parametric density estimation methods in line transect sampling. *Metron LIX*, 21-37.
- Buckland, S. T. (1985). Perpendicular distance models for line transect sampling. *Biometrics*, 41, 177-195.
- Buckland, S. T (1992). Fitting density functions with polynomials. *Applied Statistics*, 41, 63-76.
- Burnham, K. P. & Anderson, D. R. (1976). Mathematical models for nonparametric inferences from line transect data. *Biometrics*, 32, 325-336.
- Burnham, K. P., Anderson, D. R. & Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildlife Monograph*, 72.
- Chen, S. X. (1996). A kernel estimate for the density of a biological population by using line-transect sampling. *Applied Statistics*, 45, 135-150.
- Eberhardt, L. L. (1968). A preliminary appraisal of line transects. *J. Wildlife. Manage.*, 32, 82-88.
- Hayes, R. J., & Buckland, S. T. (1983). Radial distance models for line-transect method. *Biometrics*, 39, 29-42.
- Mack, Y. P. and Quang, P. X. (1998). Kernel methods in line and point transect sampling. *Biometrics*, 54, 606-619.
- Pollock, K. H. (1978). A family of density estimators for line transect sampling. *Biometrics*, 34, 475-478.
- Silverman, B. W. (1986). *Density Estimation*. Chapman & Hall, London.
- Wand, M., & Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall, London.
- Zhang, S. (2001). Generalized likelihood ratio test for the shoulder condition in line transect sampling. *Communication. in Statistics Theory and Methods*, 30, 2343-2354.

Accurate Binary Decisions For Assessing Coronary Artery Disease

Mehmet Ali Cengiz
University of Ondokuz Mayıs
Samsun, Turkey

Generalized linear models offer convenient and highly applicable tools for modeling and predicting the behavior of random variables in terms of observable factors and covariates. This paper investigates applications of a special case of generalized linear model to improve the accuracy of predictions and decisions adopting Bayesian methods, in the specific context of assessing coronary artery disease. The basic model is developed for this application using binary response. The results clearly demonstrate the potential advantages offered by this approach.

Key words: Bayesian methods, coronary artery disease

Introduction

The aim of this paper is to determine the probability of using Bayesian inference, in place of Classical inference, and to compare these two approaches and then to present new approach in assessing the probability of presence of Coronary artery disease. Multiple logistic regression was used, which is a special case of generalized linear models. This model is commonly used when the independent variables include both numerical and nominal measures and the outcome variable is binary, or dichotomous, having only two values. It requires no assumptions about the distribution of the independent variables.

Another advantage is that the regression coefficient can be interpreted in terms of relative risk in cohort studies or odds ratios in case-control studies. The Bayesian inference is based on the famous published posthumously by the Rev. Thomas Bayes in 1763. In this inference the numerical values allotted to probabilities do not relate to long-run frequencies and an attempt

is made to account for prior knowledge by quantitative measurement. The process of inference requires the evaluation of further integrals and the selection of appropriate prior.

In this paper a suitable prior distributions is presented. In some practical applications there is very little prior information available. In this case, the standard choice over recent years has been the invariant prior proposed by Jeffreys (1939). The other suitable priors may be Uniform, which is described many authors such as Bernardo and Smith (1994) and O'Hagan (1994).

The evaluation of integrals may be difficult analytically but numerical methods can overcome this difficulty. Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. Percy (1993) used this approach in the context of generalized linear models. Tierney and Kadane (1989) introduced The Laplace approximation that can be used to obtain a marginal of the posterior distribution. The above mentioned approaches were used and modified to binary data. By analyzing a set of data relating a real surgical problem (diagnosis of Coronary artery disease), several questions and suggestions arise regarding this application.

Mehmet Ali Cengiz is an Assistant Professor of Statistics at Ondokuz Mayıs University, Department of Statistics in Turkey. He graduated from the University of Salford in England with a Ph.D. in Applied Statistics in 1999. E-mail him at: macengiz@omu.edu.tr.

Coronary Artery Disease

Balcı et al (2000) previously investigated this surgical application. Their aim was to investigate the relationship between plasma insulin levels and the angiographical severity of coronary artery disease in male patients with normal glucose tolerance and unstable angina. The current work uses their data and results. Start by briefly reviewing the medical details that are relevant to the present analysis. Coronary Artery Disease is a progressive disease process that generally begins in childhood and has clinical manifestations in the middle to late adulthood.

Two decades ago, Coronary Artery Disease was considered to be a degenerative process because of the accumulation of lipid and necrotic debris in the advanced lesions. It is now recognized that it is a multifactorial process, which, if it leads to clinical sequelae, requires extensive proliferation of smooth muscle cells within the intima of the affected artery. The form and content of the advanced lesions of Coronary Artery Disease demonstrates the results of three fundamental biological processes.

These are: (1) proliferation of intimal smooth muscle cells, together with variable numbers of accumulated macrophages and T-lymphocytes; (2) formation by the proliferated smooth muscle of large amounts of connective tissue matrix, including collagen and elastic fibers (3) accumulation of lipid, principally in the form of cholesteryl esters and free cholesterol within the cells as well as in the surrounding connective tissues. The development of the concept of risk factors and their relationships to the incidence of coronary Artery Disease evolved from prospective epidemiological studies. These studies demonstrated a consistent association among characteristics observed at one point in time in apparently healthy individuals with the subsequent incidence of coronary artery disease in these individuals (Braunwald, 1992).

These associations include an increase in the concentration of plasma cholesterol, the incidence of cigarette smoking, hypertension, clinical diabetes, insulin levels, obesity, age or male sex, and occurrence of coronary artery disease. As a result of these associations, each

characteristic has been termed a risk factor and this terminology has been generally accepted and has become part of the scientific literature associated with this problem. The aim here is to develop a generalized linear model to calibrate coronary arterial stenoses against some risk factors, so that disease severity can be assessed with using some risk factors.

Bayesian Inference of Logistic Model to Binomial Data

Assuming n binomial observations of the form y_i , $i=1, \dots, n$ where $E(y_i) = p_i$ and p_i is the success probability corresponding to the i th observation, the linear logistic model for the dependence p_i on the values of the k explanatory variables $x_{1i}, x_{2i}, \dots, x_{ki}$, associated with that observation, is

$$\begin{aligned} \text{logit}(p_i) &= \log(p_i / (1 - p_i)) \\ &= \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \end{aligned} \quad (1)$$

In order to fit a linear logistic model to a given set of data, unknown parameters must be estimated first. In Classical approach, these parameters are estimated using the methods of maximum likelihood. The likelihood function is given by

$$L(\beta_i; y_i) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i} \quad (2)$$

The problem is to obtain estimations of parameters, which maximise the

$$l(\beta_i; y_i) = \sum_{i=1}^n y_i \log(p_i) + (1 - y_i) \log(1 - p_i).$$

Bayesian inference is used to obtain parameter estimations.

Assuming some training data $D = \{(Z_i, y_i); i = 1, \dots, n\}$ which consist of observed response vectors y_i and matrices of explanatory variables Z_i , typically one will observe Z_{n+1} for a new individual, and our aim is to predict the response vector y_{n+1} . The

conditional distribution of y_i given Z_i is assumed known as a function of unknown parameters contained in a vector β . The posterior predictive distribution of y_{n+1} given Z_{n+1} and the data D is given by

$$f(y_{n+1} | Z_{n+1}, D) = \int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) f(\beta | D) d\beta \quad (3)$$

In the third equation (3), $f(\beta | D) \propto L(\beta; D) \times f(\beta)$ where L is the likelihood and f represents the prior density. The likelihood function is given in equation (2). If information about parameters before observing the data is vague, the use of the uniform prior distribution for a location parameter is supported by several researchers. The task of finding logically consistent realistic representations of prior ignorance meets some difficulties. In particular the uniform distribution may not represent ignorance. Jeffreys (1967) proposed a solution using Fisher information matrix. There are different Jeffreys prior to the binomial experiments, so that posterior inference using the Jeffreys prior will violate the Likelihood Principle. So uniform and Jeffreys prior distributions for our application are used.

The required integrations in equation (3) and (5) are not feasible analytically and approximation methods are needed. Dunsmore (1976) considered an asymptotic Bayesian approach to prediction analysis. If we expand $f(y_{n+1} | Z_{n+1}, \beta)$ in equation (3) about the maximum likelihood estimate of β by Taylor's theorem, A first order approximation and second order approximation to the predictive distribution are then obtained by truncating the expanded series. The following equation for first order approximation is obtained.

$$f(y_{n+1} | Z_{n+1}, D) \approx f(y_{n+1} | Z_{n+1}, \hat{\beta}) \quad (4)$$

The Laplace approximation is useful for evaluating the multiple integral in equation (5) to predict disease severity, since the information matrix can be obtained without a lot of effort. The equation may be re-expressed (3) as

$$f(y_{n+1} | Z_{n+1}, D) = \frac{\int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) L(\beta; D) f(\beta) d\beta}{\int_{\beta} L(\beta; D) f(\beta) d\beta} \quad (5)$$

From equation (5), the posterior expectation of $f(y_{n+1} | Z_{n+1}, D)$ can be expressed as the ratio

$$E\{f(y_{n+1} | Z_{n+1}, D)\} = \frac{\int_{\beta} f(y_{n+1} | Z_{n+1}, \beta) L(\beta; D) f(\beta) d\beta}{\int_{\beta} L(\beta; D) f(\beta) d\beta} \quad (6).$$

Referring to Tierney and Kadane (1986), it may be written

$$E\{g(\beta)\} \approx \left(\frac{\det \tilde{\Sigma}}{\det \Sigma} \right)^{1/2} \exp \left[n \left\{ \tilde{l}(\tilde{\beta}) - l(\hat{\beta}) \right\} \right]$$

where $\tilde{\beta}$ and $\hat{\beta}$ maximize $\tilde{l}(\beta) = (\log g + \log f + \log L) / n$ and $l(\beta) = (\log f + \log L) / n$, respectively, and $\tilde{\Sigma}$ and Σ are minus the inverse Hessians of $\tilde{l}(\beta)$ and $l(\beta)$ evaluated at $\tilde{\beta}$ and $\hat{\beta}$, respectively and n is the sample size for which data have been observed.

Methodology & Results

The data for the analyses were collected in 1996 – 1997 and presented in Table 1, at University Hospital in Erzurum, Turkey. One hundred consecutive men undergoing elective coronary angiography formed the study population. Eligible patients met the following criteria: (1) no history of diabetes; (2) normal fasting blood glucose; (3) no treatment with lipid lowering drugs (4) no antecedent history of myocardial infarction, coronary bypass, or angioplasty. Cardiovascular medications including β blockers, calcium antagonists, nitrates, aspirin, angiotensin-converting enzyme inhibitors were not discontinued before the study. A standard oral glucose tolerance test was performed 3 days before coronary angiography. Selective coronary

angiography was performed by Judkins technique in the right and left oblique views. 3 observers unaware of the laboratory results examined angiograms.

The luminal percent diameter narrowing was estimated by a consensus of the observers or by the mean of different measurements. Diameters stenosis $\geq 50\%$ were considered significant and these patients (68 patients) were assigned to the diseased one. Stepwise logistic regression was performed to evaluate the independence of risk factor effects on presence of Coronary Artery Disease.

For patient $i = 1, \dots, n$ expert judgments were used to classify each patient as healthy ($y_i = 0$) or diseased ($y_i = 1$) as mentioned above. After performing stepwise regression, Patient i has also has three covariates:

x_{1i} : age for patient i

x_{2i} : Log fasting insulin level for patient i x_{3i} :

Log Lp(a) (Lp(a): Lipoprotein (a))

Now consider the following model with using different prior distributions and different numerical approaches.

$$\begin{aligned} \text{logit}(p_i) &= \log(p_i / (1 - p_i)) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \end{aligned}$$

where $E(y_i) = p_i$ and p_i is the success probability corresponding to the i th patient. Considered is the same model as above with following cases.

1. case: the model above with Uniform prior and First Order approximation (corresponding to Classical approach using Likelihood method).

2. case: the model above with Uniform prior and Laplace approximation.

3. case: the model above with Jeffreys prior and First Order approximation.

4. case: the model above with Jeffreys prior and Laplace approximation.

The main aim for this section is to show how Bayesian inference in Bernoulli response models can be used to improve predictive accuracy in practice. Adopting a Bayesian approach to the analysis, a vague prior is used,

which is multiple uniform, and Jeffrey's prior because no specific prior information is available. Furthermore, we are merely demonstrating the potential of this model with different approaches in this paper; the goal is to develop a suitable informative prior in the future, to judge how sensitive the predictions are to the choice of prior.

Consequently, the joint posterior distribution, on which all predictive inference is based, is proportional to the likelihood function. In particular, the posterior predictive distribution for a new patient, with ages in vector x_1 , Log fasting insulin levels in vector x_2 and Log Lp(a)'s in vector x_3 , is

$$\begin{aligned} p(y | x_1, x_2, x_3, D) &= \\ \int_{\beta} p(y | x_1, x_2, x_3, \beta) f(\beta | D) d\beta \quad (7) \end{aligned}$$

where $p(y | x_1, x_2, x_3, \beta)$ is the binomial sampling distribution defined by equation (1) and $f(\beta | D)$ is the joint posterior density, which is maximum likelihood function and prior distribution.

The assessment of diagnostic performance is now dealt with. Applied is the First order approximation and Laplace method using Fortran computer programs and subroutines from the NAG library to obtain approximate posterior predictive distributions as given by equation (7).

Two criteria to assess our predictive accuracy for each case were used. These are a binary loss function, corresponding to the percentage of correct classifications based on cross-validation of the training data set with a default classification threshold of 0.5 and the linear loss function

$$s_1 = \sum_{i=1}^{100} \{y_i(1 - \hat{p}_i) + (1 - y_i)\hat{p}_i\}$$

Table 1. Coronary Artery Disease

Patient i	y_i	x_{1i}	x_{2i}	x_{3i}	Patient i	y_i	x_{1i}	x_{2i}	x_{3i}
1	0	45	0.9542	1.6128	51	1	56	2.0334	2.0294
2	0	57	1.8261	1.3222	52	1	57	2.0792	1.7482
3	0	38	1.7404	0.4771	53	1	57	2.0128	1.2553
4	0	37	1.8129	0.3010	54	1	63	1.8921	0.9542
5	0	35	1.9031	1.2304	55	1	45	1.9031	1.7160
6	0	49	2.0934	1.9868	56	1	63	2.0000	1.2041
7	0	49	1.9912	1.0414	57	1	51	1.7634	1.0792
8	0	55	1.9243	1.4149	58	1	60	1.6812	1.5563
9	0	45	1.6021	0.8451	59	1	77	2.1461	2.2695
10	0	50	1.7634	1.2787	60	1	58	1.4771	2.4502
11	0	48	1.9031	1.0414	61	1	50	1.9031	1.6902
12	0	48	2.2355	1.2041	62	1	65	1.9031	2.0212
13	0	50	1.5315	1.5682	63	1	55	1.8062	1.9731
14	0	43	1.6021	0.7781	64	1	50	1.9031	2.0212
15	0	53	1.8325	1.7634	65	1	55	1.8325	1.5563
16	0	50	2.1703	1.6434	66	1	44	1.8195	0.4771
17	0	42	1.4472	1.3979	67	1	50	1.5798	2.4885
18	0	45	1.6021	0.9031	68	1	58	1.6021	0.9031
19	0	45	1.6021	1.6335	69	1	53	1.7482	2.5752
20	0	55	1.8129	1.9445	70	1	60	1.9243	0.6021
21	0	62	1.6021	0.8451	71	1	55	1.8062	2.4265
22	0	57	1.9031	1.2304	72	1	64	2.1461	1.8976
23	0	33	1.7404	0.6989	73	1	56	2.0492	1.4771
24	0	50	1.8129	1.3010	74	1	63	2.1399	1.5185
25	0	49	1.5051	1.5682	75	1	53	1.6532	0.0000
26	0	60	1.8195	1.7634	76	1	53	1.9345	1.3979
27	0	43	1.8976	0.3010	77	1	60	2.0086	1.3424
28	0	46	0.8451	1.5682	78	1	40	2.0253	1.6721
29	0	60	1.7781	0.6021	79	1	65	1.7781	0.4771
30	0	38	1.8261	1.3424	80	1	65	2.2878	1.2787
31	0	43	1.7324	1.2041	81	1	50	2.0792	1.6232
32	0	58	1.5563	0.7781	82	1	58	1.8195	1.1461
33	1	64	1.4771	1.2553	83	1	65	1.8808	1.5315
34	1	47	1.5315	1.6989	84	1	46	1.6812	1.6532
35	1	48	1.9777	1.6127	85	1	55	1.6021	2.4048
36	1	42	2.3617	0.4771	86	1	65	1.9138	1.5911
37	1	40	1.9445	1.9085	87	1	60	1.3424	1.3802
38	1	58	1.6021	1.1461	88	1	59	1.3424	2.0453
39	1	65	1.8062	2.1643	89	1	64	1.7634	1.6021
40	1	60	2.4232	1.4400	90	1	46	1.7324	1.2787
41	1	63	2.2041	0.0000	91	1	54	2.1367	1.5911
42	1	42	2.1732	1.5563	92	1	63	1.8808	1.6628
43	1	43	2.1614	1.1461	93	1	46	1.9031	2.3345
44	1	33	2.1987	1.2787	94	1	62	2.1987	1.2787
45	1	45	1.7243	1.2304	95	1	42	1.9138	1.8751
46	1	65	1.4771	1.4149	96	1	42	1.9031	1.4771
47	1	50	1.9542	2.2355	97	1	42	2.1461	2.4149
48	1	69	1.6989	1.59116	98	1	51	1.8808	1.9345
49	1	60	1.8451	1.7482	99	1	38	2.1004	1.6532
50	1	60	1.6532	1.9191	100	1	63	2.1367	1.4771

Table 2. Posterior predictive probabilities for the model with Laplace approximation and Jeffreys prior.

Patient i	$\hat{p}_i$	Patient i	$\hat{p}_i$	Patient i	$\hat{p}_i$	Patient i	$\hat{p}_i$
1	0.5885	26	0.0248	51	0.6241	76	0.4376
2	0.0817	27	0.4444	52	0.6095	77	0.5620
3	0.4042	28	0.6152	53	0.5062	78	0.3177
4	0.5689	29	0.1679	54	0.4961	79	0.3940
5	0.4168	30	0.3822	55	0.3533	80	0.6952
6	0.0303	31	0.3641	56	0.5775	81	0.4931
7	0.1665	32	0.2685	57	0.2610	82	0.4208
8	0.0635	33	0.3869	58	0.4703	83	0.6125
9	0.4532	34	0.2100	59	0.8436	84	0.2521
10	0.2116	35	0.4168	60	0.5185	85	0.5172
11	0.2220	36	0.2768	61	0.4316	86	0.6315
12	0.0737	37	0.3285	62	0.6848	87	0.2795
13	0.2737	38	0.3196	63	0.5217	88	0.3967
14	0.4954	39	0.6760	64	0.4949	89	0.5700
15	0.0716	40	0.6978	65	0.4558	90	0.2023
16	0.0333	41	0.4570	66	0.0890	91	0.5658
17	0.4794	42	0.3948	67	0.4463	92	0.6070
18	0.4436	43	0.3257	68	0.2722	93	0.4929
19	0.3159	44	0.2109	69	0.5749	94	0.6389
20	0.0435	45	0.1748	70	0.3984	95	0.3406
21	0.1728	46	0.4347	71	0.5974	96	0.2539
22	0.0701	47	0.5528	72	0.7214	97	0.5435
23	0.5950	48	0.6109	73	0.5434	98	0.4848
24	0.1847	49	0.5689	74	0.6651	99	0.3159
25	0.3045	50	0.5246	75	0.0854	100	0.6589

Table 3. Predictive accuracy results for the model with all cases.

Case	Binary Percentage	Linear loss
1	%79	0.3152
2	%81	0.2955
3	%81	0.2961
4	%83	0.2622

where $\hat{p}_i = P(Y_i = 1 | x_{1i}, x_{2i}, x_{3i}, D)$ from equation (7). Ultimately, the binary loss function is of most interest in diagnosing the disease, but the alternatives provide more insight into the predictive accuracy of the model with different approaches.

To illustrate the typical output from which loss functions are calculated, Table 2 presents the predictive probabilities for patients in the observed set of training data, based on the model with Jeffreys prior and Laplace approximation. The summary results for all cases investigated in Table 3.

First, is illustrated the improved predictive accuracy by adopting Bayesian inference here, over Classical approach. In case 1, uniform prior and First order approximation is used, which is the same as Classical approach (using the Likelihood function to obtain parameter estimations). Column 2 of Table 3 demonstrates this by presenting the percentage of diseased patients correctly diagnosed by each case, if costs are such that a threshold of 0.5 is appropriate. Note that, without further information, we could correctly diagnose 50 per cent of patients by chance alone, and that large values are desirable for the percentage of patients correctly diagnosed. Clearly, Bayesian approach with different priors and approximations performs consistently better than the classical approach.

Second, compared are the different priors and different approximations for the same model using two assessments criteria identified above: namely the binary and linear loss function. These results are presented in Table 3. Although large values are desirable for second column of these, small values are preferable for linear loss function. As expected, the model with Laplace approximation gives better results than the others.

Conclusion

This article described and discussed the properties and applications of multiple logistic regression models, suggesting simplifications and suitable approximations for a Bayesian analysis. Considered were different subjective priors, which are uniform, and Jeffreys, using different approximations, which are First order and Laplace approximation. It has also demonstrated how these prior distributions and approximations may be used and useful in an important application, relating to the diagnosis of coronary arterial disease.

References

- Balcı, B. et al, (2000). The relationship between plasma insulin levels and angiographical severity of coronary artery disease. *Turk Kardiyol., Dern., Ars.* 28, 617-621.
- Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian theory*. John Wiley & Sons.
- Braunwald. E. (1992). *Heart disease*, Vol. 2. W. B. Saunders Company.
- Dunsmore, I. R. (1976). Asymptotic prediction analysis. *Biometrika*, 63, (3) 627-630.
- Jeffreys. H. (1939). *Theory of probability*. Oxford: Oxford University Press.
- O'Hagan. A. (1994). *Bayesian inference*. London: Edward Arnold.
- Percy. D. F. (1993) Prediction for generalized linear models. *Journal of Applied Statistics*, 20, 285-291.
- Tierney L., & Kadane J B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82-86.

A Generalized Quasi-likelihood Model Application To Modeling Poverty Of Asian American Women

Jeffrey R. Wilson
School of Health Management and Policy
W. P. Carey School of Business
Arizona State University

A generalized quasi-likelihood function that does not require the assumption of an underlying distribution when modeling jointly the mean and the variance, is introduced to examine poverty of Asian American women living in the West coast of the United States, using data from U.S. Census Bureau.

Key words: Logistic regression, extravariation, generalized linear models

---

Introduction

All systems of social inequality create poverty. In 1998, the U.S. Census Bureau (1998) states that 12.7% of the U.S. population is poor. Racial minorities are more likely to live in poverty than whites (U.S. Census Bureau, 1999). Previous studies on poverty have focused on whites and other racial minorities and few studies have modeled the poverty for Asian Americans. This research is useful since in recent years Asian Americans have increased significantly and are very diverse in socioeconomic status and country of origin. Poverty among Asian Americans has increased rapidly as a result of a large influx of Asian immigrants from many different countries, many of whom face difficulties in economic opportunities as a result of poor English fluency and low educational attainment. Data from the 1998 Current Population Survey were examined to study the effects of certain variables on the poverty level among Asian American women living in the Western region of the United States.

Because the use of ordinary least squares regression to predict binary response would violate the assumptions of a constant variance (homoscedasticity) and normal distribution (Allison 1999), it is common practice to model binary random variables using logistic regression models. As several variables of interest in social sciences and medical research are binary, logistic regression models have been used widely in these areas. Such models require a logistic transformation on the probability in such a way that the odds is modeled and thus the predicted probabilities are not outside the bounds for probability.

However, there may be times when the fitted logistic regression model does not adequately describe the observed proportions, because of the presence of extravariation or overdispersion as it is often referred to. The presence of overdispersion results in the assumption of binomial variability to be invalid (Collett, 1991). When overdispersion occurs, it may be necessary to consider other binary models. One such approach is to consider a quasi-likelihood model thus negating the need for the binomial variation assumptions. A quasi-likelihood model does not make any distributional assumption about the random variable in the mean modeling.

Modeling the mean of a binary response model consists of several approaches. Some approaches have been proposed where the

---

Jeffery R. Wilson is a Professor of Biostatistics and Director of the School of Health Management and Policy at Arizona State University, Tempe AZ 85287-4506.

parameters of the distribution are allowed to vary based on some known distribution (Williams, 1982; Crowder, 1978; Wilson, 1989; Wilson & Koehler, 1991). Other methods have made use of a mean-variance relation (Wedderburn, 1974; McCullagh, 1983; Firth, 1987; Moore & Tsiatis, 1991) and so the knowledge of an underlying distribution is not required.

These methods assume that the variance is related to the mean through the variance function, which is a function of the mean. Neither of these approaches considered modeling the variance of the distribution. Analyzing the poverty data among Asian Americans showed that through there is sufficient extravariation that needs to be modeled. A review of a binary logistic function is follows.

Generalized linear models (Nelder & Wedderburn, 1972) encompass a wide range of models. These models include linear regression, analysis of variance, logit and probit models for binary response data, and log-linear and multinomial response models for count data. A generalized linear model has three components. The random component specifies the distribution of the response variable from the exponential family of distributions. The systematic component defines a linear predictor based on some set of known covariates and the link component combines the random component and the systematic component. The link function is a monotonic twice-differentiable function that provides a relation between the mean of the response variable and the covariates.

Generalized linear models differ in their underlying distribution and in their link function. The systematic component of these models has a linear structure. Generalized linear models reduce the problem of scaling and do not require the assumption of normality and constancy of variance. For linear regression and analysis of variance models the distribution is normal with an identity link. For logit and probit models the distribution is binomial with logistic and cumulative distribution function of normal distribution as link functions, respectively. Log-linear models have a multinomial distribution with a log link. Estimation of these regression parameters in the systematic function can be

done through maximum likelihood procedure (Finney, 1952). However, for exponential family distributions, the maximum likelihood estimation is equivalent to the weighted least squares method (Bradley, 1973). Thus, generalized linear models lead to a unified method for estimating the parameters for a wide range of models. They provide a method for modeling the mean of the distribution.

The modeling of the mean and the dispersion jointly through two sub models using a generalized linear model framework was first suggested by Pregibon (1984) and later addressed by Efron (1986), Aitkin (1987) and Smyth (1989). In the joint modeling of the mean and the variance, three components similar to the mean sub model are required for modeling the variance. The response variable for the dispersion submodel is the deviance obtained from the mean submodel. The extended quasi-likelihood function (Nelder & Pregibon, 1987; McCullagh & Nelder, 1989) and the pseudo likelihood function (Carroll & Ruppert, 1982) are useful for joint modeling of the mean and the dispersion, when only the relation between mean and variance has been specified for the mean submodel.

Extended quasi-likelihood and pseudo likelihood functions can be used for comparison of the link and the variance function. Further generalizations and modifications of the extended quasi-likelihood functions have been presented by Yanez and Wilson (1995).

Binary logistic function

Consider Y_i for $i = 1, \dots, n$; to denote the i th observation for each of the Asian women with mean p_i where p_i is the probability that an Asian woman falls below the poverty level. A linear logistic model for poverty level based on marital status, educational attainment, residence, employment status, and number of children for each of these women is

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}, \end{aligned} \quad (1)$$

where x_{ki} is the value of the k th variable on the i th woman. Thus the probability of an event is:

$$p_i = \frac{e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki})}}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki})}} \quad (2)$$

and the variance function is defined by $\text{var}(y) = \Phi \frac{p_i}{1 - p_i}$. In most cases Φ is one.

When $\Phi \neq 1$, it is usually common to use quasi-likelihood models (McCullagh & Nelder, 1989). For modeling the poverty data pertaining to Asian Americans, both the mean and variance parameters are modeled using a quasi-likelihood function.

Methodology

A generalized quasi-likelihood model (GQL) for poverty among Asian American women is now proposed. The model is simple and less restricted in that it does not require the assumption of an underlying distribution, when modeling either the mean or the variance jointly. The generalized quasi-likelihood function assumes that the distributional form for both the mean and the dispersion submodels are not known and relies on a mean-variance relation. In the dispersion submodel the mean and the variance of the response variable are ϕ_i^α and $2\phi_i^{2\alpha}$ respectively, where α is a nonlinear parameter.

Thus, the variance function is assumed to be a squared function of the mean in the dispersion submodel, with a dispersion parameter of value two. In the analysis of these data the link and variance functions used for the mean submodel is quasi and log-root, respectively, whereas the link and variance for the variance submodel is quasi and square root, respectively.

For a single observation y_i with mean μ_i $i = 1, 2, \dots, n$; a generalized quasi-likelihood function is defined as

$$Q_1^*(y_i, \mu_i, \phi_i, \alpha, \tau) = -\frac{1}{2} \left[\frac{d(y_i; \mu_i)}{\phi_i^\alpha} + \ln(\phi_i^\alpha) + \ln(2\pi V_\tau(y_i)) \right],$$

where $d(y_i, \mu_i) = -2 \int_{y_i}^{\mu_i} \frac{y_i - u_i}{V_\tau(u_i)} du_i$, ϕ_i is the

dispersion parameter for the mean submodel, $V_\tau(y_i)$ is the variance function evaluated at y_i , and α and τ are nonlinear parameters. The generalized quasi-likelihood model has a mean submodel with random, systematic, and link components as $Y_i \sim (\mu_i, \phi_i^\alpha V_\tau(\mu_i))$, $\eta_i = \mathbf{x}_i' \vec{\beta}$, and $\eta_i = g(\mu_i)$, respectively.

Its dispersion submodel has response variable $d_i \sim (\phi_i^\alpha, 2\phi_i^{2\alpha})$, $\eta_i^* = \mathbf{v}' \vec{\gamma}_i$, and $\eta_i^* = h(\phi_i, \alpha)$ as the random, systematic, and link function component, respectively. The estimating equations for the linear parameters $\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$, in the mean submodel based on the GQL function are

$$\frac{\partial Q_1^*}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{\phi_i^\alpha V_\tau(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j}.$$

Similarly, the estimating equations for the linear parameter $\vec{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \dots, \gamma_p)$ in the dispersion submodel are

$$\frac{\partial Q_1^*}{\partial \gamma_r} = \frac{\alpha}{2} \sum_{i=1}^n \frac{d_i - \phi_i^\alpha}{\phi_i^{\alpha+1}} \frac{\partial \phi_i}{\partial \gamma_r}.$$

A simultaneous iterative weighted least squares procedure is used to solve these estimating equations as $\vec{\beta}$ and $\vec{\gamma}$ are orthogonal. The orthogonality of μ_i and ϕ_i , leads to the orthogonality between $\vec{\beta}$ and $\vec{\gamma}$ which follows since the expected partial derivatives,

$$E \left[\frac{\sum \partial^2 Q_1^*}{\sum \partial \mu_i \sum \partial \phi_i} \right] = E \left[-\frac{\alpha(y_i - \mu_i)}{\mu_i^\tau \phi_i^{\alpha+1}} \right] = 0.$$

Thus holding α, τ and ϕ_i fixed, the maximum

quasi-likelihood estimator, $\tilde{\beta}$ are obtained for the function Q_1^* through $\mathbf{X}'\mathbf{W}\mathbf{X}\tilde{\beta}^{(m)} = \mathbf{X}'\mathbf{W}\mathbf{z}$,

where $\mathbf{W} = \text{diag} \left(\frac{1}{\text{var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right)$,

$\text{diag}(t)$ denotes the diagonal elements of the matrix $\mathbf{T}$ and $\mathbf{z}$ is a vector of order n with elements

$$z_i = \sum_{k=1}^p x_{ik} \beta_k^{(m-1)} + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \quad i = 1, \dots, n;$$

The maximum quasi-likelihood estimates for the regression parameters, $\tilde{\gamma}$, in the dispersion submodel are estimated from

$$\mathbf{V}'\mathbf{W}^*\mathbf{V}\tilde{\gamma}^{(m)} = \mathbf{V}'\mathbf{W}^*\mathbf{z}^*$$

where $\mathbf{W}^* = \text{diag} \left(\frac{\alpha^2}{2\phi_i^2} \left(\frac{\partial \phi_i}{\partial \eta_i^*} \right)^2 \right)$ and $\mathbf{z}^*$ is a

vector with elements

$$z_i^* = \sum_{l=1}^{p^*} v_{il} \hat{\gamma}_l^{(m-1)} + \frac{(d_i - \phi_i^\alpha) \partial \eta_i^*}{\alpha \phi_i^{(\alpha-1)} \partial \phi_i}$$

by fixing ϕ_i and $\tilde{\beta}$ and estimates of the nonlinear parameters α and τ at known value. The process is continued until convergence is achieved.

The variance of $\tilde{\beta}$ under the generalized quasi-likelihood function is

$$\text{cov}(\tilde{\beta}) = (\mathbf{I}_m' \mathbf{V}_m^{-1} \mathbf{I}_m)^{-1}, \text{ where } \mathbf{I}_m = \left[\frac{\partial \mu_i}{\partial \beta_j} \right]_{i,j} \text{ is}$$

the vector of partial derivatives and $\mathbf{V}_m = \text{diag}(\phi_i^\alpha \mathbf{V}_\tau((\mu_i)))$. Similarly for the vector of estimates $\tilde{\gamma}$, $\text{cov}(\tilde{\gamma}) = (\mathbf{I}_d' \mathbf{V}_d^{-1} \mathbf{I}_d)^{-1}$ where

$$\mathbf{I}_d = \left[\frac{\partial \phi_i}{\partial \gamma_1} \right]_{i,1} \text{ and } \mathbf{V}_d = \text{diag} \left(\frac{\alpha^2}{2\phi_i^2} \right).$$

Results

The major interest is to determine which social factors contribute if any to Asian American women living in poverty. These social factors included whether she is married, her years of schooling, residence, whether she works, and how many children she has. These data are confined to those women living in the western region of the United States (i.e. California, Washington, Oregon, Arizona, etc.). There are a total of 639 Asian American women in our sample.

Studies on poverty have focused on whites and other racial minorities and few studies examine the likelihood of poverty for Asian Americans. In this study, the definition of an Asian American living in poverty follows the definition given by the U.S. Census Bureau. A woman is considered to live in poverty if she lives on her own with family income less than \$7,500, if a woman lives with another family member with family income less than \$10,000, if a woman lives with two other family members with family income less than \$15,000, etc. This definition is based on 1998 figures and takes into account the family size. Of all the poor people eighteen and older, 62% are women and 38% are men (U.S. Census Bureau, 1999). The motivating factor that brought these data into focus is in part due to an emerging belief that there is a trend by which women represent an increasing proportion of the poor.

Previous research on other racial groups reveals that marital status, educational attainment, area of residence, employment status, and number of children are strong predictors of poverty. The increases in poverty among women are partly as a result of increases in unmarried women, and families headed by single mothers (Macionis, 2001).

Although people living in central cities are most likely to live in poverty, people living in suburban areas are least likely to live in poverty (Macionis, 2001). Asian American women living in metropolitan areas are less likely to live in poverty as compared to those living in non-metropolitan areas, although Asian Americans are least likely to live in non-metropolitan areas. Educational attainment and employment status are as expected significant

predictors: the more educated women the less likely they live in poverty; no jobs translate into more poverty (Wilson, 1996). The number of children also has a positive impact on poverty: the more children a woman has it is more likely for her to live in poverty (Wilson, 1996).

In the binary models used to model poverty, variables are coded as follows. Marital status is coded 1 if a woman is unmarried (widowed, divorced, separated, or never married) and 0 if a woman is married. Educational attainment has four categories: "1" denotes less than high school; "2" denotes high school; "3" denotes some college; and "4" denotes college graduate and above. Area of residence is coded 1 if a woman lives in metropolitan areas and 0 if a woman lives in non-metropolitan areas. Employment status is coded 1 if a woman worked for pay and 0 otherwise.

There are three categories for number of children: "1" denotes no children; "2" denotes 1 to 3 children; and "3" denotes more than 3 children. Table 1 provides a percentage distribution of women living in poverty and the tabulation between poverty and each of the predictors.

Bivariate analyses of poverty and each predictor reveal that of 639 Asian American women in the sample, 23.2% live in poverty. A higher percentage of unmarried Asian American women lived in poverty compared to married Asian American women (26.5% vs. 19.9%). Women with high school education have the highest percentage living in poverty (41.3%). Women with college education and above have the lowest percentage living in poverty. Fewer Asian American women lived in non-metropolitan areas than in metropolitan areas (56 vs. 583). Those living in metropolitan areas have higher percentage living in poverty than those living in non-metropolitan areas (37.5% vs. 21.8%). Of employed women, only 18.8% lived in poverty while 30% of unemployed women lived in poverty. The number of children is not significant at the 0.05 level.

These bivariate results are consistent with those obtained from previous literature on poverty for other racial groups. However, simultaneous effects of these predictors on poverty are more informative if one is to adequately assess the different impacts. Thus a multivariate logistic regression model suitable for a $2 \times 4 \times 2 \times 2 \times 3$ contingency table is required. The logistic regression model and the generalized quasi-likelihood function were compared in their use to analyze the data from U.S. Census Bureau's 1998 Current Population Survey.

Applications of Binomial Logistic Regression Model

A logistic regression model with a binomial distribution and a logit link function was fitted to the $2 \times 4 \times 2 \times 2 \times 3$ contingency table. This model was presented to determine the simultaneous impact of marital status, educational attainment, area of residence, employment status, and number of children on the probability that Asian American women live in poverty. Table 2 provides a summary of the results from the fit of such a maximum likelihood binomial logistic regression model. The odds ratios are obtained from the exponentiation of the parameter estimates. Unmarried Asian American women are 1.75 times as likely to be poor than married Asian American women. Educational attainment has a negative effect on poverty: It also seems that more educational years reduced the odds of living in poverty by 33.9%.

Asian American women living in nonmetropolitan areas are 1.63 times as likely to be poor than those living in metropolitan areas. Evidently, whether a woman has a job affects the likelihood of being poor: those without jobs are 1.56 times as likely to be poor than those with jobs. The impact of number of children on poverty is not significant. This could be due to the fact that poverty measure (whether a person lives in poverty) is adjusted by the family size.

Table 1. Percentage Distributions of Asian American Women Living in Poverty by Marital Status, Educational Attainment, Type of Residence, Employment, and Number of Children.

Variable	% in poverty	Number of Cases
Total	23.2%	639
<i>Marital Status**</i>		
Married	19.9%	326
Unmarried	26.5%	313
<i>Educational</i>		
Less Than High School	22.5%	111
High School	41.3%	150
Some College	21.7%	184
College Graduate and	10.8%	194
<i>Area of Residence***</i>		
Metropolitan	21.8%	583
Nonmetropolitan	37.5%	56
<i>Employed?****</i>		
Yes	18.8%	389
No	30.0%	250
<i>Number of Children</i>		
No children	21.5%	311
1-3 children	23.7%	296
4 and more children	34.4%	32

Note: **, significant at the .05 level and ***, significant at the .01 level (Pearson chi-square test).

Table 2 Parameter estimates, Standard errors, and Odds Ratios For Binomial Logistic Regression Model.

Covariate	Parameter	Standard	Odds Ratios
Intercept	-.705	.557	.494
<i>Marital Status</i>			
Unmarried	.559	.239	1.749
Married			
<i>Educational</i>			
Educational	-.273	.089	.761
<i>Area of Residence</i>			
Metropolitan	-.638	.305	.528
Nonmetropolitan			
<i>Employment Status</i>			
Employed	-.446	.201	.640
Not Employed			
<i>Number of Children</i>			
Number of Children	.487	.194	1.63

It is imperative to know, prior to accepting the odds ratios as obtained, whether or not there is a good fit with the model: the extent to which the fitted values of the response variable under the model compare with the observed values. If the agreement between the observations and the corresponding fitted values is good, the model may be acceptable (Collett, 1991). To examine the fit, the likelihood ratio

with the covariates in the model, $\hat{L}_c$, is compared with the likelihood ratio with the saturated model, $\hat{L}_f$. The deviance,

$$D = -2 \log(\hat{L}_c / \hat{L}_f) = -2[\log \hat{L}_c - \log \hat{L}_f],$$

where $\hat{L}_c$ is obtained based on the predicted probability of the event under the model with covariates while $\hat{L}_f$ is obtained based on the observed proportions of the event provides such a measure.

The deviance from the model with covariates is 138.81 with 74 degrees freedom. The ratio of the deviance to the degrees of freedom (1.87) is substantially greater than one. Thus, there is a strong likelihood that over-

dispersion is present and the assumption of the binomial variability may not be valid (Collett, 1991). Such results suggest that the data exhibit overdispersion. Thus there is a significant amount of variation unaccounted for. This indicates that Φ is greater than 1 in the variance

function where $\text{var}(y) = \Phi \frac{p_i}{1 - p_i}$. Thus, the

assumption that Φ is equal to 1 in the logistic regression model is not valid. Thus it is evident that the data are over-dispersed.

Overdispersion arises because of clustering in the population (McCullagh and Nelder, 1989). Overdispersion could be present due to the fact that unobserved heterogeneity operates at the level of groups rather than individuals (Allison, 1999). It may also be an account of the cost of living differences between metropolitan and nonmetropolitan cities.

Given the presence of such overdispersion, a quasi-likelihood model was chosen to analyze the data. The quasi-likelihood model allows us to estimate the parameters in the model and determine its significance without specifying the distribution function while accounting for the overdispersion. The model is fully determined since the link and variance

functions are sufficient for fitting the model. Once these are specified, the same iterative procedure that is used for fitting the other families can be used to estimate the linear parameters. This is readily available in SPLUS.

Applications of Generalized Quasi-Likelihood Function

Using the logistic regression model to fit the data left indication that overdispersion was present. The overdispersion may be due to the fact that some variables tend to produce clustering in poverty and thus some unobserved heterogeneity affects the fit of the model. To account for any such extra variation, a joint modeling of the mean and the variance using the generalized quasi-likelihood function was used. Quasi-likelihood estimation makes it possible to estimate relationships without fully knowing the random component of model.

The difference between a quasi-likelihood function and a maximum likelihood function is analogous to the comparison between normal-theory regression models and least squares regression estimates. As least-squares estimation and normal theory models give identical regression parameter estimates so does quasi-likelihood and maximum likelihood procedures. However, least-squares estimation relies on second moment assumptions for its variance whereas normal-theory models rely on full distributional assumptions.

Under quite general conditions, quasi-likelihood estimates are consistent and asymptotically normal (Agresti, 1990). Quasi-likelihood estimators still retain relatively high efficiency as long as the degree of overdispersion is moderate (Cox, 1983; Firth, 1987). Thus, quasi-likelihood function allows us to estimate the dispersion parameter in moderately over-dispersed regression models. We applied these principles to the present data under investigation.

The mean submodel has first and second moments as

$$E(y|x) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

and $\text{var}(y_i) = \Phi V(\mu_i)$ respectively, where Φ is the overdispersion parameter. Systematic components consist of marital status,

educational attainment, type of residence, employment status and number of children. The model was fitted to the data using several different link functions including logit, log, and complementary log-log. For the variance functions, choices were made from μ , $\mu(1 - \mu)$, and the constant.

Based on the goodness of fit statistics, the mean submodel with a log link and μ as the variance function gave the best fit. The log link corresponds to multiplicative effects of the covariates. The " μ " variance function is equivalent to Φ as the coefficient of variation of the response (McCullagh & Nelder, 1989). The regression coefficient estimates for the mean and the dispersion submodel are given in the first two columns of Table 3.

The dispersion submodel was also fitted with different link and variance functions. The choices for link functions included identity and square root and the choices for variance functions included the constant, μ , and μ^2 (the squared coefficient of variation). Based on the goodness-of-fit statistics (mostly, how much deviance relative to the degrees of freedom), the dispersion model with square root link function and μ the variance function was chosen.

Some parameter estimates from the generalized quasi-likelihood model from Table 3 are similar in value to the corresponding values of Table 2 when the binomial logistic regression model was applied. In the generalized model, there are two variables significant at the .05 level. Education has a negative effect on poverty, thus the more educated they are the less likely they are in poverty, while the more children in the household increased the odds of Asian women living in poverty. The deviance from the generalized quasi-likelihood model suggests that the overdispersion is accounted for and the model is a good fit.

The response variable of the dispersion submodel is the square of the residual. Residuals are one principal tool for assessing how well a model fits the data. They can be used to assess the importance and relationship of a term in the model as well as to search for anomalous values. For generalized linear models, residuals can also

help assess and verify the form of the variance as a function of the mean response.

There are different kinds of residuals that can be employed. First the deviance residuals,

$$r_i^D = \text{sign}(y_i - \mu_i) \sqrt{d_i}$$

where d_i is the contribution of the i th observation to the deviance. The deviance is

$$D = \sum_i (r_i^D)^2$$

These residuals are useful detecting observations with unduly large influence on the fitting process, since they reflect the same criterion as used in the fit. Secondly, there is the Pearson residuals,

$$r_r^P = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}} \text{ and } \chi^2 = \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

is the chi-square statistic.

The dispersion submodel has as its response variable the squares of the residuals (the difference between observed values and fitted values). If the mean submodel fits the model well, then there may not be a need to model the deviance and none of the parameter estimates in the dispersion model may be significant. An examination of the parameter estimates and standard errors from the dispersion submodel in Table 3 suggests that the form of the variance as a function of the mean response is appropriate in our model and there are almost no anomalous values in our model. The mean deviance for the dispersion model is 2.05.

Table 3. Parameter estimates and (standard errors) for Generalized Quasi-likelihood model.

Covariate	Mean Submodel		Dispersion Submodel	
	Parameter Estimate	Standard Errors	Parameter Estimate	Standard Errors
Intercept	-1.128*	.504	1.698**	.414
<i>Marital Status</i>				
Unmarried	.388	.220	-.034	.201
Married				
<i>Educational</i>	-.206*	.084	-.104	.069
<i>Area of Residence</i>				
Metropolitan	-.412	.264	.234	.209
Non-metropolitan				
<i>Employment Status</i>				
Employed	-.315	.191	.216	.199
Not Employed				
<i>Number of Children</i>	.338*	.172	-.345*	.138

Note: * at the .05 level, and ** at the .01 level.

Conclusion

Generalized linear models such as binomial logistic regression and Poisson regression are very widely used in social, economic, and medical research. While the binomial logistic regression is easy to use and interpret, we need to look for an alternative if there is overdispersion in our data.

When the data are over-dispersed, due to heterogeneity or the clustering effect at the group level, it is necessary to model the overdispersion. Quasi-likelihood models allow you to model such overdispersion as the estimation process assumes only a form for the functional relationship between the mean and the variance. Further they allow us to simultaneously model the mean and the variance without accounting for any distributional assumptions.

Quasi-likelihood models were used to model the data from U.S. Census Bureau's 1998 Current Population Surveys. Data pertaining to Asian American women who lived in the western region of the United States showed that covariates such as marital status, educational attainment, area of residence, employment status, and number of children are not all predictors when modeling poverty, as with other ethnic and racial groups. Use of the binomial logistic regression model showed the presence of overdispersion. Quasi-likelihood functions were used to model that overdispersion. Several link functions and variance functions were examined to identify a model with the best fit. For these data, a mean submodel with the log as the link function and μ as the variance function and a dispersion submodel with square root as the link function and μ as the variance function fit well. Thus, the binomial logistic regression models overstated the effects of the covariates, in part due to the unaccounted extravariation.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
 Allison, P. D. (1999). *Logistic Regression Using the SAS System*. Cary: SAS Institute.

Aitkin, M. (1987). Modeling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, 36, 332-339.
 Bain, L. J. & Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics*, Duxbury.
 Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT.
 Carroll, R. J. & Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, 10, 429-441.
 Collett, D. (1991). *Modeling Binary Data*. London: Chapman & Hall.
 Cox, D. R. (1983). Some Remarks on Overdispersion. *Biometrika*, 70, 269-274.
 Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81, 701-721.
 Firth, D. (1987). On the Efficiency of Quasi-Likelihood Estimation. *Biometrika*, 74, 233-245.
 Macionis, J. J. (2001). *Sociology* (8th ed). Upper Saddle River: Prentice Hall.
 McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, 11, 59-67.
 McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.
 Moore, D. F. & Tsatis, A. (1991). Robust Estimation of the Variance in Moment Methods for Extra-Binomial and Extra-Poisson Variation. *Biometrics*, 47, 383-401.
 Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, A135 370-384.
 Nelder, J. A. & Pregibon, D. (1987). An Extended quasi-likelihood function. *Biometrika*, 74, 221-232.
 Pregibon, D. (1984). Review of *Generalized Linear Models* by McCullagh and Nelder. *Annals of Statistics*, 12, 1589-1596.
 Smyth, G. K. (1989). Generalized Linear Models with Varying Dispersion. *Journal of Royal Statistical Society Ser. B* 51, 47-60.
 S-Plus 4 Guide to Statistics. (1997). Seattle: Data Analysis Products Division, MathSoft.

Stroud, T. W. F. (1971). On obtaining large sample tests from asymptotically normal estimators. *Annals of Statistics*, 42, 1412-1424.

U.S. Census Bureau. (1998). *Statistical Abstract of the United States 1998*. Washington, D.C.: U.S. Government Printing Office.

U.S. Census Bureau. (1999). *Poverty in the United States 1998*. P60-207. Washington, D.C.: U.S. Government Printing Office.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss Newton method. *Biometrika*, 61, 439-447.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, 144-148.

Wilson, J. R. (1989). Chi-square tests for overdispersion with multiparameter estimates. *Applied Statistics*, 38, 441-453.

Wilson, W. J. (1996). *When Work Disappears: The World of the New Urban Poor*. New York: Alfred A. Knopf.

Yanez, N. D. & Wilson, J. R. (1995). Comparison of quasi-likelihood models for overdispersion. *Australian Journal of Statistics*, 37, 217 -231.

Meta-Analysis Of Results And Individual Patient Data In Epidemiological Studies

Aurelio Tobías

Department of Statistics and Econometrics
Universidad Carlos III de Madrid, Spain

Marc Saez

Department of Economics
Universitat de Girona, Spain

Manolis Kogevinas

Research Unit of Respiratory and Environmental Health
Health Institut Municipal d'Investigació Mèdica, Barcelona, Spain

---

Epidemiological information can be aggregated by combining results through a meta-analysis technique, or by pooling and analyzing primary data. Common approaches to analyzing pooled studies through an example on the effect of occupational exposure to wood dust on sinonasal cancer are described. Results were combined applying a meta-analysis technique. Alternatively, primary data from all studies were pooled and re-analyzed using mixed effect models. The combination of individual information rather than results is desirable to facilitate interpretations of epidemiological findings, leading also to more precise estimations and more powerful statistical tests for study heterogeneity.

Key words: Pooled-analysis, meta-analysis, generalized linear mixed models, random effects, epidemiological methods

---

Introduction

The requirement of large samples of subjects is particularly important in studies of uncommon diseases, such as most types of cancer, and even in diseases with higher prevalence, such as asthma. Large multi-center studies, or combining information from multiple studies, are the best approaches for improving the information from, and overcoming lack of power in individual studies. Information from multiple

epidemiological studies can be aggregated either by combining results, such as summary measures (for example, odds ratios), through a meta-analysis technique, or by pooling and analysing primary data.

The combination of results, usually called meta-analysis, involves the compilation of published results from different studies (Thacker, 1988). Another option is to pool individual information from each study and to conduct an analysis for the entire data set, this being defined as meta-analysis of individual patient data (Stewart & Parmar, 1993). Meta-analysis of individual patient data was originally applied to clinical trials, although in epidemiological studies this procedure is usually known as pooled-analysis (Checkoway, 1991).

Both, meta-analysis of results and meta-analysis of individual patients have advantages and limitations (Thacker, 1988; Friedenreich, 1993). Meta-analysis of results has a relatively low cost and the appropriate statistical techniques are straightforward to understand and implement. It does not require sharing of primary data, because it can be performed from reviews of internal reports in multi-center

---

Aurelio Tobías is an Assistant Professor, and is concluding his doctoral studies. His research interests include meta-analysis and time-series regression in environmental epidemiology. Email: aurelio.tobias@madrid.org. Marc Saez is Professor of Statistics and Econometrics. Manolis Kogevinas is a Senior Scientist, and is affiliated with the International Agency for Research on Cancer, and at The Open University as an Associate Lecturer.

studies, or from reviews of published and unpublished results. In this situation, meta-analysis of results is sensitive to publication bias, since unpublished results are usually difficult to locate or obtain. This fact must be taken into account, treating results with caution (Vanderbroucke, 1988). Practices of data reporting also pose difficulties when examining specific diseases (Checkoway, 1991).

Meta-analysis of published results is limited to the information available, permitting usually only a meta-analysis of overall risks. The procedure has also been criticised since it can be conducted without full consideration of the underlying statistical assumptions and inferences required for this type of analysis (Oakes, 1990). Further, the use of a chi-square statistic for assessing heterogeneity in the original studies has been criticised due to its lack of power (Spector & Thompson, 1991).

An alternative to meta-analysis is to pool and then re-analyse individual data. Pooled analysis of epidemiological studies, defined as a combination of primary data from published and unpublished studies has become common recently. With such an approach, rare exposures can be more easily studied (Clayton, 1991), and confounding and possible interaction effects can be more accurately estimated.

Pooled analysis however, is more difficult to conduct since it is more labour and time-intensive. Common definitions for outcomes and other covariates must be used. Thus, important issues are how to accommodate differences in the populations and methods used in the original studies, and to assess their possible effect on the results. Friedenreich (1993) outlined guideline procedures on pooling of primary data for the integration of qualitative assessments of studies with quantitative estimates of the results. However, there are no clear guidelines on the statistical analysis of pooled data, especially if there is heterogeneity in the original studies.

The objective of this article is to describe and compare common statistical techniques for analysing pooled and multi-center studies. Discussed are the alternative methodologies of performing a meta-analysis of results, and of pooling and re-analysing primary data.

Methodology

Fixed effects model

The meta-analysis technique is a straightforward process of weighting results under a simplistic assumption, that the true effect (θ) is the same for each centre, or study, that is an assumption of homogeneity ($\theta_i = \theta$ for all i). Most meta-analyses use fixed effects estimates. The weighted average $\hat{\theta} = \sum(w_i \hat{\theta}_i) / \sum w_i$ is an unbiased estimate of θ , where the weight $w_i = 1/v_i$ is determined by variance (v_i) of the effect estimate, which depends on the effect size and the size of the study. This weighted average has the smallest estimated variance $\hat{v} = 1/\sum w_i$ among the weighted averages of $\hat{\theta}_i$ (Cox, 1982).

There are different versions of this estimator, differing either in the scale of the effect (log or untransformed odds ratio) or in the approximation of the variance used. The Mantel-Haenszel (Mantel & Haenszel, 1959) method weights the untransformed odds ratios approximately proportional to their sample sizes. In Woolf's method (Woolf, 1955), the log odds ratio are weighted inversely according to their estimated variances from a 2x2 table or asymptotically from a logistic regression. Finally, Peto's method (Peto et al., 1977) uses the observed minus expected values over their variances as an approximation to the log odds ratio. Among these, Woolf's method is the most frequently used. Although Peto's method has been recommended to analyse experimental studies other authors suggest using Woolf's method for any type of study (Greenland, 1987).

Testing heterogeneity

An overall test of heterogeneity of the original studies is provided by calculating $Q = \sum w_i (\hat{\theta}_i - \hat{\theta})^2$ following a χ^2_{k-1} distribution under the homogeneity assumption, where k is the number of studies to pool. The lack of power of this test has been well established (Spector & Thompson, 1991), and the absence of formal statistical significance need not imply true homogeneity. Graphically, heterogeneity can also be assessed in first instance from a Forrest plot (Light et al., 1994), although other methods

have been developed to complement this test and also to detect sources of heterogeneity. Among these, the Galbraith plot (Galbraith, 1988) has been more frequently recommended (Thompson, 1993) than others, such as the I'Abbé plot (I'Abbé et al., 1987) or the odd man-out procedure (Walker et al., 1988), which will not be discussed further.

Random effects model

An alternative method suggested by DerSimonian and Laird (1986) considers that the heterogeneity between studies is unexplained. This is known as a random effects model where, $\hat{\theta}_i \sim N(\theta_i, v_i)$, and $\theta_i \sim N(\theta, \sigma^2)$. Here the θ_i effects have some dispersion around the overall estimate θ , indicated by the between-study variance σ^2 . An estimate $\hat{\sigma}^2$ of σ^2 must therefore be derived from the results. Then, the inverse variance weights become $w_i^* = 1/(v_i + \sigma^2)$, where v_i is the variance within the i^{th} original study and σ^2 is the variance between studies. The combined estimate of the effect is defined by $\hat{\theta} = \sum(w_i^* \hat{\theta}_i) / \sum w_i^*$ with variance $\hat{v} = 1 / \sum w_i^*$.

Among the standard packages, Stata, S-Plus and SAS have available macros to perform meta-analysis, which can be downloaded from <http://www.prw.le.ac.uk/epidemiology/personal/ajs22/meta/>. However, meta-analysis formulae could be easily programmed in other standard packages or even in a simple spreadsheet.

Pooled analysis of primary data

Fixed-effects model and testing heterogeneity

The analysis of pooled data does not present any difficulty if a fixed effect model is considered, that is assuming that all the effects are fixed for study. For example, if the outcome variable is dichotomous (i.e., case-control status) standard logistic regression can be used. Test for heterogeneity by comparing the model that includes the interaction between study and the exposure of interest and the previous model without the interaction, using the likelihood ratio test. From the statistical point of view, the most important question is to consider or not the presence of heterogeneity. If statistical heterogeneity is presented mixed effects models must be used (Breslow & Clayton, 1993).

Mixed-effects model

Mixed effects model differs from conventional fixed effects model in that, as well as modelling location parameters, they also model the underlying covariance structure of the data. The simplest way to model covariance is by specifying random effects in the model. Briefly, a Normal Mixed Model is defined as $y = X\alpha + Z\beta + e$, where X is a design matrix for fixed effects and Z is a design matrix for random effects, then $\beta \sim N(0, G)$, $V(e) = R$ and $V(y) = ZGZ' + R$, where G is a diagonal matrix of variance parameters, R is the residual variance matrix, and e is the residual error.

However, when the dependent variable is non-linear, define a Generalised Linear Mixed Model as follows: $y = \mu + e$, $g(\mu) = X\alpha + Z\beta$, with $\beta \sim N(0, G)$, $V(e) = R$ and $V(y) \approx BZGZ'B + R$, where the new parameter μ are the expected values, g is the link function, and B is the diagonal matrix of variance terms. An extended notation about mixed models can be found in Brown and Prescott (1999).

There are no clear rules to define if the variables included in the model should be defined as fixed or random effects. Pooled analyses in epidemiology are usually carried out because insufficient subjects are available for the study at any one centre. Thus, there will be extra variability in the risk factor estimates, which can usually be due to differences between studies (for example different investigators, types of patients, etc.) This extra variability can be taken into account by including study and interaction between study and risk factor in the model. When study and interaction between study and risk factor are taken as random, allowance is made for variability in the magnitude of risk factor estimates between studies.

The choice will depend on whether risk factor estimates are related to the set of studies used in the pooled analysis. Thus, local risk factor estimates for the sampled set of individual studies will be obtained fitting the study and interaction between study and risk factor variables. To obtain a global risk factor estimate the study and interaction between study and risk factor should be fitted as random. When this is done the standard error of the risk factor

estimate is increased to reflect the heterogeneity across studies. Taking study as a random effect can increase the accuracy of risk factor estimates since information from the study error stratum is used in addition to that from the residual stratum (Brown & Prescott, 1999).

In a pooled analysis of epidemiological studies there are other factors that differ at the study level that can help to explain differences in results between studies. These may be sensible to be included as random effects in a mixed model and reduce the variability of the interaction between study and risk factor, leading to more precise estimates.

Mixed effects for linear models are available in standard packages: SAS (GLM and MIXED procedures), Stata (xtreg) and S-Plus (lsfit). To fit a mixed effects model for non-linear data, specific macros for Stata (gllamm) and SAS (GLIMMIX) have been recently developed. However, mixed-effects models can also be fitted in other specialised software such as MLnWin.

Analysis of 8 case-control studies on sinonasal cancer

The aim of the investigation was to reanalyse data available from eight previously published case-control studies focused on the differential effect that occupations exposed to wood dust have on the major histological types of sinonasal cancer. The reanalysis was done within each individual study, and pooling them after that to obtain a summary measure of the exposure effect. This research formed part of a wider project on occupational cancer in Europe by the International Agency for Research on Cancer.

Primary data from 8 case-control studies from Germany, Netherlands, France, Sweden, and four studies in Italy (Vigevano, Brescia, Biella, Siena) were available. These studies examined the association of occupational wood dust exposures and sinonasal cancer, taking into account histological types. A detailed description of the process for selection of the

Table 1. Description of eight published case-control studies on the association between occupational wood dust exposures and sinonasal cancer.

Study	Sex		Age		
	Male n (%)	Female n (%)	< 55 n (%)	55-65 n (%)	> 65 n (%)
Germany	59 (59.6)	40 (40.0)	25 (25.3)	23 (23.2)	51 (51.5)
Netherlands	286 (100.0)	0 (0.0)	87 (30.4)	87 (30.4)	112 (39.2)
France	487 (79.1)	129 (20.9)	216 (35.1)	191 (31.0)	209 (33.9)
Sweden	585 (100.0)	0 (0.0)	190 (32.5)	129 (22.0)	266 (45.5)
Italy					
Siena	238 (71.9)	93 (28.1)	83 (25.1)	79 (23.9)	169 (51.0)
Biella	110 (83.9)	21 (16.3)	32 (24.4)	48 (36.7)	51 (38.9)
Brescia	93 (68.4)	43 (31.6)	41 (30.1)	21 (15.4)	74 (54.4)
Vigevano	31 (77.5)	9 (22.5)	10 (21)	11 (27.5)	19 (47.5)
TOTAL	1889 (84.9)	335 (15.1)	684 (30.8)	589 (26.5)	951 (42.8)

Table 1 continued.

Smoking status			Controls	Cases
non n (%)	ex n (%)	Current n (%)	(exposed/ non-exposed)	(exposed/ non-exposed)
46 (46.5)	11 (11.1)	42 (42.4)	1/53	2/43
16 (5.6)	108 (37.8)	91 (27.6)	35/160	25/66
234 (38.0)	237 (38.5)	145 (23.5)	46/363	99/108
215 (36.6)	136 (23.2)	234 (40.0)	272/269	20/24
113 (34.1)	127 (38.4)	91 (27.5)	26/228	16/62
38 (29.0)	45 (34.3)	48 (36.6)	7/98	7/19
64 (47.1)	35 (25.7)	37 (27.2)	7/95	3/31
12 (30.0)	11 (27.5)	17 (42.5)	4/23	0/13
738 (33.2)	710 (31.9)	776 (34.9)	398/1288	176/366

studies and classification of exposures can be found elsewhere (Mannetje et al., 1999). The pooled data set includes cases includes 538 cases and 1,686 controls. The cases also includes 238 squamous cell carcinomas, 155 adenocarcinomas, 79 other histologies, and 59 unknown histology. However, the studies differed in the methods for recruitment and interview of the subjects. Table 1 presents a description of the studies by sex, age, smoking status and occupational wood dust exposures to sinonasal cancer.

Results

Meta-analysis of results

Initially the odds ratio was obtained (OR) by each study using logistic regression adjusted by age, sex and smoking status. Results for each study are showed in a Forrest plot (Figure 1). Note that the logistic regression model for the study from Vigevano (Italy) did not converge because no cases were exposed. However, a crude odds ratio for Vigevano using a Mantel-Haenzsel estimate or thorough an exact-method could be obtained, but that estimate may be seriously biased since it would be unadjusted by the potential confounding variables considered in the logistic regression models. Results are presented for 7 of the studies. The Forrest plot gives a first indication

that there is heterogeneity between studies.

Thus, as a first approach to obtain a summary measure of the exposure effect, combined the results of each study applying a meta-analysis technique, weighting by the inverse of variance (OR=2.93, 95% CI: 2.24 to 3.83). The Forrest plot gives an initial indication that there is heterogeneity between studies. Heterogeneity of effects between studies was tested using the Q-statistic (Table 2), which confirms that there is a considerable amount heterogeneity between studies ($\chi^2=45.357$, df=6, $p<0.001$). Finally, a random effects model was applied using DerSimonian and Laird's method (OR=2.43, 95% CI: 1.06 to 5.59). Analyses were done using Stata, release 7.0, statistical software.

Pooled analysis of individual data

Primary data from all studies were pooled and first analysed using a fixed effects model (Table 2). Thus, standard logistic regression was applied adjusting again by sex, age and smoking status, providing different risk estimates with a narrowness confidence interval than meta-analysis (OR=3.05, 95% CI: 2.36 to 3.95). This difference is mainly due to the fact that in the pooled-analysis the data from Vigevano study are included, while in the meta-analysis they were not, because no risk estimates can be estimated for this study.

Figure 1. Results from eight published case-control studies on the association between occupational wood dust exposures and sinonasal cancer. Odds ratios for each study are adjusted by sex, age, and smoking status.

Table 2. Results from eight published case-control studies on the association between occupational wood dust exposures and sinonasal cancer, analysed combining results (meta-analysis using fixed and random effects model), and combining individual patient data (pooled analysis using a fixed effects model).

Model	Occupational dust wood exposure			Test for heterogeneity		
	β (se)	OR	(95% CI)	χ^2	df	p-value
Meta-analysis						
Fixed effects	1.074 (0.136)	2.93	(2.24, 3.83)	45.357	6	<0.001
Random effects	0.891 (0.424)	2.43	(1.06, 5.59)			
Fixed effects pooled analysis						
Including all studies	1.116 (0.132)	3.05	(2.36, 3.95)	51.317	7	<0.001
Excluding Vigevano study	1.079 (0.131)	2.94	(2.28, 3.80)	44.374	6	<0.001

Table 3. Sensitivity analysis for mixed-effects models to combine individual patient data (pooled analysis) of eight published case-control studies on the association between occupational wood dust exposures and sinonasal cancer.

Variables defined as random effects	Deviance	Occupational wood dust exposure β (se)	OR	(95% CI)	Variance components
Study	2210.94	1.052 (0.133)	2.86	(2.21, 3.72)	0.6592
Study×wood dust exposure	2162.64	0.718 (0.369)	2.05	(0.99, 4.23)	0.3781
Study	2163.07	0.662 (0.361)	1.94	(0.96, 3.93)	0.3498
Study×wood dust exposure					0.3703
Study	2166.49	0.661 (0.356)	1.94	(0.96, 3.89)	0.3568
Study×wood dust exposure					0.3761
Sex					0.0375
Age					0.0076
Smoking status					0.0039

Pooled analysis, excluding the Vigeveno study, gives a result from a fixed effects model much closer (OR=2.94, 95% CI: 2.28 to 3.80) to those from a meta-analysis. Heterogeneity was assessed using the likelihood ratio test ($\chi^2=51.317$, $df=7$, $p<0.001$).

For the mixed-effects model the analyses were performed using the SAS macro GLIMMIX which implements the Penalised Quasi Likelihood (PQL) approach. Firstly, three different models are fit, defining study, the interaction between study and occupational dust exposure, and both study and its interaction with occupational dust exposure to be the random effects (with an unstructured covariance matrix), respectively.

Also introduced are the covariates sex, age and smoking status, as fixed effects (Table 3). In the first model, where only the variable study is defined as a random effect, occupational dust exposure is closer to the previous result using a fixed effect approach (OR=2.86, 95% CI: 2.21 to 3.72), although it increases the accuracy of the exposure estimate. However, when the interaction between study and occupational dust exposure are included as random effects the standard error of the exposure estimate is increased coming to lose the statistical significance (OR=2.05, 95% CI:

0.99 to 4.23), due to it is reflecting the heterogeneity across studies. Finally, when both study and its interaction with occupational dust exposure are included as random effects, although the standard error of the exposure estimate is again increase, results are more accurate than previous model (OR=1.94, 95% CI: 0.96 to 3.93).

However, as seen from Table 1, the effects of the covariates factors varied across studies. For this reason, it was decided to include these factors also as random effects, as a sensitivity analysis (Table 3). In this situation, occupational dust estimate do not change, although this model provides slightly more accurate result (OR=1.94, 95% CI: 0.96 to 3.89) due to inclusion of covariates sex, age and smoking status as a random effects. This fact is reflected in the variance components, being lower than those for previous models.

Conclusion

It is important to consider the differences between pooled studies using individual patient data and classical meta-analyses of results. The key point in a pooled study is to integrate accommodate in the populations and methods

used in the original studies, and to assess their possible effect on the results. Friedenreich (1993) reported useful guidelines for pooling of primary data.

The principal advantage of having individual patient data is that adjustments can be made for different covariates. However, the assumptions that are made in fitting the pooled analysis of individual patient data need to be specified and discussed. In our analysis, the confounding effects of sex, age and smoking status must be assumed to be the same across the original studies. However, if the effects of the confounding factors varied across studies, then it may be sensible to include these as random effects.

The main difference between a fixed and a random effect will depend on the intention of the analysis. If local estimates need to be provided, then a fixed effects model must be fitted. Moreover, if the aim of the analysis is to report a global estimate, then always define the study and its interaction with the risk factor as random effects. Thus random effects are sources of variation in a model due to individuals or groups over above the individual error term (Campbell, 2001). For these reasons, one should consider that combining individual patient data from different sources is complex, and in practice, various assumptions need to be made. Various models with a variety of combinations of fixed and random effects should be fitted to assess the sensitivity of the chosen model.

It is usually desirable to work with individual information rather than combined results to facilitate interpretations of epidemiological findings (Blettner et al. 1999), although others (Steinberg et al., 1997) suggested that meta-analysis of results is adequate under certain circumstances. The obvious advantages in a pooled-analysis pertain to increases in the study size, both of the overall and the reference populations used in the analysis. This leads to more precise estimations and more powerful statistical tests for heterogeneity. Furthermore, there may be studies that are difficult or impossible to incorporate into a meta-analysis because of zero counts, as was the case with the study in Vigeveno presented in the example for instance, which can be included in the pooled analysis. Their

absence from the meta-analysis produces bias that the pooled analysis does not suffer from. However meta-analysis of results is much less costly (Steinberg et al., 1997).

The accuracy with which variance components are estimated is dependent on the number of studies included in the analysis. Problem arise when only few studies are available, which means that there will be considerable uncertainty in the estimate of the between study variance. Also mixed effects models, rather than classical fixed effects models, make more assumptions.

In consequence, there could be problems of bias or lack of convergence of the model fitting process for complex models, such as fitting fixed effects within a random effect, modelling repeated measurements, or dealing with small to moderate samples (Breslow & Lin, 1995; Kuk, 1995). Nevertheless potential solutions such as bootstrapping or full Bayesian analysis are available (Brown & Prescott, 1999), but these methods require very large amount of computer power and time. The main difference between a Bayesian analysis and a maximum likelihood method (as PQL approach used in our analysis) is that techniques are used to evaluate the likelihood surface, rather than estimate the parameters that maximise it.

In absence of heterogeneity, both meta-analysis and pooled analysis produce close results, in terms of estimates and variances. This is done because the meta-analysis estimate is a weighted mean of the means by each centre, and the pooled analysis estimate from a regression model is also a weighted mean. So, both methods are estimating the same quantity. In a meta-analysis technique, a random effects model will produce same estimates as a fixed effects model, and in a pooled analysis fixed and mixed effects models will produce similar results.

Whenever heterogeneity is assessed one approach is to look for possible sources of it. Meta-analyses should incorporate a careful investigation of potential sources of heterogeneity (Thompson, 1994), because statistical tests for heterogeneity may fail to detect moderate degrees of it. Graphical techniques, like Galbraith plots, are useful in searching for sources of heterogeneity. Statistical heterogeneity may be caused by

known clinical differences between populations or by methodological characteristics between studies. Interpretation of possible sources of heterogeneity requires caution because analyses are post-hoc (Spector & Thompson, 1991).

Frequently, heterogeneity is related to unknown causes. Then the formal approach should be to fit a random effects model -in a meta-analysis-, or a mixed effects model -in a pooled analysis-. The choice between these fixed and random, or mixed, effects rarely affect the conclusions obtained (Spector & Thompson, 1991). The greater is the amount of heterogeneity, the greater will be differences between estimates from fixed and random/mixed effects models. However, variances from random, or mixed, effects model will always be higher than those from fixed effects model, because in the former models both variances, between and within studies, are taken into account. Independently of whether fixed or random/mixed effects models are used, estimates from pooled analyses are more precise than those from meta-analyses.

When dealing with pooled or multi-centre studies, results have to be evaluated for the researcher. Then, if an individual analysis for each centre, or study, is done, a meta-analysis can quickly and easily be performed. This result should be compared, as a sensitivity analysis, with the result from the model using individual data, due to conflicting results possibly being found.

For example, Harrison and Waterbor (1999) found disagreeing results in the relationship between dietary fat and breast cancer if primary study results were heterogeneous. In that way, it was seen in the study that if the two methods (meta-analysis and pooled analysis) produce marked different results then a possible source of divergence, such as absence of exposed cases, should be considered in further analysis. This implies that meta-analysis techniques are still useful; according to Spector and Thompson (1991), "Meta-analysis is here to stay. Epidemiologists, statisticians, and clinicians should all be aware the uses and limitations of the technique".

References

- Blettner, M., Sauerbrei, W., Schehofer, B., Scheuchenpflug, T., & Friedenreich, C. (1999). Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology*, 28, 1-9.
- Breslow, N. E., & Clayton, D. (1993). Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82, 81-93.
- Brown, H., & Prescott R. (1999). *Applied mixed-models in medicine*. New York: Wiley.
- Campbell, M. J. (2001). *Statistics at square two*. London: MJB Books.
- Checkoway, H. (1991) Data pooling in occupational studies. *Journal of Occupational Medicine*, 33, 1257-1260.
- Clayton, D. (1991). The EURODEM collaborative re-analysis of case control studies of Alzheimer's disease: some methodological considerations. *International Journal of Epidemiology*, 20, S62-S64.
- Cox, D. R. (1982). Combination of data. In: *Encyclopaedia of Statistical Sciences* (S. Kotz, N.L. Johnson, Eds.), p. 45-53. New York: Wiley.
- DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.
- Friedenreich, C. M. (1993). Methods for pooled analyses of epidemiologic studies. *Epidemiology*, 4, 295-302.
- Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, 7, 889-894.
- Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiological Reviews*, 9, 1-30.
- Harrison, R. A., & Waterbor, J. C. (1999). Understanding meta-analysis in cancer epidemiology: dietary fat and breast cancer. *Cancer Detection and Prevention*, 23, 97-106.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of the Royal Statistical Society, Series B*, 57, 395-407.

L'Abbé, K. A., Detsky, A. S., & O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine*, 107, 224-233.

Light, R. J., Singer, J. D., Willett, J. B., Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Mannetje, A., Kogevinas, M., Luce, D., Demers, P., Bolm-Audorf, U., Comba, P., Hardell, L., Hayes, R. B., Leclerc, A., Maganani, C., Merler, E., Tobías, A., & Boffeta, P. (1999). Differential effect of tobacco and occupation on specific histological types of sinonasal cancer. *American Journal of Industrial Medicine*, 36, 101-107.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 719-748.

Oakes, M. (1990). On meta-analysis. In *Statistical Inference*, p. 157-163. Chestnut Hill, MA: Epidemiology Resources Inc.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., & Smith P. G. (1977). Design and analysis of randomised clinical trials requiring prolonged observations of each patient. II. Analysis and examples. *British Journal of Cancer*, 35, 1-39.

Spector, T. D., & Thompson, S. G. (1991). The potential and limitations of meta-analysis. *Journal of Epidemiology and Community Health*, 45, 89-92.

Steinberg, K. K., Smith, S. J., Stroup, D. F., Olkin, I., Lee, N. C., Williamson, G. D., & Thacker, S. B. (1997). Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer. *American Journal of Epidemiology*, 145, 917-925.

Stewart, L. A., & Parmar, M. K. (1993). Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet*, 341, 418-422.

Thacker, S. B. (1988). Meta-analysis, a quantitative approach to research integration. *Journal of the American Medical Association*, 259, 1685-1689.

Thompson, S. G. (1993). Controversies in meta-analysis, the case of the trials of serum cholesterol reduction. *Statistical Methods in Medical Research*, 2, 173-192.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, 309, 1351-1355.

Vanderbroucke, J. P. (1988). Passive smoking and lung cancer, a publication bias?. *British Medical Journal*, 296, 391-392.

Walker, A. M., Martin-Moreno, J. M., & Rodríguez-Artalejo, F. (1988). Odd man out: A graphical approach to meta-analysis. *American Journal of Public Health*, 78, 961-966.

Woolf, B. (1955). On estimating the relationship between blood group and disease. *Annals of Human Genetics*, 19, 251-253.

Validation Studies: Matters Of Dimensionality, Accuracy, And Parsimony With Predictive Discriminant Analysis And Factor Analysis

David A. Walker
Educational Research and Assessment Department
Northern Illinois University

Two studies were used as examples that examined issues of dimensionality, accuracy, and parsimony in educational research via the use of predictive discriminant analysis and factor analysis. Using a two-group problem, study 1 looked at how accurately group membership could be predicted from subjects' test scores. Study 2 looked at the dimensionality structure of an instrument and if it developed constructs that would measure theorized domains.

Key words: Predictive discriminant analysis, factor analysis, dimensionality, accuracy, parsimony

---

Introduction

The first study in this article has two intentions. First, if there is an interest in the degree to which group membership, based upon a set of predictor variables, can be predicted the question posed may be: How accurately can group membership in either Average grade point average (GPA) or Above Average GPA from the subjects' Florida College Level Academic Skills Test (CLAST) scores? A second question may be: In terms of their relative contribution to classification accuracy, how well can a ranking of the predictor variables predict if a subject taking the CLAST is going to be in the Average GPA group or the Above Average GPA group?

Study 1.

The CLAST is an achievement test that was first implemented by the Florida State Board of Education (SBE) in 1984 as part of its educational accountability measures. The test is comprised of four subtests in mathematics,

reading, writing, and essay that purport to measure students' academic proficiency, by the completion of the sophomore year, in the areas of computation and communication. The CLAST is administered three times a year in October, February, and June. Students who have accrued a minimum of 18 semester hours may apply to sit for the test. Institutions may require students to pass 3 subtests before they can earn more than 60 degree credits and/or pass all 4 subtests before obtaining 96 degree credits toward a baccalaureate degree.

Subtests, however, can be taken as many times as needed until passed. To receive an associate in arts degree from any of Florida's 28 public community colleges or obtain admission to upper-division status in any of Florida's 11 public, 4-year institutions, a student must pass all subtests of the CLAST or receive one of many exemption options (Florida Atlantic University, 2002; Florida Department of Education, 2000).

Exemptions from any of the three communication subtests are predicated on attaining a 2.50 GPA in two designated college-level English courses. Exemption from the mathematics portion is based on a 2.50 GPA in two defined courses. Also, an ACT score of 21 in mathematics, a 22 in reading, a 21 in English, or an SAT score of 500 in quantitative and/or verbal are approved exemptions. A documented learning disability or physiological impairment, or if a student has already earned a Bachelor's

---

David Walker is an Assistant Professor of Educational Research and Assessment at Northern Illinois University. His research interests include structural equation modeling, effect sizes, factor analyses, predictive discriminant analysis, predictive validity, weighting, and bootstrapping. Email: dawalker@niu.edu.

degree and is seeking a second undergraduate degree, will merit an exemption (Florida International University, 2002; University of South Florida, 2002). It should be noted that such exemptions have the ability to reduce the internal and external score validity of the CLAST.

The subtests measure students' academic proficiency in lower-division course work in the general areas of mathematics, reading, writing, and essay. The mathematics subtest includes selection-type items (i.e., multiple-choice) in the following areas: algebra, arithmetic, geometry, logical reasoning, measurement, probability, and statistics. The reading subtest has multiple-choice items that measure two areas: literal comprehension and critical comprehension. The English portion of the CLAST also uses multiple-choice items and measures students' skill levels in the areas of word choice, sentence structure, grammar, spelling, capitalization, and punctuation. Scores for the mathematics, reading, and English subtests range from 200 to 400 points.

The SBE has changed the cut scores for passing these 3 subtests from a minimum score of 260 in 1984 to a present score of 295. Current mean averages for first-time examinees from the 1999-00 academic year show that mathematics had a 3 administration average of 299, reading was 305, and English was 309 (Florida Atlantic University, 2002; Florida Department of Education, 2000).

The essay test allows students to choose from two topics and write about one of these. Essay writing measures students' skills in the areas of composition, effective language use, and the dissemination of ideas. Using a holistic rubric, two trained readers rate each essay test. Essay scores range from 2 to 12 points. In 1984, the original cut score was a 4, however; the current minimal score has been changed to a 6. From academic year 1999-00, the mean average for the essay test was a 7 (Florida Department of Education, 2000; Indian River Community College, 2002).

Methodology

The four predictor variables were the subtests on the CLAST: mathematics, reading, English, and

essay. The criterion variable was undergraduate GPA, where 4.00 = A, 3.00 = B, 2.00 = C. There were no GPAs below 2.00 because to be in the sample as a recent graduate of a Florida four-year public institution, a participant needed at least a 2.00 to graduate. Thus, GPA was operationalized as a comprehensive academic performance measure of students' cognitive abilities in their entire degree program of study. GPA has been used as a criterion variable and is often influenced by many factors such as the facility or difficulty level of course content, student effort, instructor competency, and student involvement, or not, in co-curricular activities. More considerably, GPA is a variable that has been cited as a measure of students' cognitive abilities, especially in the areas of verbal and quantitative skills (Brown & Campion, 1994; Roth & Bobko, 2000; Wolfe & Johnson, 1995).

Reliability

Using the Kuder-Richardson 20 method, the reliability of the CLAST subtest scores for the 3 administrations in 1999-00, along with standard error of measurements shown in parentheses, were .83 (3.03), .84 (3.02), and .86 (3.07) for mathematics; .74 (2.74), .83 (2.38), and .77 (2.37) for reading; and .71 (.2.21), .67 (2.17), and .68 (2.21) for English. The essay subtest score reliability, pertaining to the trained readers' ratings of each of the two essay topics, was measured through inter-rater reliability (IRR) derived from a six-point holistic scoring rubric. For the 3 administrations in 1999-00, the IRR scores for the 2 essay topics were .86, .85, and .86 for topic 1 and .86, .87, and .83 for topic 2 (Florida Department of Education, 2000).

Results

Using a resampling cross-validation technique, the Leave-One-Out (L-O-O) rule or U method (Huberty, 1994; Lachenbruch & Mickey, 1968), the subset of all possible variables were analyzed for the purpose of parsimony, theoretically where "simpler hypotheses are more falsifiable," (Meehl, 1993, p. 5) and to increase the cross-validation accuracy of the proposed model (Lieberman & Morris, 2004; Morris & Meshbane, 1995). Morris and

Meshbane's FORTRAN program (Huberty, 1994, Morris & Meshbane, 1995) for an all subset analysis to yield the best L-O-O hit rate for predictor selection, or $2^p - 1$ where p are the predictors, was conducted.

Of the initial four variables considered, two predictors were deleted that did not contribute to high predictive accuracy (i.e., math and reading). Thus, only writing and essay were retained as components of a parsimonious and more credible model (i.e., in terms of the population). That is, there were 4 predictor variables for the 2-group problem, which meant that there were 15 all possible subset analyses (i.e., $2^4 - 1$). When the number of predictors in the best subset of $2^p - 1$ emerged, the maximum hit rate increased by almost 1.00% to 58.40% from the second best hit rate of 57.47% with 3 predictors (i.e., writing, essay, and math), and, thus, parsimony with increased accuracy was achieved. Other variations within the all possible subset analyses yielded a range of maximum hit rates between 52.80% and 58.40%.

With the L-O-O method, it has been noted that a minimum sample size can be calculated as $N = 3kp$ or a large sample size of $N = 5kp$, where k is the number of groups and p is the number of predictors, and the 3 or 5 derived from the n/p ratio (Huberty, Wisenbaker, & Smith, 1987). The study's sample size of 750 subjects was adequate. Multivariate normality of the data and equality of covariance matrices of the groups were met, with a normal-based rule establishing normality via a review of normal probability plots for data in each of the two groups (Huberty & Lowman, 1998).

A significant degree of discrimination separating the two groups of study was confirmed. As a classification rule, equal prior probabilities external to the sample were established at $.50 (q_1) / .50 (q_2)$, which measured the probability of population membership in either group and equal cost of misclassification for the two populations. The choice of equal priors assumed that the accuracy of this decision was based on estimated priors from the population and not the sample. It has been noted that adjusting for unequal priors based on an estimation from the group size of the sample can be misleading and potentially costly in terms of

decreased model classification accuracy (Meshbane & Morris, 1996).

The GPA for subjects classified as Average ranged between a "C" (i.e., 2.00) and "B-" (i.e., 2.99), and the GPA for subjects classified as Above Average ranged between a "B" (i.e., 3.00) and "A" (i.e., 4.00). The cut point chosen for the two groups was the median GPA for all of the subjects in the study at 3.00. Thus, those subjects with GPAs below this cut point were grouped as Average and coded as a 0, and those with GPAs equal to or above this cut were grouped as Above Average and coded as a 1 (cf. Press & Wilson, 1978).

Table 1. Predictive Discriminant Analysis: Linear External Classification.

Cross-Validation L-O-O

	Average GPA	Above Average GPA	Total
Average GPA	168 (62.92%)	99 (37.08%)	267
Above Average GPA	213 (44.10%)	270 (55.90%)	483

58.40 of cross-validated grouped cases correctly classified.

The results from Table 1 present the L-O-O rule that was established as a bias correction method for classification error rates. L-O-O took 1 subject out of the sample and developed a rule on the other 749 subjects and then took another subject out and developed a rule on the other 749, and so on. This linear, external classification rule was applied to all subjects in the sample so that rules were built on all 750 (Huberty, 1994; Lachenbruch, 1967). From an SPSS (v. 12.0) analysis, table 1 presents the accuracy of the model on cross-validation, meaning how well does this model apply to subjects from the population or its generalizability.

For Average GPA, there were 168 or 62.92% subjects (90% CI for a Binomial Parameter = .578, .678; SE = .03) classified as Average or hits and 99 or 37.08% (CI = .322,

.422) that were predicted as Above Average or misses. For the Above Average GPA group, there were 213 or 44.10% subjects ($CI = .403, .479$; $SE = .02$) misclassified as Average or misses and 270 or 55.90% ($CI = .521, .597$) that were predicted as Above Average or hits. In terms of total precision for all of the subjects, there was 58.40% accuracy ($CI = .554, .614$; $SE = .02$). The model correctly classified a little over half of the cases, with a total group error rate estimate of 41.60% ($CI = .386, .446$).

When assessing each variable's contribution to the discriminant function, the standardized canonical discriminant function coefficients (weights) indicated that writing's relative importance in predicting GPA was .716 followed by essay at .514. Predictor importance was also noted via another method when writing, for example, was taken out of the model, which produced the lowest hit rate for total group accuracy at 52.80% (cf. Huberty & Lowman, 1998). The order of the response variables' contribution toward predictive accuracy indicated how the predictor variables should be arranged. In terms of structure coefficients, the largest absolute correlation associated with the discriminant function was writing at .872, with essay at .731.

In regard to particular cases that may be fence riders, or subjects that were classified correctly, but when their probabilities were reviewed, confidence waned in terms of proper classification, the probability split between highest group and second highest group was established at .52/.48. Of the 750 subjects, 32, or 4.27%, were deemed fence riders. Outliers were determined to be cases that had typicality probabilities less than .10. That is, although a subject was classified correctly with confidence, it appeared to be atypical of that group and hence garnered a low probability. Of the 750 subjects, 35, or 4.67%, were estimated to be outliers. The fence riders and the outliers were kept in the data and analyzed because omitting them may have inflated the hit rate of the model, which potentially could have yielded a model that was more accurate than in actuality.

Using a proportional chance criterion, Huberty's (1994) Z statistic was calculated from a FORTRAN program (J. D. Morris, personal communication, March 13, 2003) to determine if

expected hit rates were exceeded.

$$Z = (o - e) / [e(n - e) / n]^{1/2} \quad (1)$$

o = observed frequency

e = expected frequency

n = number of subjects

This test is a one-tailed test because there is little interest in whether the hit rate was significantly below expectation. The null hypothesis was that the hit rate is what would be expected by chance (e.g., $.50 \times 267 + .50 \times 483 = 375$). The alternative hypothesis was that the present hit rate is better than chance expectation. With an observed hit rate of 438 (i.e., $168 + 270$), the Z of 2.34 ($p < .02$) for the total sample occurred because this hit rate was above expectation, which offers some evidence that the null should be rejected or that classification by the discriminant function resulted in more hits than random assignment by prior probabilities.

However, when the Z value for each group was examined, a different inference emerged. The Z value for the Average GPA group was very large and statistically significant at 9.32 ($p < .001$), but the Z for the Above Average group was .00 and not statistically significant ($p > .05$). The reason this model appeared to be better than chance was that it was quite good at predicting the Average GPA group, but very poor at predicting the Above Average GPA group based on subjects' CLAST scores. That is, the percentage improvement over chance for the Average GPA group was 42.42% and for the Above Average GPA group was -23.87%. The percentage of improvement over chance for the total sample was only 9.27%. Thus, the classification of the two groups was only slightly better, by 9%, than would have been accomplished by chance.

To add to this argument from a different perspective, and also to address the issue of the intermediate inequality of group sizes, the model was looked at via a maximum chance criterion ($\max(q_1, q_2)$) (Huberty, 1994). The maximum chance criterion assigned all of the subjects to the largest group for this study, the Average GPA group, as a criterion for a hit rate better than chance. The Z value was -.08, which meant that the model's hit rate was not better than chance. Further, the percent improvement over

chance for the total sample was -16.85%. Thus, this model did not have good accuracy for the two-group classification problem using either of the chance criteria proposed.

Huberty's (1994) effect size measure, the I statistic, was calculated to determine the

$$I = (1-e) - (1-o) / 1-e \quad (2)$$

$$= o-e / 1-e$$

percentage correctly classified exceeding chance. The Average group had an $I = .258$, the Above Average group had an $I = .118$, and the total model had an $I = .168$. Previous research (Huberty & Lowman, 2000) conducted on I indicated that these values should be regarded as having a low effect, except for the medium effect of the Average group, in terms of their ability to measure proportional reduction in error, meaning, for instance, that the total model had roughly 16% less misclassifications than would have occurred if just classified by chance.

Conclusion

The addition of many more exemptions on the CLAST created a problem where it was supposed that students from various colleges within a university could have opted out of the test, leaving the study with a more homogeneous sample (i.e., participants from only a few colleges who did not have as many exemption options).

For future institutional decisions related to students' academic success, the PDA model chosen for this study, which was parsimonious and contained two estimators of the CLAST subtest scores to classify students into one of two alternative populations consisting of Average GPA or Above Average GPA, was not accurate enough across all groups, or for each group, and its total sample hit rate was only slightly better than chance. Overall, the CLAST subtest scores did not estimate effectively academic success in terms of predicting GPA. In fact, the predictors' relative contribution ranged within a moderate level of ordered importance from writing (.716) to essay (.514), both of which were also rank-ordered as first and second most important using a variable deletion method, with 2 unimportant variables (i.e., math and

reading) removed because classification accuracy did not diminish without their presence in the model. Thus, CLAST score use by institutions as a general measure of educational accountability, specifically in the instance as a mode to estimate high academic success, does not appear to be an effective model.

Study 2.

The New Teacher Academy (NTA) was created as a link to Florida's A+ Plan for K-12 public schools in Broward County, which during academic year 2001-02 enrolled 260,892 students (Broward County Public Schools, 2003) making it one of the 10 largest school districts in the United States. Specifically, the NTA was initiated to assist new teachers in Broward County Public Schools with bolstering their performance levels in the classroom as a measure of accountability, but also as a means of professional development in the sense of sustained, active development (Fullan, 2000; Garet, Porter, Desimone, Birman, & Yoon, 2001).

Further, to address the challenge of hiring more non-education major teachers to educate the increasing student enrollments within Broward's K-12 system, NTA was contrived to support these new teachers' development and overall preparedness in the classroom. In this manner, the NTA could be thought of as an approach for early professional development, but also as an agent for "teacher change" (Clarke & Hollingsworth, 2002).

A cross-functional planning committee, along with survey responses from educators in various capacities throughout the Broward County Public School system, assisted in identifying critical domains that all new teachers should know and be able to practice in the classroom to promote achievement levels as outlined in Florida's A+ Plan. Ten major domains were recognized. Of those 10 domains, two were rated as high priority and dealt approximately with the following areas: instruction (Bandura, 1997; Fullan, 1991; Putnam & Borko, 1997) and classroom-based competencies (Newmann, King, & Youngs, 2000; Wenglinsky, 2002; Zeichner, 1993). These two domains were the principal emphasis of the NTA.

Methodology

There were two research questions that this study intended to answer about the instrument so that results may begin to assist in defining it for future generalizations back to the K-12 and college and university teacher training populations:

- 1) What is the dimensionality structure of the instrument?
- 2) Does the instrument develop constructs that will measure the theorized domains?

Using a four-point Likert-type scale, where 1 = Not Adequately Prepared; 2 = Somewhat Prepared; 3 = Prepared; and 4 = Very Prepared, the instrument consisted of 16 items, which asked respondents to indicate how prepared they felt to perform various classroom instructional and management tasks (Appendix A).

Reliability

Using the software Analysis of Moment Structures (AMOS) version 4.01 (Arbuckle, 1999), a model was created to obtain measurement reliability estimates based on 2000 bootstrapped samples. The reliability estimates for the instrument's scores were very high, which meant that this instrument had internal consistency and the items on the instrument shared a large percentage of the variance. For the NTA group, the estimated reliability coefficient = .920 with bootstrapped 90% lower and upper confidence limits of (.895, .937). For the Non-NTA group, the estimated reliability coefficient = .922 with bootstrapped 90% lower and upper confidence limits of (.878, .947). The small width found in both bootstrapped confidence limits indicates that there was stability in the sample measurement reliabilities and, thus, estimates based on these samples had a high probability of stability upon replication.

As a medium to allow others to implement further testing of the instrument, or produce competing models, means and standard deviations are provided pertaining to the participants' responses to the 16 items in Table 2. Pearson correlations of the 16 items are presented in Table 3. Because of the number of statistical tests performed, a Bonferroni

correction of $\alpha = .001$ was utilized to ensure that the possibility of false rejections was not too great.

Table 2. Descriptive Statistics of Participants' Responses to Questions.

Item	M	SD
1	2.53	.72
2	2.56	.84
3	2.51	.83
4	2.66	.96
5	2.39	.90
6	2.67	.88
7	2.65	.80
8	2.57	.72
9	2.58	.91
10	2.67	.70
11	2.77	.92
12	2.56	.85
13	2.68	.75
14	2.80	.65
15	2.49	.87
16	3.19	.67

The scale needed to be validated to determine if it measured the two domains and if these domains held together. Factor analysis reduces the number of original variables, 16 in this case, into a smaller set of factors to obtain parsimonious dimensionality. Thus, there will be an attempt to capture as much of the variation among the 16 variables as possible with the least amount of dimensions. However, there is a cost and benefit situation to consider. How much loss in precision of the original variables will be tolerated (i.e., the cost) for the benefit of attaining a more parsimonious solution? It was felt that a two dimensional structure would exhibit the nature of the 16 variables, and also that the variance of each variable would be captured sufficiently by the factor structure. That is, all individual variables would be well represented.

A confirmatory factor analysis (CFA), using the extraction method of maximum likelihood with oblimin rotation, was conducted to look at the total variance explained by the model.

Table 3. Correlation Matrix.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	---	.49	.60	.49	.28	.52	.52	.59	.49	.68	.40	.49	.47	.53	.43	.28
2	.49	---	.70	.52	.34	.41	.43	.34	.29	.48	.31	.37	.37	.51	.50	.19
3	.60	.70	---	.65	.46	.37	.46	.43	.35	.52	.50	.38	.46	.47	.56	.23
4	.49	.52	.65	---	.62	.21	.39	.23	.23	.44	.38	.22	.34	.33	.42	.18
5	.28	.34	.46	.62	---	.09	.17	.14	.17	.34	.45	.18	.26	.24	.26	.13
6	.52	.41	.37	.21	.09	---	.57	.59	.54	.60	.22	.66	.53	.56	.42	.41
7	.52	.43	.46	.39	.17	.57	---	.48	.47	.50	.39	.55	.59	.41	.46	.19
8	.59	.34	.43	.23	.14	.59	.48	---	.48	.56	.39	.49	.44	.53	.38	.33
9	.49	.29	.35	.23	.17	.54	.47	.48	---	.56	.42	.76	.67	.56	.37	.22
10	.68	.48	.52	.44	.34	.60	.50	.56	.56	---	.51	.54	.57	.65	.52	.38
11	.40	.31	.50	.38	.45	.22	.39	.39	.42	.51	---	.35	.49	.39	.42	.22
12	.49	.37	.38	.22	.18	.66	.55	.49	.76	.54	.35	---	.70	.52	.41	.30
13	.47	.37	.46	.34	.26	.53	.59	.44	.67	.57	.49	.70	---	.53	.54	.25
14	.53	.51	.47	.33	.24	.56	.41	.53	.56	.65	.39	.52	.53	---	.46	.30
15	.43	.50	.56	.42	.26	.42	.46	.38	.37	.52	.42	.41	.54	.46	---	.43
16	.28	.19	.23	.18	.13	.41	.19	.33	.22	.38	.22	.30	.25	.30	.43	---

Because the scores from the items on the instrument were correlated moderately, it was theorized that the underlying factors for these items were correlated as well. Therefore, oblimin rotation was used, which permits the factors to be correlated and adds to the simplicity and the generalizability of results (Pedhazur & Schmelkin, 1991).

The extent of the correlation between factors was predetermined at $\geq .350$ based on the researcher's prediction that the degree of correlation would remain in the moderate to high range. Although high correlations of the ilk $\geq$

.700 are preferred, the scholarly literature has indicated that loadings between .300 and .500 are often the norm (cf. Biggs, Kember, & Leung, 2001). The variables were reasonably multivariate normal. To determine if it was appropriate to proceed with a factor analysis, an examination of the correlation matrix established that the variables of study were sufficiently related to one another, to a degree significantly different than the identity matrix (Bartlett's Test of Sphericity $\chi^2 = 901.347(120)$; $p < .001$).

In terms of the goodness-of-fit of the model to the sample data, large values of chi

square (χ^2) mean that the model is a bad fit for the data and small values signify that the data is a good fit. The study's sample size of $n = 105$ appears to be ample enough in terms of adhering to the principle of having "... the minimum number of subjects required is 5-10 times the number of observed indicators" (Bryant & Yarnold, 1995, p. 117).

Taking sample size into account, the use of only the χ^2 statistic as a measure of fit may render uncertainty concerning the overall appropriateness of the study's model. Thus, a χ^2 change test was conducted, which compared the values for χ^2 from a one-factor solution, a two-factor solution, and a three-factor solution. Further, the ratio of χ^2 to degrees of freedom (χ^2/df ratio) was used to compare the relative fit of the three models. As the χ^2/df ratio decreases, the fit of a model is improved (Hoelter, 1983).

The one-factor solution had $\chi^2 = 261.160(104)$; $p < .001$; 2.51 χ^2/df ratio, the two-factor had $\chi^2 = 140.558(89)$; $p < .001$; 1.58 χ^2/df ratio, or a χ^2 change of 120.602, and the three-factor had $\chi^2 = 98.114(75)$; $p < .05$; 1.31 χ^2/df ratio, or a χ^2 change of 42.444. The highly statistically significant change test for the two-factor solution indicated that it fit the data better than a one-factor or three-factor solution, where the latter factor solution did not indicate a more significant change by adding a third factor to the model. Also, the χ^2/df ratio was very similar between the two-factor (1.58) and the three-factor (1.31) models. The two-factor model was preferred because of its more simple nature and the fact that the three-factor, more complex model did not appear to offer much more substantial data about model fit.

As advocated by Mulaik et al. (1989) and Tanaka (1993), various indicators of fit were utilized, beyond the χ^2 criterion of fit or no fit, to examine the multiple aspects that may encompass a model and also to determine how closely the model fits the data. Arbuckle's (1999) software AMOS was used to specify the model. As relative fit measures, the incremental fit index (IFI = .977), the comparative fit index (CFI = .977), and the Tucker-Lewis index (TLI = .969) all indicated that the proposed model compared very well to, and exceeded, a null model per the cut point fixed at $\geq .95$ (Hu &

Bentler, 1999), which was established due to lower magnitudes of a few of the factor loadings. For all fit indices, a rigid cut point was necessary to yield a rejection rate for the few instances where there were low loading circumstances.

For indices based on χ^2 , or an absolute fit measure, the root mean square error of approximation (RMSEA) ranges from 0 to 1, with scores of .05, .08, and, .10 representing the magnitude of population misfit (Browne & Cudeck, 1993). This index can also serve as a noncentrality-based fit index. For this model, the RMSEA = .104, meaning that this model was a fairly good estimation of misfit to the population correlation matrix, but did have some error. The expected cross-validation index (ECVI) was 3.049 (90% CI 2.680, 3.492), which is an approximated measure of the goodness-of-fit that the present model would attain in an additional sample of the same size.

To determine how many factors to retain, multiple decision rules were used (Thompson & Daniel, 1996). The traditional eigenvalue greater than 1.00 rule (K1) was analyzed as the lower boundary for the number of factors to be retained (Guttman, 1954; Kaiser, 1960). However, this method of extraction has been noted to both overestimate (Hakstian, Rogers, & Cattell, 1982; Zwick & Velicer, 1986) and underestimate (Cattell & Vogelmann, 1977; Hakstian et al., 1982) the number of factors retained and yield false support for classifying scales as multidimensional (Bernstein & Teng, 1989).

A second method was used with a scree plot (Cattell, 1966). In this technique, the total factors retained were based on the number of eigenvalues that fell before the last major drop on the scree plot. This method potentially could lend itself to subjectivity and poor decisions in terms of the number of factors to retain due to its variability of results and, thus, reliability (Zwick & Velicer, 1986). Yet, results indicated that the scree test produced limited accuracy (Zwick & Velicer, 1982; 1986).

In a third method, a parallel analysis (PA) was run on the data and factors were retained based on a comparison between the scree plot from the random data generated via the PA and the scree plot from the actual data.

Factors from the actual data that had eigenvalues greater than the eigenvalues produced from the PA were extracted because they exceeded chance levels of the eigenvalues from the PA and, thus, indicated that they were “authentic” factors (Horn, 1965; Thompson & Daniel, 1996). This technique has yielded fairly accurate results (Zwick & Velicer, 1986). Finally, Velicer’s (1976) Minimum Average Partial (MAP) method was utilized. Using a matrix of partial correlations from the study, the average of the partial squared correlation was determined. When the smallest average squared correlation was attained, no more factors were removed. This extraction method has been found to be very accurate, especially when compared against the traditional K1 rule (Zwick & Velicer, 1982).

Based on the implementation of multiple decision rules and splitting the data in half to determine if the number of factors extracted replicated on all of the multiple decision rules applied, it was determined that two factors should be extracted for the model. The variable (p) to factor (m) ratio was 8:1, where the number of variables was a constant at 16 and the number of factors extracted was 2. This p: m ratio has been cited as reasonable for practical usage (Zwick & Velicer, 1986). The variance of the first factor was = 7.531 and the second factor = 1.789. The two eigenvalues had a cumulative percentage = 58.247. They accounted for 58% of the variation among the 16 variables. The correlation between factor 1 and factor 2 was .526.

To name these two factors, the solution was rotated to simulate a simple structure via oblimin rotation. This will yield the relative contribution of each variable to a factor by correlating variables to factors. The pattern coefficients are standardized regression weights that account for the correlation among the two factors and the structure coefficients are bivariate correlations between the two factors and the 16 variables.

Examination of both the pattern (p) and structure (s) coefficients rendered like interpretations of the factor structure. In terms of convergent validity, how a factor primarily influenced a variable was established as both $p \geq .700$ and $s \geq .700$, while a more moderate

extent influence was established as both p and s between .350 and .699. Factor 1 appeared to influence principally X_6 ($p_6 = .807$; $s_6 = .762$), X_9 ($p_9 = .891$; $s_9 = .800$), X_{12} ($p_{12} = .933$; $s_{12} = .839$), and X_{13} ($p_{13} = .757$; $s_{13} = .780$). It influenced to a moderate degree X_1 ($p_1 = .485$; $s_1 = .679$), X_7 ($p_7 = .586$; $s_7 = .671$), X_8 ($p_8 = .630$; $s_8 = .667$), X_{10} ($p_{10} = .615$; $s_{10} = .758$), X_{14} ($p_{14} = .614$; $s_{14} = .706$), X_{16} ($p_{16} = .459$; $s_{16} = .448$), and X_{15} ($p_{15} = .375$; $s_{15} = .579$). Factor 1 had a lesser influence on X_{11} ($p_{11} = .302$; $s_{11} = .503$). Both X_{11} and X_{15} were shared with Factor 2.

Due to this result, Factor 1 should be named Classroom and Behavior Management. This incorporated in-class activities, which addressed issues that impacted both learning and instruction such as motivating students to behave, implementing techniques to accommodate various learning styles, and promoting an effective learning environment. This combination of subject matter and pedagogical knowledge has been found to enable teachers to understand and explain content-related tasks and concepts connected to student learning (Beijaard, 1995; Bennett & Carre, 1993).

Factor 2 seemed to influence primarily X_3 ($p_3 = .768$; $s_3 = .849$) and X_4 ($p_4 = .814$; $s_4 = .785$). To a moderate degree, it influenced X_2 ($p_2 = .617$; $s_2 = .714$), X_5 ($p_5 = .646$; $s_5 = .600$), X_{11} ($p_{11} = .384$; $s_{11} = .542$), and X_{15} ($p_{15} = .389$; $s_{15} = .586$), with both of the latter two variables shared with Factor 1. Factor 2 should be named Instructional Knowledge and Skills, which looked at questions that measured if teachers thought they were prepared to teach students the content standards deemed important toward achieving grade level proficiency. Teacher preparedness in terms of content knowledge has been found to inform classroom learning, which affects instructional decisions (Swafford, Chapman, Rhodes, & Kallis, 1996).

In general, there was a rotation that separated the variables in a manner in which highly correlated variables had sufficient factor pattern coefficients on one factor and very little on a second factor, or discriminant validity was established. In fact, only two variables, X_{11} and X_{15} , had factor pattern coefficients split on more than one factor.

This is important for future theoretical use and measurement of the scale, where dimension one separated classroom and behavior management items from dimension two related to instructional knowledge items.

The two dimensional structure appeared to capture the 16 variables. Now, however, were there individual variables that were not well represented in the structure? Communalities (h^2) are the proportion of each variable explained by the factor structure (i.e., akin to R^2). Extraction communalities ranged from .201 to .721. For example, X_3 had the highest $h^2 = .721$. This is the percentage of variation of this variable that is accounted for by the factor solution. X_{16} had the lowest communality at .201. If the cut point of $h^2 \geq .350$ is used, which was previously implemented in the study, to examine these communalities, all of the variables, with the exception of X_{16} , were accounted for noticeably by the factor solution.

Looking at X_{16} , its unique variance was .799 (i.e., $1 - .201$), which indicated that 80% of this variable's variance was unexplained by factor one. However, this variable's pattern and structure coefficients were acceptable, signifying that X_{16} 's factor had a moderate influence on it, but was less sufficient at predicting the amount of variance pertaining to the variable. Yet, given the high score reliability of the instrument at .920 and .922 for both groups, there appears to be little error and, therefore, the large unique variance for X_{16} should not be attributed extensively to measurement error.

Conclusion

In terms of the sample size, and admittedly a border-line size, there were a number of techniques previously-mentioned (e.g., χ^2 change test, various indicators of fit were utilized beyond the χ^2 criterion, splitting the data in half to determine if the number of factors extracted replicated on all of the multiple decision rules applied, etc.) throughout the study to monitor size to establish if it had a substantial influence on the results. It appeared that this study's sample size was within a suitable range of the number of subjects per observed indicators.

The findings of this research suggest that the NTA scale was measured as a multidimensional instrument with two distinct factors. This implied that one factor was not adequate for the entire instrument. A CFA corroborated that the instrument had construct validity by providing evidence that these two domains held together and had a set of 16 items that were relatively homogeneous. These findings assisted in answering the study's two research questions: what is the dimensionality structure of the instrument and does the instrument develop constructs that will measure the theorized domains? The preliminary findings connected to these questions are salient because they suggest that this instrument has an adept developmental foundation both in terms of measurement and substance. To be sure, more validation of scores needs to be secured across many implementations of this instrument, but early development appears promising.

Figure 1: Two-Dimensional Model

References

- Arbuckle, J. L. (1999). *AMOS* (Version 4.01). Chicago: SmallWaters Corporation.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Beijaard, D. (1995). Teachers' prior experiences and actual perceptions of professional identity. *Teachers and Teaching: Theory and Practice*, 1, 281-294.
- Bennett, N., & Carre, J. C. (1993). *Learning to teach*. London: Routledge.
- Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, 105, 467-477.
- Biggs, J., Kember, D., & Leung, D. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, 71, 133-149.
- Broward County Public Schools. (2003). *Interesting factoids*. Retrieved January 7, 2003, from <http://www.browardschools.com/about/factoids.htm>
- Brown, B., & Campion, M. A. (1994). Biodata phenomenology: Recruiters' perceptions and use of biographical information in resume screening. *Journal of Applied Psychology*, 79, 897-908.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models*, p. 136-162. Thousand Oaks, CA: Sage.
- Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm, & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics*, p. 99-136. Washington, D.C.: American Psychological Association.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, 12, 289-325.

- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18, 947-967.
- Florida Atlantic University. (2002). *Questions and answers about CLAST*. http://www.fau.edu/student/testing/clast_page2.html
- Florida Department of Education. (2000). *CLAST technical report: 1999-2000*. Tallahassee, FL: State of Florida Department of State.
- Florida International University. (2002). *CLAST*. Retrieved November 24, 2002, from <http://www.fiu.edu/~testing/clast.htm>
- Fullan, M. G. (1991). *The new meaning of educational change*. New York: Teachers College Press.
- Fullan, M. G. (2000). The return of large-scale reform. *Journal of Educational Change*, 1, 5-28.
- Garet, M. S., Porter, A. G., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, 38(4), 915-945.
- Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, 19, 149-161.
- Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, 17, 193-219.
- Hoelter, J. W. (1983). The analysis of covariance structures: Goodness of fit indices. *Sociological Methods and Research*, 11, 325-344.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.
- Huberty, C. J. (1994). *Applied discriminant analysis*. New York: John Wiley & Sons, Inc.
- Huberty, C. J., & Lowman, L. L. (1998). *Discriminant analysis in higher education research*. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 181-234). New York: Agathon Press.
- Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, 60, 543-563.
- Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. *Multivariate Behavioral Research*, 22, 307-329.
- Indian River Community College. (2002). *CLAST*. Retrieved November 24, 2002, from <http://www.ircc.cc.fl.us/educserv/admis/tests/clast.html>
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 23, 639-645.
- Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11.
- Lieberman, M. G., & Morris, J. D. (2004, April). *Selecting predictor variables in logistic regression*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Meehl, P. E. (1993). Four queries about factor reality. *History and Philosophy of Psychology Bulletin*, 5, 4-5.
- Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement*, 55, 438-441.

- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.
- Newmann, F. M., King, M. B., & Youngs, P. (2000, April). *Professional development that addresses school capacity: Lessons from urban elementary schools*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73, 699-705.
- Putnam, R., & Borko, H. (1997). Teacher learning: Implications of new views of cognition. In B. J. Biddle, T. L. Good, & I. F. Goodson (Eds.), *The international handbook of teachers and teaching*, p. 1223-1296. Dordrecht, The Netherlands: Kluwer.
- Roth, P. L., & Bobko, P. (2000). College grade point average as a personnel selection device: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology*, 85, 399-406.
- Swafford, J., Chapman, V., Rhodes, R., & Kullis, M. (1996). A literate analysis of trends in literacy education. In D. J. Leu, C. K. Kinzer, & K. A. Hinchman (Eds.), *Literacies for the 21st century: Research and practice*, p. 437-446. Chicago: National Reading Conference.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In J. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*, p. 10-39. Newbury Park, CA: Sage.
- Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, 56, 197-208.
- University of South Florida. (2002). *CLAST*. Retrieved November 24, 2002, from http://usfweb.usf.edu/ugrads/eandt/clast_state_document.htm
- Velicer, W. F. (1976). The relation between factor score estimates, image scores and principal component scores. *Educational and Psychological Measurement*, 36, 149-159.
- Wenglinsky, H. (2002, February 13). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, 10(2). <http://epaa.asu.edu/epaa/v10n12>
- Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement*, 55, 177-185.
- Zeichner, K. M. (1993). Traditions of practice in U.S. preservice teacher education programs. *Teaching and Teacher Education*, 9, 1-13.
- Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, 17, 253-269.
- Zwick, W. R., & Velicer, W. F. (1986). Factors influencing five rules for determining the number of components to retain. *Psychological Bulletin*, 99, 432-442.

Appendix A

- X*₁: Identify how individual differences and learning styles affect instructional delivery.
- X*₂: Recognize Grade Level Expectations (GLE).
- X*₃: Recognize Critical Content (CC).
- X*₄: Recognize Sunshine State Standards (SSS).
- X*₅: Recognize the Florida Comprehensive Test (FCAT).
- X*₆: Develop strategies to motivate students to learn.
- X*₇: Advance the delivery of instruction through effective organization and time management skills.
- X*₈: Identify effective teaching behaviors.
- X*₉: Develop strategies to diminish misbehavior.
- X*₁₀: Identify individual differences and learning styles.
- X*₁₁: Develop effective record keeping routines.
- X*₁₂: Acquire strategies to motivate students to behave.
- X*₁₃: Promote positive classroom behavior through effective organization and time management skills.
- X*₁₄: Demonstrate teaching and learning behaviors that promote an effective learning environment.
- X*₁₅: Develop goals that are realistic and achievable for your Professional Growth Plan (PGP).
- X*₁₆: *Work* cooperatively with students, colleagues, administrators, and parents.

A Visually Adaptive Bayesian Model In Wavelet Regression

Dongfeng Wu

Department of Mathematics and Statistics
Mississippi State University

The implementation of a Bayesian approach to wavelet regression that corresponds to the human visual system is examined. Most existing research in this area assumes non-informative priors, that is, a prior with mean zero. A new way is offered to implement prior information that mimics a visual inspection of noisy data, to obtain a first impression about the shape of the function that results in a prior with non-zero mean. This visually adaptive Bayesian (VAB) prior has a simple structure, intuitive interpretation, and is easy to implement. Skorohod topology is suggested as a more appropriate measure in signal recovering than the commonly used mean-squared error.

Key words: Wavelet regression, wavelet shrinkage, optimal, Skorohod topology, uniform distance, mean-squared error

---

Introduction

Wavelets unify many ideas from the fields of applied mathematics, signal processing, and physics (see Daubechies 1992). Wavelets are families of basis functions that can be used to approximate other functions, with powerful properties such as orthonormality, compact support, localization in time and scale, etc. Daubechies (1988) and Mallat (1989) encouraged the use of wavelets in the mathematical sciences, while Donoho and Johnstone (1994, 1995) popularized wavelets in the statistics community.

Some of the uses of wavelets for statistical problems have been developed by Donoho and Johnstone (1993, 1994) and Nason (1994) and are available in the S+ package. More recent work includes the block thresholding method of Cai (1999), which achieves adaptivity, spatial adaptivity and computational efficiency simultaneously.

When fitting wavelet-based models, shrinkage of the empirical wavelet coefficients is an effective tool for denoising the data. Shrinkage of the empirical wavelet coefficients works best in problems where the underlying set of the true coefficients of f is sparse. One natural way to obtain the shrinkage estimates of the true coefficients is via Bayesian methods.

An appealing and simple model (ABWS) using the posterior mean has been proposed by Chipman, Kolaczyk, and McCulloch (1997) who assume that an accurate estimate of the noise level σ is available. A more complete Bayesian approach that captures the uncertainty about the noise level σ was proposed by Clyde, Parmigiani, and Vidakovic (1998). Abramovich, Sapatinas and Silverman (1998) proposed the posterior median method, with almost the same set up as Clyde et.al., but using the posterior medians to estimate the true coefficients. Huang and Cressie (1999) proposed a normal prior with non-zero means for wavelet coefficients, and estimated the hyper-parameters of the prior covariance by a pseudo maximum likelihood method.

A different prior structure with non-zero means is offered. The model is simple, combining a normal prior with non-zero mean and a point mass. Explanations are provided for each hyper-parameter in addition to a specific way to choose the prior parameters.

The author wishes to thank David V. Hinkley for suggestions regarding this study. Most of the numerical work was conducted using *wavethresh* (Nason, 1994, Version 3). Contact the author at dw183@ra.msstate.edu

Methodology

The Bayesian model

Suppose the function f is sampled at $n = 2^J$ equally spaced points, but is observed with additive white noise,

$$y_i = f(i/n) + \sigma z_i, \quad i=0, 1, \dots, n-1, \quad (1)$$

where z_i , $i = 0, 1, \dots, n-1$, are iid standard normal random variables, and σ is unknown. Equivalently this observation model can be expressed in wavelet regression form,

$$v_{j,k} = w_{j,k} + \sigma z_{j,k}, \quad j = 0, \dots, J-1, \quad k = 0, \dots, 2^j, \quad (2)$$

where $v_{j,k}$ s are the discrete wavelet coefficients of noisy observation y ; $w_{j,k}$ s are the discrete wavelet coefficients of f ; and $z_{j,k}$ s are still iid $N(0,1)$ random variables.

In the Bayesian approach, a prior distribution is placed on the coefficients, and some particular prior distributions that are designed to capture the sparseness common to most wavelet applications are proposed. Most of the published works in this area have a common characteristic, that is, a prior distribution is designed such that some of the mass is concentrated on values close to zero or just being zero, while the rest of the mass is spread to accommodate the possibility of large coefficients.

Then, the posterior means or the posterior medians are used as the estimates of the true coefficients. Though appealing, this framework assumes that all of the coefficients have the same prior in each level, with zero mean, which overlooks the facts that certain coefficients are significantly departs from zero. The overall shape of the curve gives us more useful information, and accommodation of this information will ease the procedure to denoise, and hence, recover the curve.

Inspired by the work of Chipman et al. (1997), Clyde et al. (1998), and Abramovich et al. (1998), and assuming that a good estimate of the obtained noise level σ , the following prior model is proposed:

$$w_{j,k} | \gamma_{j,k} \sim \gamma_{j,k} N(a_{j,k}, \tau_j^2) + (1 - \gamma_{j,k}) \delta(0) \quad (3)$$

In this prior model, the coefficients are mutually independent, and modeled as a mixture of a normal distribution and a point mass at zero. The innovation is that assumed is that the normal prior has non-zero mean $a_{j,k}$ for each coefficient $w_{j,k}$. Also, a really small variance τ_j depends on each level j , so that each coefficient has a different prior associated with it.

This idea comes from the observation that when coefficients are changed in a small scale in each level, the function estimate won't change much, and it won't affect our visual perspective either. This means that each coefficient can change around its true value in a small scale, called its safety range, without any deleterious effects. This is captured in the form of $N(a_{j,k}, \tau_j^2)$, where $a_{j,k}$ is the prior information on the true value of the coefficient, and τ_j is the allowable perturbation on level j , so that the estimate would be close to the true function.

A point mass at zero is assumed based on the belief that the coefficients are sparse. This simple form of prior modeling has intuitive interpretations and captures the few big spikes in the coefficients. Empirical evidence shows that if $a_{j,k} = w_{j,k}$, $\forall j = 0, \dots, J-1, k = 0, \dots, 2^j$, the "recovered estimate" $\tilde{f}$ is a slight shift from the true f .

The mixture parameter $\gamma_{j,k}$ has its own prior distribution given by

$$\gamma_{j,k} \sim \text{Bernoulli}(p_{j,k}), \quad (4)$$

The prior parameters $a_{j,k}$, τ_j , $p_{j,k}$ need to be decided. A different prior is assigned for each individual coefficient, though in each level the coefficients share a common prior variance τ_j , which reflects the perturbation in level j .

Once data are observed, the wavelet coefficients of the signal y are distributed as

$$v_{j,k} | w_{j,k}, \sigma^2 \sim N(w_{j,k}, \sigma^2). \quad (5)$$

The posterior distribution on the (unobserved) true value of $w_{j,k}$, and use its expected value as the estimate. Then the inverse wavelet is applied transformation to get $\hat{f}$.

The Prior Parameters

In this section details are given on how to choose the values for each of the prior parameters. This prior seems more intuitive, and computer simulation demonstrates that it works well.

The intuitive meaning of $a_{j,k}$ is the prior mean of each coefficient. The value of a specific coefficient is not necessarily zero, but is determined by the overall shape of the signal; in other words, it is related to the first impression of the data. The *Universal* thresholding method is used to get the value $a_{j,k}$ for each coefficient. The *Universal* threshold value is generally bigger than all the other methods, and gives the overall shape of the data. Suppose a sound estimate exists of σ , say $\hat{\sigma}$, then for each level $j = 1, \dots, J$, let $t_j = \hat{\sigma} \sqrt{2 \log(2^j)}$ according to the *Universal* rule, then

$$\begin{aligned} a_{j,k} &= T_{\text{soft}}(v_{j,k}, t_j) \\ &= \text{sgn}(v_{j,k}) (|v_{j,k}| - t_j) I(|v_{j,k}| > t_j) \end{aligned} \quad (6)$$

This process mimics a visual inspection of the noisy data whereby the first impression about the shape of the function is obtained. Using the threshold value as the empirical prior information of $a_{j,k}$ makes sense. Because this estimate is close to the true curve, only small perturbations are allowed, so the τ_j will be a small number compares to the scale in the same level j . It is believed that this τ_j is largely connected with the scales of the coefficients in the same level. Chosen was $\tau_j = 10\% M_j$ based on previous empirical experience, where $M_j = \max_{0 \leq k \leq 2^j - 1} \{|v_{j,k}|\}$.

Usually, for a smaller signal-to-noise ratio, a bigger percentage is chosen to obtain τ_j ; and a bigger signal-to-noise ratio means a smaller percentage to obtain τ_j . As for $p_{j,k}$, the probability that one specific coefficient is non-zero, also depends on the scales of the

coefficients in that level. If $v_{j,k}$ is comparatively large, it is more likely that $w_{j,k} \neq 0$, and choose was $p_{j,k} = |a_{j,k} / M_j|$, which is the ratio of the absolute value of that coefficient over the largest one in that level. Now, the prior parameters for each coefficient are given.

In practice, the noise level σ is unknown and must be replaced by an estimate $\hat{\sigma}$. Used here is the slope estimate in Wu (2002), defined by

$$\hat{\sigma} = \frac{v_{(0.75n)} - v_{(0.25n)}}{z_{0.75} - z_{0.25}} \approx \frac{IQR}{0.6745 * 2}, \quad (7)$$

where $v_{(k)}$ s are the order statistics of the highest level wavelet coefficients, $z_{0.75}$ and $z_{0.25}$ are the quantiles of the standard Normal distribution; n is the total number of coefficients in the highest level $J-1$, IQR is the inter-quartile range of the observed coefficients. Simulation studies show that this estimation is accurate in the applications (Wu, 2002).

Posterior Distribution of the Coefficients

Based on this model, it is derived that the posterior mean and variance of $w_{j,k}$ given the observation of noisy data Y , where $w_{j,k}$, $v_{j,k}$, $\gamma_{j,k}$, $a_{j,k}$, $p_{j,k}$, τ_j are simplified as w , v , γ , a , p , τ .

$$\begin{aligned} &E(w | v) \gamma E_{\gamma|v}[E(w | v, \gamma)] \\ &= P(\gamma = 1 | v) E(w | v, \gamma = 1) \\ &+ P(\gamma = 0 | v) E(w | v, \gamma = 0) \\ &= P(\gamma = 1 | v) E(w | v, \gamma = 1). \end{aligned} \quad (8)$$

Because

$$w | v, \gamma = 1 \sim N\left(\frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right), \quad (9)$$

this implies

$$E(w | v, \gamma = 1) = \frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}. \quad (10)$$

Because

$$P(\gamma=1|v) = \frac{p\pi(v|\gamma=1)}{p\pi(v|\gamma=1)+(1-p)\pi(v|\gamma=0)} \equiv \frac{O}{O+1}, \quad (11)$$

where

$$O = p\pi(v|\gamma=1)/[(1-p)\pi(v|\gamma=0)], \quad (12)$$

and because

$$\pi(v|\gamma=1) \sim N(a, \sigma^2 + \tau^2), \quad (13)$$

$$\pi(v|\gamma=0) \sim N(0, \sigma^2), \quad (14)$$

when plugged into (12), the following

$$O = \frac{p}{1-p} \sqrt{\frac{\sigma^2}{\tau^2 + \sigma^2}} \exp\left\{\frac{v^2}{2\sigma^2} - \frac{(v-a)^2}{2(\sigma^2 + \tau^2)}\right\}, \quad (15)$$

and

$$E(w|v) = \frac{O}{O+1} \frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}. \quad (16)$$

This is, the posterior mean of the coefficient. Then, apply the inverse wavelet transformation to obtain the function.

The posterior variance of a coefficient can be calculated similarly,

$$\begin{aligned} \text{var}(w|v) &= E_{\gamma|v}[\text{var}(w|v, \gamma)] + \text{var}_{\gamma|v}[E(w|v, \gamma)] \\ &= E_{\gamma|v}[\text{var}(w|v, \gamma)] + E_{\gamma|v}[E(w|v, \gamma)^2] \\ &\quad - [E_{\gamma|v}E(w|v, \gamma)]^2 \end{aligned} \quad (17)$$

where

$$\begin{aligned} E_{\gamma|v}[\text{var}(w|v, \gamma)] &= P(\gamma=1|v) \text{var}(w|v, \gamma=1) \\ &= \frac{O}{O+1} \cdot \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}, \end{aligned} \quad (18)$$

$$\begin{aligned} E_{\gamma|v}[E(w|v, \gamma)^2] &= P(\gamma=1|v)E(w|v, \gamma=1)^2 \\ &= \frac{O}{O+1} \cdot \left(\frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}\right)^2, \end{aligned} \quad (19)$$

and

$$[E_{\gamma|v}E(w|v, \gamma)]^2 = [E(w|v)]^2 = \left(\frac{O}{O+1} \cdot \frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}\right)^2. \quad (20)$$

Hence,

$$\text{var}(w|v) = \frac{O}{O+1} \cdot \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} + \frac{O}{(O+1)^2} \cdot \left(\frac{a\sigma^2 + v\tau^2}{\sigma^2 + \tau^2}\right)^2. \quad (21)$$

Results

Presented are some simulation results of different shrinkage methods. For estimation of f , the usual L_2 norm is used to evaluate performance. Let $f = \{f(x_i)\}_{i=1}^n$ and $\hat{f} = \{\hat{f}(x_i)\}_{i=1}^n$ be the vectors of true and estimated function values where x_i are equally sampled. Performance is measured by the average mean-squared error

$$R(\hat{f} - f) = \frac{1}{n} \|\hat{f} - f\|_{2,n}^2 = \frac{1}{n} \sum_{i=1}^n [\hat{f}(x_i) - f(x_i)]^2. \quad (22)$$

A smaller $R(\hat{f}, f)$ means a better estimation.

The optimal thresholding value is the value t that minimizes

$$M(t) = \sum_{i=1}^n [\hat{f}_t(x_i) - f(x_i)]^2 = \sum_{j,k} (\hat{w}_{j,k} - w_{j,k})^2, \quad (23)$$

where $\hat{f}_t$ is the t -threshold estimator using soft-thresholding. The optimal value is an ideal that is not available in a practical problem because f is unknown; however, it is a benchmark.

To simplify the presentation, the following abbreviations are used for the several thresholding methods, as follows:

OPT: the level-dependent optimal thresholding method. ABWS: the adaptive Bayesian wavelet shrinkage method in Chipman et al. (1997).

MethodS: the multiple shrinkage MethodS in Clyde et al. (1998). VAB: the visually Adaptive Bayesian method presented here.

Eight testing functions were used as in Figure 1. The add iid $N(0, \sigma^2)$ noise to each function to generate 1000 simulated noisy data sets, and run the ABWS, multiple shrinkage (MethodS) and the new method in Section 4 on these data sets. The parameters θ and c in MethodS is $\theta = (0.90, 0.90, 0.90, 0.90, 0.90, 0.50, 0.50, 0.50, 0.50, 0.05)$ and $c=1048561$ according to Clyde et al. (1998). The resulting L_2 deviations from the true function are summarized in Table 1.

In these eight simulations, ABWS performs best in the PIECEWISE polynomial and CORNER case, method S performs best in HEAVISINE and BUMPS, and our new VAB method performs best in the remaining four cases. In fact, in the case of BUMPS and SMOOTH signal, the performances of method S and our method are very close to each other; in the case of CORNER, the performances of ABWS and method S are very close to each other. Notice that in the case of DOPPLER, our VAB method slightly outperformed the level-dependent optimal soft-thresholding. There are a few other cases in which Bayes shrinkage is very close to the optimal soft-thresholding, such as, ABWS in the PIECEWISE polynomial case, VAB in the SMOOTH signal and CHIRP case, method S and ABWS in the CORNER case.

Simulation examples are plotted in Figures 2-9. In each figure, upper left is the noisy data; upper right is the signal recovered by ABWS, with real signal in dotted line; lower left is the signal recovered by method S with real signal in dotted line; lower right is the signal recovered by VAB, with real signal in dotted line.

An inspection of Figures 2-9 reveals some facts. ABWS tends to over-smooth the data, sometimes this over-smooth will cause a big departure from the original signal, as in the case of CHIRP and DOPPLER. MethodS and VAB both capture the coarse shape of the curve very effectively.

The L_2 norm might not be an appropriate value to measure performance. It is easy to find two estimates $\hat{f}_1$ and $\hat{f}_2$, such that $\|\hat{f}_1 - f\|_2 < \|\hat{f}_2 - f\|_2$, but visually $\hat{f}_2$ is preferred. It is not uncommon in our simulation

study, because only a slight left or right shift of f will lead to this result.

This created a motivation to do more investigation to determine a measure that better reflects our visual system. Clearly distance plays a very important role in pattern recognition. Many books and papers on pattern recognition try to define picture similarity without success. In fact it is not understood what is truly meant by cognitive similarity. That is the underlying intuition. However, it was found that Skorohod topology might be a good choice.

Let $D[0,1] = \{f; f:[0,1] \rightarrow \mathbb{R}^1, \text{ with properties 1) to 3) }\}$, where properties 1) to 3) are defined as follows:

- 1) $\lim_{u \downarrow t} f(u) = f(t+) = f(t), \forall 0 \leq t < 1,$
- 2) $\lim_{u \uparrow t} f(u) = f(t-), \forall 0 < t \leq 1,$ (24)
- 3) $f(1-) = f(1).$

Denote $\Lambda = \{\lambda; \lambda : [0,1] \mapsto [0,1], \text{ is a 1-1 monotone continuous mapping }\}$, and denote $\Lambda_\varepsilon = \{\lambda \in \Lambda; \sup_{t \in [0,1]} |\lambda(t) - t| \leq \varepsilon\}$, then for any $f, g \in D[0,1]$, define

$$Sk(f, g) = \inf \left\{ \varepsilon > 0; \exists \lambda \in \Lambda_\varepsilon, \sup_{t \in [0,1]} |f(t) - g(\lambda(t))| \leq \varepsilon \right\}. \quad (25)$$

The Skorohod distance considers the distance between two functions after translating or revolving them, and describes the similarity of functions very well. For details, see Billingsley (1968).

The Skorohod distance is more reasonable in describing the difference between broken functions by considering the uniform distance between two functions after doing a monotone continuous lengthening or shortening to the independent variables of the functions. It introduces a certain level of invariance to distortions and translations.

Table 1: Results of 1,000 simulations using L_2 – the mean-squared error.

Applications	σ	$(1/1000) \sum_{i=1}^{1000} \ \hat{f}_i - f\ _2^2$			
		ABWS	methodS	new method	OPT
SMOOTH	0.1	38.81(10.30)	9.90(3.23)	9.41(2.94)	7.14(2.24)
PIECEWISE	0.1	9.09(1.97)	12.36(2.39)	12.23(2.39)	8.76(1.64)
CHIRP	0.1	748.2(55.3)	45.48(5.96)	36.47(4.31)	35.92(2.83)
CORNER	0.1	11.23(2.26)	11.65(2.23)	12.26(2.29)	9.41(1.74)
BLOCKS	0.2	930.4(103.)	187.3(21.5)	154.1(16.0)	137.7(10.4)
BUMPS	0.3	1325.(118.)	360.3(35.5)	360.6(37.4)	275.8(20.1)
DOPPLER	0.2	1075.(140.)	106.1(14.8)	99.83(11.7)	101.1(9.32)
HEAVISINE	0.3	409.8(73.2)	97.09(13.8)	102.4(14.0)	82.42(11.3)

The unit is 10^{-4} ; standard deviations are inside ().

Table 2: Uniform distance of the same 1,000 simulations.

Applications	σ	$(1/1000) \sum_{i=1}^{1000} d(\hat{f}_i, f)$			
		ABWS	methodS	new method	OPT
SMOOTH	0.1	15.56(2.35)	9.41(4.07)	8.39(1.79)	7.21(1.45)
PIECEWISE	0.1	28.00(2.57)	25.50(3.10)	24.09(1.91)	23.37(2.66)
CHIRP	0.1	94.49(1.63)	51.29(9.12)	40.70(5.61)	35.73(4.61)
CORNER	0.1	22.42(4.44)	18.62(3.98)	21.84(3.72)	17.58(3.35)
BLOCKS	0.2	179.7(8.13)	92.75(16.0)	77.93(11.0)	70.08(9.46)
BUMPS	0.3	346.9(12.9)	125.3(18.8)	150.2(21.3)	107.0(14.8)
DOPPLER	0.2	136.8(6.20)	83.94(14.6)	77.11(11.6)	64.56(8.67)
HEAVISINE	0.3	103.8(5.64)	104.5(13.1)	94.42(8.66)	89.81(12.6)

The unit is 10^{-2} ; standard deviations are inside ().

Figure 1: The test functions

Figure 2: Smooth signal data, with $\sigma = 0.1$ Figure 3: Piecewise polynomial data, with $\sigma = 0.1$.

Figure 4: Chirp data, with $\sigma = 0.1$. Figure 5: Corner data, with $\sigma = 0.1$.

Figure 6: Blocks data, with $\sigma = 0.2$. Figure 7: Bumps data, with $\sigma = 0.3$.

Figure 8: Doppler data, with $\sigma = 0.2$. Figure 9: Heavisine data, with $\sigma = 0.3$.

It is well known that $C[0,1] \subset D[0,1]$ (Billingsley 1968), which means that the uniform topology is equivalent to the Skorohod topology for continuous functions. It is easy to show that in discrete cases such as in computer simulation, the uniform topology is equivalent to the Skorohod topology, where the uniform topology is defined as

$$d(f, g) = \sup_{0 < x < 1} |f(x) - g(x)|. \quad (26)$$

Convergence in the uniform topology implies convergence in the L_2 norm, but convergence in L_2 norm can not guarantee convergence in the uniform topology. In this sense, Uniform topology seems to be a better candidate to serve as the measurement of the performance.

Table 2 summarizes the uniform topology in the same simulation study. Notice that in the case of PIECEWISE polynomial, CORNER and HEAVISINE, the pedigree of the uniform topology and the L_2 are very controversial. Our visual impression seems to prefer the uniform topology. In the other cases, the two measurements are compatible.

Conclusion

This article presents and implements a new VAB method to recover signals from noisy data. The VAB method was compared with existing Bayesian methods. The results support the notion that many methods are serviceable when iid Normal noise are added.

The appealing part of this model is that it can capture the few big spikes in the coefficients effectively, thereby preserving the coarse shape of the picture. The simplicity of the model is also an advantage. Compared with other prior models, VAB uses less CPU time. In simulation studies, VAB performs best in four out of the eight cases when using the mean-squared error, and it performs best in six out of the eight cases studied when using the uniform distance.

References

- Abramovich, F., Sapatinas, T., & Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of Royal Statistical Society, 60 (Series B)*, 725-749.
- Billingsley, P. (1968). *Convergence of probability measures*, NY: Wiley.
- Bruce, A., & Gao, H. (1996a). *Applied wavelet analysis with S-Plus*. Berlin: Springer.
- Bruce, A., & Gao, H. (1996b). Understanding WaveShrink: Variance and bias estimation, *Biometrika*, 83, 727-745.
- Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics*, 27, 898-924.
- Chipman, Kolaczyk, & McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92, 1413-1421.
- Clyde, M., Parmigiani, G., & Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85, 391-401.
- Clyde, M., DeSimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91, 1197-208.
- Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.
- Daubechies, I. (Ed.) (1993). *Different Perspectives on Wavelets*. Applied Mathematics Society.
- Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1200-1224.
- Donoho, D. L., & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26, 879-921.
- Gao, H., & Bruce, A. (1997). WaveShrink with firm shrinkage. *Statistica Sinica*, 7, 855-874.
- Huang, H. C., & Cressie, N. (1999). Empirical Bayesian spatial prediction using wavelets, *Bayesian inference in wavelet-based models*, Springer-Verlag, New York.

Johnstone, I. M., & Silverman, B. M. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society*, 59 (Serial B), 319-351.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674-693.

Nason, G. P., & Silverman B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical statistics*, 3, 163-191.

Nason, G. P. (1994). Wavelet regression by cross-validation. Technical Report 447.

Nason, G. P. (1995). Choice of the threshold parameter in wavelet function estimation: *Wavelets and statistics*, Lecture Notes in Statistics 103, 261-280. NY: Springer-Verlag.

Strang, G. (1993). Wavelet transforms versus Fourier transforms. *Bulletin (New Series) of the American Mathematical Society*, 28 (2), 288-305.

Wu, D. (2002). NORM thresholding method in wavelet regression. *Journal of Statistical Computation and Simulation*, 72 (No.3), 233-246.

BRIEF REPORTS

A Test-Retest Transition Matrix: A Modification of McNemar's Test

J. Wanzer Drane

Schools of Public Health and Medicine
University of South Carolina

W. Gregory Thatcher

Department of Health, Leisure and Exercise Science
University of West Florida

McNemar introduced what is known today as a test for symmetry in a two by two contingency tables. The logic of the test is based on a sample of matched pairs with a dichotomous response. In our example, the sample consists of the scores before and after an education program and the responses before and after the program. Each pair of scores is from only one person. The pretest divides the group of responders according to their answers to a dichotomous question. The posttest divides the two groups into two groups of like labels. The result is a two by two table. We construct a test of homogeneity, where the proportion of initially partitioned subjects will be equally distributed over the same partition after the program is completed, conditioned on the initial distribution.

Key words: McNemar's test, test of homogeneity, contingency table

---

Introduction

McNemar (1947) introduced what is known today as a test for symmetry in contingency tables, although his table was only a two by two. What is more, his table is often illustrated via matched pairs and the joint classification of a dichotomy applied to each of the pair. Let zero (0) represent the absence and 1 the presence of the characteristic thus dichotomized.

Table 1 illustrates such a dichotomy. Since they were matched by some criterion, a zero response from a case ought to be matched with a zero from its control, but that does not always happen. The numbers $N(0,1)$ and $N(1,0)$ measure any departure from perfect correlation. McNemar's hypothesis was that these two numbers ought to be equal, or $P(0,1) = P(1,0)$. In our illustration this hypothesis needs to be changed.

Consider a pre-test and a post-test or a pre-program and post-program situation. A simple random sample of subjects is asked a question about a certain characteristic such as "Do you smoke?" There are $N(0)$ who do not smoke and $N(1)$ who do smoke prior to the application of a program on smoking cessation. Six months after the program is completed they are again asked the same question. A table such as Table 1 results, except that "Case" is now replaced by "Pre-Program" and "Controls" is replaced with "Post-Program".

Unless $N(0) = N(1)$, $N(0,1)$ cannot be expected to equal $N(1,0)$. $N(0,1)$ is the number of people who did not smoke, but six months after the program they were observed to be smoking. $N(1,0)$ people were smoking before the program, and six months later they were not smoking. The correct null hypothesis is $P(1|0) = P(0|1)$. That is, the proportion of prior non-smokers who changed to smokers is equal to the proportion of prior smokers who changed to non-smokers.

The application of the question prior to the program stratifies the sample into two strata that cannot be expected to be the same size. If the program is expected a priori to work, the one-sided alternative should be used, namely $P(1|0) < P(0|1)$. That is, the proportion of non-smokers who changed to smokers should be

---

J. Wanzer Drane is Professor in Department of Epidemiology & Biostatistics, School of Public Health, University of South Carolina, Columbia, SC, 29208. Email: wdrane@sph.sc.edu. W. Gregory Thatcher is Assistant Professor Department of Health, Leisure and Exercise Science, University of West Florida. Email: wthatcher@uwf.edu

significantly and clinically smaller than the proportion of smokers who changed to non-smokers. Table 2 is now rearranged and the people are relabeled as “Stayers” and “Movers.” Stayers are non-smokers who remain non-smokers, and similarly for smokers. Quitting and Beginning after the program label those who change and are called “Movers” in Table 3.

With this rearrangement, the hypothesis of homogeneity can be tested with the usual chi-squares, Pearson or Likelihood Ratio, and also with Fisher’s exact test. A significant chi-square at 0.5α coupled with $N(1,0)/N(1) > N(0,1)/N(0)$ signals a working program because a smaller fraction of non-smokers became smokers than the fraction of smokers who became non-smokers. The Fisher’s exact test would be one-tailed in the direction supporting the alternative hypothesis.

The first questionnaire revealed 142 non-smokers and 58 smokers in a teen smoking cessation project. Of the 142, after six months 11 had begun smoking, while 25 of the 58 smokers had quit smoking. Filling in Table 2 yields Table 3. Analyzing this table gives rise to $X^2 = 34.9$, $G^2 = 31.9$, $DF = 1$, $p < 0.0001$. Fisher’s exact test gives $p = 2.4 \times 10^{-8}$ with proportions moving $11/142 = 0.077 < 25/58 = 0.431$. Therefore there is statistical significance. Because only 7.7% moved from non-smoker to smoker while the reverse was true for 43.1% of the smokers, this is apparently clinically significant. Therefore the program works. If Odds Ratio is the measure of choice, the Odds of Quitting given the person was a smoker is 9.02 times the Odds of Beginning given that the person was a non-smoker with a 95% CI = (4.03, 20.2).

Table 1: A dichotomy applied to cases and their respective matched controls.

	Control: N(0)	Control: N(1)
Case: N(0)	N(0,0)	N(0,1)
Case: N(1)	N(1,0)	N(1,1)

Table 2: Stayers and Movers.

Stayers	Movers
N(0,0)	N(0,1)
N(1,1)	N(1,0)

Table 3: Stayers and Movers.

Stayers	Movers	Total
N(0,0) = 131	N(0,1) = 11	N(0) = 142
N(1,1) = 33	N(1,0) = 25	N(1) = 58
N-Stayers = 164	N-Movers = 36	N = 200
$X^2 = 34.9$, $G^2 = 31.9$, $DF = 1$, $p < 0.0001$		

References

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, 12, 153 – 157.

Estimation Of Multiple Linear Functional Relationships

Amjad D. Al-Nasser
Department of Statistics
Yarmouk University, Jordan

This article deals with multiple linear functional relationships models. Two robust estimations procedure are proposed to estimate the model, based on Generalized Maximum Entropy and Partial Least Square. They are distribution free and do not rely (so much) on classical assumptions. The experiments showed that the GME approach outperforms the PLS in terms of mean squares of errors (MSE). Empirical examples are studied.

Key words: Multiple linear functional relationships, generalized maximum entropy, partial least square

Introduction

Consider a set of mathematical variables $(\xi_1, \xi_2, \dots, \xi_K)$ that connected by the relation $\eta = \alpha + \beta_1 \xi_1 + \beta_2 \xi_2 + \dots + \beta_K \xi_K$ and the observed values

$$(x_{ik}, y_i), i = 1, 2, \dots, n, k = 1, 2, \dots, K, n > K > 1,$$

are such that

$$\begin{aligned} y_i &= \eta_i + \varepsilon_i \\ x_{ik} &= \xi_{ik} + \delta_{ik} \end{aligned} \quad (1)$$

where

$$\eta_i = \alpha + \beta_1 \xi_{i1} + \beta_2 \xi_{i2} + \dots + \beta_K \xi_{iK}$$

which can be defined as MLFR, where α is the intercept, β is the slope vector $K \times 1$, ξ unknown, unobservable latent variable with dimension $K \times n$, ε and δ , are mutually independent unobservable disturbances terms with zero means and finite variances. They are assumed to be independent of (ξ_{ik}, η_i) .

Methodology

This article considers two estimation approaches. One is the PLS developed by Wold (1975). The algorithm is presented in Helland (1990) and Geladi (1986). The second is a theoretically attractive estimation approach based on Shannon's (1948) entropy, which is called the Generalised Maximum Entropy. It was developed by Golan et al. (1996) for limited data. Subsequently, Al-Nasser et al. (2000), Abdullah (2000), and Al-Nasser (2003) used it for estimating Errors-in-Variables models. The GME method is a nonlinear programming approach to determine the most random probability distribution subject to the given information (see (1)). Without loss of generality, by assuming the intercept is equal to zero, the model can be rewritten as

$$y_i = \sum_{k=1}^K \beta_k (x_{ik} - \delta_{ik}) + \varepsilon_i, i = 1, 2, \dots, n \quad (2)$$

Because the unknowns are not in probability distribution form, their possible outcomes values are reparametrized as convex combination presented as expected value of discrete random variable:

$$\beta_k = \sum_{h=1}^H z_{hk} p_{hk}, \sum_{h=1}^H p_{hk} = 1, k = 1, 2, \dots, K, H \geq 2 \quad (3)$$

Amjad D. Al-Nasser: Department of Statistics, Faculty of Science, Yarmouk University, Irbid Jordan. Email him at amjadn@yu.edu.jo.

The restriction imposed on the parameter space through z reflects à priori knowledge about these parameters. However, if the possible values of the parameters are known from the theory, then z is specified accordingly. If not, then z is specified to be uniformly symmetric around 0 with high - low upper bounds. For example, $z = (-c, 0, c)$, c large value.

For the simulation study here, assume $z_{hk} = z_h \forall k = 1, 2, \dots, K$. In addition, the unobservable ξ_{ik} can be obtained from the observed data values x_{ik} and the disturbance δ_{ik} can be treated in similar fashion:

$$\delta_{ik} = \sum_{t=1}^T v_{ikt}^* w_{ikt}^*, \sum_{t=1}^T w_{ikt}^* = 1, k = 1, 2, \dots, K, i = 1, 2, \dots, n, T \geq 2 \quad (4)$$

$$\varepsilon_i = \sum_{j=1}^J v_{ij} w_{ij}, \sum_{j=1}^J w_{ij} = 1, i = 1, 2, \dots, n, J \geq 2 \quad (5)$$

The actual bounds for v_t^* and v_j depend on the observed sample as well as any conceptual or empirical information about the underlying error. However, if such conceptual or empirical information does not exist, then v_t^*, v_j may be specified to be uniformly and symmetrically distributed around zero. Chebychev's inequality or 3-sigma rule Pukelsheim (1994) may be used as conservative means of specifying sets of error bounds. Under these reparameterizations, the statistical MLFR model (2) can be rewritten as

$$y_i = \sum_{k=1}^K \left[\left\{ x_{ik} - \sum_{t=1}^T v_{ikt}^* w_{ikt}^* \right\} \cdot \sum_{h=1}^H z_{hk} p_{hk} \right] + \sum_{j=1}^J v_{ij} w_{ij} \quad (6)$$

Given the reparameterization inverse problem (6) involving the unknown and the unobservable p, w^*, w , the GME problem as maximizing the dual objective function may be stated in scalar summations with three nonnegative probability components:

Maximize $H(p, w^*, w)$

$$= - \sum_{h=1}^H \sum_{k=1}^K p_{hk} \log(p_{hk}) - \sum_{i=1}^n \sum_{j=1}^J w_{ij} \log(w_{ij}) - \sum_{i=1}^n \sum_{k=1}^K \sum_{t=1}^T w_{ikt}^* \log(w_{ikt}^*)$$

Subject to the consistency constraints (data)

$$y_i = \sum_{k=1}^K \left[\left\{ x_{ik} - \sum_{t=1}^T v_{ikt}^* w_{ikt}^* \right\} \cdot \sum_{h=1}^H z_{hk} p_{hk} \right] + \sum_{j=1}^J v_{ij} w_{ij}, i = 1, 2, \dots, n$$

and adding-up normalization constraints

$$(I) \quad \sum_{h=1}^H p_{hk} = 1, k = 1, 2, \dots, K$$

$$(II) \quad \sum_{j=1}^J w_{ij} = 1, i = 1, 2, \dots, n$$

$$(III) \quad \sum_{t=1}^T w_{ikt}^* = 1, i = 1, 2, \dots, n, k = 1, 2, \dots, K$$

Solving the optimization problem of the (HK+nJ+nKT) equations yields the optimal and unique solution:

$$\hat{p}_{hk} = \frac{\exp \left[- \sum_{k=1}^K z_{hk} \left\{ \sum_{i=1}^n \hat{\gamma}_i \left(x_{ih} - \sum_{t=1}^T \hat{v}_{iht}^* \hat{w}_{iht}^* \right) \right\} \right]}{\sum_{h=1}^H \exp \left[- \sum_{k=1}^K z_{hk} \left\{ \sum_{i=1}^n \hat{\gamma}_i \left(x_{ih} - \sum_{t=1}^T \hat{v}_{iht}^* \hat{w}_{iht}^* \right) \right\} \right]} \quad (7)$$

$$\hat{w}_{ij} = \frac{\exp[-\hat{\gamma}_i v_{ij}]}{\sum_{i=1}^n \exp[-\hat{\gamma}_i v_{ij}]} \quad (8)$$

$$\hat{w}_{iht}^* = \frac{\exp \left[- \hat{\gamma}_i v_t^* \sum_{k=1}^K z_{hk} \hat{p}_{hk} \right]}{\sum_{i=1}^n \sum_{h=1}^H \exp \left[- \hat{\gamma}_i v_t^* \sum_{k=1}^K z_{hk} \hat{p}_{hk} \right]} \quad (9)$$

where $\hat{\gamma}_i, i=1,2,\dots,n$ are Lagrangian multipliers. The estimated parameters can be obtained by substitute (7) in (3);

$$\hat{\beta}_k = \sum_{h=1}^H z_{hk} \hat{p}_{hk} \quad k=1,2,\dots,K$$

Similarly, by substitution of (8) and (9) in (4) and (5) respectively, the estimated values of the unobservable variables may be obtained.

Sampling Experiments

Now, consider the performance of the GME and PLS method in estimating the parameters of the MLFR models in (2). Some experiments are carried out to choose the supported weights of the unknown elements under the following conditions:

- (i) Generate 500 samples each of size $n = 15$ observations with $K = 3$ parameters.
- (ii) ξ_{ik} are initialised to be in the interval $\left[-\left\lceil \frac{nK}{2} \right\rceil, \left\lceil \frac{nK}{2} \right\rceil \right]$ with increment 1, where $[*]$ is the greater number less than or equal to $*$.
- (iii) Assume that $\beta_k = 1, \forall k = 1,2,\dots,K$.
- (iv) The disturbance δ_{ik} and ε_i are generated from Standard Normal distribution.
- (v) Simulate the observed values from the following equations:

$$x_{ik} = \xi_{ik} + \delta_{ki}, i=1,2,\dots,n, k=1,2,\dots,K$$

$$y_i = \xi_{ik} + \varepsilon_i, i=1,2,\dots,n, k=1,2,\dots,K.$$

Choice of the index parameters space (Experiment 1)

The aim of this experiment is to chose the disturbance index by varying its values as $J, T = 3, 4, 5, 6, 7$. The support space of the parameters β_k are fixed to be indexed into 5 values in the interval $[-c, c]$ for $c = 10$. The results of this experiment are tabulated as follows:

Table 1. Choice of residual indices for GME-MLFR model.

Method	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)
PLS	0.0672	0.0659	0.0668
GME (J & T =3)	0.0242	0.0224	0.0203
GME (J & T =4)	0.0243	0.0223	0.0203
GME (J & T =5)	0.0246	0.0222	0.0202
GME (J & T =6)	0.0248	0.0221	0.0201
GME (J & T =7)	0.0250	0.0220	0.0201

In the same way, an experiment is carried out to choose the index value of the parameter space z . The values of z fixed to be located in the interval $[-10,10]$. The results are tabulated as follows:

Table 2. Choice of parameters index in GME-MLFR model.

Method	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)
PLS	0.0672	0.0659	0.0668
GME (M=3)	0.0249	0.0221	0.0201
GME (M=4)	0.0249	0.0221	0.0201
GME (M=5)	0.0247	0.0222	0.0202
GME (M=6)	0.0245	0.0222	0.0203
GME (M=7)	0.0244	0.0223	0.0203

This experiment suggests that regardless of the index values of parameter supports space the GME is superior than PLS estimates and more robust.

Choice of parameter space values (Experiment 2)

Under the above conditions and by fixing five index values for the parameters in the interval $[-10,10]$, this experiment is carried out to find the suitable parameter space of the disturbance part. The choice of the supported values was obtained depending on Chebyshev's inequality, and the following table shows the simulated results:

Table 3 Choice of residuals space values for GME-MLFR model.

Method	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)
PLS	0.0672	0.0659	0.0668
GME	0.0246	0.0222	0.0202
{[-3S,3S]}			
GME	0.0241	0.0224	0.0205
{[-4S,4S]}			
GME	0.0237	0.0226	0.0208
{[-5S,5S]}			
GME	0.0235	0.0226	0.0210
{[-6S,6S]}			
GME	0.0233	0.0227	0.0213
{[-7S,7S]}			

Table.4 Choice of parameter space values of GME-MLFR model.

Method	MSE($\hat{\beta}_1$)	MSE($\hat{\beta}_2$)	MSE($\hat{\beta}_3$)
PLS	0.0672	0.0659	0.0668
GME	0.0234	0.0226	0.0211
{z : [-5,5]}			
GME	0.0246	0.0222	0.0202
{z: [-10, 10]}			
GME	0.0288	0.0203	0.0194
{z: [-50, 50]}			
GME	0.0293	0.0201	0.0194
{z: [-100, 100]}			
GME	0.0294	0.0200	0.0193
{z: [-500, 500]}			
GME	0.0294	0.0200	0.0193
{z: [-1000, 1000]}			

This experiment confirms the superiority of the GME over the PLS in all choices of the support values. Note that the 3-sigma choice gives better estimates with respect to both parameters as shown in Table 3.

Increasing the sample size and the number of parameters (Experiment 3)

Consider increasing the sample size $n = 15, 20, 30, 40$, and varying number of parameters, i.e., $K = 2, 3, 4$. The parameters index is fixed to be three, with support values $[-3S, 0, 3S]$ for the disturbance parts and $[-10, 10]$ for the slopes. The results were obtained and are displayed in Tables 5 and 6. The tabulated results demonstrate

that the GME approach performs better than the PLS with respect to their MSE for any sample size and any number of unknown parameters.

Empirical Example

The data were collected via interviews with seventeen customers in order to gain a better understanding of service loyalty and identify salient factor affecting its development Al-Nasser (2000). Two types of services, i.e., cellular phone and TV stations, were selected for the study. For this example, the model is

$$\eta_i = \beta_1 \xi_{i1} + \beta_2 \xi_{i2}, i = 1, 2, \dots, 17$$

such that

$$x_{i1} = \xi_{i1} + \delta_{i1}$$

$$x_{i2} = \xi_{i2} + \delta_{i2}$$

and

$$y_i = \eta_i + \varepsilon_i$$

where η_i = Customer Loyalty, ξ_{i1} = Service Image, and ξ_{i2} = Customer Satisfaction.

This model is solved by using both approaches (GME and PLS). The results are shown in Table 7. In addition, Figure 1 depicts a distinction between the observed error of the customer loyalty by using both methods with the observed errors closely centred around zero. This appears to support that the notion that GME is a more robust alternative than PLS.

TABLE.7 Estimated values of the parameters using the PLS and GME for customer satisfaction data.

Method	$\hat{\beta}_1$	$\hat{\beta}_2$
PLS	0.3523	0.3682
GME	1.8746	1.0876

Table. 5 Performance of the GME-MLFR for various sample sizes and number of parameters.

N		15	20	30	40
K					
2	$MSE(\hat{\beta}_1)$	0.0267	0.0199	0.0105	0.0075
	$MSE(\hat{\beta}_2)$	0.0197	0.0144	0.0120	0.0092
3	$MSE(\hat{\beta}_1)$	0.0236	0.0157	0.0114	0.0083
	$MSE(\hat{\beta}_2)$	0.0226	0.0168	0.0107	0.0080
4	$MSE(\hat{\beta}_3)$	0.0208	0.0183	0.0110	0.0086
	$MSE(\hat{\beta}_1)$	0.0229	0.0140	0.0113	0.0079
	$MSE(\hat{\beta}_2)$	0.0221	0.0162	0.0096	0.0071
	$MSE(\hat{\beta}_3)$	0.0232	0.0186	0.0142	0.0116
	$MSE(\hat{\beta}_4)$	0.0215	0.0188	0.0102	0.0077

Table.6 Performance of the PLS MLFR for various sample sizes and number of parameters.

N		15	20	30	40
K					
2	$MSE(\hat{\beta}_1)$	0.0651	0.0448	0.0325	0.0244
	$MSE(\hat{\beta}_2)$	0.0651	0.0448	0.0326	0.0244
3	$MSE(\hat{\beta}_1)$	0.0672	0.0448	0.0333	0.0249
	$MSE(\hat{\beta}_2)$	0.0659	0.0449	0.0334	0.0250
4	$MSE(\hat{\beta}_3)$	0.0668	0.0449	0.0333	0.0250
	$MSE(\hat{\beta}_1)$	0.0651	0.0449	0.0325	0.0244
	$MSE(\hat{\beta}_2)$	0.0651	0.0449	0.0236	0.0244
	$MSE(\hat{\beta}_3)$	0.0651	0.0449	0.0325	0.0244
	$MSE(\hat{\beta}_4)$	0.0651	0.0449	0.0325	0.0244

FIGURE 1 Comparison of the observed errors of customer loyalty using GME and PLS.

Conclusion

The results presented in this article provide strong evidence about the robust performance of the GME compared with the traditional methods that have been used to deal with the MLFR. In fact, the PLS that has become the conventional approach in measurement error model fails to outperform the GME in all the cases studied.

From the simulation study, it is evident that with various sample sizes and number of parameters, the MSE of GME estimates are smaller than those of the PLS. Therefore, the GME should provide a better alternative for the MLFR compared with the other methods.

References

- Mokhtar A., Al-Nasser. D. A., & Nooreha, H. (2000). Evaluating functional relationship between image, customer satisfaction, and loyalty using general maximum entropy. *Total Quality Management*, 11, (6), 826-829.
- Geladi, P., & Kowalski, R. B. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.
- Golan, A., Judge, G. & Miller, D. (1996). *A maximum entropy econometrics: robust estimation with limited data*. New York: Wiley.

Helland, S. I. (1990). Partial least squares regression and statistical model. *Scandinavian Journal of Statistics*, 17, 97-114.

Al-Nasser, D. A., Abdullah, B. N. & Wan Endut, W. J. (2000). On robust estimation of error in variables models by using generalized maximum entropy. International conference on mathematics and its applications in the new millennium. Dept. of Mathematics. Universiti Putra Malaysia. 18-21 July, 279-287.

Al-Nasser, D. A., (2000). Generalised maximum entropy estimation approach to errors-in-variables models. Unpublished Doctoral dissertation, National University of Malaysia.

Amjad, D. Al-Nasser, D. A. (2003). Customer satisfaction measurement models: generalized maximum entropy approach. *Pakistan Journal of Statistics*, 19, (2), 213-226.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, 48, (2), 88-91.

Shannon C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-423.

Wold, H. (1975). Path models with latent variables: The NIPLAS approach. In H. M. Blalock (Ed). *Quantitative sociology: International perspective on Mathematical and Statistical Modeling*. New York

Teaching Random Assignment: Do You Believe It Works?

Shlomo S. Sawilowsky
Educational Evaluation & Research
Wayne State University

Textbook authors admonish students to check on the comparability of two randomly assigned groups by conducting statistical tests on pretest means to determine if randomization worked. A Monte Carlo study was conducted on a sample of $n = 2$ per group, where each participant's personality profile was represented by 7,500 randomly selected and assigned scores. Independent samples t tests were conducted and the results demonstrated that random assignment was successful in equating the two groups on 7,467 variables. The students' focus is redirected from the ability of random assignment to create comparable groups to the testing of the claims of randomization schemes.

Key words: Random assignment, Monte Carlo, comparable groups

Introduction

Random assignment is one of the more difficult concepts in designing experiments. Researchers harbor considerable distrust in the ability of random assignment to create comparable groups. Interestingly, the seeds of distrust in random assignment are sown in statistics and research textbooks. For example, in a pretest-posttest treatment vs control group design, Tuckman (1994) noted, "It is not uncommon to assign S_s randomly to groups and then to check on the distribution of control variables by comparing the groups to assess their equivalence on these variables" (p. 130). Students are told to check on the comparability of the two groups by conducting statistical tests on the pretest means, as Krathwohl (1993) stated, "The pretest tells us whether randomization worked and the groups are really comparable" (p. 452).

This article is based on a presentation delivered in 1999 to the American Educational Research Association, Special Interest Group Educational Statisticians, Montreal, Canada. The author gratefully acknowledges discussions with Drs. Lori Rothenberg and Randy Lattimore on earlier versions of this article. Email the author at: shlomo@wayne.edu.

This problem is exacerbated when researchers consider the typical small samples available for research in applied fields. For example, Gall, Borg, and Gall (1996) stated,

The probability that random assignment will produce initially equivalent treatment groups increases as the size of the sample in each group increases. For example, equivalent groups will more likely result if 100 individuals are randomly assigned to two treatment groups ($n = 50$ per group) than if 10 individuals are assigned to those to groups ($n = 5$ per group). p. 489.

Similar statements are found in Cook and Campbell (1979), Crowl (1996), Vockell and Asher (1995), and others.

A Previous Demonstration

Strube (1991) noted "small samples cause other problems that argue against their routine use". Indeed, small samples present difficulties with regard to the generalizability of results. Strube (1991) endeavored to show that "the probability of an erroneous inference" is "generally no greater than the nominal Type I error rate" (p. 346). In this respect, Strube's (1991) article was convincing.

However, Strube's (1991) demonstration may not have been the most

effective approach in convincing researchers of the ability of random assignment to produce baseline equality among two groups. First, Strube (1991) used a relatively complex design: a 2 x 2 (treatment vs control × nuisance variable) with samples sizes from $N = 8$ to 100. Second, Strube (1991) modeled the presence of effect size from .25 to 4.00, such as Cohen's d , where

$$d = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p}, \quad (1)$$

and s_p is the pooled standard deviation. These effect sizes were treated as small to very large nuisance parameters. They were deliberately introduced as the error terms in the simulation, as opposed to studying the behavior of random assignment on random fluctuations. A simpler demonstration is clearly warranted.

Methodology

In order to explicate the effects of random assignment, and to demonstrate to researchers that it indeed works with a sample as small as $N = 4$ or $n = 2$ per group, a Monte Carlo study was conducted. A program was written in Fortran 90/95. Façade, a personality profile, was created by dimensioning four arrays. Each of the four façade arrays, representing a participant's profile, contained 7,500 values.

These values were comprised of 1,250 scores obtained from each of six real data sets described by Micceri (1989) as being representative of the most prolific shapes of data set distributions in psychology and education research. (For histograms and descriptive statistics on these data sets, see Sawilowsky & Blair, 1992). The six data sets were:

- smooth symmetric (from an achievement instrument)
- extreme asymmetry (from a psychometric instrument)
- extreme asymmetry (achievement)
- digit preference (achievement)

- discrete mass a zero with gap (achievement)
- multimodal lumpy (achievement)

The personality profile for each participant was created as follows. Scores were sampled of size $N = 4$, independently and with replacement from the data sets. Next, the scores were randomly assigned to two groups, with $n_1 = n_2 = 2$. This process was repeated 1,250 times for each data set. Then, an independent samples t test was conducted on each of these variables for a total of 7,500 t tests.

The t test is a widely used procedure for the statistical comparison of the means of two groups. The null hypothesis is $H_0: \mu_1 = \mu_2$, which is tested against the alternative hypothesis $H_a: \mu_1 \neq \mu_2$, by the formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sum x_1^2 + \sum x_2^2}{n_i(n_i - 1)}}}, \quad (2)$$

where $n_i = 2$ in this example. Essentially, the difference in the means of the raw scores of the two groups are standardized when divided by an estimate of the pooled population variance, which is the error term. Then, the obtained statistic is compared with the critical value given in t tables (as found in most statistics textbooks or statistics software packages) for the nominal α level of 0.01 and the degrees of freedom (df) of $(n_1 + n_2 - 2)$, or 2 df in the current example.

The α level indicates that if the obtained t statistic exceeds the tabled value, the difference in means between the two groups is likely to have occurred by chance with a probability of less than one out of 100 under the truth of the null hypothesis. Thus, the proposition that the two groups are equal on that construct of the personality profile for the four participants (and random assignment equalized the two groups) would be rejected. However, if the obtained t statistic is less than the critical value, then the hypothesis that the two groups are equal in term of their respective means for

that façade variable (and random assignment equalized the two groups) would be retained.

Results

Table 1 contains a compilation of façade variables where statistically $\bar{x}_1 \neq \bar{x}_2$ at the $\alpha = 0.01$ level, despite random assignment. The variable numbers refer to different characteristics presented by each participant in the experiment. For example, Variable 373 from the Extreme Asymmetry (Psychological Scale) data set might refer to a score from a standardized measure of depression. Indeed, there were 15 variables where $\bar{x}_1 > \bar{x}_2$ and 18 variables where $\bar{x}_2 > \bar{x}_1$, for a total of only 33 variables out of 7,500 where random assignment failed to make the two groups comparable when sample size was as small as $n_1 = n_2 = 2$.

The failure rate of random assignment in producing a comparable group depends on nominal α . Setting nominal α to 0.05 (probability of one out of twenty) will produce more variables where statistically $\bar{x}_1 \neq \bar{x}_2$, and setting nominal α to 0.001 (probability of one out of 1,000) will eliminate many of the variables listed in Table 1. A tangential statistical issue is discussed in the Appendix.

A Classroom Experiment

An experiment was conducted with three sections of a graduate level introductory research course to assess the effectiveness of the methodology in this article for teaching random assignment. The number of participants was $N = 56$ ($n_1 = 20$, $n_2 = 18$, $n_3 = 18$). Informed consent was not required of the participants because this was part of the regular curriculum.

The students were surveyed at the beginning of the semester with the following question: "Do you believe that random assignment of subjects in an experiment into a treatment and a control group can produce comparable groups?". The forced response format was "Yes", "Maybe", or "No". If students answered "Maybe", they were asked to explain under what conditions they believed that random assignment does not work

Two of the three classes were arbitrarily selected to receive the material in this article as part of their course pack (Treatment One), without identifying the author of the article as their instructor. Later in the semester, at the usual point in the curriculum where random assignment was assigned to be discussed, students in the Treatment One classes were referred to the materials in the course pack. (There was no reading assignment for the textbook.) The students in the Treatment Two class, who did not have this article in their course pack, were directed to their version of the syllabus which assigned the textbook chapter on random assignment. The textbook is a current, popular offering with a discussion similar to that found in many research textbooks.

After the students completed the reading assignment one week later, but prior to class discussion on random assignment, they were asked to respond again to the survey question. The pretest (i.e., beginning of the semester) and posttest (i.e., after reading this article or the textbook chapter) responses are recorded in Table 2. An analysis of the posttest scores for the Treatment One classes and Treatment Two class were conducted with a stratified 2×3 singularly ordered categorical design, with the pretest scores serving as the covariate. The data analysis was conducted with *StaxXact* (Mehta & Patel, 1999).

The Mann-Whitney statistic for the data in Table 2 was 979.5, and the exact one-sided p-value = 0.0011. An inspection of the entries in the table indicates that there was a statistically significant difference between the two curricular approaches for these 56 students. The material in this article was superior to the discussion in a typical graduate level research textbook in persuading students on the effectiveness of random assignment in research and experimental design.

Table 1. Situations Where $\bar{x}_1 \neq \bar{x}_2$ Despite Randomization For 1,250 Variables From Each Of 6 Real Achievement and Psychology Populations, $n = 2$, $\alpha = 0.01$.

Population	Variable	$\bar{x}_1 > \bar{x}_2$	$\bar{x}_2 > \bar{x}_1$
Smooth Symmetric (Achievement Scale)	370		
Smooth Symmetric (Achievement Scale)	1066		✓
Smooth Symmetric (Achievement Scale)	1100		✓
Discrete Mass At Zero (Achievement Scale)	625		✓
Discrete Mass At Zero (Achievement Scale)	831	✓	
Discrete Mass At Zero (Achievement Scale)	959	✓	
Extreme Asymmetry (Achievement Scale)	291		✓
Extreme Asymmetry (Achievement Scale)	336	✓	
Extreme Asymmetry (Achievement Scale)	667	✓	
Extreme Asymmetry (Achievement Scale)	701	✓	
Extreme Asymmetry (Psychological Scale)	190		✓
Extreme Asymmetry (Psychological Scale)	373		✓
Extreme Asymmetry (Psychological Scale)	1089	✓	
Digit Preference (Achievement Scale)	17		✓
Digit Preference (Achievement Scale)	45		✓
Digit Preference (Achievement Scale)	156	✓	
Digit Preference (Achievement Scale)	172		✓
Digit Preference (Achievement Scale)	492		✓
Digit Preference (Achievement Scale)	641		✓
Digit Preference (Achievement Scale)	693	✓	
Digit Preference (Achievement Scale)	810		✓
Multimodal Lumpy (Achievement Scale)	23	✓	
Multimodal Lumpy (Achievement Scale)	281		✓
Multimodal Lumpy (Achievement Scale)	301	✓	
Multimodal Lumpy (Achievement Scale)	323		✓
Multimodal Lumpy (Achievement Scale)	441	✓	
Multimodal Lumpy (Achievement Scale)	504		✓
Multimodal Lumpy (Achievement Scale)	564	✓	
Multimodal Lumpy (Achievement Scale)	835	✓	
Multimodal Lumpy (Achievement Scale)	841		✓
Multimodal Lumpy (Achievement Scale)	851		✓
Multimodal Lumpy (Achievement Scale)	929	✓	
Multimodal Lumpy (Achievement Scale)	1025	✓	
Total/7,500		15/7,500	18/7,500

Table 2. Responses (Percent) Of 56 Students To The Question, "Do you believe that random assignment of subjects in an experiment into a treatment and a control group can produce equal groups?"

<u>Response</u>	<u>Pretest Scores</u>		<u>Posttest Scores</u>	
	<u>Intervention</u>			
	<u>This Article</u>	<u>Textbook Chapter</u>	<u>This Article</u>	<u>Textbook Chapter</u>
Yes	2 (5.3%)	1 (5.5%)	29 (76.3%)	3 (16.7%)
Maybe	13 (34.2%)	7 (38.9%)	7 (18.4%)	8 (44.4%)
No	23 (60.5%)	10 (55.6%)	2 (5.3%)	7 (38.9%)
Total	38	18	38	18

An interesting topic of classroom discussion centered on the reasons why some students responded "Maybe" or "No". At the pretest stage, the reasons given by the students for "Maybe" were random assignment only worked if (a) there was a large sample size, (b) the data collection instruments were reliable, or (c) the researcher was lucky. These reasons were maintained by the students in Treatment Two class at the posttest stage.

It was also interesting to note that the two respondents in the Treatment One class who responded "No" at the posttest stage indicated that, as members of an ethnic minority, they remained suspicious of any methodology that purports to equalize the characteristic or traits of participants assigned to two groups in an experiment.

Conclusion

Return to the initial question on the advice of textbook authors to check on random assignment to see if randomization worked with a statistical test on the pretest scores. The current study is a demonstration of the ability of randomization to create comparable groups. Therefore, the focus of the researcher's concern should not be on the ability of random assignment. Instead, it should pertain to the validity of the scheme implemented by the researcher to randomly assign participants to groups.

For example, consider the well-documented tumult over the 1970 United States military draft lottery conducted by the Selective Service under the auspices of Executive Order No. 11497 to Part 1631.5 of the Selective Service Regulations signed on November 26, 1969 by President Richard M. Nixon. Fienberg (1971; see also Notz, Staw, & Cook, 1971) raised questions regarding the process of that lottery, where slips of paper containing birth dates were placed in capsules and subsequently into a box. There was a proclivity for dates from December ($\mu = 121.5$), November ($\mu = 148.7$), October ($\mu = 182.5$), and September ($\mu = 157.3$) to be selected from the box, rather than January ($\mu = 201.2$), February ($\mu = 203.0$), March ($\mu = 225.8$), and April ($\mu = 203.7$).

Perhaps, this occurred because capsules bearing these dates were placed in the box last. Alternatively, the capsules for the earlier months were well mixed in the box because there was room to do so. However, as the capsules for the latter months were placed in the box, the lack of room limited the ability to mix the capsules. In either case, the slips of paper were not sufficiently mixed in the box, and hence, birth dates at the end of the year were more likely to be selected. The lack of non-randomness of this scheme would have been easily detected if a statistical test been conducted.

As noted by Cook and Campbell (1979), "the equivalence achieved by random

assignment is probabilistic. Thus it is not inevitable that a correctly implemented randomization procedure will result in groups that do not differ" (p. 341). Indeed, this study showed that for 33 of the 7,500 variables, random assignment resulted in differences between the two groups. Random assignment is probabilistic; it is not a guarantee. However, "Without randomization, the possibility of bias due to prior differences on an uncontrolled third variable can seldom, if ever, be ruled out as an alternative explanation of the results" (Linn, 1986). Textbook authors should more clearly distinguish between the probabilistic nature of randomization and the limitations or failure of some schemes to achieve randomization, because poorly conceived randomization schemes do create distrust in the ability of random assignment.

References

- Blair, R. C. (1987). *RANGEN*. Boca Raton, FL: IBM Corporation.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally.
- Crowl, T. K. (1996). *Fundamentals of educational research*. (2nd ed.). Madison, WI: Brown & Benchmark.
- Fienberg, S. E. (1971). Randomization and social affairs: The 1970 draft lottery. *Science*, 171, 255-261.
- Gall, M. D., Borg, W. R., Gall, J. P. (1996). *Educational research: An introduction*. (6th ed.). White Plains, NY: Longman.
- Krathwohl, D. R. (1993). *Methods of educational and social science research: An integrated approach*. NY: Longman.
- Linn, R. L. (1986). Quantitative methods in research on teaching. (3rd ed.). In M. C. Wittrock (Ed.), *Handbook of research on teaching*. NY: Macmillan.
- Mehta, C., & Patel, N. (1999). *StatXact 4 for windows: statistical software for exact nonparametric inference: User's manual*. Cambridge, MA: Cytel Software Corporation.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Notz, W. W., Staw, B. M., & Cook, T. D. (1971). Attitude toward troop withdrawal from Indochina as a function of draft number: Dissonance or self-interest? *Journal of Personality and Social Psychology*, 20, 118-126.
- Sawilowsky, 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.
- Strube, M. J. (1991). Small sample failure of random assignment: A further examination. *Journal of Consulting and Clinical Psychology*, 59, 346-350.
- Tuckman, B. W. (1994). *Conducting educational research*. (4th ed.). Fort Worth, TX: Harcourt Brace.
- Vockell, E. L., & Asher, J. W. (1995). *Educational research*. (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Appendix

Theoretically, there should have been 75 Type I errors, instead of the 33 obtained in the study. Nevertheless, these results are consistent with the literature, as Monte Carlo studies (e. g., Sawilowsky & Blair, 1992) noted that the t test generally becomes conservative when sample sizes are low and the underlying assumption of normality is violated. In fact, data sampled from the deMoivre (normal) distribution produced 37 variables where $\bar{x}_1 > \bar{x}_2$, and 35 variables where $\bar{x}_2 > \bar{x}_1$, for a total of 72 Type I errors, which is excellent agreement with the theoretical value.

This article relates to the validity of statistical findings, but not the statistical power of a test or the generalizability of results. The purpose of this demonstration is to show random assignment works even if $n_i = 2$. The use of the randomized two group experimental design with only $N = 4$ is not suggested. It should also be noted that the t test is the only statistic available that can be used with $N = 4$ and $\alpha = 0.01$.

Early Scholars
**A Comparison Of Bayesian And Frequentist Statistics
As Applied In A Simple Repeated Measures Example**

Jan Perkins
Program in Applied Experimental Psychology
Central Michigan University

Daniel Wang
Department of Mathematics
Central Michigan University

Clinicians see Bayesian and frequentist analysis in published research papers, and need a basic understanding of both. A repeated measures data set was analyzed using both approaches. Assumptions underlying each method and conclusions reached were contrasted. The Bayesian approach is a viable alternative to frequentist statistical analysis for many clinical projects.

Key words: Bayesian statistics; frequentist statistics, clinical research

---

Introduction

Classical or frequentist statistics is the standard method of analysis in clinical research. There is another statistical option, Bayesian analysis, with advocates arguing that it can be equally or more suited to the analysis of clinical research problems. In recent years increasing numbers of studies have appeared using Bayesian analysis or a combination of Bayesian and frequentist analyses, making it likely that health care clinicians will encounter papers written using a Bayesian approach, and that students will need some exposure to both methods. The purpose of this article is to compare the analysis and interpretation of a simple clinical data set using Bayesian and frequentist approaches as a simplified introduction to the Bayesian approach for clinicians without a background in statistics.

Jan Perkins is a doctoral student in Applied Experimental Psychology, and a Physical Therapy faculty member at Central Michigan University. E-mail at: jan.perkins@cmich.edu. Daniel Wang is an Assistant Professor of Statistics at Central Michigan University. He graduated from the University of Alabama with a Ph.D. in Applied Statistics in 1999.

Methodology

Bayesian analysis has developed from the work of Thomas Bayes, an eighteenth century British Presbyterian minister with an interest in probability theory (Brooks, 2001). His theorem is used in predicting probability. In itself, it is uncontroversial and commonly used in areas such as Mendelian genetics and computerized diagnosis (Lilford & Braunholtz, 1996). For such purposes it is used by statisticians of all backgrounds (Lee, 1989/1992). The application of Bayesian analysis in a broader sense is the source of debate and controversy. An explanation of some of the basic assumptions in these cases may help clarify why there is such heated debate.

Bayesian methods essentially construct probability distributions for unknown quantities of interest given the data, for example the probability that a particular Treatment A is superior to Treatment B given data from a trial. This probability is termed the posterior distribution and then used to reach conclusions about the research question. But in Bayesian analysis researchers are required to estimate a prior distribution for the event of interest in order to run the analysis of a data set. This prior distribution may be based on a variety of external evidence that includes controlled and uncontrolled studies, case reports, and expert

opinions. When comparing the two treatments mentioned above, the prior distribution is the probability that Treatment A is superior to Treatment B based on available information before data is collected. The actual data gathered in the study is considered the likelihood.

To state the application of Bayes Theorem in simplistic terms, the posterior probability distribution is proportional to the likelihood of the collected data multiplied by the prior distribution. The likelihood function and the prior function are combined into a distribution summing to 1 to create the posterior probability. All inferences about treatment difference are based on the posterior distribution. With continued data collection, it is possible later to revise the analysis and determine a new (and hopefully more precise) posterior distribution to use in conclusions regarding the superiority of Treatment A. Logically, accumulating evidence would ultimately also change the prior - moving it to a more realistic representation of reality. This updating of the prior distribution occurs as understanding of the phenomenon of interest changes in light of the evidence gathered.

Described in these terms the Bayesian approach has a commonsense appeal: it is possible to give probabilities, integrate information from multiple sources, and revise conclusions in light of new information. The process follows the classical model of scientific thinking and experimentation and is consequently attractive to those trained in the scientific method. Proponents of Bayesian analysis in clinical trials have argued that this makes it flexible and ethical, well suited for subgroup analysis, and offers a good option for ongoing analysis over the course of a trial (Spiegelhalter, Myles, Jones, & Abrams, 1999).

But an acceptable determination of prior distribution is one of the hardest things to do in complex situations, for example when there are conflicting opinions or studies or multiple subgroups to be considered. The incorporation of prior distributions is simultaneously considered the greatest flaw and greatest strength of Bayesian analysis, depending on one's perspective (O'Hagen, Luce & Fryback, 2003; Spiegelhalter et al. 1999). Bayesian calculations have also typically required

complex statistical computation power not readily accessible to most clinical researchers.

Bayesian statisticians are working on guidelines for weighting the prior distribution, with skeptical priors being useful if there are important reasons for caution (such as risks or costs of the new treatment), weak priors used when little is known, and optimistic priors being used at selected other times. Guidelines for prior specification are beginning to appear (Kadane, & Wolfson, 1996; Spiegelhalter et al. 1999). It is also possible to use a non-informative or uniform prior which essentially lets the data speak for itself (Box & Taio, 1973; Lee, 1989/1992). The data can of course be analyzed with a variety of priors for subsequent decision making, and indeed data can be collected before knowing the prior, but this demonstrates somewhat sloppy and unscientific thinking. If well done the process should follow the scientific model - the different priors resemble competing hypotheses which are to be tested by examining the data.

The approach in frequentist statistics is philosophically quite different. Probability is viewed as "a limiting ratio in a sequence of repeatable events . . . the ratio becoming ever more exact as the series is extended" (Howie, 2002, p. 1). Data is interpreted using statistical models based on frequencies, with the p-value being a measure of "discrepancy between the data and the null hypothesis" (Goodman, 1999, p. 997). This is very different from the Bayesian view of probability being a degree of belief or knowledge about the unknown. Contrary to common misinterpretations, the p-value does not give the probability of Treatment A being superior to Treatment B, but instead a predetermined level of significance test, set by balancing Type I and Type II errors, that allows acceptance or rejection of the data set based on its compatibility with the null hypothesis. The data are analyzed independently, without the influence of previous knowledge in the analysis, although previous knowledge is of use in planning the data collection. In other words, the classical inference methods treat parameters as constants, while Bayesian methods treat them as random variables.

A frequentist statistician would argue that the introduction of the prior in Bayesian

analysis introduces an element of subjectivity that is unacceptable. A Bayesian may counter that the decision to rank Type II errors as less important than Type I and to arbitrarily select a significance level is unscientific. A frequentist statistician may weight multiple tests of variables to reduce the risk of Type I error, selecting a technique for this from a variety of more or less accepted methods. A Bayesian would view an analysis that is more skeptical of Treatment A because you are also looking at other treatments or subgroups as ridiculous. It also could be argued that in frequentist analysis based on sampling, the analysis is only of value if the researcher has chosen the appropriate statistical model and if the data set fits all the assumptions of the chosen model.

If these conditions are not met, classical analysis can act to distort interpretation and the restrictions imposed by the model can exclude relevant information. Goodwin (1999) gave an excellent summary and explanation of issues relating to the use of frequentist and Bayesian analyses in health research.

The result of either type of analysis in uncomplicated situations where model assumptions are similar and where non-informative priors are used often leads to conclusions that are not much different, but at other times this may not be true. It is possible to reach very different conclusions from the same data set. For a general discussion of Bayesian and frequentist statistics with an emphasis on medical research see Matthews (2001a) and related discussion and response (Berger, 2001; Lindley, 2001; Matthews, 2001b; Sasieni, 2001), and an editorial and related articles in the *Annals of Internal Medicine* (Davidoff, 1999; Goodman, 1999). Specific illustrations of how Bayesian analysis can be useful in clinical trials are also readily located (Johns & Anderson, 1999; Lilford, 1999; Simon, 1999).

Problem to be Analyzed

The data set used in this article was generated as part of a student research project. As such it has been analyzed conventionally and prepared for journal submission. This exercise will not give study details but merely use the data set to illustrate Bayesian and frequentist approaches to data analysis and interpretation.

The study examined the short-term effect of a single stretching session on joint range of motion (ROM). Fifteen experimental group subjects were given the treatment (stretch). Measurements were taken at baseline, and at 1, 3, 6, 15 and 30 minutes post-stretch. Fifteen control group subjects were measured at similar time periods but not subjected to the treatment.

The question of interest was whether the stretching procedure altered the ROM at each of these time points and, if so, whether the stretch altered it more than the process of being measured. It was expected that the six measurements of ROM required in the protocol would affect ROM of the control group, but to a lesser extent. A comparison of the control and experimental groups would therefore be expected to show whether the stretch had any additional effect on ROM. Although Bayesian analysis has analogs to frequentist tests that produce p-values, it was decided to examine 90% confidence intervals and their analogous Bayesian probability intervals.

Results

SPSS for Windows, version 10.1 was used. For each group the baseline was used as an initial reference point with subsequent measures expressed as the difference from this point with the baseline measured being zero. A repeated measures General Linear Model (GLM) analysis was performed with time of measurement as the within subjects factor and group assignment (control vs. experimental) as between subject factor. This analytical model assumes that the measurements are drawn from a normally distributed population and that the different groups have homogeneous variances.

The p-value for testing no difference in the mean change of ROM between the control and experimental groups is 0.000, which leads to the conclusion that there is a difference. Based on the 90% confidence intervals for the estimated mean changes over the time, it is clear that the experimental group performs better than the control group because none of the 90% intervals overlap between groups. These are expressed in Tables 1 and 2, and illustrated graphically in Figure 1.

Table 1: Control group change in measurement from Baseline (Frequentist).

Time of Measurement	Mean	90% Confidence Interval	
		Lower Bound	Upper Bound
Baseline	0	-	-
One minute	1.33	-0.55	3.22
Three minutes	2.27	0.48	4.06
Six minutes	2.67	1.08	4.26
Nine minutes	2.67	0.94	4.39
Fifteen minutes	2.13	0.4	3.86
Thirty minutes	1.87	-0.46	4.2

Table 2: Experimental group change in measurement from Baseline (Frequentist).

Time of Measurement	Mean	90% Confidence Interval	
		Lower Bound	Upper Bound
Baseline	0	-	-
One minute	10.27	8.25	12.28
Three minutes	12.33	10.08	14.59
Six minutes	14.93	12.42	17.44
Nine minutes	14.33	11.88	16.79
Fifteen minutes	12.53	9.66	15.4
Thirty minutes	13.73	10.95	16.52

It is possible to state with reasonable confidence that the data gathered represent the underlying state of affairs. Thus, it could be concluded that the stretch produced an alteration in range of motion that is greater than that caused by the measurement technique.

There is, however, reason to be concerned about the analysis. The statistical model rests on a number of assumptions. If these assumptions are violated there is less faith in the conclusions. It is assumed subjects are a random sample from a pool of suitable subjects and that

the raw scores for them (and so the error terms) are normally distributed. It is also assumed that that error terms have a mean of zero and a common variance, and that error terms between and within the groups are not related. These assumptions are based on random assignment of subjects.

Figure 1: 90% Confidence Intervals Using Frequentist Analysis.

In addition there is a complex assumption known as the sphericity assumption related to variances in the fixed factor of the design. The general linear model procedure tests for sphericity using Mauchley's test. In this sample, the test concluded that the assumption was not met. This interpretation was based on using the conservative Greenhouse-Geiser adjustment.

Analysis with Bayesian Statistics

As with most Bayesian analyses, the choice of a prior distribution was critical. In this case there was limited previous evidence to use in creating a prior distribution. Published studies using the particular technique studied did not use the same joints, protocol or exact technique. Clinical experience suggested that there would be a modest increase in range in the experimental group that might or might not decline over the 30 minute period. Experienced clinicians could not offer more specific ideas about the effect of this single stretch treatment.

This limited evidence made it appropriate to use a non-informed prior distribution in the analysis.

The analysis was done using Gibbs sampling, a technique commonly used in Bayesian analysis. Gibbs sampling is a variant of Markov chain Monte Carlo (MCMC) analyses. This computer intensive technique provides researchers with repeated random data points drawn from the statistical distribution of interest. Parameters of interest are estimated by repeated iterations of the process until estimates converge. Gibbs sampling helps compensate for small data sets such as those generated in this experiment. For additional information on the Gibbs sampling technique see Casella and George (1992).

WinBUGS version 1.3 was used in the analysis adapting a dynamic model used in repeated measure research and described in Congdon (2001). The software program is available through the Bayesian inference Using Gibbs Sampling (BUGS) project <http://www.mrc-bsu.cam.ac.uk/bugs/>). Again, baseline measures were converted to zero and subsequent measures to differences from baseline. Bayesian means and 90% probability intervals were calculated. These are presented in Tables 3 and 4 and Figure 2.

Table 3: Control group change in measurement from Baseline (Bayesian).

Time of Measurement	Mean	90% Probability Interval	
		Lower Bound	Upper Bound
Baseline	0.21	-1.25	1.64
One minute	1.28	0.13	2.47
Three minutes	2.07	0.9	3.26
Six minutes	2.45	1.32	3.62
Nine minutes	2.45	1.37	3.67
Fifteen minutes	2.19	1.04	3.32
Thirty minutes	1.96	0.53	3.38

Table 4: Experimental group change in measurement from Baseline (Bayesian).

Time of Measurement	Mean	90% Probability Interval	
		Lower Bound	Upper Bound
Baseline	0.6	-1.57	2.9
One minute	9.35	7.25	11.36
Three minutes	12.64	10.76	14.56
Six minutes	14.76	12.92	16.7
Nine minutes	14.29	12.47	16.09
Fifteen minutes	12.96	11.08	14.76
Thirty minutes	13.61	11.35	15.73

Figure 2: 90% Probability Intervals Using Bayesian Analysis.

Interpretation was done through examination of the plots and data. Again, there is no overlap in the intervals except at the baseline, where this is expected. The Gibbs sampling technique used produces a baseline estimation, making it possible to give a probability interval for this as well as for the repeated measurement points. The results for estimating the mean change of ROM are very similar to the GLM results but the probability intervals are smaller than the confidence intervals and none of the probability intervals, other than the baseline, contains zero. The Bayesian analysis, like the general linear model,

assumes random sampling, normally distributed raw scores for the subjects and a linear relationship between scores and group and time variables. It also makes the important assumption of a non-informative prior.

Comparison of Analyses

In this simple example the conclusions reached with both analytical techniques appear quite similar in terms of clinical interpretation of results and related treatment planning. With this data set and a non-informative prior this is not surprising. Both types of analyses would lead to the practical clinical conclusion that the stretch altered range of motion for at least thirty minutes. In addition there was the expected observation that the repeated measurements did have an effect on ROM, albeit a smaller effect than stretch and measurement combined.

There are, however, some key differences in the interpretation of the results. In the frequentist analysis, the null hypotheses that were no differences in the mean change of ROM is rejected. This conclusion can be reached through the 90% confidence intervals for the mean change without considering any previous information about the mean change. On the other hand, in Bayesian analysis, the distribution of the mean change was estimated and the likelihood of the mean change in terms of the probability intervals calculated. With Bayesian analysis, it is allowed to utilize the previous knowledge about the distribution of parameters.

There are a few differences apparent that may lead to a preference for the Bayesian analysis for this study. The data set is small and does violate some of the assumptions behind the general linear model with repeated measures used in the frequentist analysis. The effect of this is to weaken faith in the conclusions.

The 90% confidence intervals with the general linear model are also wider in all but the one minute measurement in the experimental group analysis than the corresponding Bayesian 90% probability intervals. The width of confidence intervals in conventional analysis gives an estimate of precision with narrower widths desirable (Brooks, 2003). None of the post-baseline Bayesian probability intervals includes zero while two of the frequentist confidence intervals do in the control group,

despite an anticipated measurement effect. The smaller intervals in the Bayesian analysis reflect the strength of the sampling procedure and its ability to deal with small data sets. The Bayesian results are more compatible with clinical expectations based on muscle stretching theories. For these reasons the authors conclude that the Bayesian analysis seems to be the better analysis option with this particular data set.

There are additional advantages for a clinician who wants to continue data collection on stretching techniques but lacks facilities for large scale experimentation. The posterior distributions determined from this study could be used as informed priors in subsequent research, refining estimates and improving accuracy with additional data collection. This approach mimics the classic model of scientific reasoning. Assuming the clinician has access to computing resources and programs for Bayesian analysis, a series of small clinical studies could incrementally add to the body of research on the subject. The reasoning process in Bayesian analysis also has its attractions. Ashby and Smith (2000) argue that the Bayesian approach is the natural one for use in evidence-based practice where information must be synthesized and used in individual decision-making.

Conclusion

As computing power increases and statistical packages become more readily available and usable, Bayesian analysis may be seen more often in the medical and health literature used to guide clinical practice. It is now not uncommon to see articles in clinical journals that use Bayesian analysis either alone or in combination with frequentist analysis. This article gives an illustration of Bayesian and classical analysis applied to a simple clinical problem and the interpretation of results. In the example used, the authors concluded that they would prefer the Bayesian approach for analysis. Future studies such as simulating the power of two types of analysis in detecting the mean change of ROM would help clinicians understand the advantage of using classical statistics and Bayesian statistics.

Whatever approach is used in data analysis, it is important to recognize that there is

more than one approach. Bayesian analysis is being used in clinical studies to guide practice. In this paper Bayesian and frequentist statistical approaches are used to analyze a sample data set in order to contrast the two approaches and make clinicians aware of different approaches to data analysis.

References

- Ashby, D., & Smith, A. F. (2000). Evidence based medicine as Bayesian decision-making. *Statistics in Medicine*, 19, 3291-3205.
- Berger J. (2001). Why should clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference*, 94, 65-67.
- Box, G. E. P., & Taio, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.
- Brooks, E. B. (2001). Thomas Bayes. Retrieved November 5, 2003, from <http://www.umass.edu/wsp/acquiring/tales/bayes.html>
- Brooks, G. (2003). Advantages of confidence intervals in clinical research. *Cardiopulmonary Physical Therapy Journal*, 14 (3), 12-14.
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician*, 46, 167-174.
- Congdon, P. (2001). *Bayesian statistical modeling*. New York: John Wiley & Sons.
- Davidoff, F. (1999). Standing statistics right side up. *Annals of Internal Medicine*, 130, 1019-1021.
- Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine*, 130, 995-1004.
- Goodman S. N. (1999b). Toward evidence-based medical statistics. The Bayes factor. *Annals of Internal Medicine*, 130, 1005-1013.
- Howie, D. (2002). *Interpreting probability. Controversies and development in the early twentieth century*. Cambridge, England: Cambridge University.
- Johns, D., & Anderson, J. S. (1999). Use of predictive probabilities in Phase II and Phase III clinical trials. *Journal of Biopharmaceutical Statistics*, 9, 67-79.
- Kadane, J. B., & Wolfson, L. J. (1996). Priors for the design and analysis of clinical trials. In D. A. Berry & D. K. Stangl (Eds.), *Bayesian biostatistics* (pp. 157-184). New York: Marcel Dekker.
- Lee, P. M. (1989/1992). *Bayesian statistics: An introduction*. New York: Halsted Press.
- Lilford, R. J. (1999). Clinical trials and rare diseases: a way out of a conundrum. *BMJ*, 311, 1621-1625.
- Lilford, R. J., & Braunholtz D. (1996). For debate: The statistical basis of public policy: A paradigm shift is overdue. *BMJ*, 313, 603-607.
- Lindley, D. V. (2001). Why should clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference*, 94, 59-60.
- Matthews, R. A. J. (2001a). Why should clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference*, 94, 43-58.
- Matthews, R. A. J. (2001b). Author's response *Journal of Statistical Planning and Inference*, 94, 69-71.
- O'Hagen, A., Luce B. R., & Fryback, D. G. (2003). *A primer on Bayesian statistics in health economics and outcomes research*. Bethesda, MD: Medtab International.
- Sasieni, P. (2001). Why should clinicians care about Bayesian methods. *Journal of Statistical Planning and Inference*, 94, 61-63.
- Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics*, 55, 484-487.
- Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (1999). An introduction to Bayesian methods in health technology assessment. *BMJ*, 319, 508-512.

On The Reporting Of Reliability In Content Analysis

Patric R. Spence

Department of Communication
Wayne State University

This article explores one type of misreporting of reliability that has been seen in recent conference papers and articles using the method of content analysis. The reporting of reliability is central to the validity of claims made using this method. A brief overview of content analysis is offered, followed by the exploration of one type of misreporting of reliability. Suggestions are offered to address the problem.

Key words: Content analysis, intercoder reliability

Introduction

Though many definitions of content analysis have been offered over the years (Berelson, 1952; Weber, 1990; Berger, 1991), a complete and concise contemporary definition is offered by Neuendorf (2002), who defines it as “summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a prior design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited to the types of variables that may be measured or the context in which the messages are created or presented” (p. 10).

Content analysis is used for numerous purposes in several fields of study. Examples include; settling disputed authorships (Berelson, 1952), during World War II, the technique was employed to gather information from enemy literature (George, 1959), rule making among jury members (Seibold, 1998), interactions in adolescent peer groups (England & Petro, 1998), advertising in children's television (Stern & Harmon, 1984), the role of face in organizational relationships (Redding & Ng, 1982), minority

advertising in children's television (Stern & Harmon, 1984), the role of face in organizational relationships (Redding & Ng, 1982), minority representation on television (Tamborini et al., 2000) and several accounts of media topics.

More specifically, in education the method has been used to research issues such as, the press as a resource for teaching science (Dimopoulos, Koulaidis & Sklaveniti, 2003), the treatment of gender in teacher education textbooks (Zittleman & Sadker, 2002), and materials in specific textbooks (Harmon, Hedrick & Fox, 2000; Plucker & Beghetto, 2000).

Content analysis is a popular method used in the behavior sciences because of its ability to be utilized for both written and oral communication as well as its ability to compare data across time and context. The method allows the researcher to identify particular words, phrases or concepts within the text(s) being examined. The text(s) that are used can be transcripts of communication, classroom interactions, historical documents, newspaper articles, magazine articles, books, interviews, essays, speeches, and almost any behavioral event that is recorded in some manner.

The importance of intercoder reliability is of central concern when content analysis is used. Intercoder reliability is “the extent to which independent coders evaluate a characteristic of a message or artifact and reach the same conclusion” (Lombard et al., 2002, p. 589). This provides a validation to the coding

---

Patric R. Spence (M. A., Michigan State University) is a graduate student Department of Communication at Wayne State University. Email: spencepa@wayne.edu.

scheme. Thus, intercoder reliability demonstrates that more than one person can use the coding scheme and obtain similar results. The validity of the data and any subsequent interpretations are suspect if intercoder reliability is not established or reported. Further, not only does the establishment of intercoder reliability help ensure validity, but it also allows the work of coding to be distributed among multiple coders (Neuendorf, 2002).

Much of the concern within the method is whether separate coders achieve agreement on the values assigned to an examined data point. The simplest method of assessing reliability between coders is a percent agreement. This statistic represents the number of between coder agreements divided by the total measures observed. Percent agreement is the most common measure of intercoder reliability; however, while it is intuitively appealing and simple to calculate, it is a misleading measure that overestimates the true score. The statistic has a range from .00 (no agreement) to 1.00 (perfect agreement).

$$PA_o = A / n \quad (1)$$

PA_o concerns the proportion agreement, observed, where A is the number of agreements between the two coders and n represents the total number of units the coders have coded (Neuendorf, 2002).

Cohen's *kappa* (1968) is the most popular reliability assessment used (Zwick, 1988), particularly because of its accessibility in SPSS. The kappa accounts for the role of chance in agreements in coding which the percent agreement does not. However, it is only used for nominal level variables. The kappa's range is from .00 (agreement at chance level) to 1.00 (perfect agreement), a value that is less than .00 illustrates an agreement that is less than chance.

$$\frac{PA_o - PA_E}{1 - PA_E} \quad (2)$$

PA_o concerns the proportion agreement, observed, and PA_E refers to the proportion agreement that is expected by chance (Neuendorf, 2002).

Some other measures of reliability include Krippendorff's *alpha* (Krippendorff, 1980), Scott's *pi* (1955) and Lin's concordance correlation coefficient (Lin, 1989), each of which have their own advantages and disadvantages.

Although there "is no simple right way to do content analysis" (Weber, 1990, p. 13) most have the following elements in common. After the research question is asked a decision needs to be made on what will be analyzed or what social artifacts will be studied. Then a decision needs to be made on the unit of analysis. Following this a categorical system needs to be developed in which the responses can be filled. Next, it needs to be determined how the data will be coded. It is a good idea to take a sample or even do a pilot study to determine if the coding structure needs to be modified.

Methodology

Recently, some researchers have used a more uncommon coding scheme that entails multiple steps in coding. In the scheme coders first code a variable in a context for its presence (variable A). If in the experimental condition variable A exists the coders then look for or categorize a next variable (variable B). The process can either stop at this point or continue. Therefore, the process of coding the second variable is contingent upon the existence of the first variable.

Consider a hypothetical study examining aggressive behaviors of children in a classroom. Variable A is a particular instance, and can be observed through video taping, in class observation or vignette. In this situation there are two coders examining the interactions (coder 1 and 2). The coders either code the behavior as (1) not aggressive or (2) aggressive. After the experiment the results of the coders are compared for intercoder reliability. This is demonstrated in table 1 (C1va and C2va). Using Cohen's kappa, the intercoder reliability is .83; no problems exist in the reporting thus far.

Consider that the next behavior coded is dependent (contingent) upon whether or not the first behavior was identified as aggressive. Thus, if the behavior in the condition was (2)

aggressive, was it (1) physical aggression or (2) verbal aggression? The coding process continues but the analysis is dependent upon the first code. It is at this point in reporting the results that a reliability reporting bias can occur. In table 1 (C1va and C1v2) the reporting of the behavior can be seen. There are 47 agreements between the coders and 4 disagreements, producing a kappa of .83. This represents excellent agreement beyond the role of chance (Banerjee et al. 1999).

Three instances exist however, where one of the coders moved on to coding the type of physical aggression (variable B) while the second coder did not. When reporting the reliability of variable B the researcher must include the non-agreements from variable A in order to give the reader an accurate assessment of the intercoder reliability. This does not always happen. Increasingly authors report the reliability without the addition of the non-agreements from the first variable under examination, which inflates reliability.

Consider in table 1 (C1vb, C2vb, C1vb2 and C1vb2) the reporting of variable B (type of physical aggression). In this situation there are 21 instances of aggressive behavior coded from the first condition. Coder 1 and 2 agreed on the type of the aggression in 18 of the 21 instances. If the researcher fails to include the non-agreements from the examination of variable A the reliability in the condition is .70: still a good measure of reliability beyond chance.

Compare those results to table 1 (C1vb3 and C2vb3) where the researcher includes the first wave of reliability assessments. The agreement is 18 out of 25 cases, producing a kappa of .49. This is considerably lower, and it is considered poor agreement beyond chance (Banerjee et al. 1999). Moreover, consider what

would be the case if this continued in a study and the author failed to include the non-agreements for variables C, D, and E. In reporting the reliabilities for variable E, the reported score would be far removed from the true value.

Conclusion

A few suggestions follow concerning this problem. The first and simplest is for the researcher to report the reliability with the inclusion of all coded responses as was done in table 1 (C1vb3 and C2vb3). When this is done the reader has an accurate assessment of the true score concerning the reliability and can have more confidence in the conclusions the data support.

If a researcher believes that due to some aspect of the research design the inclusion of the non-agreements from the first condition is unwarranted, then he or she should outline the reason behind the exclusion of the non-agreements in the results section of the article or paper. Accompanying this should be the scores from each coder and an explanation indicating that the previous condition produced X number of agreements that is not calculated in the reliability kappa. Another alternative is for the researcher may include both reliability scores within the results.

The above example used only a few instances of disagreement between the coders. In a study that has more disagreement the reporting bias can be larger. Although there are no rules explaining exactly how a researcher should report reliability, care needs to be taken in reporting and the author needs to justify the use of any reporting scheme.

References

- Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27, 2-23.
- Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL: Free Press.
- Berger, A. A. (1991). *Media research techniques*. Newbury Park, CA: Sage.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, 70, 213-220.
- Dimopoulos, K., Koulaidis, V., & Sklaveniti, S. (2003). Towards an analysis of visual images in school science textbooks and press articles about science and technology. *Research in Science Education*, 33, 189 – 216.
- Harmon, J. M., Hedrick, W. B., & Fox, E. A. (2000). A content analysis of vocabulary instruction in social studies textbooks for grades K-8. *The Elementary School Journal*, 100, 253-271.
- Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting intercoder reliability. *Human Communication Research*, 28, 587-604.
- England, E. M., & Petro, K. D. (1988). Middle school students' perceptions of peer groups: Relative judgments about group characteristics. *The Journal of Early Adolescence*, 18, 349-373.
- George, A. L. (1959). *Propaganda analysis*. Evanston: Row and Peterson.
- Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. CA: Sage.
- Neuendorf, K.A. (2001). *The content Analysis Guidebook*. London: Sage Publications.
- Plucker, J. A., & Beghetto, R. A. (2000). Needles in haystacks or field of plenty? A content analysis of popular creativity texts. *Gifted Child Quarterly*, 44, 135-138.
- Redding, S. G., & Ng, M. (1982). The role of 'face' in the organizational perceptions of Chinese managers. *Organizational Studies*, 3, 201-219.
- Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.
- Seibold, D. R. (1998). Jurors intuitive rules for deliberation: A structural approach to communication in jury decision making. *Communication Monographs*, 65, 282-307.
- Stern, B. L., & Harmon, R. R. (1998). The incidence and characteristics of disclaimers in children's television advertising. *Journal of Advertising*, 13, 12-17.
- Tamboini, R., Mastro, D. E., Chory-Assad, R. M., Huang, R. H. (2000). The color and crime of the court: A content analysis of minority representation. *Journalism and Mass Communication Quarterly*, 77, 639-653.
- Weber, R. P. (1990). *Basic content analysis*. CA: Sage.
- Zittleman, K., & Sadker, D. (2002). Gender bias in teacher education texts. *Journal of Teacher Education*, 53 (2), 168-180.
- Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374-378.

JMASM Algorithms and Code JMASM10: A Fortran Routine For Sieve Bootstrap Prediction Intervals

Andrés M. Alonso
Department of Mathematics
Universidad Autónoma de Madrid

A Fortran routine for constructing nonparametric prediction intervals for a general class of linear processes is described. The approach uses the sieve bootstrap procedure of Bühlmann (1997) based on residual resampling from an autoregressive approximation to the given process.

Key words: Sieve bootstrap, prediction, time series

Introduction

When studying a time series, one of the goals is the estimation of forecast confidence intervals based on an observed trajectory of the process. The traditional approach of finding prediction intervals for a linear time series assumes that the distribution of the error process is known. Thus, these prediction intervals could be adversely affected by departures from the true underlying distribution.

Some bootstrap approaches have been proposed as a distribution free alternative to compute prediction intervals. Stine (1987) proposed a bootstrap method to estimate the prediction mean squared error of the estimated linear predictor of an $AR(p)$ where p is known. Also, for an $AR(p)$ process with known p , and relaxing the assumptions of Stine (1987), Thombs and Schucany (1990) propose a backward then forward bootstrap method to estimate prediction intervals. Cao *et al.* (1997) study a conditional bootstrap method alternative to Thombs and Schucany's proposal, which is computationally faster. Pascual *et al.* (2001) generalize this conditional bootstrap to $ARMA(p, q)$ processes with known p and q and

also include the parameter estimation variability.

This article describes a bootstrap method to construct nonparametric prediction intervals for a class of linear processes that can be written as a one-sided infinite-order moving average process with at most a polynomial decay of the coefficients $\{\psi_j\}_{j=0}^{+\infty}$. This class includes the stationary and invertible $ARMA(p, q)$ processes. This approach uses the sieve bootstrap of Bühlmann (1997) based on residual resampling from a sequence of approximating autoregressions for $\{X_t\}_{t \in \mathbb{Z}}$ with order $p = p(n)$ that increases as a function of the sample size n .

This sieve bootstrap has a nice nonparametric property, being model-free within the considered class of linear processes. Thus, the proposed bootstrap prediction intervals could be applied to this more general class of linear models without specifying a finite dimensional model as in previous bootstrap proposals. Alonso *et al.* (2002) and (2003) studied the consistency and the finite sample properties of this sieve bootstrap.

Methodology

Let $\{X_t\}_{t \in \mathbb{Z}}$ be a real valued, stationary process with expectation $E[X_t] = \mu_X$ that admits a $MA(\infty)$ representation with $\sum_{j=0}^{+\infty} \psi_j^2 < \infty$. Under the additional assumption of invertibility $\{X_t\}_{t \in \mathbb{Z}}$ can be represented as a one-sided infinite-order autoregressive process:

Andrés M. Alonso is Assistant Professor of Statistics. Department of Mathematics, Universidad Autónoma de Madrid, 28049 Madrid, Spain. His areas of research interest are statistical computing, resampling methods and biostatistics. E-mail at: andres.alonso@uam.es.

$$\sum_{j=0}^{+\infty} \phi_j (X_{t-j} - \mu_X) = \varepsilon_t, \quad \phi_0 = 1, t \in \mathbf{Z} \quad (1)$$

with coefficients $\{\phi_j\}_{j=0}^{+\infty}$ satisfying $\sum_{j=0}^{+\infty} \phi_j^2 < \infty$. This AR(∞) representation motivates Bühlmann's sieve bootstrap. The method proceeds as follows:

1. Given a sample $\{X_1, X_2, \dots, X_n\}$, select the order $p = p(n)$ of the autoregressive approximation by AICC criterion: $\text{AICC} = -n \log(\sigma^2) + 2(p+1)n/(n-p-2)$, (cf. Section 9.3 of Brockwell & Davis, 1991).

The AICC criterion is a bias-corrected version of AIC (Akaike, 1973), and it has a more extreme penalty for large-order models which counteracts the overfitting nature of AIC. Other order selection criteria (such as BIC) could be used, but AICC is preferred assuming the view that the true model is complex and not of finite dimension, and also because the AICC is asymptotically efficient for autoregressive models, i.e., it chooses an AR model which achieves the optimal rate of convergence of the mean-square prediction error.

2. Construct some estimators of the autoregressive coefficients: $(\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p)$. Following Bühlmann (1997) the Yule-Walker estimates are taken.

3. Compute the residuals for $t \in (p+1, p+2, \dots, n)$ by:

$$\hat{\varepsilon}_t = \sum_{j=0}^p \hat{\phi}_j (X_{t-j} - \bar{X}), \quad \hat{\phi}_0 = 1. \quad (2)$$

4. Define the empirical distribution function of the centered residuals:

$$\hat{F}_{\tilde{\varepsilon}}(x) = (n-p)^{-1} \sum_{t=p+1}^n 1_{\{\tilde{\varepsilon}_t \leq x\}}, \quad (3)$$

where $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - \hat{\varepsilon}^{(\bullet)}$ and

$$\hat{\varepsilon}^{(\bullet)} = (n-p)^{-1} \sum_{t=p+1}^n \hat{\varepsilon}_t.$$

5. Draw a resample ε_t^* of i.i.d. observations from $\hat{F}_{\tilde{\varepsilon}}$.

6. Define X_t^* by the recursion:

$$\sum_{j=0}^p \phi_j (X_{t-j}^* - \bar{X}) = \varepsilon_t^*, \quad (4)$$

where the starting p observations are equal to $\bar{X}$.

In practice an AR(p) resample is generated using (4) with sample size equal to $n + 100$ and then discard the first 100 observations. Up to this step, the resampling plan coincides with the sieve bootstrap, and is valid for bootstrapping some statistics defined as a functional of a m -dimensional distribution function (see details in Section 3.3 of Bühlmann, 1997). However, it is not effective for bootstrap prediction, because it does not replicate the conditional distribution of X_{n+h} given the observed data. But, proceeding as do Cao *et al.* (1997) by fixing the last p observations resamples of the future values can be obtained

$$X_{n+h}^* \text{ given } X_{n-p+1}^* = X_{n-p+1}, \\ X_{n-p+2}^* = X_{n-p+2}, \dots, X_n^* = X_n.$$

7. Compute the estimation of the autoregressive coefficients: $(\hat{\phi}_1^*, \hat{\phi}_2^*, \dots, \hat{\phi}_p^*)$ as in step 1.

8. Compute the future bootstrap observations by the recursion:

$$X_{n+h}^* = \bar{X} - \sum_{j=1}^p \hat{\phi}_j^* (X_{n-j}^* - \bar{X}) + \varepsilon_t^*, \quad (5)$$

where $h > 0$, and $X_t^* = X_t$, for $t \leq n$.

Finally, $F_{X_{n+h}^*}^*(x)$ the bootstrap distribution of X_{n+h}^* is used to approximate the unknown distribution of X_{n+h} given the observed sample. As usual, a Monte Carlo estimate $\hat{F}_{X_{n+h}^*}^*(x)$ is obtained by repeating the steps 5 to 8 B times. The $(1-\alpha)\%$ prediction interval for X_{n+h} is given by $[Q^*(\alpha/2), Q^*(1-\alpha/2)]$, where $Q^*(\cdot) = \hat{F}_{X_{n+h}^*}^*(\cdot)$ are the quantiles of the estimated bootstrap distribution.

Fortran routines

Module TimeSeriesRoutines

In the module TimeSeriesRoutines are presented some routines required for the sieve bootstrap procedure: subroutine AutoCovarianceVector, subroutine YuleWalker, and subroutine AICCSelction.

```
SUBROUTINE AutoCovarianceVector(
  ACVector, XSeries, MaxLag, Positions)
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(0:),
  INTENT(OUT) :: ACVector
  REAL (KIND=8), DIMENSION(:),
  INTENT(IN) :: XSeries
  INTEGER, INTENT(IN) :: MaxLag
  INTEGER, DIMENSION(:), INTENT(IN),
  OPTIONAL :: Positions
```

This routine estimates the autocovariances of the XSeries for the orders from 0 to MaxLag. Notice that the implementation allows possible missing observations in the specified Positions. The expression for the autocovariance estimates is given by:

$$\hat{\gamma}_k = \frac{1}{n-m} \sum_{t=1}^{n-k} w_t w_{t+k} (X_t - \bar{X})(X_{t+k} - \bar{X}), \quad (6)$$

where m is the number of missing observations, $\bar{X} = (n-m)^{-1} \sum_{t=1}^n w_t X_t$ and w_t is equal 0 if the observation t is missing and otherwise is equal to 1.

```
SUBROUTINE YuleWalker(XSeries,
  ACMatrix, YWPhi, Residuals)
  USE Msims1
  USE Imslf90
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(:),
  INTENT(IN) :: XSeries
  REAL (KIND=8), DIMENSION(:,:),
  INTENT(IN) :: ACMatrix
  REAL (KIND=8), DIMENSION(:),
  INTENT(OUT) :: YWPhi
  REAL (KIND=8), DIMENSION(:),
  INTENT(OUT) :: Residuals
```

This routine calculates the Yule-Walker estimates of the autoregressive coefficient required in the steps 2 and 7 of sieve bootstrap procedure. It also calculates the residuals for the estimated model. The Yule-Walker estimators can be obtained from the following relation (cf. Section 8.1 of Brockwell and Davis (1991)):

$$\hat{\Gamma}_p \hat{\phi}_p = \hat{\gamma}_p, \quad (7)$$

where $\hat{\Gamma}_p$ is the estimated autocovariance matrix $[\hat{\gamma}_{i-j}]_{i,j=1}^p$, $\hat{\gamma}_p = (\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_p)'$ and $\hat{\phi}_p = (\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p)'$ is the coefficients vector. Using (2), the estimated residuals were obtained.

```
SUBROUTINE AICCSelction(XSeries,
  ACVector, PMax, PHat)
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(:),
  INTENT(IN) :: XSeries
  REAL (KIND=8), DIMENSION(0:),
  INTENT(IN) :: ACVector
  INTEGER, INTENT(IN) :: PMax
  INTEGER, INTENT(OUT) :: PHat
```

This routine implements the AICC method for selecting the order of the autoregressive model for XSeries. It considers models from $p = 0$ to $p = PMax$. Instead of using the subroutine YuleWalker for the different values of p , it uses the Durbin-Levinson algorithm (cf. Section 8.2 of Brockwell and Davis (1991)) which avoids the matrix inversion required in the direct computation of $\hat{\phi}_p$. The Durbin-Levinson algorithm uses the following recursions:

$$\hat{\phi}_{m,m} = (\hat{\gamma}_m - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} \hat{\gamma}_{m-j}) / \hat{v}_{m-1}, \quad (8)$$

$$\begin{bmatrix} \hat{\phi}_{m,1} \\ \hat{\phi}_{m,2} \\ \vdots \\ \hat{\phi}_{m,m-1} \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{m-1,1} \\ \hat{\phi}_{m-1,2} \\ \vdots \\ \hat{\phi}_{m-1,m-1} \end{bmatrix} - \hat{\phi}_{m,m} \begin{bmatrix} \hat{\phi}_{m-1,m-1} \\ \hat{\phi}_{m-1,m-2} \\ \vdots \\ \hat{\phi}_{m-1,1} \end{bmatrix}, \quad (9)$$

and

$$\hat{v}_m = \hat{v}_{m-1} (1 - \hat{\phi}_{m,m}^2), \quad (10)$$

with the following initial values: $\hat{\phi}_{1,1} = \hat{\gamma}_1 / \hat{\gamma}_0$
and $\hat{v}_1 = \hat{\gamma}_0(1 - \hat{\phi}_{1,1}^2)$.

Notice that the subroutine AICCSelection can be easily modified in order to use other information criterion as AIC or BIC. Only the two following sentences required some minor changes:

```
MinimumAIC =
RXSize*LOG(ACVector(0)) +
2.0D0*(REAL(I,KIND=8)+1.0D0)
*RXSize/(RXSize - REAL(I+2,KIND=8))

WorkAIC =
RXSize*LOG(VarianceVector(I))
+2.0D0*(REAL(I,KIND=8)+1.0D0)
*RXSize/(RXSize - REAL(I+2,KIND=8))
```

Routine FESieves

Here are described the subroutine FESieves which implements the steps 2 to 8 of the sieve bootstrap procedure. Notice that the step 1 is implemented by subroutine AICCSelection.

```
SUBROUTINE
FESieves(EDF,XSeries,PHat)
USE Msimsl
USE Imsl90
USE TimeSeries
IMPLICIT NONE
REAL (KIND=8), DIMENSION (:,:),
INTENT(OUT) :: EDF
REAL (KIND=8), DIMENSION (:),
INTENT(IN) :: XSeries
INTEGER, INTENT(IN) :: PHat
```

The inputs of subroutine FESieves are: the sample $XSeries = \{X_1, X_2, \dots, X_n\}$ and the selected order, PHat. The output is a $MaxLag \times B$ matrix, where MaxLag is the maximum prediction horizon to be considered and B is the number of resamples.

Step 2 and 7 are implemented by the following sentences:

```
CALL YuleWalker(XSeries,
ACMatrix(1:PHat+1, 1:PHat+1),
YWPhi, Residuals)
```

```
CALL YuleWalker(WSeries(101:XSize +
100), WACMatrix(1:PHat+1,
1:PHat+1), WYWPhi, YWResiduals)
```

where the WSeries are the resample obtained using recursion (4). The estimates YWPhi are used in recursion (4) and the bootstrap estimates WYWPhi are used in recursion (5). Also, in the first call to subroutine YuleWalker, the step 3 is performed. As mentioned in the previous section, a bootstrap resample was generated using (4) with sample size equal to XSize+100 and then discard the first 100 observations by WSeries(101:XSize + 100).

The resamples of step 5 are obtained by sampling with replacement from the vector of centered residuals, WResiduals = WResiduals - SUM(WResiduals) / REAL(XSize - PHat, KIND=8):

```
DO I = 1, XSize + 100 + MaxLag ! a
resample of centered residual
CALL RNUND(1, XSize-PHat,
RandomIndex)
RResiduals(I)=
WResiduals(RandomIndex)
END DO
```

Because recursions (4) and (5) are similar, here, it is only described the prediction recursion:

```
DO I = XSize+101, XSize+100+MaxLag
WSeries(I) = RResiduals(I)
DO Ip = 1, PHat
WSeries(I) = WSeries(I) +
WYWPhi(Ip)*WSeries(I-Ip)
END DO
END DO
EDF(1:MaxLag,J)=WSeries(XSize+101:
XSize+100+MaxLag) + XMean
```

Table 1. Simulation results for Model 1.

Lag	Sample size	Method	Coverage (se)	Cov. (below /above)	Length (se)
h	n	Theoretical	95%	2.50% / 2.50%	3.92
1	25	Bootstrap	89.03 (0.82)	4.44 / 6.53	3.74 (0.07)
	50		92.59 (0.52)	4.25 / 3.16	3.86 (0.05)
	100		93.77 (0.33)	3.25 / 2.98	3.90 (0.04)
h	n	Theoretical	95%	2.50% / 2.50%	4.92
3	25	Bootstrap	87.50 (0.86)	5.41 / 7.09	4.30 (0.08)
	50		92.08 (0.49)	3.97 / 3.95	4.69 (0.05)
	100		93.21 (0.38)	3.53 / 3.26	4.77 (0.05)

Table 2. Simulation results for Model 2.

Lag	Sample size	Method	Coverage (se)	Cov. (below /above)	Length (se)
h	n	Theoretical	95%	2.50% / 2.50%	3.93
1	25	Bootstrap	89.53 (0.85)	5.72 / 4.75	4.12 (0.08)
	50		92.06 (0.62)	3.63 / 4.31	3.98 (0.06)
	100		93.31 (0.43)	3.49 / 3.20	3.96 (0.04)
h	n	Theoretical	95%	2.50% / 2.50%	4.93
3	25	Bootstrap	89.19 (0.79)	5.15 / 5.66	4.52 (0.09)
	50		91.50 (0.58)	3.85 / 4.65	4.62 (0.06)
	100		92.49 (0.39)	3.19 / 4.32	4.68 (0.05)

Table 3. Simulation results for Model 3.

Lag	Sample size	Method	Coverage (se)	Cov. (below /above)	Length (se)
h	n	Theoretical	95%	2.50% / 2.50%	3.79
1	25	Bootstrap	89.45 (0.66)	4.73 / 5.82	3.54 (0.06)
	50		92.44 (0.45)	4.19 / 3.37	3.62 (0.04)
	100		93.77 (0.36)	3.38 / 2.85	3.74 (0.04)
h	n	Theoretical	95%	2.50% / 2.50%	3.93
3	25	Bootstrap	89.20 (0.65)	4.90 / 5.90	3.58 (0.06)
	50		92.79 (0.39)	3.68 / 3.53	3.75 (0.05)
	100		93.84 (0.34)	3.03 / 3.13	3.88 (0.04)

Notice that in (5) the bootstrap autoregressive coefficient is used, WYWPhi, this allows us to incorporate the parameter estimation variability in the prediction intervals.

Finally, the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the empirical density of forecasts, EDF, constitutes the prediction interval.

Results

In this section are briefly described the results of a simulation experiment using the Fortran subroutine presented in the previous section. The following models are used:

- Model 1: $X_t = 0.75 X_{t-1} - 0.5 X_{t-2} + \varepsilon_t$, where ε_t are i.i.d. $N(0,1)$.
- Model 2: $X_t = \varepsilon_t - 0.3 \varepsilon_{t-1} + 0.7 \varepsilon_{t-2}$, where ε_t are i.i.d. $N(0,1)$.

- Model 3: X_t is a Gaussian process with autocovariance generating function equal to $G(z) = \sum_{k=-\infty}^{+\infty} \gamma_k z^k$, where $\gamma_k = (|k| + 1)^{-3}$.

The autoregressive model was considered by Cao *et al.* (1997), the moving average model 2 by Pascual *et al.* (2001) and the model 3 by Alonso *et al.* (2002). Notice that neither model 2 nor model 3 admit a finite AR representation. Moreover, model 3 does not have an ARMA representation.

To evaluate the prediction intervals, their mean coverage and length are used and the proportions of observations lying out to the left and to the right of the interval. These quantities are estimated as follows:

a) For a combination of model, sample size and error distribution, simulate a series, and generate $R = 1000$ future values X_{n+h} .

b) For the bootstrap procedure obtain the $(1-\alpha)$ prediction interval based on $B = 1000$ bootstrap resamples.

c) The coverage is estimated as $C = \# \{Q^*(\alpha/2) \leq X_{n+h}^r \leq Q^*(1-\alpha/2)\} / R$,

where X_{n+h}^r with $r = 1, 2, \dots, R$ are the R future values generated in step a).

In steps a) and b) the “theoretical” and bootstrap interval lengths are obtained using

$$L_T = X_{n+h}^{\lceil R(1-\alpha/2) \rceil} - X_{n+h}^{\lceil R\alpha/2 \rceil}$$

and

$$L_B = Q^*(1-\alpha/2) - Q^*(\alpha/2),$$

respectively. Finally, the steps a) – c) are repeated 100 times.

The results are presented in Tables 1 – 3, using three sample sizes $n = 25, 50$ and 100 , nominal coverage 95% and the prediction lag $h = 1$ and 3 . Essentially, similar results are obtained in all cases. Sieve bootstrap performs reasonably well in all considered models since the mean coverage and length tend to the nominal values as the sample size grows. Notice that for models 2 and 3 the sieve bootstrap never uses the correct model. The running time for these three experiments (using a Pentium 4, running at 2.66GHz) was 22.92, 24.40 and 27.82 seconds, respectively.

Conclusion

It has been shown by Alonso *et al.* (2002) and (2003) that, for general linear process, if an AR approximation that grows with the sample size is used, it can derive a bootstrap for building prediction intervals that has the two following properties: first, the procedure is consistent, that is, it generates as prediction a random variable that converges in conditional distribution to the concerning variable; second, Monte Carlo simulations show that the proposed procedure provides better coverage results than previous methods in general cases. This article describes a Fortran routine that implement this sieve bootstrap prediction procedure. Additional simulation experiments confirm the correct behavior of the proposed procedure in finite samples.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In: *2nd International Symposium on Information Theory*, B.N. Petrov, and F. Csaki, Eds., Akademiai Kiado, Budapest, 267-281.
- Alonso, A. M., Romo, J. & Peña, D. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, 100, 1-11.
- Alonso, A. M., Romo, J. & Peña, D. (2003). On sieve bootstrap prediction intervals. *Statistics and Probability Letters*, 65, 13-20.
- Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Theory and Methods*, Springer-Verlag, New York.
- Bühlmann, P. (1997). Sieve bootstrap for time series, *Bernoulli*, 3, 123-148.
- Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J. M. & García-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes, *Communications in Statistics. Theory and Methods*, 26, 961-978.
- Pascual, L., Romo, J. & Ruiz, E. (2001). Effects of parameter estimation on prediction densities: a bootstrap approach, *International Journal of Forecasting*, 17, 83-103.

Stine, R. A. (1987). Estimating properties of autoregressive forecasts, *Journal of the American Statistical Association*, 82, 1072-1078.

Thombs, L. A. & Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression, *Journal of the American Statistical Association*, 85, 486-492

Appendix I – Module TimeSeriesRoutines

```

MODULE TimeSeriesRoutines
  REAL (KIND=8), DIMENSION(:), ALLOCATABLE :: ZSeries
  INTEGER :: p, d, q, ps, ds, qs, season

CONTAINS

SUBROUTINE AutoCovarianceVector(ACVector,XSeries,MaxLag,Positions)
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(0:), INTENT(OUT) :: ACVector
  REAL (KIND=8), DIMENSION(:), INTENT(IN) :: XSeries
  INTEGER, INTENT(IN) :: MaxLag
  INTEGER, DIMENSION(:), INTENT(IN), OPTIONAL :: Positions
  ! Local variables
  INTEGER :: K, I, J, XSize, NMissings
  REAL (KIND=8) :: RXSize, XMean
  REAL (KIND=8), DIMENSION(:), ALLOCATABLE :: Weights

  ! First executable statement

  XSize = SIZE(XSeries, 1)
  ALLOCATE(Weights(XSize))
  Weights = 1.0D0

  IF (PRESENT(Positions)) THEN
    Weights(Positions) = 0.0D0
    NMissings = SIZE(Positions, 1)
    RXSize = REAL(XSize - NMissings, KIND=8)
  ELSE
    RXSize = REAL(XSize, KIND=8)
  END IF

  XMean = SUM(XSeries*Weights)/RXSize
  DO K = 0, MaxLag
    ACVector(K) = DOT_PRODUCT(&
      (XSeries((K+1):XSize) - XMean)*Weights((K+1):XSize), &
      (XSeries(1:(XSize-K)) - XMean)*Weights(1:(XSize-K)))/RXSize
  END DO

  DEALLOCATE(Weights)
END SUBROUTINE AutoCovarianceVector

SUBROUTINE AutoCovarianceMatrix(ACMatrix,XSeries,MaxLag,MSize)
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(:, :), INTENT(OUT) :: ACMatrix
  REAL (KIND=8), DIMENSION(:), INTENT(IN) :: XSeries
  INTEGER, INTENT(IN) :: MaxLag, MSize
  ! Local variables
  INTEGER :: K, I, J, XSize
  REAL (KIND=8) :: RXSize, XMean

```

```

REAL (KIND=8), DIMENSION(0:MSize) :: ACVector

! First executable statement

XSize = SIZE(XSeries, 1)
RXSize = REAL(XSize, KIND=8)
XMean = SUM(XSeries)/RXSize
DO K = 0, MaxLag+1
    ACVector(K) = DOT_PRODUCT(XSeries((K+1):XSize) - XMean, &
        XSeries(1:(XSize-K)) - XMean)/RXSize
END DO
DO I = 1, MaxLag+1
    DO J = 1, MaxLag+1
        ACMatrix(I,J) = ACVector(ABS(I-J))
    END DO
END DO
END SUBROUTINE AutoCovarianceMatrix

SUBROUTINE YuleWalker(XSeries,ACMatrix,YWPhi,Residuals)
    USE Msimsl
    USE Imsl90
    IMPLICIT NONE
    REAL (KIND=8), DIMENSION(:), INTENT(IN) :: XSeries
    REAL (KIND=8), DIMENSION(:,:), INTENT(IN) :: ACMatrix
    REAL (KIND=8), DIMENSION(:), INTENT(OUT) :: YWPhi
    REAL (KIND=8), DIMENSION(:), INTENT(OUT) :: Residuals
    ! Local variables
    INTEGER :: MSize, XSize, I, J
    REAL (KIND=8), DIMENSION(:,:), ALLOCATABLE :: A
    REAL (KIND=8), DIMENSION(:), ALLOCATABLE :: B
    INTEGER :: M, N, IERR, IOPT, IA, IB

    ! First executable statement

    MSize = SIZE(ACMatrix, 1)
    XSize = SIZE(XSeries)

    ! Initializing LSLDS variables
    ALLOCATE(A(MSize-1, MSize-1), B(MSize-1))
    A = ACMatrix(1:(MSize-1), 1:(MSize-1))
    B = ACMatrix(2:MSize, 1)
    M = MSize-1

    ! Solving the Yule-Walker equations
    CALL DLSLDS (M, A, M, B, YWPhi)

    ! Calculating the YW residuals
    Residuals = 0
    DO I = (MSize+1), XSize
        Residuals(I) = XSeries(I)
        DO J = 1, MSize-1
            Residuals(I) = Residuals(I) - YWPhi(J)*XSeries(I-J)
        END DO
    END DO

    DEALLOCATE(A, B)
END SUBROUTINE YuleWalker

```

```

SUBROUTINE AICCSelction(XSeries,ACVector,PMaX,PHat)
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION(:), INTENT(IN) :: XSeries
  REAL (KIND=8), DIMENSION(0:), INTENT(IN) :: ACVector
  INTEGER, INTENT(IN) :: PMaX
  INTEGER, INTENT(OUT) :: PHat
  ! Local variables
  REAL (KIND=8), DIMENSION(:), ALLOCATABLE :: VarianceVector
  REAL (KIND=8), DIMENSION(:,:), ALLOCATABLE :: PPhi
  REAL (KIND=8) :: VWork, WorkAIC, MinimumAIC, RXSize
  INTEGER :: XSize, WorkP, I, J

  ! First executable statement

  ALLOCATE(VarianceVector(PMaX))
  ALLOCATE(PPhi(PMaX, PMaX))
  XSize = SIZE(XSeries)
  RXSize = REAL(XSize, KIND=8)

  ! Durbin-Levinson Algorithm

  PPhi = 0.0D0
  PPhi(1, 1) = ACVector(1)/ACVector(0)
  VarianceVector(1) = ACVector(0)*(1.0D0 - PPhi(1, 1)**2)

  DO I = 2, PMaX
    VWork = 0
    DO J = 1, I-1
      VWork = VWork + PPhi(I-1, J)*ACVector(I-J)
    ENDDO
    PPhi(I, I) = (ACVector(I) - VWork)/VarianceVector(I-1)
    DO J = 1, I-1
      PPhi(I, J) = PPhi(I-1, J) - PPhi(I, I)*PPhi(I-1, I-J)
    ENDDO
    VarianceVector(I) = VarianceVector(I-1)*(1.0D0 - PPhi(I, I)**2)
  ENDDO

  I = 0
  MinimumAIC = RXSize*LOG(ACVector(0))+2.0D0*(REAL(I, KIND=8)+1.0D0)* &
    RXSize/(RXSize - REAL(I+2, KIND=8))
  WorkP = 0
  DO I = 1, PMaX
    WorkAIC = RXSize*LOG(VarianceVector(I))+2.0*(REAL(I, KIND=8) &
      +1.0)*RXSize/(RXSize - REAL(I+2, KIND=8))
    IF (WorkAIC < MinimumAIC) THEN
      MinimumAIC = WorkAIC
      WorkP = I
    END IF
  END DO
  PHat = WorkP

  DEALLOCATE(PPhi, VarianceVector)
END SUBROUTINE AICCSelction

END MODULE TimeSeriesRoutines

```

Appendix II – Routine FESieves

```

SUBROUTINE FESieves(EDF, XSeries, PHat)
  USE Msimsl
  USE Imslf90
  USE TimeSeries
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION (:,:), INTENT(OUT) :: EDF
  REAL (KIND=8), DIMENSION (:), INTENT(IN) :: XSeries
  INTEGER, INTENT(IN) :: PHat
  ! Local variables
  INTEGER :: XSize, MaxLag, B
  REAL (KIND=8), DIMENSION (:), ALLOCATABLE :: WSeries
  REAL (KIND=8), DIMENSION (:,:), ALLOCATABLE :: ACMatrix
  REAL (KIND=8), DIMENSION (:), ALLOCATABLE :: Residuals
  REAL (KIND=8), DIMENSION (:), ALLOCATABLE :: YWPhi
  REAL (KIND=8) :: XMean

  ! First executable statement

  XSize = SIZE(XSeries, 1)
  MaxLag = SIZE(EDF, 1)
  B = SIZE(EDF, 2)
  ALLOCATE(WSeries(XSize), Residuals(XSize))
  ALLOCATE(ACMatrix(PHat+1, PHat+1), YWPhi(PHat))
  XMean = SUM(XSeries)/REAL(XSize, KIND=8)
  WSeries = XSeries - XMean

  ! Steps 2 - 3
  CALL AutoCovarianceMatrix(ACMatrix(1:PHat+1, 1:PHat+1), WSeries, &
    PHat, PHat+1)
  CALL YuleWalker(WSeries, ACMatrix(1:PHat+1, 1:PHat+1), YWPhi, &
    Residuals)

  ! Steps 4 - 8
  CALL ESievesBootstrap(EDF,XSeries,YWPhi, Residuals, PHat, MaxLag, B)

  DEALLOCATE(ACMatrix, YWPhi, WSeries, Residuals)

CONTAINS

SUBROUTINE ESievesBootstrap(EDF,XSeries,YWPhi,Residuals,PHat,MaxLag,B)
  USE Msimsl
  USE Imslf90
  USE TimeSeries
  IMPLICIT NONE
  REAL (KIND=8), DIMENSION (:,:), INTENT(OUT) :: EDF
  REAL (KIND=8), DIMENSION (:), INTENT(IN) :: XSeries
  REAL (KIND=8), DIMENSION (:), INTENT(IN) :: YWPhi
  REAL (KIND=8), DIMENSION (:), INTENT(IN) :: Residuals
  INTEGER, INTENT(IN) :: PHat, MaxLag, B
  ! Local variables
  INTEGER :: XSize, I, J, Ip, RandomIndex, NOUT, ISEED
  REAL (KIND=8), DIMENSION (:), ALLOCATABLE :: WResiduals, RResiduals
  REAL (KIND=8), DIMENSION (:), ALLOCATABLE :: WSeries, WYWPhi
  REAL (KIND=8), DIMENSION (:,:), ALLOCATABLE :: WACMatrix

```

```

REAL (KIND=8), DIMENSION(:), ALLOCATABLE :: YWResiduals
REAL (KIND=8) :: XMean

! First executable statement

XSize = SIZE(XSeries, 1)

ALLOCATE(WSeries(XSize+100+MaxLag))
XMean = SUM(XSeries)/REAL(XSize, KIND=8)
WSeries(1:XSize) = XSeries - XMean

ALLOCATE(WResiduals(XSize - PHat))
WResiduals = Residuals(PHat+1:XSize)
WResiduals = WResiduals - SUM(WResiduals)/REAL(XSize - PHat, KIND=8)

ALLOCATE(RResiduals(XSize+100+MaxLag), WYWPhi(PHat), &
          WACMatrix(PHat+1, PHat+1), YWResiduals(XSize))

CALL UMACH (2, NOUT)
CALL RNGET (ISEED)
CALL RNSET (ISEED)

DO J = 1, B
  ! Steps 4 - 5
  DO I = 1, XSize+100+MaxLag
    CALL RNUND(1, XSize - PHat, RandomIndex)
    RResiduals(I) = WResiduals(RandomIndex)
  END DO

  ! Step 6
  WSeries = RResiduals
  DO I = PHat+1, XSize+100
    DO Ip = 1, PHat
      WSeries(I) = WSeries(I) + YWPhi(Ip)*WSeries(I-Ip)
    END DO
  END DO

  ! Step 7
  CALL AutoCovarianceMatrix(WACMatrix(1:PHat+1, 1:PHat+1), &
    WSeries(101:XSize+100), PHat, PHat+1)
  CALL YuleWalker(WSeries(101:XSize + 100), &
    WACMatrix(1:PHat+1, 1:PHat+1), WYWPhi, YWResiduals)

  ! Prediction. Step 8
  WSeries(101:XSize+100) = XSeries - XMean
  DO I = XSize+101, XSize+100+MaxLag
    WSeries(I) = RResiduals(I)
    DO Ip = 1, PHat
      WSeries(I) = WSeries(I) + WYWPhi(Ip)*WSeries(I-Ip)
    END DO
  END DO
  EDF(1:MaxLag, J) = WSeries(XSize+101:XSize+100+MaxLag) + XMean
END DO
DEALLOCATE(WSeries, Residuals, RResiduals, WYWPhi, YWResiduals, &
  WACMatrix)
END SUBROUTINE ESievesBootstrap
END SUBROUTINE FSieves

```

JMASM11: Comparing Two Small Binomial Proportions

James F. Reed III
Research Institute
St. Luke's Hospital and Health Network

A large volume of research has focused on comparing the difference between two small binomial proportions. Statisticians recognize that Fisher's Exact test and Yates chi-square test are excessively conservative. Likewise, many statisticians feel that Pearson's Chi-square or the likelihood statistic may be inappropriate for small samples. Viable alternatives exist.

Key words: Fisher's Exact Test, Fisher's Mid-p, Scaled Chi-square test

---

Introduction

A large volume of research spanning nearly half a century has focused on comparing the difference between two small binomial proportions. The validity of various testing procedures remains clouded by controversy (Hirji, 1991). Fisher's exact test (FET) (Fisher, 1958) and Yates continuity corrected chi-square test (Yates, 1934) (X^2_y) have numerous criticisms based on theoretical and empirical considerations. Critiques of FET and X^2_y conclude that they are excessively conservative when used with small to moderate sample sizes leading to an implied loss of power which diminishes their utility (Berkson, 1978; Dupont, 1986; D'Agostino, 1988; Haviland, 1990).

D'Agostino showed that even with small sample sizes Pearson's chi-square test (X^2) and the Student t-test based on binary data generally provide observed significance levels not far from the postulated levels (D'Agostino, 1988). FET is also extremely sensitive to minor variations in data, even when the minimum expected cell size is fairly large (Dupont, 1986).

James F. Reed III, Ph.D., is the Director in the Research Institute at St. Luke's Hospital & Health Network in Bethlehem, PA., 18015. Email: ReedJ@slhn.org.

Upton (Upton, 1982) and Overall (Overall, 1987) evaluated a wide variety of test procedures for comparing the difference between two small binomial proportions. They concluded that in both ease of computation and the average or median actual significance level, one should use one of three tests - the X^2 , Student's t-test for binomial data (BST), or the scaled chi-square test (X^2_s) for almost all sample sizes. Others have advocated the use of Fisher's Mid-p (MP) based procedure in connection with FET (Miettinen, 1974; Plackett, 1984). Barnard (Barnard, 1989; Barnard, 1990) recommends reporting both the traditional p-value and the MP p-value when using FET.

Among applied statisticians a casual attitude towards using these tests has emerged. Practice appears to be guided by what has been described as "conventional wisdom" (D'Agostino, 1988). When the two sample sizes are large, applied statisticians generally use the X^2 or the likelihood statistic (G^2) and compare with the χ^2_{k-1} distribution. With small to moderate sample sizes, either FET or X^2_y are favored. This strategy is reinforced and seldom questioned as evidenced by the content of statistics texts used in colleges and universities.

Many statisticians recognize that FET and the X^2_y tests for comparing two independent binomial proportions are excessively conservative but continue their use. Likewise, many feel that X^2 and G^2 are inappropriate for small sample sizes. Viable alternatives to X^2

include X^2_s , Fisher's MP (14), or BST. None of these have made an appearance in statistics texts. The objective was to provide the practicing statistician with an executable code that produces these alternatives.

Methodology

The following notation was used to describe the comparison of two independent binomial proportions. Let A and B represent the number of successes in independent samples from two binomial populations (n_1, π_1) and (n_2, π_2). Let $n = n_1 + n_2$. Then the joint probability of a particular outcome is:

$$\Pr(A = a, B = b) = [n_1! / a! (n_1 - a)!] [n_2! / b! (n_2 - b)!] \pi_1^a (1 - \pi_1)^{n_1 - a} \pi_2^b (1 - \pi_2)^{n_2 - b},$$

$$\text{for } a = 0, 1, \dots, n_1 \text{ and } b = 0, 1, \dots, n_2, \\ c = n_1 - a \text{ and } d = n_2 - b.$$

Pearson and Scaled chi-square tests (X^2 and X^2_s)

For an observed pattern of (a, b), Pearson's chi-square statistic is

$$X^2 = (ad - bc)^2 n / \{n_1 n_2 (a+b)(c+d)\}$$

The scaled chi-square statistic is derived from the mean and variance of the conditional hypergeometric distribution and is defined as

$$X^2_s = X^2(n - 1)/n$$

X^2 and X^2_s are compared with a χ^2 statistic with 1 degree of freedom. X^2 and X^2_s tests are approximate tests because the distributions of X^2 and X^2_s approach χ^2 with 1 degree of freedom only when the sample sizes are large.

Student t-test (BST)

BST uses the means and variances of the two binomial distributions to compute the usual two independent sample t-statistic on a pooled estimate of the variances. BST is then compared with to the Student t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Fisher's Exact test (FET)

FET is constructed using the conditional distribution of A given A+B. FET is defined:

$$f(a, s, \phi) = \Pr(A = a | A+B = s; \phi) \text{ and } S(a, s, \phi) \\ = \Pr(A > a | A+B = s; \phi), \text{ where } \phi = \pi_1(1 - \pi_2) / (\pi_1(1 - \pi_1))$$

For a one-sided α -level test, reject H_0 if $f(a, s, \phi) + S(a, s, 1) \leq \alpha$.

Mid-P (MP)

MP tests the mean of two probabilities obtained by inclusion and exclusion of the observed point in a discrete distribution. This is equivalent to inclusion of half the probability of the observed point in each tail. MP is found by modifying the above one-sided procedure by the following:

For a one-sided α -level test, reject if $0.5 f(a, s, 1) + S(a, s, 1) \leq \alpha$.

Conclusion

A summary of the literature based on intuitive and theoretical grounds argues in favor of the use of a MP test (Barnard, 1989; Lancaster 1961). The computational effort required for the MP test is no more than that needed for FET. Further, the basis for MP is a natural adjustment for discreteness; and the test easily generates to $r \times c$ contingency tables and other discrete data problems (Hirji, 1991). It is strongly agreed upon that the Yates-corrected chi-square statistic in analyses of 2×2 contingency tables are overly conservative and that the Pearson chi-square generally provides adequate control over type I error probabilities (Haviland, 1990).

The two-tailed FET p-value is highly sensitive to small variations in 2×2 contingency tables. This sensitivity raises doubts about the utility of the FET as a measure of the relative strength of evidence provided by different tables (Dupont, 1986). Pearson's chi-square statistic generally provides adequate control over type I error probabilities without the severe conservative bias produced by Yates' correction for continuity.

When the analytic problem of comparing two independent binomial proportions the classical FET and the X^2_y chi-square tests are too conservative for practical use. A recommend analytical algorithm is: (1) When the two samples are nearly equal, and when the underlying true binomial value is near 0.5, use one of three statistics: $\{X^2, X^2_s, \text{BST}, \text{MP}\}$ for all sample sizes, and (2) in case of unequal sample sizes, or when the common binomial parameter is near 0 or 1, use MP statistic.

An executable Fortran program that produces the statistics outlined in the previous section and sample data is provided in the appendix. A literature search did not produce any references related to public domain software that produces these statistics. The program as written may not be optimal. Any suggestion for refinements to the program would be gratefully accepted.

References

- Barnard, G. A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine*, 8, 1469-1477.
- Barnard, G. A. (1990). Comment. *Statistics in Medicine*, 9, 373-375.
- Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56, 223-234.
- Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistical Inference and Planning*, 2, 27-42.
- D'Agostino, R. B., Chase, W., & Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *American Statistician*, 42, 198-202.
- Dupont, W. D. (1986). Sensitivity of Fisher's exact test to minor perturbations in 2 x 2 contingency tables. *Statistics in Medicine*, 5, 629-635.
- Fisher, R. A. (1958). *Statistical Methods for Research Workers*. (13th ed.). New York: Hafner.
- Haviland, M. G. (1990). Yate's correction for continuity and the analysis of 2 x 2 contingency tables. *Statistics in Medicine*, 9, 363-367.
- Hirji, K. F., Tan, S. J., & Elashoff, R. M. (1991). A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine*, 10, 1137-1153.
- Miettinen, O. S. (1974). Discussion of Conover's "Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables". *Journal of the American Statistical Association*, 69, 374-382.
- Overall, J. E., Rhoades, H. M., & Starbuck, R. R. (1987). Small-sample tests for homogeneity of response probabilities in 2 x 2 contingency tables. *Psychological Bulletin*, 102, 307-314.
- Plackett, R. L. (1984). Discussion of Yate's "Tests of significance for 2 x 2 contingency tables". *Journal of the Royal Statistical Society, Series A*, 147, 426-463.
- Upton, G. J. G. (1982). A comparison of alternative tests for the 2 x 2 comparative trial. *Journal of the Royal Statistical Society, Series A*, 145, 86-105.
- Yates F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society (Supplement)*, 1, 217-235.

Appendix

An executable Fortran program that produces the statistics outlined in this article follows:

```

INTEGER O(2,2),ROWS(2),COLS(2)
INTEGER ROW,COL
COMMON A,B,C,D

C
OPEN(8,FILE='PKIN')
READ(8,*) ROW,COL
READ(8,*) ((O(I,J),J=1,COL),I=1,ROW)
CLOSE(8)
A=O(1,1)
B=O(1,2)
C=O(2,1)
D=O(2,2)
ROW1=A+B
ROW2=C+D
COL1=A+C
COL2=B+D
TOT =A+B+C+D

C
CLOSE(8)
OPEN(8,FILE='PKOUT')

CALL CHISQ
CALL BST
CALL FISH (O)

STOP
END
C=====
SUBROUTINE CHISQ
COMMON A,B,C,D

ROW1=A+B
ROW2=C+D
COL1=A+C
COL2=B+D
TOT=A+B+C+D

PI1=A/(A+C)
PI2=B/(B+D)

CHI=((A*D-B*C)**2*TOT)/(COL1*COL2*ROW1*ROW2)

WRITE(8,101)
WRITE(8,102) A, B
WRITE(8,103) C,D
WRITE(8,104) PI1, PI2

IDF = 1
CALL CHIP (CHI,IDF,XPROB)

```

```

      IF (CHI .EQ. 0) XPROB = 1.0D0
      WRITE(8,105) CHI,XPROB
C
      SCS = CHI*(TOT-1)/(TOT)
      CALL CHIP (SCS,IDF,YPROB)
      IF (SCS .EQ. 0) YPROB = 1.0D0
      WRITE(8,106) SCS,YPROB
C
101  FORMAT(14X,'TRT A',15X,'TRT B',/)
102  FORMAT(10X,'    a = ',F6.1,7X,'    b = ',F6.1)
103  FORMAT(10X,'    c = ',F6.1,7X,'    d = ',F6.1,/)
104  FORMAT(10X,'ã(1) = ',F6.5,7X,'ã(2) = ',F6.3,/)
105  FORMAT(/,10X,'XȲ    = ',F6.2,'    , p-value = ',F7.4)
106  FORMAT(/,10X,'XȲ(S)= ',F6.2,'    , p-value = ',F7.4)
C
      RETURN
      END
C=====
SUBROUTINE BST
      COMMON A,B,C,D
      INTEGER DF

      ROW1=A+B
      ROW2=C+D
      COL1=A+C
      COL2=B+D
      TOT =A+B+C+D

      P1 = A/COL1
      P2 = B/COL2
      TN = ABS(P1-P2)
      PI = (A+B)/TOT
      D1 = PI*(1-PI)/COL1
      D2 = PI*(1-PI)/COL2
      TD = SQRT(D1+D2)
      T  = TN/TD
      FT =T*T

      DF = TOT-2
      CALL FAPPROX (FT,1,DF,QX)
      CALL NPROB (QX,TPROB)

      WRITE(8,101) T,TPROB
101  FORMAT(/,10X,'BST  = ',F6.2,'    , p-value = ',F7.4)
      RETURN
      END
C=====
SUBROUTINE GTEST (O)
      INTEGER O(2,2),ROWS(2),COLS(2)

      COMMON A,B,C,D

      ROWS(1)=A+B
      ROWS(2)=C+D
      COLS(1)=A+C
      COLS(2)=B+D

```

```

      TOT=A+B+C+D
C
      DO 10 I = 1,2
      DO 10 J = 1,2
        IF (O(I,J).NE.0) G = G + O(I,J)*LOG(REAL(O(I,J)))
100  CONTINUE
      DO 20 I = 1,2
        IF (ROWS(I).NE.0.0) G = G - ROWS(I)*LOG(REAL(ROWS(I)))
20   CONTINUE
      DO 30 J = 1,2
        IF (COLS(J).NE.0.0) G = G - COLS(J)*LOG(REAL(COLS(J)))
30   CONTINUE

      G = G + TOT*LOG(TOT)
C
      IF (G.LT.0.0) G = 0.0
      IF (G.GE.0.0) G = 2.0 * G
      IDF = 1
      CALL CHIP(G,IDF,PROB)
      IF (G.EQ.0.0D0) PROB = 1.0
C
      WRITE(8,100)G,PROB
100  FORMAT(/,10X,'G      = ',F6.2,' , p-value = ',F7.4)
      RETURN
      END
C=====C
FISHER'S EXACT TEST
C
C      1) One-tail computations
C      2) Two-tail computations
C
C-----C
      SUBROUTINE FISH (O)
      INTEGER O(2,2)
      INTEGER A,B,C,D,S
      REAL    P(9),T(20),MFPROB1,MFPROB2
C
      K = 0
      A=O(1,1)
      B=O(1,2)
      C=O(2,1)
      D=O(2,2)

1    K = K+1
      XL = 0.0D0
      CALL EPROB (A,PR)
      P(2)=PR
      CALL EPROB (B,PR)
      P(3) = PR
      CALL EPROB (C,PR)
      P(4) = PR
      CALL EPROB (D,PR)
      P(5) = PR
      S = A+B
      CALL EPROB (S,PR)
      P(6) = PR
      S = C+D

```

```

CALL EPROB (S,PR)
P(7) = PR
S = B+D
CALL EPROB (S,PR)
P(8) = PR
S = A+C
CALL EPROB (S,PR)
P(9) = PR
S = A+B+C+D
CALL EPROB (S,PR)
P(1) = PR
DO 20 J = 6,9
20   XL = XL + P(J)
DO 30 J = 1,5
30   XL = XL - P(J)
T(K) = EXP(XL)
C
IF((A.EQ.0).OR.(B.EQ.0).OR.(C.EQ.0).OR.(D.EQ.0))THEN
C
C   CONDITIONAL PROBABILITY
C
      FPROB1 = 0.0D0
      DO 40 I = 1,K
40      FPROB1 = FPROB1 + T(I)
C
C   FISHERS EXACT TEST PROBABILITY (Two-Tail)
      PROB2 = 1 - FPROB1 + T(1)
      FPROB2 = 2*MIN(FPROB1,PROB2)
      IF (FPROB2 .GE. 1) FPROB2=1
C
C   FISHER MID-P
      T(1) = 0.5D0*T(1)
      MFPROB1 = 0.0D0
      DO 50 I = 1,K
50      MFPROB1 = MFPROB1 + T(I)
      MFPROB2 = 2*MFPROB1
      IF(MFPROB2 .GE. 1) MFPROB2=1
      IF (A.LT.C) MFPROB = 1.0D0 - MFPROB

      WRITE(8,100) FPROB1
C   WRITE(8,101) FPROB2
      WRITE(8,102) MFPROB1
C   WRITE(8,103) MFPROB2
100  FORMAT(/,10X,'FET (One-tail), p-value = ',F7.4)
C 101  FORMAT(/,10X,'FET (Two-tail), p-value = ',F7.4)
102  FORMAT(/,10X,'MP (ONE-tail), p-value = ',F7.4)
C 103  FORMAT(/,10X,'MP (TWO-tail), p-value = ',F7.4)

      RETURN
ENDIF
C
IF (A*D - B*C .LT. 0) THEN
      A = A-1
      D = D-1
      B = B+1
      C = C+1
ELSE

```

```

      A = A+1
      D = D+1
      B = B-1
      C = C-1
ENDIF
C
      GO TO 1

      RETURN
      END
C=====
SUBROUTINE EPROB (S,PR)
      INTEGER S
      REAL PR
      PR = 0
      DO 10 I = 1,S
10      PR = PR + LOG(REAL(I))
      RETURN
      END
C=====
SUBROUTINE FAPPROX (F,N1,N2,QX)
      REAL F,V1,V2,XNUM,XDEN,QX
      V1 = REAL(N1)
      V2 = REAL(N2)
      XNUM = F**(1.0/3.0)*(1.0-2.0/(9.0*V2))-(1.0-2.0/(9.0*V1))
      XDEN = 2.0/(9.0*V1)+F**(2.0/3.0)*(2.0/(9.0*V2))
      QX = XNUM/SQRT(XDEN)
      RETURN
      END
C=====
SUBROUTINE NPROB (X,PROB)
      REAL D,PROB,X
      DATA D1,D2,D3/0.0498673470,0.0211410061,0.0032776263/
      DATA D4,D5,D6/0.0000380036,0.0000488906,0.0000053830/
      PROB = 1.0/(2.0*(1.0+D1*X+D2*X*X+D3*X*X*X
&          +D4*X*X*X*X+D5*X*X*X*X*X+D6*X*X*X*X*X*X)**16)
      IF (PROB .GE. 1.0) PROB = 1.0
      RETURN
      END
C=====C
ABRAMOWITZ & STEGUN
C
C          *****
C          Q(X2|DF)
C          PG 941, EQ 26.4.14
C          PG 941, EQ 26.4.15
C          PG 932, EQ 26.2.19
C-----
      SUBROUTINE CHIP (STAT,IDF,P)
C
      DATA D1/0.0498673470/,D2/0.0211410061/,D3/0.0032776263/
$      D4/0.0000380036/,D5/0.0000488906/,D6/0.0000053830/
C
C  CUBE ROOT APPROXIMATION
C
      DF = REAL(IDF)
      X = ((STAT/DF)**(1.0D0/3.0D0)-(1.0D0-(2.0D0/(9.0D0*DF))))/
$      (SQRT(2.0D0/(9.0D0*DF)))

```

```

C
C IMPROVED CUBE ROOT APPROXIMATION
C
      IF ((X.GE.-3.5).AND.(X.LE.-3.0))
$      X=X+0.6060606*(-0.0067+0.0102*(-3.0-X))
      IF ((X.GT.-3.0).AND.(X.LE.-2.5))
$      X=X+0.6060606*(-0.0033+0.0068*(-2.5-X))
      IF ((X.GT.-2.5).AND.(X.LE.-2.0))
$      X=X+0.6060606*(-0.0010+0.046*(-2.0-X))
      IF ((X.GT.-2.0).AND.(X.LE.-1.5))
$      X=X+0.6060606*(0.0001+0.0022*(-1.5-X))
      IF ((X.GT.-1.5).AND.(X.LE.-1.0))
$      X=X+0.6060606*(0.0006+0.0005*(-1.0-X))
      IF ((X.GT.-1.0).AND.(X.LE.-0.5))
$      X=X+0.6060606*(-0.0006)
      IF ((X.GT.-0.5).AND.(X.LE.0.0))
$      X=X+0.6060606*(-0.0002+0.0008*X)
      IF ((X.GT.0.0).AND.(X.LE.0.5))
$      X=X+0.6060606*(-0.0003-0.001*(0.5-X))
      IF ((X.GT.0.5).AND.(X.LE.1.0))
$      X=X+0.6060606*(-0.0006-0.006*(1.0-X))
      IF ((X.GT.1.0).AND.(X.LE.1.5))
$      X=X+0.6060606*(-0.0005+0.0002*(1.5-X))
      IF ((X.GT.1.5).AND.(X.LE.2.0))
$      X=X+0.6060606*(0.0002+0.0014*(2.0-X))
      IF ((X.GT.2.0).AND.(X.LE.2.5))
$      X=X+0.6060606*(0.0017+0.003*(2.5-X))
      IF ((X.GT.2.5).AND.(X.LE.3.0))
$      X=X+0.6060606*(0.0043+0.0052*(3.0-X))
      IF ((X.GT.3.0).AND.(X.LE.3.5))
$      X=X+0.6060606*(0.0082+0.0078*(3.5-X))
C
      X2 = X*X
      X3 = X2*X
      X4 = X3*X
      X5 = X4*X
      X6 = X5*X
      P = 0.5*(1.0+D1*X+D2*X2+D3*X3+D4*X4+D5*X5+D6*X6)**(-16.0)
C
C ERROR CHECKS
C
      IF (P.GT.1.0) P=0.999
      IF (P.LT.0.0) P=0.001
      RETURN
      END

```

Statistical Pronouncements III

“It is a curious circumstance that a science so profoundly mathematical as the theory of probability should have originated in the games of chance which occupy the thoughtless and the profligate” – Robert S. Woodward (1906), *Probability and theory of errors*, NY: Wiley, p. 7.

“Since the beginning of the eighteenth century almost every mathematician of note has been a contributor to or an expositor of the theory of probability” – *ibid*, p. 8.

“Of all the applications of the doctrine of probability none is of greater utility than the theory of errors” – *ibid*, p. 9.

“After all, is not faith at the bottom of all scientific knowledge?” - J. V. Uspensky (1937), *Introduction to mathematical probability*, NY: McGraw-Hill, p. 10.

“The intuition, like the conscience, must be trained” – William Edwards Deming (1950), *Some theory of sampling*, NY: Wiley, p. viii.

“Statistics are a basis for action” – William Edwards Deming, *ibid*, p. 4.

“The only excuse for taking a survey is to enable a rational decision to be made on some problem” - William Edwards Deming, *ibid*, p. 545.

“Consult a statistician at an early stage of your planning” – D. H. Finney (1953), *An introduction to statistical science in agriculture*, Copenhagen: Ejnar Munksgaard, p. 173.

“Do not reject data merely because they seem extraordinary” – D. H. Finney, *ibid*, p. 174.

“When you are experienced enough to make your own statistical analyses, be sure you choose the right technique and not merely any one that you can remember!” – D. H. Finney, *ibid*.

“To put too many entries into one table or too many curves on one diagram ensures that few will read and fewer understand it” – D. H. Finney, *ibid*.

“Express your conclusions in terms of the subject under investigation and its behaviour, with a minimum of statistical jargon” – D. H. Finney, *ibid*.

“Remember that no amount of statistical manipulation can make a bad experiment give good results” – D. H. Finney, *ibid*.

“In order to optimize my expenditure of effort, I divide intellectual difficulties into two classes: those which worry me and those which do not” - Maurice G. Kendall, (1961) Presidential Address, *Journal of the Royal Statistical Society*, 124(A), p. 11.

“Not only can choice mimic chance, but chance can mimic choice” - Maurice G. Kendall, *ibid*, p. 12.

“The typical inference of the detective, historian, or conjecturing mathematician and the clever inferences of science are not statistical inferences” – Leonard J. Savage (1962), *The foundations of statistical inference*, London: Methuen, p. 11.

“I have a natural sympathy with anyone who is trying to thrash out better ways of handling the problems of statistical inference” – Egon Sharpe Pearson, *ibid*, p. 53.

“Through the lack of close contact with my partner during the last twenty years, it would be a little difficult to say where precisely the Neyman and Pearson theory stands today” – Egon Sharpe Pearson, *ibid*.

“Significance tests, in their usual form, are not compatible with a Bayesian attitude” – C. A. B. Smith, *ibid*, p. 60.

“What I, and many other statisticians, call the Neyman-Pearson view may, for all I know, never have been held by Professor Neyman or by Professor Pearson” - Leonard J. Savage, *ibid*, p. 62.

“Probability theory as such presents ideological difficulties for communism” - R. Syski, *ibid*, p. 86.

“The objectives of randomization: to make the experiment useful to others and to guard against one’s own subconscious” - Leonard J. Savage, *ibid*, p. 88.

“Such rumours as that artists can make more random-looking designs than random number generators can are a little disquieting” - Leonard J. Savage, *ibid*, p. 89.

“Ignoring Baye’s Theorem has put much of modern statistics out of gear with scientific thinking” - G. M. Jenkins, *ibid*, p. 94.

“I regard the separation between statistician and client as an accidental detail of real life that we should try to overcome”, Leonard J. Savage, *ibid*, p. 99.

“Cambridge University”, said Fisher, “should never appoint a professor who is older than 39. If they do, then by the time his proposal for his teaching program has been approved by the university, he will have reached retirement age” - attributed to Ronald A. Fisher by William G. Cochran (1967), Footnote, *Science*, 156(3781), p. 1460.

“I have often wondered, as I suppose does Neyman, why Fisher seems not to have regarded the power of the test as relevant, although he developed the power functions of most of the common tests of significance” - William G. Cochran, *ibid*, p. 1461.

“The *l*-statistics are called ‘polykays’ by some authors, but we feel that there are limits to linguistic miscegenation which should not be exceeded” - Maurice G. Kendall, (1969), *The advanced theory of statistics*, (3rd ed.), London: Charles Griffin & Co., p. 303.

“They have not converted me to thinking factor analysis is worth the time necessary to understand it and carry it out” - M. Hills (1977), Book review, *Applied Statistics*, 26, p. 339-340.

“Mathematics... is purely an abstract enterprise that need have nothing to do with the real world... Thus the statement *iggle wug drang flous* could be a legitimate mathematical statement in a set of rules stating that when any *iggle* is *wugged* it *drang* a *flous*... Of course... [this] might not be of any practical use” - Jum C. Nunnally (1978), *Psychometric theory*, (2nd ed.). NY: McGraw-Hill, p. 9-10.

“A random sample is random because of the sampling procedure used to select it, not because of the composition of the sample” - Eugene S. Edgington, (1980), *Randomization tests*, NY: Dekker, p. 2.

“Ninety per cent all the mathematics we know has been discovered (or invented) in the last hundred years” - George Temple (1981), *100 Years of mathematics: A personal viewpoint*. NY: Springer-Verlag, p. xv.

“The great advances in mathematics have not been made by logic but by creative imagination” - George Temple, *ibid*, p. 3.

“Why does the appalling ignorance of statistics persist in spite of the great number of books written on the subject? Simple - the books are written in a *foreign language*” - Myles Hollander & Frank Proschan (1984), *The statistical exorcist: Dispelling statistics anxiety*. NY: Dekker, p. v.

“The rotation of factors is not intrinsically subjective in nature” - Alexander Basilevsky (1994), *Statistical factor analysis and related methods*, NY: Wiley, p. xii.

“Does the Monte Carlo method help one win at roulette? The answer is *No*” - Ilya A. Sobol (1994), *A primer for the Monte Carlo method*. Boca Raton, FL: CRC Press, p. vi.

PASS 2002

Power Analysis and Sample Size Software from NCSS

PASS performs power analysis and calculates sample sizes. Use it before you begin a study to calculate an appropriate sample size (it meets the requirements of government agencies that want technical justification of the sample size you have used). Use it after a study to determine if your sample size was large enough. *PASS* calculates the sample sizes necessary to perform all of the statistical tests listed below.

A power analysis usually involves several "what if" questions. *PASS* lets you solve for power, sample size, effect size, and alpha level. It automatically creates appropriate tables and charts of the results.

PASS is accurate. It has been extensively verified using books and reference articles. Proof of the accuracy of each procedure is included in the extensive documentation.

PASS is a standalone system. Although it is integrated with *NCSS*, you do not have to own *NCSS* to run it. You can use it with any statistical software you want.

PASS comes with two manuals that contain tutorials, examples, annotated output, references, formulas, verification, and complete instructions on each procedure. And, if you cannot find an answer in the manual, our free technical support staff (which includes a PhD statistician) is available.

System Requirements

PASS runs on Windows 95/98/ME/NT/2000/XP with at least 32 megs of RAM and 30 megs of hard disk space.

PASS sells for as little as **\$449.95**.

PASS Beats the Competition!

No other program calculates sample sizes and power for as many different statistical procedures as does *PASS*.

Specifying your input is easy, especially with the online help and manual.

PASS automatically displays charts and graphs along with numeric tables and text summaries in a portable format that is cut and paste compatible with all word processors so you can easily include the results in your proposal.

Choose *PASS*. It's more comprehensive, easier-to-use, accurate, and less expensive than any other sample size program on the market.

Trial Copy Available

You can try out *PASS* by downloading it from our website. This trial copy is good for 30 days. We are sure you will agree that it is the easiest and most comprehensive power analysis and sample size program available.

Analysis of Variance

Factorial AOV
Fixed Effects AOV
Geisser-Greenhouse
MANOVA*
Multiple Comparisons*
One-Way AOV
Planned Comparisons
Randomized Block AOV
New Repeated Measures AOV*

Regression / Correlation

Correlations (one or two)
Cox Regression*
Logistic Regression
Multiple Regression
Poisson Regression*
Intraclass Correlation
Linear Regression

Proportions

Chi-Square Test
Confidence Interval
Equivalence of McNemar*
Equivalence of Proportions
Fisher's Exact Test
Group Sequential Proportions
Matched Case-Control
McNemar Test
Odds Ratio Estimator
One-Stage Designs*
Proportions – 1 or 2
Two Stage Designs (Simon's)
Three-Stage Designs*

Miscellaneous Tests

Exponential Means – 1 or 2*
ROC Curves – 1 or 2*
Variances – 1 or 2

T Tests

Cluster Randomization
Confidence Intervals
Equivalence T Tests
Hotelling's T-Squared*
Group Sequential T Tests
Mann-Whitney Test
One-Sample T-Tests
Paired T-Tests
Standard Deviation Estimator
Two-Sample T-Tests
Wilcoxon Test

Survival Analysis

Cox Regression*
Logrank Survival - Simple
Logrank Survival - Advanced*
Group Sequential - Survival
Post-Marketing Surveillance
ROC Curves – 1 or 2*

Group Sequential Tests

Alpha Spending Functions
Lan-DeMets Approach
Means
Proportions
Survival Curves

Equivalence

Means
Proportions
Correlated Proportions*

Miscellaneous Features

Automatic Graphics
Finite Population Corrections
Solves for any parameter
Text Summary
Unequal N's

*New in PASS 2002

PASS 2002 adds power analysis and sample size to your statistical toolbox

WHAT'S NEW IN PASS 2002?

Thirteen new procedures have been added to *PASS* as well as a new home-base window and a new Guide Me facility.

MANY NEW PROCEDURES

The new procedures include a new multi-factor repeated measures program that includes multivariate tests, Cox proportional hazards regression, Poisson regression, MANOVA, equivalence testing when proportions are correlated, multiple comparisons, ROC curves, and Hotelling's T-squared.

TEXT STATEMENTS

The text output translates the numeric output into easy-to-understand sentences. These statements may be transferred directly into your grant proposals and reports.

GRAPHICS

The creation of charts and graphs is easy in *PASS*. These charts are easily transferred into other programs such as MS PowerPoint and MS Word.

NEW USER'S GUIDE II

A new, 250-page manual describes each new procedure in detail. Each chapter contains explanations, formulas, examples, and accuracy verification.

The complete manual is stored in PDF format on the CD so that you can read and printout your own copy.

GUIDE ME

The new *Guide Me* facility makes it easy for first time users to enter parameter values. The program literally steps you through those options that are necessary for the sample size calculation.

NEW HOME BASE

A new home base window has been added just for *PASS* users. This window helps you select the appropriate program module.

COX REGRESSION

A new Cox regression procedure has been added to perform power analysis and sample size calculation for this important statistical technique.

REPEATED MEASURES

A new repeated-measures analysis module has been added that lets you analyze designs with up to three grouping factors and up to three repeated factors. The analysis includes both the univariate F test and three common multivariate tests including Wilks Lambda.

RECENT REVIEW

In a recent review, 17 of 19 reviewers selected *PASS* as the program they would recommend to their colleagues.

PASS calculates sample sizes for...

Please rush me my own personal license of *PASS 2002*.

Qty

- ___ **PASS 2002 Deluxe** (CD and User's Guide): \$499.95.....\$ _____
- ___ **PASS 2002 CD (electronic documentation):** \$449.95 .....\$ _____
- ___ **PASS 2002 5-User Pack (CD & 5 licenses):** \$1495.00.....\$ _____
- ___ **PASS 2002 25-User Pack (CD & 25 licenses):** \$3995.00.....\$ _____
- ___ **PASS 2002 User's Guide II (printed manual):** \$30.00.....\$ _____
- ___ **PASS 2002 Upgrade CD for PASS 2000 users:** \$149.95 .....\$ _____

Typical Shipping & Handling: USA: \$9 regular, \$22 2-day, \$33 overnight. Canada: \$19 Mail. Europe: \$50 Fedex.....\$ _____

Total: .....\$ _____

FOR FASTEST DELIVERY, ORDER ONLINE AT

WWW.NCSS.COM

Email your order to sales@ncss.com

Fax your order to (801) 546-3907

NCSS, 329 North 1000 East, Kaysville, UT 84037

(800) 898-6109 or (801) 546-0445

My Payment Options:

- ___ Check enclosed
- ___ Please charge my: ___VISA ___MasterCard ___Amex
- ___ Purchase order enclosed

Card Number _____ Expires _____

Signature _____

Please provide daytime phone:

( ) _____

Ship my PASS 2002 to:

NAME

COMPANY

ADDRESS

CITY/STATE/ZIP

COUNTRY (IF OTHER THAN U.S.)

NCSS

329 North 1000 East
Kaysville, Utah 84037

Announcing NCSS 2004 Seventeen New Procedures

NCSS 2004 is a new edition of our popular statistical **NCSS** package that adds seventeen new procedures.

New Procedures

Two Independent Proportions
Two Correlated Proportions
One-Sample Binary Diagnostic Tests
Two-Sample Binary Diagnostic Tests
Paired-Sample Binary Diagnostic Tests
Cluster Sample Binary Diagnostic Tests
Meta-Analysis of Proportions
Meta-Analysis of Correlated Proportions
Meta-Analysis of Means
Meta-Analysis of Hazard Ratios
Curve Fitting
Tolerance Intervals
Comparative Histograms
ROC Curves
Elapsed Time Calculator
T-Test from Means and SD's
Hybrid Appraisal (Feedback) Model

Documentation

The printed, 330-page manual, called *NCSS User's Guide V*, is available for \$29.95. An electronic (pdf) version of the manual is included on the distribution CD and in the Help system.

Two Proportions

Several new exact and asymptotic techniques were added for hypothesis testing (null, noninferiority, equivalence) and calculating confidence intervals for the difference, ratio, and odds ratio. Designs may be independent or paired. Methods include: Farrington & Manning, Gart & Nam, Conditional & Unconditional Exact, Wilson's Score, Miettinen & Nurminen, and Chen.

Meta-Analysis

Procedures for combining studies measuring paired proportions, means, independent proportions, and hazard ratios are available. Plots include the forest plot, radial plot, and L'Abbe plot. Both fixed and random effects models are available for combining the results.

Curve Fitting

This procedure combines several of our curve fitting programs into one module. It adds many new models such as Michaelis-Menten. It analyzes curves from several groups. It compares fitted models across groups using computer-intensive randomization tests. It computes bootstrap confidence intervals.

Tolerance Intervals

This procedure calculates one and two sided tolerance intervals using both distribution-free (nonparametric) methods and normal distribution (parametric) methods. Tolerance intervals are bounds between which a given percentage of a population falls.

Comparative Histogram

This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups.

Random Number Generator

Matsumoto's Mersenne Twister random number generator (cycle length > 10**6000) has been implemented.

Binary Diagnostic Tests

Four new procedures provide the specialized analysis necessary for diagnostic testing with binary outcome data. These provide appropriate specificity and sensitivity output. Four experimental designs can be analyzed including independent or paired groups, comparison with a gold standard, and cluster randomized.

ROC Curves

This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUC's and partial AUC's for paired and independent sample designs.

Hybrid (Feedback) Model

This new edition of our hybrid appraisal model fitting program includes several new optimization methods for calibrating parameters including a new genetic algorithm. Model specification is easier. Binary variables are automatically generated from class variables.

Statistical Innovations Products

Through a *special arrangement* with Statistical Innovations (S.I.), NCSS customers will receive \$100 discounts on:

Latent GOLD® - latent class modeling

SI-CHAID® - segmentation trees

GOLDMineR® - ordinal regression

For demos and other info visit:

www.statisticalinnovations.com

Please rush me the following products:

Qty _____

_____ NCSS 2004 CD upgrade from NCSS 2001, \$149.95 ..... \$ _____

_____ NCSS 2004 User's Guide V, \$29.95 ..... \$ _____

_____ NCSS 2004 CD, upgrade from earlier versions, \$249.95 ..... \$ _____

_____ NCSS 2004 Deluxe (CD and Printed Manuals), \$599.95 ..... \$ _____

_____ PASS 2002 Deluxe, \$499.95 ..... \$ _____

_____ Latent Gold® from S.I., \$995 - \$100 NCSS Discount = \$895..... \$ _____

_____ GoldMineR® from S.I., \$695 - \$100 NCSS Discount = \$595 ..... \$ _____

_____ CHAID® Plus from S.I., \$695 - \$100 NCSS Discount = \$595.... \$ _____

Approximate shipping--depends on which manuals are ordered (U.S: \$10 ground, \$18 2-day, or \$33 overnight) (Canada \$24) (All other countries \$10) (Add \$5 U.S. or \$40 International for any S.I. product) ..... \$ _____

Total..... \$ _____

TO PLACE YOUR ORDER

CALL: (800) 898-6109 **FAX:** (801) 546-3907

ONLINE: www.ncss.com

MAIL: NCSS, 329 North 1000 East, Kaysville, UT 84037

My Payment Option:

_____ Check enclosed

_____ Please charge my: ☐ VISA ☐ MasterCard ☐ Amex

_____ Purchase order attached _____

Card Number _____ Exp _____

Signature _____

Telephone:

( ) _____

Email:

Ship to:

NAME _____

ADDRESS _____

ADDRESS _____

ADDRESS _____

CITY _____ STATE _____

ZIP/POSTAL CODE _____ COUNTRY _____

Forest Plot of Odds Ratio

Statistical and Graphics Procedures Available in NCSS 2004

Analysis of Variance / T-Tests

Analysis of Covariance

Analysis of Variance

Barlett Variance Test

Crossover Design Analysis

Factorial Design Analysis

Friedman Test

Geiser-Greenhouse Correction

General Linear Models

Mann-Whitney Test

MANOVA

Multiple Comparison Tests

One-Way ANOVA

Paired T-Tests

Power Calculations

Repeated Measures ANOVA

T-Tests – One or Two Groups

T-Tests – From Means & SD's

Wilcoxon Test

Time Series Analysis

ARIMA / Box - Jenkins

Decomposition

Exponential Smoothing

Harmonic Analysis

Holt - Winters

Seasonal Analysis

Spectral Analysis

Trend Analysis

Plots / Graphs

Bar Charts

Box Plots

Contour Plot

Dot Plots

Error Bar Charts

Histograms

Histograms: Combined*

Percentile Plots

Pie Charts

Probability Plots

Poisson Curves*

Scatter Plots

Scatter Plot Matrix

Surface Plots

Violin Plots

Experimental Designs

Balanced Inc. Block

Box-Behnken

Central Composite

D-Optimal Designs

Fractional Factorial

Latin Squares

Plackett-Burman

Response Surface

Screening

Taguchi

Regression / Correlation

All-Possible Search

Canonical Correlation

Correlation Matrices

Cox Regression

Kendall's Tau Correlation

Linear Regression

Logistic Regression

Multiple Regression

Nonlinear Regression

PC Regression

Poisson Regression

Response-Surface

Ridge Regression

Robust Regression

Stepwise Regression

Spearman Correlation

Variable Selection

Quality Control

Xbar-R Chart

C, P, NP, U Charts

Capability Analysis

Cusum, EWMA Chart

Individuals Chart

Moving Average Chart

Pareto Chart

R & R Studies

Survival / Reliability

Accelerated Life Tests

Cox Regression

Cumulative Incidence

Exponential Fitting

Extreme-Value Fitting

Hazard Rates

Kaplan-Meier Curves

Life-Table Analysis

Lognormal Fitting

Log-Rank Tests

Probit Analysis

Proportional-Hazards

Reliability Analysis

Survival Distributions

Time Calculator*

Weibull Analysis

Multivariate Analysis

Cluster Analysis

Correspondence Analysis

Discriminant Analysis

Factor Analysis

Hottelling's T-Squared

Item Analysis

Item Response Analysis

Loglinear Models

MANOVA

Multi-Way Tables

Multidimensional Scaling

Principal Components

Curve Fitting

Bootstrap C.I.'s*

Built-In Models

Group Fitting and Testing*

Model Searching

Nonlinear Regression

Randomization Tests*

Ratio of Polynomials

User-Specified Models

Miscellaneous

Area Under Curve

Bootstrapping

Chi-Square Test

Confidence Limits

Cross Tabulation

Data Screening

Fisher's Exact Test

Frequency Distributions

Mantel-Haenszel Test

Nonparametric Tests

Normality Tests

Probability Calculator

Proportion Tests

Randomization Tests

Tables of Means, Etc.

Trimmed Means

Univariate Statistics

Meta-Analysis*

Independent Proportions*

Correlated Proportions*

Hazard Ratios*

Means*

Binary Diagnostic Tests*

One Sample*

Two Samples*

Paired Samples*

Clustered Samples*

Proportions

Tolerance Intervals*

Two Independent*

Two Correlated*

Exact Tests*

Exact Confidence Intervals*

Farrington-Manning*

Fisher Exact Test

Gart-Nam* Method

McNemar Test

Miettinen-Nurminen*

Wilson's Score* Method

Equivalence Tests*

Noninferiority Tests*

Mass Appraisal

Comparables Reports

Hybrid (Feedback) Model*

Nonlinear Regression

Sales Ratios

*New Edition in 2004

Qualitative research has come a long way...

from this...

to this!

THE LATEST PRODUCTS
HAVE ARRIVED
www.qsrinternational.com

WORLD LEADING PRODUCTS FROM THE
NUD*IST LINE OF SOFTWARE

Read more about QSR software in this edition of JMASM.

Two Years in the Making...

Intel® Visual Fortran 8.0

The next generation of Visual Fortran is here! Intel Visual Fortran 8.0 was developed jointly by Intel and the former DEC/Compaq Fortran engineering team.

Visual Fortran Timeline

- 1997** DEC releases
Digital Visual Fortran 5.0
- 1998** Compaq acquires DEC
and releases DVF 6.0
- 1999** Compaq ships CVF 6.1
- 2001** Compaq ships CVF 6.6
- 2001** Intel acquires CVF
engineering team
- 2003** Intel releases
Intel Visual Fortran 8.0

Intel Visual Fortran 8.0

- CVF front-end +
Intel back-end
- Better performance
- OpenMP Support
- Real*16

Performance

Outstanding performance on Intel architecture including Intel® Pentium® 4, Intel® Xeon™ and Intel Itanium® 2 processors, as well as support for Hyper-Threading Technology.

Compatibility

- Plugs into Microsoft Visual Studio® .NET
- Microsoft PowerStation4 language and library support
- Strong compatibility with Compaq® Visual Fortran

Support

1 year of free product upgrades and Intel Premier Support

"The Intel Fortran Compiler 7.0 was first-rate, and Intel Visual Fortran 8.0 is even better. Intel has made a giant leap forward in combining the best features of Compaq Visual Fortran and Intel Fortran. This compiler... continues to be a 'must-have' tool for any Twenty-First Century Fortran migration or software development project."

—Dr. Robert R. Trippi
Professor Computational Finance
University of California, San Diego

FREE trials available at:
programmersparadise.com/intel

Programmer's Paradise®

To order or request additional information call:
800-423-9990
Email: intel@programmers.com

"Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away."

- Antoine de Saint Exupery

F is a carefully crafted subset of the most recent version of Fortran, the world's most powerful numeric language.

Using F has some very significant advantages:

- Programs written in F will compile with any Fortran compiler
- F is easier to use than other popular programming languages
- ***F compilers are free*** and available for Linux, Windows, and Solaris
- Several books on F are available
- F programs may be linked with C, Fortran 95, or older Fortran 77 programs

F retains the modern features of Fortran—modules and data abstraction, for example—but discards older error-prone facilities of Fortran.

It is a safe and portable programming language.

F encourages Module-Oriented Programming.

It is ideal for teaching a programming language in science, engineering, mathematics, and finance.

It is ideal for new numerically intensive programs.

The Fortran Company

11155 E. Mountain Gate Place, Tucson, AZ 85749 USA

+1-520-256-1455 +1-520-760-1397 (fax)

<http://www.fortran.com> info@fortran.com

GGUM2000

Item Response Theory Models for Unfolding

The GGUM2000 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests

that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2000 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected a posteriori (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. The software is accompanied by a detailed user's manual. **GGUM2000 is free** and can be downloaded from:

<http://www.education.umd.edu/EDMS/tutorials>

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2000 today!

Are you involved in Data Modeling or Data Mining?

Are you spending a large percentage of your time dealing with data issues?

If so, you will be happy to know that we have developed a tool that specifically addresses the data prep tasks associated with data modeling and data mining. The tool is called the Digital Excavator from Digital Archaeology (www.digarch.com). Data modelers are well aware of the time-consuming and sometimes frustrating nature of data set-up. In many cases data preparation can represent 60%-80% of the data mining project length. With Digital Archaeology's Digital Excavator, data preparation tasks are streamlined, results are more accurate, and the modeler has more time to focus on finding the appropriate mathematical solution--rather than wasting time with painful data issues. Digital Archaeology's software is intuitive, visual, self-documenting, and deploys what a number of analysts and customers have termed the "most elegant" user interface for data analysis and exploration ever conceived. It's the only tool specifically designed for the data prep tasks of data modeling.

Visit our website and see for yourself! >>>> www.digarch.com

Functions have been created which perform the following:

- Frequency Distributions
- Categorical Variable Profile
- Continuous Variable Profile
- Histograms
- De-duping
- Find and Replace Missing Values
- Find and Split Out Outliers
- Binning
- Correlation Matrix
- Cross-Tabs
- Panel Variables (Occupancy Map)
- Lag functions
- Decimal Scaling
- Rank and Sample Variables
- Recency, Frequency, Monetary Analysis
- N-Tile Distributions
- Gains Charts
- Many others

15721 COLLEGE BOULEVARD
LENEXA, KS 66219
1-877-DIGARCH (344-2724)
WWW.DIGARCH.COM

Version
2.0
See our Website for details

MuPAD® Pro

The Open Computer Algebra System

Mackichan Software is proud to bring you MuPAD Pro, a full-featured computer algebra system in an integrated and open environment for symbolic and numeric computing. The MuPAD language has a Pascal-like syntax and allows imperative, functional, and object-oriented programming. Its domains and categories are like object-oriented classes that allow over-riding and overloading methods and operators, inheritance, and generic algorithms. A comfortable notebook interface includes a graphics tool for visualization, an integrated source-level debugger, a profiler, and hypertext help.

Scientific WorkPlace, the proven solution for mathematical publishing, is an excellent companion to MuPAD Pro. It integrates with MuPAD so that you can evaluate, simplify, solve, and plot from inside your document, and evaluate functions that you have defined with MuPAD. Best of all, you typeset in LaTeX with just the click of a button.

Mackichan
SOFTWARE, INC.

Tools for Scientific Creativity Since 1981

Toll Free: 877-724-9673 • Email: info@mackichan.com

Go to our homepage for free trial versions of all our products.

www.mackichan.com/jmsm

Announcing the highly-anticipated new Numerical Recipes products

Numerical Recipes in C++

The Art of Scientific Computing
Second Edition

William H. Press, Saul A. Teukolsky,
William T. Vetterling, and
Brian P. Flannery

"This monumental and classic work is beautifully produced and of literary as well as mathematical quality. It is an essential component of any serious scientific or engineering library."

—*Computing Reviews*

This new version incorporates completely new C++ versions of the more than 300 *Numerical Recipes Second Edition* routines widely recognized as the most accessible and practical basis for scientific computing, in addition to including the full mathematical and explanatory contents of *Numerical Recipes in C*.

Key Features:

- Includes linear algebra, interpolation, special functions, random numbers, nonlinear sets of equations, optimization, eigensystems, Fourier methods and wavelets, statistical tests, ODEs and PDEs, integral equations, and inverse theory.
- The routines, in ANSI/ISO C++ source code, can be used with almost any existing C++ vector/matrix class library, according to user preference

0-521-75033-4, Hardback, \$70.00

Visit us.cambridge.org/numericalrecipes for more information on the complete line of Numerical Recipes products.

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

800-872-7423

us.cambridge.org/mathematics

Other new Numerical Recipes products for your library...

Numerical Recipes Example Book [C++]

0-521-75034-2, Paperback, \$35.00

Numerical Recipes in C and C++ Source Code CDROM with Windows, DOS, or Macintosh Single Screen License

0-521-75037-7, CD-ROM, \$50.00

Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras
0-521-75036-9, CD-ROM, \$150.00

Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras
0-521-75035-0, CD-ROM, \$90.00

Numerical Recipes in Fortran from Cambridge University Press

Numerical Recipes in Fortran 77

Volume 1 of Fortran Numerical Recipes
Second Edition

*William H. Press, Saul A. Teukolsky,
William T. Vetterling, and Brian P. Flannery*

"This reviewer knows of no other single source of
so much material of this nature. Highly recommended."

—*Choice*

"...a valuable resource for those with a specific need for
numerical software. The routines are prefaced with lucid, self-
contained explanations....highly recommended for those who
require the use and understanding of numerical software."

—*SIAM Review*

1992 992 pp. 0-521-43064-X Hardback \$70.00

Highlights include:

- A chapter on integral equations and inverse methods
- Multigrid and other methods for solving partial differential equations
- Improved random number routines
- Wavelet transforms
- The statistical bootstrap method
- A chapter on "less-numerical" algorithms including compression coding and arbitrary precision arithmetic.

Numerical Recipes in Fortran 77 Example Book

Second Edition

William T. Vetterling, Saul A. Teukolsky, William H. Press, and Brian P. Flannery

1992 256 pp. 0-521-43721-0 Paperback \$35.00

Numerical Recipes in Fortran 90

The Art of Parallel Scientific Computing
Volume 2 of Fortran Numerical Recipes
Second Edition

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery

"This present volume will contribute decisively to a significant breakthrough, as it provides models not only of the numerical algorithms for which previous editions are already famed, but also of an excellent Fortran 90 style."

—*From the Foreword by Michael Metcalf, one of Fortran 90's original designers and author of FORTRAN 90 Explained*

"This book is a classic and is essential reading for anyone concerned with the future of numerical calculation. It is beautifully produced, inexpensive for its content, and a must for any serious worker or student."

—*Computing Reviews*

Contains a detailed introduction to the Fortran 90 language and to the basic concepts of parallel programming, plus source code for all routines from the second edition of Numerical Recipes.

1996 576 pp. 0-521-57439-0 Hardback \$50.00

Numerical Recipes Multi-Language Code CDROM with LINUX or UNIX Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75036-9 CD-ROM \$150.00

Numerical Recipes Multi-Language Code CDROM with Windows, DOS, or Macintosh Single Screen License

Source Code for Numerical Recipes in C, C++, Fortran 77, Fortran 90, Pascal, BASIC, Lisp and Modula 2 plus many extras

2002 0-521-75035-0 CD-ROM \$90.00

Visit us.cambridge.org/numericalrecipes for more information on the complete line of *Numerical Recipes* products.

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

800-872-7423

us.cambridge.org/mathematics

NEW! XML PLUG-IN FOR sas®

IMPORT XML-FORMATTED DATA INTO SAS EXPORT SAS DATA AS XML-FORMATTED DATA

Now, an easy way to move CDISC and other XML-formatted data into or out of your SAS-based systems. You don't have to know perl, XSLT, Xpath, Java®, or exotic languages. The remarkable Tekoa™ XML plug-in does it all for you.

Provided free of charge to any SAS user currently wrestling with XML.
Developed by Zurich Biostatistics, the pioneer in SAS/XML integration.

FREE. EASY. AND IT WORKS.

Just e-mail Michael Palmer (mcpalmer@zbi.net) and receive the fully-functional, proven Tekoa XML plug-in by e-mail.

No charge. No obligation. No hassle. (We even support the tool. Imagine.)

Zurich Biostatistics, Inc.

45 Park Place South, PMB 178, Morristown, NJ 07960 973 727-0025 www.zbi.net

XML is easy if you know how. And we do.

Tekoa XML Technology is a service mark of Zurich Biostatistics, Inc.
SAS is a registered trademark of SAS Institute Inc.
Java is a registered trademark of Sun Microsystems, Inc.

Resampling Stats for Excel

Select the data you want to resample, select "resample" or "shuffle," then specify an output range for the resampled data. Calculate a statistic of interest, select "Repeat & Score," and the resampling operation will be repeated thousands (or tens of thousands) of times, and each time the value of your statistic of interest will be recorded. Does not use Excel's random number generator.

**View complete user guide and download free
30-day trial at <www.resample.com>**

\$249 commercial • \$149 personal/academic • \$89 student

612 N Jackson St., Arlington, VA 22201
Tel 703-522-2713 • Fax 703-522-5846
stats@resample.com

DataMinelt^{SM} announces Permutelt^{TM} v2.0

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

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

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

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

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

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

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

JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to *Psychological Methods* or *Psychological Assessment* (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional \$18)
- *The Score* – the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: \$38 (**APA membership not required**); student membership is only \$8

For further information, please contact the Division's Membership Chair, Yossef Ben-Porath (ybenpora@kent.edu) or check out the Division's website:

<http://www.apa.org/divisions/div5/>

ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the **Special Interest Group - Educational Statisticians** of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.

We also publish a twice-yearly electronic newsletter.

Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at <http://orme.uark.edu/edstatsig.htm>

To join SIG-ES you must be a member of AERA. Dues are \$5.00 per year.

For more information, contact Joan Garfield, President of the SIG-ES, at jbg@umn.edu.

**DEPARTMENT CHAIR
DEPARTMENT OF
MEASUREMENT, STATISTICS & EVALUATION
University of Maryland
College Park, Maryland**

The University of Maryland invites nominations and applications for the position of Chair. We seek a Chair who would understand the research mission and accomplishments of the department, as well as its teaching and service missions in the College of Education and the University. Responsibilities include:

- Facilitating and enhancing the premier graduate education experience offered by EDMS as well as leading in the development of new programs
- Continuing the Department's tradition of national and international prominence within the professional community
- Fostering the research productivity of faculty
- Serving as a strong advocate for the Department's mission in the College and the University
- Coordinating the Department's research centers and managing fiscal planning

The Department has existed within the College for more than 30 years and comprises seven tenure-track faculty members with expertise in educational statistics, applied measurement, psychometrics, evaluation, and assessment. The Department operates three research centers (CSAVE, MARCES and PADI) with which faculty members are associated, and one faculty member currently has a Career Grant from NSF. Several students have held prestigious fellowships including ETS and Fulbright Fellowships.

The Department offers Masters and Doctoral degrees in the general areas of measurement, statistics and evaluation. In addition, a Fifth Year Masters/Bachelors program is available for selected undergraduates. Jointly with the National Center for Education Statistics, we offer a Certificate Program in Large Scale Education Assessment and recently have had a certificate program approved in Measurement, Statistics and Evaluation for non-major graduate students. The proximity to Washington DC and the many agencies there make our location very desirable. More information about the Department can be found at www.education.umd.edu/EDMS.

Qualifications for the position include:

- Earned doctorate in educational statistics, educational measurement, psychometrics or related field
- Record of publications and external funding commensurate with appointment at rank of Professor
- Strong communication and leadership skills
- Experience mentoring faculty at all levels
- Administrative experience at some level is desirable but not mandatory
- Managing fiscal planning

We encourage applications from minority candidates, women and persons with disabilities. The University of Maryland is an equal opportunity, affirmative action employer, and is dedicated to increasing diversity of its faculty through hiring and retention of minorities, women, and persons with disabilities.

Candidates should submit a letter of application addressing the qualifications and expectations noted above, a current curriculum vitae, and names and addresses of three persons from whom letters of reference may be requested. Requests for confidentiality until the final phase of the search will be honored.

All nominations and applications should be forwarded to: Philip J. Burke Ph.D., Department of Special Education, Room 1308 Benjamin Building, College of Education, University of Maryland, College Park, Maryland 20742. Voice: 301-405-6515, Email: pjburke@umd.edu

Review of applications will begin immediately and continue until the position is filled. For fullest consideration, please submit by October 1, 2004. Appointment will be effective July 1, 2005.

Position Available: Top bio-tech company seeks a seasoned statistical manager to hire, develop and lead a team of applied statisticians. Primary role is to integrate statistical methodology and practice into product/process development, manufacturing operations and quality. This key leader will provide linkage between manufacturing, engineering, development and biostatistics. MS in statistics or related field.

Research Statistician: Established clinical group adding staff to provide dedicated preclinical support to a development center. Interact and support scientists with formulation, stability testing, bioanalytics and bio assays. PhD w/3 yrs or MS w/ 6 years industry experience required along with expertise in complicated design methods. Northeast location.

Contact Information: Eve Kriz, Smith Hanley Associates, 99 Park Avenue, New York, NY 10016, 212-687-9696 ext. 228, ekriz@smithhanley.com.

Statistics Through Monte Carlo Simulation With Fortran

Shlomo S. Sawilowsky and Gail F. Fahoome
Copyright © 2003 ISBN: 0-9740236-0-4

Purchase Email, CD, & Softcover Versions Online Via Secure Paypal At

<http://tbf.coe.wayne.edu/jmasm>

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@edstat.coe.wayne.edu. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are not acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font. If the technical expertise is available, submit the manuscript in two column format.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs “in-line”, not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified, indent optional.
8. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
9. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use “&” instead of “and” in multiple author listings.
10. *Suggestions for style*: Instead of “I drew a sample of 40” write “A sample of 40 was selected”. Use “although” instead of “while”, unless the meaning is “at the same time”. Use “because” instead of “since”, unless the meaning is “after”. Instead of “Smith (1990) notes” write “Smith (1990) noted”. Do not strike spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professions is US \$60 per year; graduate students is US \$30 per year; and libraries, universities, and corporations is US \$195 per year. Subscribers outside of the US and Canada pay a US \$10 surcharge for additional postage. Online access is currently free at <http://tbf.coe.wayne.edu/jmasm>. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@edstat.coe.wayne.edu.

Notice To Advertisers

Send requests for advertising information to jmasm@edstat.coe.wayne.edu.

SOFTWARE SOLUTIONS
for Science & Engineering

Lahey/Fujitsu Fortran

The standard for Fortran programming
from the leader in Fortran language systems

LF95 Fortran for Linux and Windows

Full Fortran 95/90/77 support
Unsurpassed diagnostics
Intel and AMD optimizations

IMSL compatible
Fujitsu SSL2 math library
Wisk graphics package

LF Fortran for the Microsoft® .NET Framework - Coming Soon !

Visual Studio integration
Windows / Web Forms designer
Project and code templates

On-line integrated help
XML Web services
ADO.NET support

Visit www.lahey.com for more information

```
ELSE
  poly_coefficient
END IF
ELSE
  poly_coefficient
END IF
END FUNCTION poly_c
SUBROUTINE poly_init
  TYPE(poly), INTENT
  REAL(fpkind), INTE
  IF ( .NOT. PRESENT
    NULLIFY ( p%coef
  ELSE
    m = UBOUND(v,l)
    IF ( max_degree
      ALLOCATE ( p%
      p%coeffs
    ELSE
      ALLOCATE ( p%
      p%coeffs
    END IF
  END IF
```

Lahey Computer Systems, Inc.
865 Tahoe Blvd - P.O. Box 6091
Incline Village, NV 89450 USA
1-775-831-2500
www.lahey.com

NEW IN 2004

significance

statistics making sense

The new magazine of the
Royal Statistical Society

Edited by Helen Joyce

Significance is a new quarterly magazine for anyone interested in statistics and the analysis and interpretation of data. It aims to communicate and demonstrate, in an entertaining and thought-provoking way, the practical use of statistics in all walks of life and to show how statistics benefit society.

Articles are largely non-technical and hence accessible and appealing, not only to members of the profession, but to all users of statistics.

As well as promoting the discipline and covering topics of professional relevance, **Significance** contains a mixture of statistics in the news, case-studies, reviews of existing and newly developing areas of statistics, the application of techniques in practice and problem solving, all with an international flavour.

Special Introductory Offer:
25% discount on a new personal subscription
Plus **Great Discounts for Students!**

**Blackwell
Publishing**

Further information including submission guidelines, subscription information and details of how to obtain a free sample copy are available at

www.blackwellpublishing.com/SIGN