11-1-2004

Vol. 3, No. 2 (Full Issue)

JMASM Editors

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

Recommended Citation

Editors, JMASM (2004) "Vol. 3, No. 2 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 29.

DOI: 10.22237/jmasm/1099267200

Available at: <http://digitalcommons.wayne.edu/jmasm/vol3/iss2/29>

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.

STATISTICIANS

HAVE YOU VISITED THE

Mathematics Genealogy Project?

The Mathematics Genealogy Project is an ongoing research project tracing the intellectual history of all the mathematical arts and sciences through an individual's Ph.D. advisor and Ph.D. students. Currently we have over 80,000 records in our database. We welcome and encourage all statisticians to join us in this endeavor.

Please visit our web site

<http://genealogy.math.ndsu.nodak.edu>

The information which we collect is the following:

The full name of the individual, the school where he/she earned a Ph.D., the year of the degree, the title of the dissertation, and, MOST IMPORTANTLY, the full name of the advisor(s). E.g., Fuller, Wayne Arthur; Iowa State University; 1959; *A Non-Static Model of the Beef and Pork Economy*; Shepherd, Geoffrey Seddon

For additions or corrections for one or two people a link is available on the site. For contributions of large sets of names, e.g., all graduates of a given university, it is better to send the data in a text file or an MS Word file or an MS Excel file, etc. Send such information to:

harry.coonce@ndsu.nodak.edu

The genealogy project is a not-for-profit endeavor supported by donations from individuals and sales of posters and t-shirts. If you would like to help this cause please send your tax-deductible contribution to: Mathematics Genealogy Project, 300 Minard Hall, P. O. Box 5075, Fargo, North Dakota 58105-5075E

Journal Of Modern Applied Statistical Methods

Invited Articles

261 – 270	Juliet P. Shaffer	Confidence Intervals on Subsets may be Misleading
271 – 287	James Algina,
H. J. Keselman	Assessing Treatment Effects in Randomized Longitudinal Two-Group Designs with Missing Observations
288 – 304	S. James Press,
Judith M. Tanur	An Overview of the Respondent-Generated Intervals (RGI) Approach to Sample Surveys

Regular Articles (Theory)

305 – 321	Hakan Demirtas	Modeling Incomplete Longitudinal Data
322 – 332	Du Feng,
Norman Cliff	Monte Carlo Evaluation of Ordinal d with Improved Confidence Interval
333 – 344	Lisa M. Lix,
Aynsle M. Hinds	Multivariate Contrasts for Repeated Measures Designs Under Assumption Violations
345 – 356	Pingfu Fu,
J. Sunil Rao	On a Simple Method for Analyzing Multivariate Survival Data Using Sample Survey Methods
357 – 369	Victor M. Guerrero,
Rafael Perera	Variance Stabilizing Power Transformation for Time Series
370 – 385	Ghazi Shukur,
Panagiotis Matalos	Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach
386 – 398	Amjad Al-Nasser,
Ayman Baklizi	Interval Estimation for the Scale Parameter of Burr Type X Distribution Based on Grouped Data
399 – 405	Vladimir Gercsik,
Mark Kelbert	On Comparison of Hypothesis Tests in the Bayesian Framework Without Loss Function

Early Scholars

406 – 416	Miguel A. Padilla,
James Algina	Type I Error Rates for a One Factor Within-Subjects Design with Missing Values
417 – 431	LiPing Chu,
S. James Press,
Judith M. Tanur	Confidence Elicitation and Anchoring in the Respondent-Generated Intervals (RGI) Protocol
432 – 442	Xinyue Zhou,
Joel R. Levin	A Note on Extending Scheffé's Modified Multiple-Comparison Procedure to Other Analysis Situations

443 – 461	Livio Corain, Luigi Salmaso	Multivariate and Multistrata Nonparametric Tests: The NonParametric Combination Method
462 – 475	Xiaosheng Lei, Janet K. Holt, T. Mark Beasley	Aligned Rank Tests as Robust Alternatives for Testing Interactions in Multiple Group Repeated Measures Designs with Heterogeneous Covariances
<i>JMASM Algorithms and Code</i>		
476 – 484	Jann-Huei Jinn	JMASM12: The President's Problem
485 – 497	Hakan Demirtas	JMASM13: Pseudo-Random Number Generation in R for Commonly Used Multivariate Distributions
498 – 503	Sikha Bagui, Subhash Bagui	JMASM14: An Algorithm and Code for Computing Exact Critical Values for the Kruskal-Wallis Nonparametric One-Way ANOVA
504 – 517	Maria A. Barceló, Marc Saez	JMASM15: A Modification of the EM Algorithm to Estimate an Andersen-Gill Gamma Frailty Model for Multivariate Failure Time Data
<i>Special Section: Commemorating the Retirement of R. Clifford Blair</i>		
518 – 566	Shlomo Sawilowsky	A Conversation with R. Clifford Blair on the Occasion of his Retirement
567 – 575	James J. Higgins	Statistics and Technology: Reflections on 35 Years of Change
576 – 577	Bruce W. Hall	Mentoring Doctoral Students: A Personal Perspective
578 – 580	Howard Stoker	"Teaching" in honor of Cliff Blair
581 – 593	John H. Neel	A New Goodness-of-Fit Test for Item Response Theory

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

Production Staff
Christina Gase
Jack Sawilowsky

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

Entire Reproductions and Imaging Solutions	248.299.8900 (Phone)	e-mail:
Internet: www.entire-repro.com	248.299.8916 (Fax)	sales@entire-repro.com

Editorial Board of Journal of Modern Applied Statistical Methods

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

Chris Barker
Oncology Biostatistics
South San Francisco, 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 Houghton
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

Confidence Intervals On Subsets May Be Misleading

Juliet Popper Shaffer
University of California, Berkeley

A combination of hypothesis testing and confidence interval construction is often used in social and behavioral science studies. Sometimes confidence intervals are computed or reported only if a null hypothesis is rejected, perhaps to see whether the range of values is of practical importance. Sometimes they are constructed or reported only if a null hypothesis is accepted, in order to assess the range of plausible nonnull values due to inadequate power to detect them. Even if always computed, they are interpreted differently, depending on whether the null value is or is not included. Furthermore, many studies in which the null hypothesis is not rejected are never published (the “file drawer” problem). This article discusses the coverage probability of nominal $1 - \alpha$ confidence intervals when examining intervals that do or do not cover some specified null value, usually zero. A briefer treatment considers interval coverage when undesirable results are suppressed. The coverage probability of such conditional confidence intervals may be very far from the nominal value. The magnitude of the effect of selection on interval coverage probability and possible resultant biases in inference are illustrated, and discussed in relation to effect sizes of importance in social and behavioral science research and to estimation of effect sizes.

Keywords: Hypothesis tests, selected confidence intervals, censored studies

Introduction

There has been an enormous amount of literature, much of it in the social sciences, recommending that confidence intervals always be constructed, either in addition to or instead of p-values or other information related to testing hypotheses. The purpose of this article is to

point out a problem in interpreting confidence intervals when they are pertinent to a hypothesis of interest.

The correct interpretation of $1 - \alpha$ confidence intervals is that these randomly-chosen intervals have probability $1 - \alpha$ of covering the true values of the parameter being estimated. Given a set of intervals, on the average $1 - \alpha$ proportion should cover the true values. However, it is often true that special attention is paid to intervals depending on their coverage. Often there is special interest in a particular value of the parameter involved, either zero (often in comparing two groups) or some specified nonzero value. This article will consider the situation in which zero is of special interest; the results generalize to any other value with only obvious changes.

In such cases of selective interest, special attention may be paid to intervals that

Juliet Popper Shaffer is Senior Lecturer Emeritus of Statistics. Her research is primarily in the area of linear models--regression and analysis of variance, multivariate analysis, and simultaneous inference. She also has a background in psychology and in educational measurement and methodology. Email: shaffer@stat.berkeley.edu.

don't include zero, in order to estimate the size of plausible parameter values. There may be special interest in intervals that are far from zero. Or on the contrary, special attention may be paid to intervals that do include zero, to see whether there might be differences of substantial interest that could be verified by more powerful studies. Usually the direction of departure from the null hypothesis is of special interest, and intervals in one or the other direction may be especially scrutinized.

Furthermore, it is well known that studies with insignificant results often are not reported, and therefore not known, as is sometimes true of studies with results in a direction opposite to that of the desires or expectations of the sponsoring organization. Then only some selected intervals are available to be considered.

As soon as there is special consideration of a subset of intervals based on the values they include, the probability that they cover the true parameter value, in other words their conditional coverage probability, may be considerably different from the nominal $1 - \alpha$ probability that applies to the whole set.

Such conditional considerations apply to all situations in which confidence intervals are obtained. This article will give detailed quantitative results for the comparison of the means of two distributions, assuming independent, normally-distributed observations with equal variance and equal sample sizes. All quantitative results reported here for known variance apply also to the case of matched pairs of observations with variances of the matched differences known, given the appropriate one-sample test in that case, provided the tabled effect sizes are divided and tabled sample sizes multiplied by the square root of 2.

Section 1 will give a general overview of conditional probability coverage both when the intervals do and when they do not cover the value zero, with most attention on the former. The coverage depends on the noncentrality parameter, a function of the sample size and the effect size. Sections 2 and 3 will examine the coverage for effect sizes and sample sizes that are frequently encountered in social and behavioral science research: Section 2 primarily when zero is not covered, and Section 3 when

intervals in one direction are not calculated or reported. Section 4 will discuss effect size estimation issues as they relate to conditional coverage. Section 5 discusses and summarizes the issues raised.

Comparing the means of two distributions:
Conditional on coverage or noncoverage of a specified value

Consider two groups of independent, normally-distributed observations with equal, known variance, and of equal sample size. With unknown variance, the standard test of equality of the means is the two-sample t test. With known variance, the known value σ is used in place of the estimate s ; the test statistic then has a normal distribution, and the resulting test will be referred to as the two-sample z test. Since the t distribution tends to the normal distribution as the number of degrees of freedom tends to ∞ , the properties of the z test hold approximately for the t test when the variance is estimated with large degrees of freedom.

Suppose a $1 - \alpha$ confidence interval is constructed for a difference between the means of the two groups, where $\alpha = .05$ is assumed throughout the paper. Consider separately the probability of covering the correct value for confidence intervals that do not include the value zero, and the same probability for confidence intervals that do include zero. Figure 1 gives the conditional coverage of those intervals, as a function of the noncentrality parameter, which is the standard effect size measure $(\mu_1 - \mu_2)/\sigma$ (Cohen, 1962, 1988), multiplied by the square root of $n/2$, where n is the sample size of each group. Given the known sample size n of each group, the noncentrality parameter, and therefore the conditional coverage, is a function of the unknown true effect size.

What Figure 1 illustrates is the well-known fact that intervals that do not cover zero also have very small conditional probabilities (given that fact) of covering values close to zero (see, e.g., Olshen, 1973). Correspondingly, intervals that do cover zero are also more likely than the nominal confidence coefficient to cover values close to zero. These properties are true for intervals of fixed length as in this case, when

the standard deviation is known. For the same true effect size, the coverage probabilities depart even further from the nominal values when the standard deviation is unknown and must be estimated, so that the size of the conditional intervals varies with the estimated standard deviation. In that case, for a given effect size, intervals that don't cover zero are likely to be shorter than intervals that do, so both location and interval length affect the conditional coverage. Figure 2 gives the correlation between the interval length and the probability that the interval includes the correct value, for t intervals with varying degrees of freedom.

Relation of conditional true value coverage and non-coverage to effect sizes and sample sizes frequently encountered in social-behavioral science research.

The noncentrality parameter that determines the coverage probability is a function of the known sample size n of each group and of the effect size. Thus, consideration of effect sizes is crucial in examining conditional confidence interval coverage. Of course, there is no direct way of making use of the quantitative information in a particular case, since the true effect size is unknown. However, many studies in the social sciences, as noted in Cohen's (1962) pioneering paper, support the assumption that effect sizes in these fields are often between .1 and .5. Cohen suggested the now-standard terminology of small effects = .2, medium effects = .5, and large effects = .8. He stated "Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined..." (Cohen, 1988, p. 13).

Examples of estimated effect sizes in the literature show many around .2 or less. For instance, Fukkuk and Glopper (1998), in a meta-analysis of studies of learning of word-meaning from context, found out of 22 effect size estimates that nine were smaller than .20, ten were between .21 and .40, and only three were greater than .40. Grissmer (1999), in a meta-analytic study of the effects of class size

reduction on achievement, found effect sizes between .15-.25 for grades K-3, and .11-.20 for grades 4-7. Although researchers carrying out meta-analytic studies try to find as many studies as possible, it seems clear that it is easier to locate studies with significant effects, and thus probably larger real or apparent effect sizes, than those with insignificant effects, which may never have been reported. Furthermore, in the former study (Fukkuk and Glopper), it was noted that the data for some studies, even though the studies were found, could not be obtained. Thus, the obtained values reported above are likely to represent an upwardly-biased sample. It follows that even when all reported confidence intervals are considered equally, the available studies are likely to include an overabundance of intervals that do not include zero.

In summary, a small effect size of .2 or smaller is likely to be a feature of many studies of this kind, and furthermore, the reported values may be upwardly biased. Since the conditional coverage probability of confidence intervals is a function of the effect size, an examination of effect sizes in the range assumed to be common in social and behavioral science research, and their relation to conditional coverage, is called for.

Table 1 gives the coverage probability for the two-sample z Test, equal sample sizes, with sample sizes ranging from 5 per group to 50 per group, assuming effect sizes of .1 to .5, and assuming the null hypothesis is rejected, so that the intervals do not include zero. The values in parentheses are the probabilities of rejecting the null hypothesis for the associated sample size-effect size combination. All values hold approximately when variances are estimated with large degrees of freedom.

If the variance must be estimated from the information in the two sets of observations, the confidence coverage results are still further from the nominal values. When there are 5 observations per group, so that t is based on 8df, the first row entries in Table 1 would be

Figure 1: Conditional coverage probability, two-sample z test, n per group

Figure 2: Correlation between length of t-interval and probability of correct confidence interval coverage

Table 1: True conditional probability that the nominal .95 confidence interval based on the z test covers the correct value, given rejection of the null hypothesis (values in parentheses are probabilities of rejection).

Sample size	Effect size				
	.1	.2	.3	.4	.5
5	.20(.05)	.41(.06)	.57 (.07)	.69(.10)	.77 (.12)
10	.29(.05)	.55(.07)	.72(.10)	.81(.15)	.87(.20)
20	.41(.06)	.69(.09)	.83(.16)	.90(.24)	.93(.35)
30	.49(.07)	.77(.11)	.88(.21)	.93(.34)	.95(.48)
40	.55(.07)	.81(.14)	.91(.26)	.94(.43)	.96(.60)
50	.60(.08)	.84(.17)	.92(.32)	.95(.51)	.96(.70)

replaced by .18(.05), .36(.06), .52(.07) .64(.09), and .73(.11), respectively, and when there are 10 observations per group (18df), the second row entries would be replaced by .28(.06), .53(.07), .69(.10), .80(.14), and .86(.18), respectively. For larger df, the differences are very small, so the results for known variance are approximately correct.

For an effect size of .1, the probability of rejecting the null hypothesis, i.e. the probability that the interval does not include zero, is quite small, even for samples of size 50 in each group. However, if these cases are the ones that get attention, perhaps the only ones that get published, the extreme departure from the nominal coverage probability of the associated confidence intervals means that incorrect quantitative inferences are highly

likely. Even for effect sizes larger than .1, the under-coverage of the intervals can be non-negligible, and the probability that the intervals don't contain zero becomes much larger. As noted above, effect sizes within the range .1 to .3 are very common in social-behavioral science research.

Values that are covered when the true value is not covered

When intervals that do not include zero also do not include the true values, they will include either only values in the wrong direction from the true effect, smaller than the true effect in the correct direction, or, more likely with small effect sizes, values in the correct direction but farther away from zero than the true values. When the true effect is barely different from

zero, clearly the probability of a range of values more extreme in the right direction and a range in the wrong direction will each be approximately .50. When the true effect is extremely large, the probability of ranges of values more extreme in the right direction and less extreme in the right direction will each be approximately .50. For the effect sizes and sample sizes in Table 1, the probabilities of intervals covering only smaller values in the correct direction are all equal to zero. Table 2 gives the conditional probabilities that the results do not cover the true values; the entries in parentheses are the expected proportion of these non-covering intervals that are in the right direction but more extreme. Subtracting these proportions from one gives the conditional probabilities of confidence intervals with ranges entirely in the wrong direction.

Note that for the smaller effect sizes and/or sample sizes in this table, the probability that the intervals do not cover the true values can be quite substantial, as can the probabilities that they cover values in the correct direction but larger. In some cases, the probability of intervals entirely in the wrong direction is non-negligible. Thus, the calculated intervals may lead to either incorrect directional inferences or unwarranted optimism about the true sizes of the effects under study.

It has been noted that when studies with insignificant effects are not reported, many studies in the literature claim real differences when in fact the null hypotheses are true. However, it is shown here that even when the null hypotheses are false, the confidence intervals are likely to indicate that the effect sizes are larger than they really are. This is true if special attention is paid to confidence intervals that do not include zero, even when there is no withholding of studies showing insignificant effects.

Suppose, however, that confidence intervals including zero are specially noted, in order to estimate the range of plausible nonzero values. When the true value is small, these intervals are likely to have probability higher than the nominal probability of covering true values, and thus also to give falsely optimistic

impressions of possible null hypothesis departures.

If the variance must be estimated from the two samples themselves, the first row would be replaced by the values for t with 8 df : .82(.59), .67(.33), .48 (.74), .36 (.81), and .27 (.87), respectively, while the entries in the second row, replaced by the values for t with 18 df , would be .72(.63), .47(.75), .31(.84), .20(.91), and .14 (.95), respectively. For larger degrees of freedom, the values are very close to those for known variance. As for known variance, the probability of coverage in the correct direction but smaller than the true value is zero for the sample sizes and effect sizes in the table.

Conditioning when significant results in one direction only are noted

According to an Associated Press article in the September 9, 2004 *San Francisco Chronicle*, and also reported in other places, editors of 11 medical journals are adopting a policy requiring the results of all clinical studies to be made public, noting that “drug company-sponsored studies with negative results rarely are submitted to medical journals” (Tanner, 2004). In this case, “negative” means results contrary to the desires of the company. This can be interpreted in two ways, noted by (a) and (b) below.

(a) The results may be reported only if significant and in the direction desired by the company. If the results are significant, but the true value is in the direction that is not reported, then reported confidence intervals will have probability zero of including the correct value, and from Table 2 it is possible to calculate the probability of results in the false direction being reported (multiply the probability of rejection by the conditional probability of intervals in the incorrect direction, given rejection). If the true value is in the direction that is reported, the values in Table 1 are the probabilities that the reported intervals cover the true values.

Table 2: Conditional probability of noncoverage (of true values) of the nominal .95 confidence interval, and (in parentheses) the proportion of noncovering intervals containing larger values in the correct direction.

Sample size	Effect size				
	.1	.2	.3	.4	.5
5	.80(.59)	.59(.69)	.43(.77)	.31(.84)	.23(.89)
10	.71(.63)	.45(.76)	.28(.85)	.19(.92)	.13(.96)
20	.59(.69)	.31(.84)	.17(.93)	.10(.98)	.07(.99)
30	.51(.73)	.23(.89)	.12(.97)	.07(.99)	.05(1.00)
40	.45(.76)	.19(.92)	.09(.98)	.06(1.00)	.04(1.00)
50	.40(.78)	.16(.94)	.08(.99)	.05(1.00)	.04(1.00)

Table 3: True conditional probability that the nominal .95 confidence interval covers the correct value, as a function of effect size and sample size per group, given that the results are not significant in the true direction, for a two sample z test (values in parentheses are probabilities that the interval is reported).

Sample size	Effect size				
	.1	.2	.3	.4	.5
5	.94(.96)	.92(.95)	.91(.93)	.88(.91)	.85(.88)
10	.93(.96)	.91(.93)	.88(.90)	.83(.86)	.78(.80)
20	.92(.95)	.88(.91)	.82(.84)	.73(.76)	.62(.65)
30	.92(.94)	.86(.88)	.76(.79)	.63(.66)	.48(.51)
40	.91(.93)	.83(.86)	.71(.73)	.54(.57)	.37(.39)
50	.90(.93)	.81(.83)	.65(.68)	.46(.48)	.27(.29)

(b) Suppose results are reported if either nonsignificant or significant in the desired direction, i.e. suppressed only when the results are significant in the less-favored direction, as might be the case if some studies suggested undesirable side effects of a medication. If the favored direction happens to be the true one, the confidence interval coverage will be equal to the nominal coverage, .95 in the example, regardless of the true effect size. Table 3 gives the probabilities that the confidence intervals cover the true values, variance known, when the true values are in the less-favored direction: This is the probability that the null hypothesis is accepted and contains the true values. The probability that the interval is reported is given in parentheses.

Coverage probabilities and effect size estimation

Given the type of conditioning, conditional confidence interval coverage depends on the noncentrality parameter, which is a function of the sample size (known) and the effect size (unknown). Thus, if the effect size were known, the conditional coverage probability would be known, and vice versa. It would appear, then, that estimating the effect size would be helpful in estimating the confidence interval coverage. The relation between effect size estimation and confidence interval coverage, however, is complex.

If the variance were known, estimation of effect size would be equivalent to estimation of the mean difference. With unknown variance, however, estimation of the effect size, which requires an estimate of the unknown standard deviation in the denominator, is considerably more difficult and less robust than estimation of the mean difference. In either case, estimation of effect size is unlikely to be helpful in estimating confidence coverage of the true mean difference. Although the confidence interval coverage when the variance is estimated with small degrees of freedom is not drastically different from the coverage with known variance, estimation of the effect size is very much poorer in the former case.

Note that problems with effect size estimation exist even if there is equal information on and attention to any outcome, while in that case confidence interval coverage is equal to the nominal level, given the assumptions of the model. Calculation of the confidence interval is straightforward, while there are a number of different estimates of effect size even in this simplest case (Hedges & Olkin, 1985).

Hedges (1984) studied the theoretical properties of effect size estimation when only significant effect sizes are observed; see also Hedges and Olkin (1985). The standard estimate

$$g = (\bar{X}_E - \bar{X}_C) / s$$

an estimate of

$$\delta = (\mu_E - \mu_C) / \sigma,$$

where E is the experimental group mean and C is the control group mean, is biased towards more extreme absolute values even with no censoring, and is also biased when such censoring occurs. The exception is for $\delta = 0$, in which case neither is biased. Note that in this case, with censoring, the confidence interval coverage is zero. The variance of g when $\delta = 0$ is much larger under censoring than without censoring, and is bimodal, so highly nonnormal, for small sample sizes and/or effect sizes. Thus, under the conditions for which confidence interval coverage is far from optimal coverage, estimation of effect sizes is no help in trying to estimate the non-coverage probability.

Even under known censoring conditions, effect size estimation for single studies is of little value when the noncentrality parameter is small. The value of effect size estimation comes through meta-analysis, when a series of estimates of the same effect size are available. One of the problems, even in that case, is that there is almost certainly some censoring, but the type and extent are usually unknown.

Conclusion

It is often claimed that confidence interval or confidence set procedures give more useful information than hypothesis-testing procedures, since they indicate not only whether the hypothesis of a specified value for a parameter would be accepted or rejected, but indicate plausible values of that parameter. This article points out some difficulties in interpreting confidence intervals when there is a parameter value of special interest, as is true when hypotheses are tested. Confidence intervals conditional on covering or especially on not covering that particular value may have coverage probabilities considerably different from their nominal probabilities, under conditions frequently encountered in research in the social and behavioral sciences.

Although the conditional considerations are obviously important when confidence intervals are computed or known only for selected cases, they are also important when confidence intervals are calculated in all cases, if subsets of intervals are examined for different purposes. As noted, intervals that do not include zero are often examined to see whether the range of plausible values is of practical importance, while intervals that do include zero may be examined to see whether studies with greater power would be worth carrying out. Under the conditions reported in this paper, the true coverage probability of each of these subsets of intervals may be very different from their nominal coverage probabilities.

This article has dealt only with inference concerning a single test or interval. Additional problems arise in multiple testing or estimation. A recent paper on this subject (Benjamini & Yekutieli, 2004, with discussion) notes that the conditional problem is insoluble when there are no plausible assumptions about the possible effect sizes; thus, the conditional coverage properties noted in this paper are relevant for the ranges of sample sizes and effect sizes covered. The authors suggest an alternative interpretation involving unconditional aspects that allows some bounds on the probability of non-covering intervals.

References

- Benjamini, Y., & Yekutieli, D. (2004, in press). False discovery rate adjusted multiple confidence intervals for selected parameters (with discussion). *Journal of the American Statistical Association*.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition. NY: Academic Press.
- Fukkuk, R. G. & de Glopper, K. (1998). Effects of instruction in deriving word meaning from context: A meta-analysis. *Review of Educational Research*, 68, 450-469.
- Grissmer, D. (1999). Class size effects: Assessing the evidence, its policy implications, and future research agenda. *Educational Evaluation and Policy Analysis*, 21, 231-248.
- Hedges, L. V., (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, 9, 61-85.
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. NY: Academic Press.
- Olshen, R. A. (1973). The conditional level of the F test. *Journal of the American Statistical Association*, 68, 692-698.
- Tanner, L. (2004, September 9). Medical editors stress better access to data. *San Francisco Chronicle*, C4.

Assessing Treatment Effects in Randomized Longitudinal Two-Group Designs with Missing Observations

James Algina
University of
Florida

H. J. Keselman
University of Manitoba

SAS's PROC MIXED can be problematic when analyzing data from randomized longitudinal two-group designs when observations are missing over time. Overall (1996, 1999) and colleagues found a number of procedures that are effective in controlling the number of false positives (Type I errors) and are yet sensitive (powerful) to detect treatment effects. Two favorable methods incorporate time in study and baseline scores to model the missing data mechanism; one method was a single-stage PROC MIXED ANCOVA solution and the other was a two-stage endpoint analysis using the change scores as dependent scores. Because the two-stage approach can lack sensitivity to detect effects for certain missing data mechanisms, in this article we examined variations of the single-stage approach under conditions not considered by Overall et al., in order to assess the generality of the procedure's positive characteristics. The results indicate when and when not it is beneficial to include a baseline score as a covariate in the model. As well, we provide clarification regarding the merits of adopting an endpoint analysis as compared to the single-stage PROC MIXED procedure.

Keywords: Randomized designs, repeated measurements, missing data, PROC MIXED

Introduction

Overall and his colleagues (Ahn, Tonidandel, & Overall, 2000; Overall, Ahn, Shivakumar, & Kalburgi 1999, Overall, Ghasser, & Fiore, 1996) have provided very valuable information to biopharmaceutical researchers regarding the analysis of data from randomized longitudinal

two-group designs. In particular, they compared various formulations of SASs (SAS, 1995) PROC MIXED program for analyzing effects in repeated measures designs when data are missing over time, finding that many formulations did not provide effective Type I error control, while others lacked power to detect treatment effects. In their studies, they found that a number of analysis of covariance (ANCOVA) analyses, using baseline scores and time in study as covariates, provided effective Type I error control and were, among the procedures compared, relatively powerful to detect treatment effects. In particular, they found that a single-stage PROC MIXED (1999, p. 208) and several two-stage analyses (1999, pp. 205-209) provided good results. Among the two-stage analyses, the endpoint analysis had the largest estimated power.

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.

Algina and Keselman (2003) however, compared the single-stage PROC MIXED analysis and endpoint analysis, as well as others presented in the literature, and found that though Overall et al.'s (1999) two-stage endpoint procedure had power similar to that of the other procedures when data were missing completely at random, it was lacking in power to detect treatment effects when data were not missing completely at random (See discussion below).

For example, the two-stage power value in one condition was .26, while the other procedures investigated had values clustered around .60. On the other hand, Overall et al.'s single-stage procedure controlled rates of Type I error and was the most powerful (or second next most powerful in one case) procedure among those procedures that were never liberal. Moreover, with regard to bias and sampling variability its values were not very different from bias and sampling variability for the other procedures that did not exhibit liberal rates of Type I error. Thus, in the investigation reported herein, we only examined modifications of the single-stage PROC MIXED procedure enumerated by Overall et al. (1999) as well as another method to be described.

The variations of the Overall et al. (1999) PROC MIXED procedure that we investigated are based on their acknowledgement that there was some concern regarding "the propriety of ... including the baseline scores as *both* linear covariate and as one of the repeated measurements to which a linear regression model was fitted" (p. 267). Given the very positive operating characteristics of their approach to the analysis of longitudinal data with missing observations we thought it important to further investigate their method of analysis by comparing PROC MIXED models that do and do not include a baseline score as both a covariate and repeated measurement in the analysis. In addition, we vary other conditions such as drop out mechanism, number of repeated measurements, and pattern of parameters in order to assess the operating characteristics of their procedure over conditions not yet examined in order to assess the generality of their findings.

Missing Data Mechanisms

To set the stage for our investigation we first discuss conditions under which data may be missing in randomized longitudinal two-group designs.

Consider a design in which N participants are randomly assigned to $K = 2$ treatments. The researcher plans to observe each participant J times on the dependent variable, with the first observation prior to initiating a treatment and the remaining $J - 1$ observations following initiation of a treatment. The effect of primary interest, typically, is whether there are differential rates of change over time, that is, whether there is a group by time interaction.

Let Y_{ijk} denote a random variable underlying the score, in treatment k ($k = 1, 2$), for participant i ($i = 1, \dots, n_k$), on occasion j ($j = 1, \dots, J$). A possible model for the subject-specific regression of the dependent variable on time of measurement is

$$y_{ik} = \mathbf{X}\beta_{ik} + \varepsilon_{ik}$$

where $\mathbf{y}'_{ik} = (Y_{i1k}, \dots, Y_{iJk})$, β_{ik} is an unobservable r -dimensional random vector, ε_{ik} is a J -dimensional random vector,

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{r-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_J & t_J^2 & \cdots & t_J^{r-1} \end{bmatrix},$$

and $t_1, \dots, t_J$ indexes time of measurement. We assume $\varepsilon_{ik} \sim N(0, \sigma^2 \mathbf{I}_J)$.

In this paper we focus on situations in which it is reasonable to assume that the subject-specific regressions are well described by a linear trend. Therefore

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_J \end{bmatrix}$$

and $\beta'_{ik} = (\beta_{i0k} \ \beta_{i1k})$. The between-subjects model for β_{ik} is

$$\begin{bmatrix} \beta_{i0k} \\ \beta_{i1k} \end{bmatrix} = \begin{bmatrix} 1 & z & 0 & 0 \\ 0 & 0 & 1 & z \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{bmatrix} + \begin{bmatrix} u_o \\ u_1 \end{bmatrix} \quad (1)$$

where $z=0$ for the first treatment and 1 for the second treatment. More compactly

$$\beta_{ik} = \mathbf{W}\gamma + \mathbf{u}.$$

We assume that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$.

In randomized longitudinal two-group designs, participants may not be observed on all occasions. In general, the correct method of analysis depends on the missing data mechanism. Using an incorrect method can result in inconsistent estimates of the parameters. The design considered in this paper is a special case of the longitudinal design considered by Little (1995). Little presented his review in the context of monotone missing data patterns, a context we adopt here. That is, we assume that if a participant is not observed on a particular occasion, the participant is not observed on any subsequent occasion.

In order to clarify missing data mechanisms, we employ a random coefficients selection model perspective to the analysis of missing data in longitudinal data. Let J_{ik} denote the last occasion at which participant i in group k was observed and $t_{J_{ik}}$ the value of t for this time point and let $\mathbf{y}_{ik}$ be partitioned as $\mathbf{y}'_{ik} = (\mathbf{y}'_{obs,ik} \ \mathbf{y}'_{miss,ik})$, $R_{ik} = J_{ik}$ if the participant has complete data, and $R_{ik} = J_{ik}$, otherwise. According to Little (1995), in this approach the joint distribution of $\mathbf{y}_{ik}$, β_{ik} , and R_{ik} is factored as

$$f(\mathbf{y}_{ik}, \beta_{ik}, R_{ik} | \mathbf{X}, \mathbf{W}) = f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \beta_{ik}) f(\beta_{ik} | \mathbf{W}) f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \beta_{ik}).$$

In our context, the model for $f(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \beta_{ik})$ is

$$(\mathbf{y}_{ik} | \mathbf{X}, \mathbf{W}, \beta_{ik}) \sim N(\mathbf{W}\gamma + \mathbf{X}\mathbf{u}, \sigma^2 \mathbf{I}_J)$$

and

$$(\beta_{ik} | \mathbf{W}) = \mathbf{u} \sim N(\mathbf{0}, \mathbf{D}).$$

The model for $f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \beta_{ik})$ is the model for the missing data mechanism. The data are referred to as missing completely at random (MCAR) if

$$f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{ik}, \beta_{ik}) = f(R_{ik}).$$

(See, e.g., Rubin, 1976; Little, 1995; Little & Rubin, 1987). That is, the data are MCAR if the probability of a particular data point being missing does not depend on either $\mathbf{y}_{ik}$, β_{ik} , $\mathbf{X}$ or $\mathbf{W}$. The missing data mechanism is called covariate dependent (CD) if the probability of a particular data point being missing does not depend on either

$\mathbf{y}_{ik}, \beta_{ik} :$

$$f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{obs,ik}, \mathbf{y}_{miss,ik}, \beta_{ik}) = f(R_{ik} | \mathbf{X}, \mathbf{W}).$$

The missing data mechanism is called missing at random (MAR) if

$$f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{obs,ik}, \mathbf{y}_{miss,ik}, \beta_{ik}) = f(R_{ik} | \mathbf{X}, \mathbf{W}, \mathbf{y}_{obs,ik}),$$

that is, the probability of a particular data point being missing does not depend on either $\mathbf{y}_{miss,ik}$ or β_{ik} . Following Verbeke and Molenberghs (2000, p. 213), a missing data mechanism that does not meet any of these criteria can be referred to as missing not at random (MNAR). Consistent estimates for γ can be obtained from the likelihood for $\mathbf{y}_{obs,ik}$ and R_{ik} . However if the data are MCAR, CD, or MAR (and if the parameters of the missing data mechanism are distinct from the parameters for the data), consistent estimates can be obtained by maximizing the likelihood for $\mathbf{y}_{obs,ik}$, a process that is called ignoring the missing data mechanism. Thus, for the purposes of estimating the fixed effects, the missing data mechanism is ignorable if the mechanism is MCAR, CD or MAR, but the missing data mechanism is non-ignorable if the mechanism is MNAR.

Hedeker and Gibbons (1997) noted that, frequently, missing data are related to performance or other characteristics of participants. (See Schafer, 1997, Ch. 2, for other examples of studies where MAR is a reasonable

model of “missingness”). Accordingly, MAR may very well be a reasonable process to presume for the missing data in a study. It should be noted that to legitimately ignore the missing data mechanism for estimation purposes, not only must the data be missing at random, but also, the parameters of the missing data mechanism must be independent of the parameters of the data model (Schafer, 1997). This independence or distinctness of parameters is quite realistic in many contexts (e.g., Schafer, 1997, p. 11-15). When the missing data mechanism is ignorable, numerical results can easily be obtained with commercially available software, e.g., the SAS (1995) PROC MIXED program (see Littell et al., 1996).

Overall et al.’s (1999) Approach

Overall and his colleagues (See Overall et al., 1999) investigated an ANCOVA approach using the baseline score on Y (Y_{i1k}) and the number of available measurements for participant i as covariates. Their model is

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

$$\beta_{0ik} = \lambda_{00} + \lambda_{01}J_{ik} + \lambda_{02}Z + \lambda_{03}Y_{i1k} + u_{i0}$$

$$\beta_{1ik} = \lambda_{10} + \lambda_{12}Z + u_{i1}.$$

PROC MIXED code (See Overall et al., 1999, p. 208) for the model is

```
proc mixed method=ml;
class id group;
model score=nrm scr1 group time
time*group/solution;
random intercept time/type=un subject=id;
```

The variable *nrm* is the number of measurements (time in study) available for a participant. The variable *scr1* is the baseline score. As Overall et al. (1999, p. 193) note “The covariates entered the PROC MIXED model statement in numeric form by being excluded from the class statement.”

In this article, we compare Type I error and power for the test of equality of average slopes, bias in the difference in the average slopes, and the variability in estimating this

difference as a function of the covariates included in the model.

Methodology

Four methods of examining the group by time interaction effect in a randomized longitudinal two-group design were examined. Specifically, the methods (with their acronyms) were:

- (1) PROC MIXED analysis that presumes the data are missing at random (PMMAR),
- (2) Overall et al.’s (1999) PROC MIXED analysis that uses *scr1* as a covariate (SCR1),
- (3) Overall et al.’s (1999) PROC MIXED analysis that uses *nrm* as a covariate (NRM),
- (4) Overall et al.’s (1999) PROC MIXED analysis that uses *scr1* and *nrm* as covariates (SCR1&NRM).

It should be noted that PMMAR is Overall et al.’s procedure without any covariates.

We investigated three factors in our study: number of equally spaced levels of the repeated measures variable (5 and 9), missing data mechanism (MCAR, MAR and MNAR), and covariance structure for the repeated measures. (The variations on the covariance structure are presented when we describe the model we used to simulate the data.) Overall and his colleagues (See Ahn, Tonidandel and Overall, 2000; Overall et al., 1999; Overall et al., 1996) examined the group by time interaction effect in a design containing a baseline score and eight additional repeated measurements; thus, for comparative purposes we had nine levels for one of our cases of number of repeated measurements. For examining generality of results, we also included a smaller case, that is, five levels.

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:

$$Y_{ijk} = \beta_{0i} + \beta_{1i}t_j + \varepsilon_{ijk} \quad (2).$$

In each treatment group, data were simulated for 100 participants. The variable t_j was coded (0,

0.23077, 0.46154, 0.69231, 0.92308, 1.15385, 1.38462, 1.61538, 1.84615). To get the codes for conditions with five time points we eliminated the last four codes.

The mean for β_{0i} was 50 in both groups, implying that both treatment groups had the same population pretest mean. For Type I error data, the mean for the slope was 4.5 in treatment 1 and treatment 2 [$\gamma_{11} = 0$, where γ_{11} is defined in equation (1)], indicating identical average rates of increase over time, hence, a null condition. For our power comparisons, the slope was 9.0 in treatment 2 and 4.5 in treatment 1 ($\gamma_{11} = 4.5$) when there were nine occasions and 12.5 in treatment 2 and 4.5 in treatment 1 ($\gamma_{11} = 8$) when there were five occasions. The slopes for treatment 2 were selected to provide similar power for both levels of the number of occasions factor. 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. The variance for the residuals, conditional on time, was 240. In half of the conditions 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 less rapidly. In the other half of the conditions we changed the covariance to 12.42 from -12.42 and retained all other features of the design. Changing the covariance for the slope and intercept changes the covariance structure for the repeated measures. It should be noted that Overall et al. assumed that there was no between-subjects variation in the subject-specific slopes. The correlation in their repeated measurements was due to the random intercept and correlation in the residuals in equation (2). Thus, we investigated the performance of Overall et al.'s procedure, as well as the alternatives, for different correlation structures than Overall et al. employed.

Overall et al. (1999) investigated three variations on the missing data mechanism, which they called completely random, treatment dependent, and treatment and baseline dependent. In each 30% of the simulated participants dropped out of the study. In the completely random condition dropping out was not related to scores on the repeated measures, time, or the treatment indicator. Thus the completely random condition meets the requirements for a MCAR mechanism. In the treatment dependent condition, dropping out was not related to scores on the repeated measures or time but was related to the treatment indicator: two-thirds of the dropouts came from the treatment group. The treatment dependent condition meets the requirements for a CD mechanism.

In the treatment and baseline dependent condition, missing data were related to the random effects for the intercept with dropouts from the treatment group coming from those that had a subject-specific intercept above the mean and dropouts from control group coming from those that had a subject-specific intercept below the mean. Thus, the treatment and baseline dependent condition employed a MNAR missing data mechanism.

In our study, once the data were generated, data were eliminated according to a MCAR, a MAR, or one of two MNAR missing data mechanisms. As indicated in our introduction, when the missing data mechanism is MNAR, ignoring the mechanism can result in inconsistent estimates of the unknown parameters. To select missing observations we used the following model

$$Z_{ijk} = \theta_{1j} + \theta_2 \beta_{0i} + \theta_3 \beta_{1i} + \theta_4 Y_{i(j-1)k} + \theta_5 Y_{ijk}.$$

An observation was set as missing if $U_{ijk} < \phi(Z_{ijk})$ where U_{ijk} is a uniformly distributed random variable and ϕ is the standard normal distribution. The missing data mechanism is MCAR if $\theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$, MAR if $\theta_2 = \theta_3 = \theta_5 = 0$ and MNAR if θ_2 , θ_3 , or θ_5 is not equal to zero. In one MNAR mechanism only θ_2 and θ_3 were not equal to

zero (MNAR-SI). In the other MNAR mechanism, only θ_5 was not equal to zero (MNAR-Y). The values of θ_{1j} were selected to give cumulative missing data rates between 30% and 40% at the ninth occasion. In all conditions missing data conformed to a monotone drop out pattern. That is, if a simulated respondent had missing data on occasion j , the respondent had missing data on all subsequent occasions. Thus we investigated the performance of Overall et al.'s procedure, as well as variations, for different missing data mechanisms than Overall et al. employed. In particular our MAR and MNAR conditions were different than those employed by Overall et al. (1999).

Figure 1 shows estimated proportions of participants remaining in the study at each occasion in the non-null condition with a negative correlation between the slope and intercept and nine time points under the MCAR, MAR, MNAR-SI and MNAR-Y mechanisms. To obtain these estimates, 100,000 data points were generated for each treatment group. (For the MCAR mechanism, a total of 100,000 data points were generated since in our MCAR condition the drop out rate was the same in both treatments.) For our MAR condition the probability of dropping out at occasion j was positively related to the participant's score at occasion $j-1$. For our MNAR-SI condition the probability of dropping out at occasion j was positively related to the participant's intercept and slope. For our MNAR-Y condition the probability of dropping out at occasion j was positively related to the score the participant would have attained at occasion j if the participant had not dropped out.

Thus in all panels of Figure 1, except the top right, drop out rates are higher for the treatment group with the average slope equal to 9 (treatment 2). Drop out rates vary across type of missing data mechanism; however, because we will compare methods for a particular mechanism, and not the performance of a method across mechanisms, this variation in drop out rates across mechanisms is not problematic. Each condition was replicated 2500 times. When there were five time points, the drop out rates for the j th time point ($j = 1, \dots, 5$) were equal to the drop out rates for the j th time

point in the design with nine time points. All hypothesis tests were conducted with a nominal alpha of .05.

Results

Type I error rates and power are reported in Table 1 for the MCAR and MAR conditions and in Table 2 for the MNAR conditions. All procedures exhibited adequate control of the Type I error rate. Power differences were negligible in the MCAR conditions and in the MAR conditions when the correlation between the slope and intercept was positive and very small in the MAR conditions when the correlation between the slope and intercept was negative. Larger differences emerged in the MNAR conditions and clearly indicated lower power for the PMMAR procedure than for the other procedures. Among the remaining procedures NRM is the most powerful, though the advantage is fairly small, ranging from about .004 to .061.

Table 3 contains means and standard deviations (empirical standard errors) of the estimates for the MCAR and MAR conditions when $\gamma_{11} = 0$. Table 4 contains the same information for the MNAR conditions. When $\gamma_{11} = 0$ none of the procedures had an average estimate that was significantly different from zero and, across all conditions, empirical standard errors were fairly similar.

Table 5 contains means and standard deviations of the estimates for the MCAR and MAR conditions when $\gamma_{11} \neq 0$. As expected from theory PMMAR produced unbiased estimators under the MCAR and MAR missing data mechanisms. The other procedures produced unbiased estimators in the MCAR conditions and in the MAR conditions when the correlation between the slope and intercept was positive. When the correlation between the slope and intercept was negative with MAR data, all procedures, except PMMAR (No/No) produced slightly biased estimators. The estimator for the NMR (No/Yes) procedure was less biased, although the difference was small. (Biased estimators are delineated in the tables in bold face type.)

Figure 1. Percent of Data that is Not Missing by Occasion and Missing Data Mechanism

Table 1. Type I Error and Power Rates for MCAR and MAR Conditions.

Missing Data	Correlation	Covariates		5-levels		9-levels	
		SCR1	NRM	Type I Error	Power	Type I Error	Power
MCAR	Positive	No	No	.046	.623	.056	.613
		Yes	No	.046	.624	.056	.607
		No	Yes	.046	.624	.056	.613
		Yes	Yes	.046	.625	.055	.608
	Negative	No	No	.052	.628	.054	.614
		Yes	No	.055	.623	.055	.601
		No	Yes	.052	.630	.054	.614
		Yes	Yes	.054	.623	.055	.601
MAR	Positive	No	No	.048	.592	.048	.604
		Yes	No	.052	.592	.049	.602
		No	Yes	.050	.594	.048	.604
		Yes	Yes	.053	.592	.047	.604
	Negative	No	No	.055	.616	.049	.607
		Yes	No	.052	.625	.045	.615
		No	Yes	.056	.632	.044	.631
		Yes	Yes	.052	.622	.044	.616

Table 2. Type I Error and Power Rates for MNAR Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Correlation	SCR1	NRM	Type I Error	Power	Type I Error	Power
MNAR-SI	Positive	No	No	.056	.236	.061	.356
		Yes	No	.056	.363	.061	.414
		No	Yes	.055	.383	.060	.423
		Yes	Yes	.055	.363	.061	.414
	Negative	No	No	.052	.237	.056	.303
		Yes	No	.052	.418	.047	.453
		No	Yes	.049	.474	.049	.525
		Yes	Yes	.053	.420	.047	.464
MNAR-Y	Positive	No	No	.055	.507	.051	.553
		Yes	No	.055	.552	.051	.562
		No	Yes	.053	.556	.051	.571
		Yes	Yes	.054	.550	.051	.560
	Negative	No	No	.053	.514	.053	.519
		Yes	No	.050	.581	.052	.600
		No	Yes	.046	.604	.055	.622
		Yes	Yes	.050	.577	.052	.597

Table 3. Mean and Standard Deviation of the Difference between the Control and Treatment Group ($\gamma_{11} = 0$): MCAR and MAR Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Correlation	SCR1	NRM	MEAN	SD	MEAN	SD
MCAR	Positive	No	No	0.068	3.544	0.044	2.056
		Yes	No	0.069	3.557	0.045	2.073
		No	Yes	0.067	3.544	0.044	2.056
		Yes	Yes	0.069	3.557	0.045	2.074
	Negative	No	No	0.047	3.518	0.023	2.018
		Yes	No	0.051	3.548	0.027	2.050
		No	Yes	0.048	3.518	0.023	2.019
		Yes	Yes	0.051	3.548	0.027	2.049
MAR	Positive	No	No	0.003	3.543	-0.016	2.066
		Yes	No	-0.017	3.579	-0.009	2.067
		No	Yes	-0.006	3.541	-0.016	2.037
		Yes	Yes	-0.015	3.568	-0.012	2.047
	Negative	No	No	-0.047	3.622	-.004	1.973
		Yes	No	-0.088	3.610	-.019	2.006
		No	Yes	-0.078	3.558	-.021	1.954
		Yes	Yes	-0.084	3.600	-.017	1.994

Table 4. Mean and Standard Deviation of the Difference between the Control and Treatment Group ($\gamma_{11} = 0$): MNAR Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Correlation	SCR1	NRM	MEAN	SD	MEAN	SD
MNAR-SI	Positive	No	No	0.024	3.809	0.009	2.049
		Yes	No	0.022	3.810	-0.007	2.053
		No	Yes	0.024	3.765	0.004	2.039
		Yes	Yes	0.020	3.807	-0.006	2.049
	Negative	No	No	0.058	3.667	-0.010	2.015
		Yes	No	0.071	3.706	0.002	2.023
		No	Yes	0.054	3.610	-0.004	1.946
		Yes	Yes	0.071	3.702	0.004	2.018
MNAR-Y	Positive	No	No	-0.040	3.608	-0.054	1.929
		Yes	No	-0.035	3.601	-0.054	1.937
		No	Yes	-0.039	3.582	-0.053	1.928
		Yes	Yes	-0.036	3.599	-0.056	1.935
	Negative	No	No	-0.087	3.564	-0.022	2.037
		Yes	No	-0.097	3.506	-0.026	1.990
		No	Yes	-0.094	3.435	-0.020	1.954
		Yes	Yes	-0.098	3.503	-0.026	1.985

Table 5. Mean and Standard Deviation of the Difference between the Control and Treatment Group ($\gamma_{11} \neq 0$): MCAR and MAR Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Correlation	SCR1	NRM	MEAN	SD	MEAN	SD
MCAR	Positive	No	No	8.080	3.523	4.573	1.996
		Yes	No	8.073	3.527	4.569	2.012
		No	Yes	8.081	3.523	4.573	1.996
		Yes	Yes	8.073	3.527	4.570	2.012
	Negative	No	No	8.080	3.513	4.500	2.073
		Yes	No	8.064	3.544	4.506	2.106
		No	Yes	8.083	3.514	4.501	2.073
		Yes	Yes	8.064	3.544	4.505	2.105
MAR	Positive	No	No	7.989	3.595	4.546	2.058
		Yes	No	8.037	3.630	4.530	2.060
		No	Yes	7.992	3.595	4.499	2.039
		Yes	Yes	8.018	3.620	4.501	2.040
	Negative	No	No	8.109	3.657	4.493	2.053
		Yes	No	8.279	3.670	4.597	2.054
		No	Yes	8.199	3.625	4.550	2.001
		Yes	Yes	8.261	3.661	4.574	2.043

Table 6 contains means and standard deviations of these estimates for the MNAR conditions when $\gamma_{11} \neq 0$. In these conditions the PMMAR (No/No) estimator was clearly more biased than were the other estimators. In most conditions the NRM estimator was less biased than were the SCR1 (Yes/No) and SCR1&NRM (Yes/Yes) estimator, though in many conditions the differences among the three procedures were negligible.

Discussion

We compared the performance of four data analysis procedures, which varied in terms of the covariates employed: no covariates, SCR1, NRM, and SCR1 and NRM. As expected from theory, when the missing data mechanism was MCAR or MAR there was no advantage to including SCR1, NRM, or both in the model. However, including SCR1 and/or NRM did not have a negative impact on the results. For the MNAR missing mechanisms including SCR1 and/or NRM improved power and reduced bias relative to the analysis without covariates. However, including SCR1 in addition to NRM did not enhance power or reduce bias relative to including only NRM as a covariate. And in some conditions including only NRM did enhance power and or reduce bias relative to the analyses that included SCR1 in addition to or in place of NRM.

Additional Results

Given our results and the fact that Overall et al. (1999) used both NRM and SCR1 as covariates and that Ahn et al. (2000) used only SCR1 as a covariate in their PROC MIXED analyses that included a random statement, but not a repeated statement, the question arises as to when is it necessary to employ SCR1 or SCR1 and NRM as covariates. To explore this question we simulated data using the treatment and baseline dependent missing data mechanism employed by Overall and his colleagues. In Overall et al. and Ahn et al. there was no between-subject random variation in the subject-specific slopes; accordingly, we included conditions like those studied by Overall and his

colleagues as well as conditions in which there was between-subject random variation in the subject-specific slopes. For the latter analyses

$$\mathbf{D} = \begin{bmatrix} 15.21 & 0.00 \\ 0.00 & 82.81 \end{bmatrix}$$

and accordingly the treatment and baseline dependent missing data mechanism does not result in indirect selection on the slope, as would occur if the covariance between the subject-specific slopes and intercepts were non-zero. We refer to the treatment and baseline dependent conditions without slope variation as MNAR-I.NSV since the missing data mechanism is MNAR; that is, the missing data indicator is dependent on the intercept, and, there is no slope variations. The other conditions are referred to as MNAR-I.SV. Type I error rates and power results are presented in Table 7. The results indicate that when the probability of missing data depends on the subject-specific intercept, but not the slope, it is essential to control for SCR1 (baseline score) and the addition of NRM (number of repeated measurements) does not enhance control of the Type I error rate or power.

Results in Overall et al. (1999) suggest that a two-stage endpoint analysis is more powerful than the PROC MIXED analysis that includes SCR1 and NRM. Results in Ahn et al. (2000) suggest that the endpoint analysis is more powerful than the PROC MIXED analysis that includes SCR1 only. As noted in our introduction, Algina and Keselman (2003) did not find the endpoint analysis to be more powerful than the other procedures in the study. However, as noted above, Algina and Keselman simulated data with random variation in the subject-specific slopes, but Overall and his colleagues did not. To determine whether random variation in the subject-specific slopes accounts for the results with regard to the endpoint analysis, we estimated Type I error rates and power under the two MNAR-I missing data mechanisms. In all endpoint analyses both SCR1 and NRM were included as covariates. Results, shown in Table 7, indicate that the endpoint analysis controlled the Type I error rate

Table 6. Mean and Standard Deviation of the Difference between the Control and Treatment Group ($\gamma_{11} \neq 0$): MNAR Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Correlation	SCR1	NRM	MEAN	SD	MEAN	SD
MNAR-SI	Positive	No	No	5.005	4.000	3.279	2.090
		Yes	No	6.469	3.988	3.587	2.083
		No	Yes	6.644	3.933	3.610	2.066
		Yes	Yes	6.480	3.984	3.575	2.078
	Negative	No	No	4.779	4.079	3.041	2.164
		Yes	No	6.972	4.122	3.909	2.190
		No	Yes	7.392	4.001	4.077	2.086
		Yes	Yes	7.033	4.116	3.939	2.180
MNAR-Y	Positive	No	No	7.218	3.679	4.084	1.972
		Yes	No	7.532	3.659	4.143	1.970
		No	Yes	7.563	3.644	4.133	1.964
		Yes	Yes	7.520	3.657	4.123	1.969
	Negative	No	No	7.303	3.661	4.049	2.060
		Yes	No	7.839	3.605	4.398	2.030
		No	Yes	7.950	3.546	4.434	1.985
		Yes	Yes	7.846	3.602	4.394	2.028

Table 7. Type I Error and Power Rates for MNAR-I Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Data Analysis	SCR1	NRM	Type I Error	Power	Type I Error	Power
MNAR-I.NSV	Proc Mixed	No	No	.072	.521	.092	.714
		Yes	No	.053	.698	.058	.909
		No	Yes	.070	.525	.091	.717
		Yes	Yes	.053	.698	.059	.909
	Endpoint	Yes	Yes	.056	.522	.050	.557
MNAR-I.SV	Proc Mixed	No	No	.068	.477	.080	.508
		Yes	No	.054	.628	.056	.666
		No	Yes	.067	.480	.079	.510
		Yes	Yes	.055	.628	.056	.665
	Endpoint	Yes	Yes	.064	.461	.057	.383

Table 8. Type I Error and Power Rates for CD Conditions.

Missing Data		Covariates		5-levels		9-levels	
Mechanism	Data Analysis	SCR1	NRM	Type I Error	Power	Type I Error	Power
CD.NSV	Proc Mixed	No	No	.053	.676	.046	.882
		Yes	No	.054	.688	.049	.879
		No	Yes	.054	.680	.046	.883
		Yes	Yes	.054	.686	.049	.878
	Endpoint	Yes	Yes	.050	.602	.051	.616
CD.SV	Proc Mixed	No	No	.052	.635	.058	.605
		Yes	No	.052	.626	.060	.598
		No	Yes	.052	.634	.057	.603
		Yes	Yes	.052	.626	.060	.598
	Endpoint	Yes	Yes	.051	.568	.054	.494

in all conditions, but was not more powerful than the PROC MIXED analysis that controlled the Type I error rate.

Algina and Keselman (2003) also did not include a treatment dependent drop-out condition like that included in Overall et al. (1999) and Ahn et al. (2000) and this may account for differences in terms of the endpoint analysis. Table 8 contains Type I error rates and power for a treatment dependent drop out condition like that included in Overall et al. and Ahn et al. We refer to this condition as CD. In CD conditions with between-subject random variation in the subject-specific slopes (CD.SV) reported in Table 8

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

However, we also conducted simulations with $D_{12} = 12.42$ and $D_{12} = 0.00$ and the general pattern of results was the same as those reported in Table 8: the endpoint analysis controls the Type I error rate, but was not more powerful than the PROC MIXED analyses.

Conclusion

In summary, we believe our results provide some clarification to the findings reported by Overall et al (1999) and Ahn et al. (2000), clarification, we believe, that adds to the importance of their contributions to the literature regarding the analysis of missing data in randomized longitudinal two-group designs.

First, with the regard to the controversy of including baseline scores as both independent and dependent variables in the analysis, our results show that it is not always necessary to include the baseline score as a covariate. Except when the distribution of the missing data depended exclusively on the subject-specific intercept, neither Type I error control nor power to detect effects was enhanced by including baseline as a covariate in addition to specifying the number of repeated measurements as a covariate in the model. However, it is also true that including the baseline as an additional covariate did not detract from control of the Type I error rate and detracted noticeably from

power only when the probability of missing data depended on the subject-specific slopes and intercepts.

When the probability of a missing value depended on the subject-specific intercept, our results, along with those reported by Overall et al. (1999) indicate that the baseline score should be specified as a covariate in the model. In this case, however, no additional gains in terms of Type I error control or power to detect effects, is acquired by including as a second covariate number of repeated measurements.

Lastly, the findings from our study indicate that an endpoint analysis need not be more powerful than the single-stage PROC MIXED analysis presented by Overall et al. (1999). That is, like Overall et al., we also found the endpoint analysis to be effective in controlling the rate of Type I error and to have power similar to that for the single-stage procedures when data are not missing at random. However, our findings indicate that the single-stage strategy was never less powerful than the two-stage endpoint analysis and was in some cases substantially more powerful. In general, our recommendation for analysis would be to include both covariates in Overall et al.'s single-stage procedure.

Reference

- 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, 265-286.
- Algina, J. Keselman, H. J, & Othman, A. R. (2003). Analyzing group by time effects in longitudinal two-group randomized trial designs with missing data. *Journal of Modern Applied Statistical Methods*, 2, 50-72.
- Hedeker, D., & Gibbons, R. D (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64-78.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Inc.
- 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. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125-134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrics*, 81, 471-483.

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*, 1, 377-388.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SAS Institute. (1995). *Introduction to the MIXED procedure: Course Notes*. Cary, NC: Author.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall/CRC.

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

An Overview Of The Respondent-Generated Intervals (RGI) Approach To Sample Surveys

S. James Press
University of
California, Riverside

Judith M. Tanur
State University of New
York, Stony Brook

This article brings together many years of research on the Respondent-Generated Intervals (RGI) approach to recall in factual sample surveys. Additionally presented is new research on the use of RGI in opinion surveys and the use of RGI with gamma-distributed data. The research combines Bayesian hierarchical modeling with various cognitive aspects of sample surveys.

Keywords: anchoring, Bayesian, confidence scale, recall, surveys

I. Introduction

This work provides an overview of the research to date on the Respondent-Generated Intervals, or RGI, protocol for asking questions in sample surveys. It brings together a body of research that started in 1996 with some theoretical ideas about how survey questionnaire design might be improved by asking respondents for more than just a basic answer to a question, but by also trying to elicit information about how certain the respondents might be about their answers. Over the years we developed various theoretical models for analyzing such RGI data from a survey, culminating in the current Bayesian hierarchical model detailed in Section II. With the development of a theoretical model came the

need to explore how well the model might work in practice, with real people and real empirical data.

We examined pencil-and-paper classroom surveys, and a telephone survey using Census data. We have thought about possible internet surveys, but have not yet fielded this type of survey. Various surveys we carried out under the different survey protocols are described in Section III. Our conclusions so far can be found in Section IV.

The RGI protocol was originally developed to deal with survey questions requiring recall of numerical facts; it has since been extended to address questions of opinion as well. This extension will be discussed below. In its original form, the RGI protocol for asking questions in sample surveys involves asking each respondent not only for a basic answer to a recall-type question (an answer we call a “usage quantity”) but also, for a smallest value his/her true answer could be, and a largest value his/her true answer could be. We’ll refer to these values as the lower and upper bounds. The result of the RGI protocol is that the respondents themselves generate the intervals in which their true beliefs

Dr. S. James Press is a Distinguished Professor of Statistics, at the University of California, Riverside. Email: jpress@ucr.edu. Dr. Judith M. Tanur is a Distinguished Teaching Professor at the State University of New York, Stony Brook. Email: jtanur@notes.cc.sunysb.edu.

lie, instead of having their quantitative beliefs forced into intervals pre-assigned by the survey designer, as is often done in other survey protocols. (For a discussion of other survey protocols using intervals or brackets, see Press, 2004).

Interval-Response Surveys

Survey protocols that permit the respondent to give answers in intervals, self-determined, or pre-assigned by the survey designer, are often preferred by respondents for sensitive questions because the respondent need not be specific about the exact value being requested. Interval response protocols are also often preferred by respondents for questions for which the answers are not very well known. By responding in intervals for such questions, respondents need not be precise about the exact answer (see Lusinchi, 2003). Respondents prefer the RGI technique because it allows them to have control over their disclosures, and RGI allows respondents to feel confident about the accuracy of the information they provide. The intervals RGI respondents provide tend to be narrower than pre-defined intervals (see Schwartz and Paulin, 2000).

Genesis of RGI

The RGI protocol for questionnaire design has its origins in Bayesian assessment procedures. In that context, for a specific individual, we might assess an entire prior distribution about an unknown parameter. That prior distribution represents the individual's degrees of uncertainty about that unknown parameter. In certain contexts, we might assess many points on the individual's subjective probability distribution for that parameter by means of a sequence of elicitation questions, and then connect those points by a smooth curve that purports to represent the underlying distribution. In the RGI protocol, because of concern for respondent burden in surveys, we ask for only three points on the recall distribution.

For example, using some purely hypothetical numbers, suppose an individual has a normal subjective probability distribution representing " θ_0 ", the true (but unknown) change in the number of doctor visits he/she

believes he/she made last year, compared with the previous year, so that $\theta_0 \sim N(4,1)$. (We use "change" in doctor visits as our illustrative variable in order to provide for both positive and negative values of the variable; thus we make the assumption of normality more plausible.) In such a case, the individual believes that it is most likely that he/she visited a doctor 4 more times last year than the previous year, with a standard deviation of 1.

So this individual equivalently believes that there is a 99.7% chance that he/she visited a doctor between 1 and 7 more times last year, or that there is really almost no chance that the true number of additional times was less than 1 or greater than 7. This probability distribution is subjective, in that it represents a specific individual's degrees of belief about his/her uncertainty about the underlying quantity, in the case of this example, the individual's uncertainty about how many more visits he/she believes he/she truly made to the doctor last year compared with the previous year.

We postulate that: in a factual survey each respondent has a distinctive recall distribution, and in an attitude or opinion survey he/she has an underlying probability distribution for his/her opinion or attitude about some issue. In the case of a recall-type question, we assume that the respondent knew the true value at some time in the past (or knew enough to construct an accurate answer) but because of imperfect recall, he/she is not now certain of the true value. He/she may feel confident that he/she knows the true value (but may be wrong in spite of high confidence), or he/she may be quite uncertain of the true value (and conceivably could be correct about the true value, but not realize it). We furthermore assume that the respondent is not purposely trying to deceive. In the case of opinion or attitude questions, the respondent may have a very fuzzy idea of his/her attitude about an issue, or he/she may feel quite strongly and specifically about it.

II. Theoretical Developments

A. Normal Data for Recall Questions

Suppose respondents answer independently and 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. We'll refer to y_i as respondent i 's "usage quantity" (the term "usage quantity" was introduced originally to reflect estimated frequency of a behavior). The random quantities (y_i, a_i, b_i) are jointly distributed. 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. In other situations the gamma or another sampling distribution might be more appropriate. In such a case, we assume the y_i 's (and the (a_i, b_i)) have been pre-transformed, so that after the transformation, the resulting variables are approximately normally distributed. 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 mean value is θ_i (e.g., each respondent is attempting to recall his/her particular 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 $\underline{y} = (y_i)$, and the column vector of means by $\underline{\theta} = (\theta_i)$. Let $\underline{\sigma}^2 = (\sigma_i^2)$ denote the column vector of data variances. The joint density of the y_i 's is given in summary form by:

$$p(\underline{y} | \underline{\theta}, \underline{\sigma}^2) \propto \exp \left\{ -\frac{1}{2} \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 \right\}. \quad (A3)$$

The joint density of the θ_i 's is given by:

$$p(\underline{\theta} | \theta_0, \tau^2) \propto \exp \left\{ -\frac{1}{2} \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2 \right\}. \quad (A4)$$

So the joint density of $(\underline{y}, \underline{\theta})$ is given by:

$$p(\underline{y}, \underline{\theta} | \theta_0, \tau^2, \underline{\sigma}^2) = p(\underline{y} | \underline{\theta}, \underline{\sigma}^2) p(\underline{\theta} | \theta_0, \tau^2)$$

or, multiplying eqn. (A3) and eqn. (A4), gives:

$$\begin{aligned} p(\underline{y}, \underline{\theta} | \theta_0, \tau^2, \underline{\sigma}^2) &\propto \exp \left(-\frac{1}{2} \left[\sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 \right] \right) \\ &\quad \cdot \exp \left(-\frac{1}{2} \left[\sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2 \right] \right) \\ &\propto \exp \left\{ \left(-\frac{A(\underline{\theta})}{2} \right) \right\}, \end{aligned} \quad (A5)$$

$$\text{where: } A(\underline{\theta}) \equiv \sum_1^n \left(\frac{y_i - \theta_i}{\sigma_i} \right)^2 + \sum_1^n \left(\frac{\theta_i - \theta_0}{\tau} \right)^2. \quad (A6)$$

Expand eqn. (A6) in terms of the θ_i 's by completing the square. This takes some algebra. Then:

$$A(\underline{\theta}) = \sum_1^n \left\{ \alpha_i \left[\left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2 + \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2} \right) \right] \right\}, \quad (A7)$$

$$\alpha_i = \frac{1}{\sigma_i^2} + \frac{1}{\tau^2}, \quad \beta_i = \frac{y_i}{\sigma_i^2} + \frac{\theta_0}{\tau^2}, \quad \gamma_i = \frac{\theta_0^2}{\tau^2} + \frac{y_i^2}{\sigma_i^2}. \quad (A8)$$

Now find the marginal density of $\underline{y}$ by integrating eqn. (A5) with respect to $\underline{\theta}$. Then:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto J(\theta_0) \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\},$$

where

$$J(\theta_0) \equiv \int \exp \left\{ \left(-\frac{1}{2} \right) \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2 \right\} d\theta, \quad (A9)$$

$$\delta_i = \left(\frac{\gamma_i}{\alpha_i} - \frac{\beta_i^2}{\alpha_i^2} \right).$$

Rewriting eqn. (A9) in vector and matrix form, to simplify the integration, it is found that if

$$f \equiv \left(\frac{\beta_i}{\alpha_i} \right), \quad K^{-1} \equiv \text{diag}(\alpha_1, \dots, \alpha_n),$$

$$(\theta - f)' K^{-1} (\theta - f) = \sum_1^n \alpha_i \left(\theta_i - \frac{\beta_i}{\alpha_i} \right)^2. \quad (A10)$$

Carrying out the (normal) integration gives:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto \frac{1}{|K^{-1}|^{1/2}} \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (A11)$$

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:

$$p(y|\theta_0, \tau^2, \sigma^2) \propto \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (A12)$$

Note that the proportionality constant in eqn. (A12) does not depend upon θ_0 . Now apply Bayes' theorem to θ_0 in eqn. (A12).

$$p(\theta_0|y, \tau^2, \sigma^2) \propto p(\theta_0) \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}, \quad (A13)$$

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 upper bound. So adopt a uniform prior distribution on (a_0, b_0) . To be fully confident of covering all possibilities, however, adopt an (improper) prior density. Therefore adopt a prior density of the form:

$$p(\theta_0) \propto \text{constant}, \quad (A14)$$

for all θ_0 on the entire real 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 we could improve on our estimator by using a proper prior distribution for θ_0 instead of the one given in eqn. (A14).) The development for a normal (rather than a vague) prior distribution on the population mean is simple and analogous.

Inserting (A14) into (A13), and noting that $p(\theta_0) \propto \text{constant}$, gives:

$$p(\theta_0|y, \tau^2, \sigma^2) \propto \exp \left\{ \left(-\frac{1}{2} \sum_1^n \alpha_i \delta_i \right) \right\}. \quad (A15)$$

Next substitute for δ_i and complete the square in θ_0 to get, after some algebra, the final result that if:

$$\lambda_i \equiv \frac{\left(\frac{1}{\sigma_i^2 + \tau^2} \right)}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}, \quad \sum_1^n \lambda_i = 1, \quad 0 \leq \lambda_i \leq 1, \quad (A16)$$

the conditional posterior density of θ_0 is seen to be expressible as:

$$(\theta_0 | y, \tau^2, \sigma^2) \sim N(\tilde{\theta}, \omega^2), \quad (\text{A17})$$

where:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i, \quad (\text{A18})$$

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}. \quad (\text{A19})$$

Thus, the mean, $\tilde{\theta}$, of the conditional posterior density of the population mean, θ_0 , 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 mean 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. We must still assess the variances $(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2, \tau^2)$.

Assessing the Variances

Suppose that in addition to eqn. (A1):

$$a_i | a_{i0}, \psi_{ai}^2 \sim N(a_{i0}, \psi_{ai}^2); \quad (\text{A1.1})$$

$$b_i | b_{i0}, \psi_{bi}^2 \sim N(b_{i0}, \psi_{bi}^2), \quad (\text{A1.2})$$

where θ_i in eqn. (A.1) denotes the true population value for the mean usage for respondent i ; a_{i0} , b_{i0} denote the true population values for respondent i 's lower and upper bounds, respectively; and $(\sigma_i^2, \psi_{ai}^2, \psi_{bi}^2)$ denote the corresponding population variances, respectively. Next, using the structure of the normal distribution, assume the approximate bounds for all subjects in the population are

approximately 2 standard deviations on either side of the respective means. Accordingly, take approximately, $4\sigma_i \doteq b_i - a_i$, $i = 1, \dots, n$, as our assessments for the σ_i 's.

Then, define:

$$\begin{aligned} a^* &= \frac{1}{N} \sum_1^N a_{i0}; & b^* &= \frac{1}{N} \sum_1^N b_{i0}; \\ \bar{a} &= \frac{1}{n} \sum_1^n a_i; & \bar{b} &= \frac{1}{n} \sum_1^n b_i, \end{aligned}$$

where: a^*, b^* are averages of the *true* (unobserved) values of these bounds over the entire population; $\bar{a}, \bar{b}$ are the averages of the *observed* values of the bounds over the sample.

Assume approximately:

$$\psi_{a1}^2 = \psi_{a2}^2 = \dots = \psi_a^2; \quad \psi_{b1}^2 = \psi_{b2}^2 = \dots = \psi_b^2.$$

Then,

$$\bar{a} \sim N(a^*, \frac{\psi_a^2}{n}); \quad \bar{b} \sim N(b^*, \frac{\psi_b^2}{n}). \quad (\text{A20})$$

Next note that the true population mean value for respondent i must be between its bounds,

$$a^* \leq \theta_0 \leq b^*. \quad (\text{A21})$$

Case 1—Extended Average Estimator

For 95% credibility on a^* with respect to a vague prior we have (approximating 1.96 by 2, here and throughout, for convenience):

$$\bar{a} - 2 \frac{\psi_a}{\sqrt{n}} \leq a^* \leq \bar{a} + 2 \frac{\psi_a}{\sqrt{n}}; \quad (\text{A22})$$

for 95% credibility on b^* with respect to a vague prior we have:

$$\bar{b} - 2\frac{\psi_b}{\sqrt{n}} \leq b^* \leq \bar{b} + 2\frac{\psi_b}{\sqrt{n}}. \quad (\text{A23})$$

From eqns. (A21), (A22) and (A23) we get:

$$\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \leq a^* \leq \theta_0 \leq b^* \leq \bar{b} + 2\frac{\psi_b}{\sqrt{n}},$$

or:

$$\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \leq \theta_0 \leq \bar{b} + 2\frac{\psi_b}{\sqrt{n}}. \quad (\text{A24})$$

From the normality and 95% credibility,

$$\begin{aligned} 4\tau &= \left(\bar{b} + 2\frac{\psi_b}{\sqrt{n}} \right) - \left(\bar{a} - 2\frac{\psi_a}{\sqrt{n}} \right) \\ &= (\bar{b} - \bar{a}) + \frac{2}{\sqrt{n}}(\psi_a + \psi_b). \end{aligned} \quad (\text{A25})$$

But ψ_a and ψ_b are unknown. Estimate them by their sample quantities:

$$\begin{aligned} s_a^2 &\equiv \hat{\psi}_a^2 \equiv \frac{1}{n} \sum_1^n (a_i - \bar{a})^2; \\ s_b^2 &\equiv \hat{\psi}_b^2 \equiv \frac{1}{n} \sum_1^n (b_i - \bar{b})^2. \end{aligned} \quad (\text{A26})$$

Then, the assessment procedure for τ becomes:

$$4\tau \doteq (\bar{b} - \bar{a}) + \frac{2}{\sqrt{n}}(s_a + s_b). \quad (\text{A27})$$

There is a Minitab 13 macro for computing the Bayesian RGI extended average estimator (See Miller, 2003).

Case 2—Extended Range Estimator

From eqn. (A24), since $a_0 < \bar{a}$, and $\bar{b} < b_0$, we can consider for an alternative assessment procedure,

$$a_0 - 2\frac{\psi_a}{\sqrt{n}} \leq \theta_0 \leq b_0 + 2\frac{\psi_b}{\sqrt{n}}. \quad (\text{A28})$$

Then, (A25) becomes:

$$\begin{aligned} 4\tau &= \left(b_0 + 2\frac{\psi_b}{\sqrt{n}} \right) - \left(a_0 - 2\frac{\psi_a}{\sqrt{n}} \right) \\ &= (b_0 - a_0) + \frac{2}{\sqrt{n}}(\psi_a + \psi_b). \end{aligned} \quad (\text{A29})$$

Using eqn. (A26) gives:

$$4\tau \doteq (b_0 - a_0) + \frac{2}{\sqrt{n}}(s_a + s_b). \quad (\text{A30})$$

Note that the second term in (A27), and in (A30) disappear for large sample sizes, leaving us with just the average or range of the bounds, but for smaller sample sizes, the second term can have a substantial effect.

B. Non-Normal Data for Recall Questions

Suppose the usage quantity data, the y_i 's, follow a 2-parameter gamma distribution instead of the normal distribution assumed in Section IIA. Adopt the probability density structure:

$$\begin{aligned} f(y|\alpha, \beta) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-\frac{y}{\beta}} \quad y > 0, \\ \alpha > 0, \quad \beta > 0, \end{aligned} \quad (\text{B1})$$

with:

$$\begin{aligned} E(Y) &= \alpha\beta \equiv \mu, \quad \text{mode}(Y) = (\alpha - 1)\beta, \\ \text{var}(Y) &= \alpha\beta^2. \end{aligned} \quad (\text{B2})$$

Define a new transformation parameter μ by:

$$\beta = \frac{\mu}{\alpha}$$

We can rewrite the gamma distribution in terms of μ as:

$$f(y|\alpha, \mu) = \frac{1}{\Gamma(\alpha)\left(\frac{\mu}{\alpha}\right)^\alpha} y^{\alpha-1} e^{-\frac{\alpha y}{\mu}}, \quad (\text{B3})$$

with: mean $E(Y) = \mu$,

$$\text{mode}(Y) = \mu - \frac{\mu}{\alpha},$$

$$\text{var}(Y) = \frac{\mu^2}{\alpha}.$$

Now make the normalizing transformation (see McCullagh and Nelder, 1983, Chapter 7.2):

$$Z = 3 \left[\left(\frac{y}{\mu} \right)^{1/3} - 1 \right], \quad (\text{B4})$$

so that now, the transformed variable is approximately a standard normal variable; i.e., $Z \sim N(0,1)$. Under this transformation the precision parameter $\alpha = \sigma^{-2}$ is assumed constant for all observations. Applying this transformation to all the variables creates a new set of standard normal variables. Modifying their locations and scales, as shown below, reduces the problem, approximately, to the one discussed in IIA.

Applying the transformation in (B4) to all the usage quantities gives:

$$Z_i = 3 \left[\left(\frac{y_i}{\theta_i} \right)^{1/3} - 1 \right]. \quad (\text{B5})$$

Now, the Z_i 's are independent, and approximately,

$$Z_i \sim N(0,1). \quad (\text{B6})$$

Next define the new variables, Z_i^* by:

$$Z_i^* \equiv \theta_i + \sigma_i Z_i. \quad (\text{B7})$$

Now we have the Z_i^* 's mutually conditionally independent, and

$$(Z_i^* | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2), \quad i = 1, \dots, n. \quad (\text{B8})$$

Suppose the θ_i 's are exchangeable, with

$$\theta_i \sim N(\theta_0, \tau^2). \quad (\text{B9})$$

Assume

$$p(\theta_0) \propto \text{constant}. \quad (\text{B10})$$

We would like to find a Bayesian estimator of the population mean, θ_0 .

We already know that for given $(\sigma_1^2, \dots, \sigma_n^2, \tau^2)$, by Bayes' theorem,

$$(\theta_0 | Z_1^*, \dots, Z_n^*, \sigma_1^2, \dots, \sigma_n^2, \tau^2) \sim N(\tilde{\theta}, \omega^2), \quad (\text{B11})$$

where the posterior mean of θ_0 is given by:

$$\begin{aligned} \tilde{\theta} &= \sum_1^n \lambda_i Z_i^*, \\ \lambda_i &\equiv \frac{\left(\frac{1}{\sigma_i^2 + \tau^2} \right)}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}, \quad \sum_1^n \lambda_i = 1, \\ 0 &\leq \lambda_i \leq 1, \end{aligned} \quad (\text{B12})$$

and

$$\omega^2 = \frac{1}{\sum_1^n \left(\frac{1}{\sigma_i^2 + \tau^2} \right)}. \quad (\text{B13})$$

Now we substitute approximations for the unknown parameters.

C. RGI And Opinion Questions

Suppose there is a population of opinions about some issue, say, "Issue A". Perhaps the analyst would like to establish the mean of the opinions of all people living in the City of New York about Issue A. There is no "correct" answer for an opinion or for an attitude for a given respondent, as there would be for a person answering a recall-type of question. Similarly, response bias does not have the same meaning as in recall. (With a recall-type of question, one of the reasons for response bias arises out of faulty memory.)

When using RGI for attitudes or opinions we can find both point and interval estimators. The RGI point estimator provides some information about the intensity of opinions of New Yorkers about “Issue A”, more so than would a mere traditional sample mean that includes some people with very fuzzy opinions, and some people who have very firm opinions. RGI can provide various measures of strength-of-opinion. One such is the average range of the bounds supplied by all respondents, $(\bar{b} - \bar{a})$. It can also supply a credibility interval measure of belief. Of course, a confidence interval can also supply an interval measure of belief, but the confidence interval only reflects sampling uncertainty, whereas the RGI credibility interval also reflects individual fuzziness of opinion. The range-of-belief also available with RGI, $(b_0 - a_0)$, is somewhat different in that it measures the distance between the extremes of opinion.

Another measure of strength-of-opinion is one we call “fuzziness.” There is certainly no unique way to define such a quantity. One way might be to measure it using the following scale. Recall that the i th respondent’s bounds are given by (a_i, b_i) , and the usage quantity for respondent i is given by y_i . Now define the fuzziness of respondent i ’s opinion as:

$$f_i = (b_i - a_i) \left[1 - \exp \left\{ -\frac{(b_i - a_i)}{|y_i|} \right\} \right]. \quad (C1)$$

As y_i varies, this measure varies between 0 and $(b_i - a_i)$. It is a monotone increasing function of the range, $(b_i - a_i)$. So the greater the range, the greater the degree of fuzziness, and conversely. Moreover, when $y_i = 0$, $f_i = (b_i - a_i)$. This definition is driven by the need to avoid mathematical difficulties using $(b_i - a_i)/y_i$ when y_i is near the origin.

III. Empirical Studies of RGI

During the time that we have worked on RGI, our thinking has evolved in several directions. We have improved our modeling, the way we assess parameters (the population variance and the prior mean), and the form of our questioning. These changes are reflected in the design, analyses, and findings of our empirical work.

In our very first empirical effort we ran parallel record-check surveys on our campuses, asking students questions about their life on campus. If the student-respondents gave their consent, their answers were verified through the appropriate campus offices. On both campuses we asked about the number of credits the student had earned (CREDITS), about his/her SAT math and verbal scores (SATM, SATV), his/her GPA, the number of grades of C or below s/he had received (Cs), and the number of parking tickets s/he had been given (TICKETS). At the University of California at Riverside (UCR) we also asked about the registration fee (REGFEE) and the recreation center fee (RECFEE) the student had paid at the start of the quarter. At the State University of New York at Stony Brook (SUNY-SB) we also asked about the student activities fee (SAFEE) and the health fee (HEALTH) the student had paid at the start of the semester, as well as the amount s/he had spent on food via the food plan (FOOD) and the number of library fines (FINES) s/he had been assessed.

In the campus surveys there were two versions of the questionnaire, both asking about the same usage quantities in the same order. In one version the first half of the items also asked the respondent to provide an answer for the item (such as credits earned) in the following form:

a) Please fill in the blank – “I would be surprised if I had earned more than _____ credits by the beginning of the quarter”.

We refer to this question form as the “surprise form.” The second half of the items on this ballot asked the respondent to answer a question of the form:

b) Please fill in the blanks – “There is almost no chance that the number of credits that I had earned by beginning of this quarter was less than _____ and almost no chance that it was more than _____.”

We refer to this form of the bounds question as the “interval form.” In the other version of the questionnaire the interval form was used for the first half of the items and the surprise form was used for the second half of the items, hence counterbalancing to control for any order effects.

By assuming normality of the responses, and by defining what we mean by “surprise” (which fractile of the recall distribution corresponds to “surprise”?), a complete recall distribution would therefore be defined for each respondent from the surprise form. Again assuming normality of the responses, and defining what is meant by “almost no chance” (which fractile corresponds to “almost no chance”?), we could also generate a complete recall distribution for each respondent from the interval form. But which of these two approaches, “surprise” or “interval,” was a better way to elicit the desired recall distribution information?

At the time we designed the survey instrument we knew that both methods would give us the respondent’s recall distribution (as described in Section II), and we wanted to compare the efficacies of the two forms. When it came time to analyze the data, however, we realized that the interval form, was preferable, *a priori*. It offered us a direct measure of the location of the respondent’s usage quantity, in case the respondent had not given an answer to the usage question, either as a midpoint of the interval given by the bounds, or as some weighted average of the bounds. This information was not available from the surprise form of the question. Also, lack of symmetry of the responses to the two questions required for the interval form immediately would signal the non-normality of the recall distribution. Hence we only analyzed the data from the interval form questions in this experiment, and only used that form in later experiments.

In an attempt to estimate the population mean, our initial estimation procedure for these experiments compared:

- (1) the usual sample mean;
- (2) the average of the midpoints of the intervals given by the respondents, designated the midpoint estimator; as well as
- (3) a Bayesian point estimator.

That Bayesian estimator was the mean of the posterior distribution of the true population mean value obtained from a two-stage hierarchical model using an assumed normal likelihood, exchangeable normal priors for the means of each respondent’s data distribution, and an exponential distribution for the common precision parameter of the respondents’ exchangeable normal priors. In addition, we adopted a normal prior for the population mean, centered at the mean of the averages of the bounds provided by the respondents (this was called the midpoint estimator, $(\bar{a} + \bar{b}) / 2$).

The posterior distribution for the population mean was complicated (the ratio of multiple integrals), but was evaluated numerically by Gibbs sampling Markov Chain Monte Carlo (MCMC). (See Press, 1997 for a derivation of the estimator, and Press and Tanur, 2000, for further details about the campus experiments.) The results given by these estimators were compared in terms of their closeness to the true means found in record checks.

For the 18 items tested in the two campus experiments, this initial analysis found that the posterior mean was always very close to the midpoint estimator. This similarity was not surprising as we chose deliberately to use a sharp (non-vague) prior. The Bayesian estimator looked relatively good; but it was difficult to compute. Of the three estimates, the Bayesian estimate was least accurate for just one item, the midpoint estimate least accurate 7 times, and the sample mean of the usage quantities least accurate 10 times.

Our next empirical study was carried out during a fellowship at the Bureau of the Census held by S. J. Press. Census Bureau interviewers carried out telephone interviews with respondents from 500 households, asking questions about the household's economic situation. Respondents were asked questions about their income from salary and wages for the most recent calendar year and the year previous to that and about the change in their income from these sources over the previous 5 years. They were also asked similar questions about their income from interest and dividends. This study involved extensive cognitive testing of the question form (see Marquis and Press, 1999), and finally settled on asking 25% of the respondents the usage quantity first, followed by questions about the bounds (e.g.):

- a) What is your best estimate of your household's income from salary and wages in 1997?
- b) What is the lowest the correct value could be?
- c) What is the highest the correct value could be?

Thus the bounds question was broken into two separate questions. In addition, for the remaining 75% of the respondents, the form of the bounds questions shown above was asked before the usage quantity question, rather than in the reverse order, to see whether the order of the questions would make any difference. Bayesian estimation was again carried out using normal priors, and MCMC, as described above. In this work, however, two versions of the estimation were carried out. One used the sample median of the usage quantities as the mean of the prior distribution and the other used the midpoint estimator as was done in the campus experiments.

Because of the split ballot nature of the experiment, there were 12 comparisons possible between the sample mean and the two Bayesian estimators. Of these comparisons with the sample mean, the sample mean was closest to the truth 4 times, the Bayesian estimator using the median closest to the truth 4 times, the Bayesian estimator using the midpoint estimator

closest to the truth 3 times, and there was one tie between the Bayesian estimators. In a "head to head contest" between the two Bayesian estimators, the one using the median as the prior mean was closer to truth 5 times, the one using the midpoint estimator as the prior mean closer to the truth 6 times, and there was one tie. The order in which the usage and bounds questions were asked did not seem to make any difference in the accuracy of estimation. See Press and Marquis (2001) for more details on the Census experiment.

Meanwhile, other progress was being made. Schwartz and Paulin (2000) did a study at the Bureau of Labor Statistics comparing several techniques using bounds/interval questions. They found that respondents liked the RGI technique because they felt it gave them some control over their disclosures of income. They also found that the intervals offered by respondents tended to be smaller than those generated by the investigators themselves in another condition of the experiment. And intervals generated by the respondents had been used in several other contexts. Earlier rounds of the Survey of Consumer Finances used interval estimates to elicit answers from reluctant respondents (Kennickell, 1997) and the 2004 round was planning to put more emphasis on letting respondents who can't or won't give exact amounts determine their own ranges--rather than falling back on a range card or a decision tree (Kennickell, 2004).

Further, Lusinchi (2003) had encouraged respondents on a web survey to use such intervals when they were not sure of their answers. We ourselves (Press and Tanur, 2001) showed that in the early campus experiments up to 41% of respondents who did not choose to give a point estimate of a usage quantity did give a set of bounds. If we use the midpoint of the bounds as an approximation of what the respondent might have answered for the usage quantity, we see that the RGI protocol has the potential to reduce item nonresponse considerably. Clearly, RGI was useful, but we needed to work on the estimation strategy and the question format.

As our thinking evolved, we went on to develop a new model that allowed a closed form solution rather than the MCMC computer

intensive numerical evaluation. That new model was presented in Section II above. The new modeling develops results for a vague prior for the population mean, but results for a proper (normal) prior for the population mean are analogous. We tried this model out on the data from the campus experiments described above. In order to assess the hyperparameters for a proper prior distribution we needed demographic information about respondents. (For example, we needed to know the composition of the sample in terms of year in school in order to derive a prior mean of the number of credits students would have earned. For a description of the how the prior means were derived see Press and Tanur, 2004, p. 272.)

Unfortunately, over the years some of those demographic data for the campus experiments became separated from respondents' reports on the items using the RGI questioning protocol. Hence our reanalysis of the campus experiments could use only 6 variables at SUNY-SB and only 4 at UCR. These results appear in Table 1. We see that the posterior mean, using a proper prior and the range of the bounds to estimate the population variance was closer to truth than the sample mean for 8 of the 10 items. Moreover, the Bayesian credibility interval covered truth for all 10 items, while the traditional confidence interval covered truth only for 6 of the 10.

Table 1 – Comparing Sample and RGI Posterior Means for Estimating Population Means in Campus Experiments Using Normal Priors and Range Estimator

Boldface point estimates denote “winners;” boldface interval estimates denote intervals that cover truth.

SUSB

	Truth	x-bar	Conf. Int.	Post-Mean	Cred. Int.
CREDITS	67.53	63.13	(56.12, 70.14)	63.69	(55.54, 71.84)
GPA	2.91	2.99	(2.89, 3.09)	2.97	(2.85, 3.09)
SATM	570.80	593.72	(572.40, 615.00)	591.97	(553.15, 630.79)
SATV	503.20	526.00	(503.80, 548.20)	519.01	(478.52, 559.50)
TICKETS	0.53	0.92	(.56, 1.28)	0.95	(.32, 1.58)
FINES	1.52	2.25	(0, 5.41)	1.00	(.03, 1.96)

UCR

	Truth	x-bar	Conf. Int.	Post-Mean	Cred. Int.
GPA	3.05	3.10	(3.00, 3.20)	3.04	(2.88, 3.21)
SATM	574.08	572.60	(549.50, 595.60)	574.05	(537.54, 610.56)
SATV	485.40	503.00	(481.20, 524.80)	500.38	(463.74, 537.02)
TICKETS	0.21	0.51	(.27, .75)	0.63	(.09, 1.16)

Clearly, the closed form estimation procedure was doing better than the MCMC procedure, but there was still room for improvement. We turned to issues of assessing the hyperparameters and to the questioning format to attempt further improvement.

We moved to expressing the hyperparameter τ according to Eqns. A27 (for the extended average estimator) and A30 (for the extended range estimator). (Earlier we had taken 4τ to be equal to the difference between the sample means of the bounds for the average estimator or equal to the difference between the highest sample upper bound and the lowest sample lower bound for the range estimator.)

From Equation. A16 it is clear that sample usage quantities that are coupled with narrow intervals receive greater weight in the

Bayesian estimation than do sample usage quantities that are coupled with wide intervals. Hence it would improve estimation if respondents who give accurate usage quantities also gave narrow intervals and respondents who give inaccurate usage quantities gave wide intervals. We had found earlier that there is indeed a correlation between interval length and accuracy (see Press and Tanur, 2003); we set out to improve that correlation via our questioning.

To test these hypotheses we designed a new UCR classroom survey which was administered to a large undergraduate statistics class in spring, 2003. We worked through respondents' confidence, having earlier found a correlation between confidence and accuracy (see Press and Tanur, 2002).

Figure 1. Confidence Scale for RGI Protocol

- 1) What is your best guess as to what your score was on your first exam in this class? (Please don't answer if you've missed the first exam)._____.
- 2) How confident are you about your answer to Question 1? Please answer on the following confidence scale. (Place a check in the first column next to the answer you prefer.)

Confidence Scale

Place a check somewhere in this column	Numerical Score	Interpretation of confidence rating	Which question should I answer next?
	0	I have absolutely no idea what my exam score was	Go to Question 3b
	2.5	I am uncertain what my exam score was	Go to Question 3b
	5.0	I might be right and I might be wrong about what my exam score was	Go to Question 3b
	7.5	I think that I know what my exam score was	Go to Question 3a
	10.0	I am absolutely certain what my exam score was	Go to Question 3a

In the questionnaires, we encouraged/prompted confident respondents to give narrow intervals and less confident respondents to give wide intervals. As in our earlier campus experiments, we asked students about everyday facts of their life on campus that we could verify – we asked for the score the respondent had earned in the midterm for that class, the score on the second homework, and again we asked about the registration fee paid at the beginning of the quarter (for details about this experiment, see Chu, Press, and Tanur, 2004). But before the respondent answered each question, s/he responded to a confidence scale we devised, as shown in Figure 1. The questions the respondent was directed to varied somewhat in format, but essentially they resembled the form shown in Figure 2.

Because we varied the amount of guidance we gave the respondent on how wide or narrow the intervals should be, we had 3 conditions for each of the 3 items we inquired about. Thus we had 9 chances to measure the accuracy of the extended range and extended

average estimators against the accuracy of the sample mean. Using a vague prior, we found that in 6 of these cases the extended average estimate was closest to truth (and in all these cases, the extended range estimate was in second place), in one case the extended range estimate was closest to truth, and in the remaining two cases the sample mean “won.” Using a normal prior (see Chu, in progress) the results are even more encouraging. For the question about the midterm grade the extended average estimate was closest to truth in all 3 cases, and for the other 2 questions the extended range estimate was closest to truth in all 6 cases.

In both this classroom experiment and in another that followed some months later (and is described just below), we varied the amount of guidance we gave the respondents about how wide their intervals should be if they were not confident about the accuracy of their recall. This manipulation worked in that those instructed to give a wider interval did indeed give a wider one on average than those who were instructed to give a less wide interval. Thus, the results given

Figure 2. Classroom Experiment – Form of RGI Question

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. Please fill in:

The smallest my exam score could have been is % _____,

The largest my exam score could have been is % _____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score

Please fill in:

The smallest my exam score could have been is % _____,

The largest my exam score could have been is % _____.

above and those to be presented below use only “obedient” respondents – those who followed our guidelines on how wide their intervals should be. For details on these guidelines and results for all respondents, see Chu, Press, and Tanur (2004).

Because the sample sizes in the classroom experiment of spring, 2003 were small, we ran a similar experiment later (Nov. 5, 2003; see Chu, in progress). The questions were asked in the same form as in the spring, 2003 experiment (including the confidence scale), with the exception that instead of asking about scores on homework the student-respondents were asked for the number of movie videos they owned. (Verification data consisted of an earlier report these students had given to the professor in a questionnaire designed to acquaint the professor with the students’ interests and given as part of regular classroom routine.) In this case, we again used both a vague prior and a normal prior and the extended range and extended average estimators. Again we had 9 cases for which we could compare the estimators. Using a vague prior we found that the extended average estimate was closest to truth in 3 cases, the extended range estimate closest once, and the sample mean closest for 5 cases. When we used a normal prior, the results

were somewhat more encouraging, with the extended average estimate, the extended range estimate, and the sample mean each being closest to truth in 3 cases.

In the November, 2003 survey, almost exactly one year before the 2004 US presidential election we also asked our student respondents an opinion question: “In your opinion, what percentage of the total vote will Mr. George W. Bush receive in the 2004 presidential election (0-100%)?”

We found that the modal response was 40%, in contrast to the actual percent of the popular vote achieved by President Bush on November 2, 2004 of 51%. A graph of the respondents’ bounds plotted against their usage quantities is shown in Figure 3, in which respondents have been ordered first by their usage quantity, then by their lower bounds within values of the usage quantity, and then by their upper bounds within values of the usage quantity and of the lower bounds to smooth the graph as much as possible.

Nevertheless, the many spikes in the graphs, and the wide variations in bounds from one respondent to another, in Figure 3, shows that about a year before the actual presidential election of 2004, these respondents were very uncertain (fuzzy) about how strong or weak the

Figure 3. Opinion about Percent for Bush, 2004 Election. (Group 1, N=80)

support for President Bush would be. It is also interesting to note from Figure 3 that as usage increase beyond about 40%, the spikiness of the graphs tends to decrease, and the lower and upper bounds tend to get closer.

For the opinion data in this example, we have calculated f_i (see Eqn. C1) for all respondents, and present a histogram of the distribution of the f_i 's in Figure 4.

The mean fuzziness for this group of respondents on this question = 18.37; the corresponding standard deviation is 18.31. Note that these data are not available in a traditional survey of opinion where bounds information is not available. So there is an additional "intensity of belief" (or degree of fuzziness of belief) that is being provided by an RGI survey.

The data from this more recent classroom experiment presented an opportunity to refine our modeling. Note that the derivation in Section II assumes that the recall distribution for each respondent is normal. Of course this assumption is untestable, but evidence of possible violations of the normality assumption for the recall distributions might be reflected in a lack of normality in the sample distribution of recall quantities. Chu (in progress) studies the sample distributions for each of the items in the questionnaire. In particular, she finds for one treatment group the distribution of usage

quantities for the question about the midterm examination seems to follow a gamma distribution. Applying the Wilson-Hillferty transformation (Eqn. B4, see McCullagh and Nelder, 1983), should transform the distribution of these data to approximate normality. Work is continuing on applying the gamma transformation to our data sets and on exploring the usefulness of other transformations of the data that will improve the normality of the sample distributions, and we are very hopeful that improving the conformity of the data to the assumptions of the model will improve our estimation results.

IV. Conclusions

The RGI approach to sample surveys has several advantages over more conventional methods of fielding and analyzing surveys.

1) It provides a method for getting respondents to give an answer to sensitive questions which they might not otherwise answer. Respondents generally feel that providing merely bounds to a question that has a numerical answer is less revealing to the interviewer than is answering a question that requires a specific point of estimate for an answer. Hence, RGI can be useful in reducing item nonresponse.

Figure 4. Fuzziness Histogram

- 2) Many respondents feel more comfortable giving their own point estimate and range than their true value could possibly be than merely giving a point estimate, because they feel it is more accurate.
- 3) Respondents to questions that use the RGI protocol are able to provide bounds for their responses as long as the bounds questions are carefully worded, and respondents are prompted with examples.
- 4) It is helpful to have respondents provide confidence scores for how sure they are of their answers.
- 5) Providing respondents with guidance in the width of intervals to use is an approach that can be used for the analyst to focus attention on the answers of those respondents who are most confident of their responses.
- 6) To improve accuracy it is helpful to study a measure of the distribution of the sample data. If the data are non-normal it is likely that a transformation of the data to approximate normality followed by an RGI estimation of the transformed data will generate accurate point and interval estimates of the population parameter.
- 7) When the RGI protocol is used with opinion questions it can provide various measures of intensity-of-belief in the opinions of a group.

References

- Chu, LP. (2005, In progress). Robustness of RGI Survey Estimates. Ph.D. Dissertation, Department of Statistics, University of California, Riverside.
- Chu, LP., Press, S. J., & Tanur, J. M. (2004). Anchoring Questions in the Respondent-Generated Intervals Protocol. *Journal of Modern Applied Statistical Methods*, 3, 2, 417-431.
- Kennickell, A. B. (1997, January). *Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances*. Board of Governors of the Federal Reserve System, Washington, D.C. 20551.
- Kennickell, A. B. (2004). Personal Communication, September 9, 2004.
- Lusinchi, D. (2003). Respondent-Generated Intervals: Do They Help in Collecting Quantitative Data? Paper presented at the Annual Conference of the American Association for Public Opinion Research.
- Marquis, K. H., & Press, S. J. (1999). Cognitive Design and Bayesian Modeling of a Census Survey of Income Recall, in *Federal Committee on Statistical Methodology Research Conference*, Washington, DC, pp. 51-64.
- McCullagh, P., & Nelder, A. J. (1983). *Generalized Linear Models*, London: Chapman and Hall, Chapter 7.2, p. 152.
- Miller, D. (2003). Minitab macros that simplify computing the numerical results required for RGI Bayesian estimates. For information about these macros contact her at diane.m.miller@ngc.com.
- Press, S. J. (1997). *Bayesian Recall: A Cognitive Bayesian Modeling Approach to Surveying a Recalled Quantity*. Technical Report No. 239, Jan. 1997, Department of Statistics, University of California, Riverside, Revision of Tech. Rept. #236.
- Press, S. J. (2004). Respondent-Generated Intervals for Recall in Sample Surveys. *Journal of Modern Applied Statistical Methods*, 3, 1, 104-116.
- Press, S. J., & Marquis, K. H. (2001). Bayesian Estimation in a U.S. Census Bureau Survey of Income Recall Using Respondent-Generated Intervals. *Research in Official Statistics*, 4, 1, 151-168.
- Press, S. J., & Tanur, J. M. (2000). Experimenting with Respondent-Generated Intervals in Sample Surveys, with discussions, in *Survey Research at the Intersection of Statistics and Cognitive Psychology*, Working Paper Series #28, Monroe G. Sirken, Editor, National Center for Health Statistics, Center for Disease Control and Prevention, pp. 1-18.
- Press, S. J., & Tanur, J. M. (2001). Can Respondent-Generated Interval Estimation in Sample Surveys Reduce Item-Nonresponse? Pages 39-49 in M. Ahsanullah, J. Kenyon, and S.K. Sarkar (Eds.), 2001, *Applied Statistical Science V*. Huntington, NY: Nova Science Publishers, Inc.

Press, S. J., & Tanur, J. M. (2002). Decision Making of Survey Respondents. *Proceedings of the Annual Meetings of the American Statistical Association, Survey Research Methods Section, 2002*. Washington, D. C.: American Statistical Association.

Press, S. J., and. Tanur, J. M. (2003). The Relationship between Accuracy and Interval Length in the Respondent-Generated Interval Protocol. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Nashville, TN, May 2003 and *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Washington, D.C.: American Statistical Association.

Press, S. J., & Tanur, J. M. (2004). Relating RGI Questionnaire Design to Survey Accuracy and Response Rate. *Journal of Official Statistics*, 20, 2, 265-287. Special issue on Questionnaire Development, Evaluation, and Testing Methods.

Schwartz, L. K., & Paulin, G. D. (2000). Improving Response Rates to Income Questions. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 965-969. Washington, D.C.: American Statistical Association.

Regular Articles Modeling Incomplete Longitudinal Data

Hakan Demirtas
School of Public Health
University of Illinois at Chicago

This article presents a review of popular parametric, semiparametric and ad-hoc approaches for analyzing incomplete longitudinal data.

Key words: longitudinal data, missing data, multiple imputation, ignorability, estimating equations, selection models, pattern-mixture models

Introduction

Missing observations are common in longitudinal studies. This article focuses on attrition, where responses are available for a subject until a certain occasion, and missing for all subsequent occasions. In the presence of incomplete data, the risk of reaching incorrect decisions is higher, because missing data may degrade the performance of confidence intervals, bias parameter estimates and reduce statistical power. Handling incomplete data generally requires special techniques and inferential tools. In this article, commonly used ad-hoc methods, semiparametric methods and likelihood-based models for incomplete repeated-measures data were reviewed and these approaches were applied to a real dataset.

The real data example pertains to a psychiatric trial in which dropout behavior appears to be quite different in the treatment and control groups. Data were obtained from the National Institute of Mental Health Schizophrenia Collaborative Study, where patients were randomly assigned to receive one of three anti-psychotic medications or a placebo.

As noted by Hedeker and Gibbons (1997), performance of the three drugs was quite similar; following their approach, the subjects from the three drug treatments were collapsed into a single group. The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which was treated as continuous. Measurements were planned for weeks 0, 1, 3, and 6, but missing values occurred primarily due to dropout. A few patients had missing measurements and subsequently returned; for simplicity these have been removed. A small number of measurements were also taken at intermediate time points (weeks 2, 4, and 5) which were also ignored. These exclusions reduced the sample from 1,603 subject-observations to 1,500.

With these exclusions, the sample contains 312 patients who received a drug and 101 who received a placebo. In the drug group, 3 patients dropped out immediately after week 0, 27 dropped out after week 1, 34 dropped out after week 3, and 248 completed the study. In the placebo group, no patients dropped out after week 0, 18 dropped out after week 1, 19 dropped out after week 3, and there were 64 completers. In this trial, the mean profile for placebo group is slightly declining, indicating mild improvement over time, but the drug group declines more dramatically. Dropout affects the two groups differently. If patients are classified as dropouts or completers, the dropouts in the placebo group appear to be more severely ill

Hakan Demirtas is an Assistant Professor of biostatistics at the University of Illinois at Chicago. His research interests are the analysis of incomplete longitudinal data, multiple imputation and Bayesian computing. E-mail address: demirtas@uic.edu

than the completers and show less improvement. In the drug group, however, the opposite occurs: dropouts appear to be less severely ill than completers and improve more rapidly. Mean profiles for dropouts and completers in the two groups are shown in Figure 1. One plausible explanation is that those receiving the placebo who experience little or no improvement may be leaving the study to seek treatment elsewhere. On the other hand, those in the drug group who improve dramatically may be dropping out because they feel that treatment is no longer necessary.

Figure 1. Mean observed response in psychiatric trial by treatment group (placebo, drug) and dropout status (dropout, completer), plotted versus T = square root of week.

Organization of this article is as follows: An overview is provided with background information on incomplete longitudinal data and ignorability. Popular longitudinal modeling techniques such as linear and nonlinear mixed models, semiparametric marginal approaches and their weighted versions, single imputation and its variants, multiple imputation, selection and pattern-mixture models are presented along with the implications of missing data for these commonly used methods. In the portion dealing with application, most of the mentioned methods were applied to the psychiatric trial dataset and findings were compared. Conclusions include remarks and discussion stressing the importance of sensitivity analyses and robustness studies.

Overview

Mechanisms for missing data and dropout

The properties of missing-data methods depend on the manner in which data became missing; every missing-data technique makes implicit or explicit assumptions about the missing-data mechanism. In this section, major classes of missing-data mechanisms were discussed, emphasizing the taxonomy introduced by Rubin (1976).

Many missing-data procedures in use today assume that missing values are missing at random (MAR) (Rubin, 1976). Let Y denote the complete set of responses for all subjects, and suppose that the distribution of Y depends on a set of unknown parameters of interest θ . Let R be the associated set of missing-value indicators. The elements of R take the values 1 or 0, indicating whether the corresponding elements of Y are observed or not. The conditional distribution of R given Y depends on the set of parameters δ . Let $Y = (Y_{obs}, Y_{mis})$ denote the partition of the data into the respective sets of observed and missing values. Finally, let (y_{obs}, r) be the realized value of (Y_{obs}, R) .

The missing values are said to be MAR if

$$P(R = r | Y_{obs} = y_{obs}, Y_{mis}, \delta) = P(R = r | Y_{obs} = y_{obs}; \delta)$$

holds for all possible δ . Under MAR, the probability distribution of the indicators of missingness may depend on the observed data but must be functionally independent of the missing data. Intuitively speaking, MAR means that once appropriate account is taken of what have been observed, there remains no dependence of the missingness on unobserved quantities. A simple example is a two-occasion study of blood pressure where subjects are called back for the second measurement if the first measurement is high. This example is MAR because missingness on the second measurement depends only on the value of the first measurement which is always observed.

An important special case of MAR is missing completely at random (MCAR). Under MCAR, $P(R = r | Y_{obs} = y_{obs}, Y_{mis}; \delta) = P(R = r; \delta)$ for all possible δ . In this case, the response probabilities are independent of both the observed and unobserved parts of the dataset.

Suppose, for example, in a two-occasion study of blood pressure, a randomly chosen subset of subjects is called back for a second measurement. In this case, the missing-data mechanism is MCAR, because the probability that the second measurement is missing does not depend on blood pressure at either occasion.

If MAR is violated, the response probabilities depend on unobserved data; in this case, the missing values are said to be missing not at random (MNAR). MNAR situations require special care; to obtain correct inferences, one must specify a joint probability model for the complete data and the indicators of missingness.

Types of dropout

When missing data arise only through dropout, R can be summarized in a single variable that records the first time at which a value is missing or the time of a subject's last observed measurement. Special terminology has evolved for dropout, and this terminology is best understood by its relationship to MAR, MCAR and MNAR (Diggle & Kenward, 1994; Little 1995; Verbeke & Molenberghs, 2000).

Under MAR dropout, the probability of dropout may depend on observed covariates and past responses. Nonignorable dropout (ND) is used interchangeably with MNAR. Under ND, the dropout probability may depend on unobserved covariates, current and future unobserved responses. Little (1995) clarified the role of covariates in this classification scheme. He used "covariate-dependent dropout" (CDD) for the situation where dropout may depend on completely observed covariates. Under CDD,

$$P(R = r | Y_{obs} = y_{obs}, Y_{mis}, x; \delta) = P(R = r | x; \delta)$$

where x is the realized value of fully observed covariates X . A clinical trial where dropout rates differ among treatment groups, but otherwise unrelated to responses, would be an example of this type. Diggle and Kenward (1994) use the terms *random dropout* (RD) for MAR dropout, *informative dropout* (ID) for nonignorable dropout, and *completely random dropout* (CRD) if the dropout does not depend on responses or covariates. Little's terminology is more consistent with the literature on general missing-

data problems, because the term completely random has historically been reserved for situations where missingness does not depend on any variables at all.

Ignorability

An important concept in the theory of missing data, closely related to MAR, is ignorability. A missing-data mechanism is ignorable if (a) the missing data are MAR and (b) the parameters δ and θ are distinct (Little & Rubin, 2002). From a frequentist perspective, distinctness means that the joint parameter space of (δ, θ) is the Cartesian cross-product of the individual parameter spaces for δ and θ . From a Bayesian perspective, it means that the joint prior distribution of (δ, θ) factors into independent priors for δ and θ (Schafer, 1997a).

The term ignorable suggests that the missing-data mechanism can, in some sense, be ignored when performing statistical analyses. Rubin (1976) precisely explained what it means to ignore the missing-data mechanism, both from frequentist and likelihood/Bayes standpoints, and provided conditions under which ignoring the missing-data mechanism is valid for inferences about θ . In the frequentist case, ignoring the missing-data mechanism means fixing R at its realized value and using $P(Y_{obs} | R = r; \theta, \delta)$ as a repeated-sampling distribution. That is, it is pretended that Y_{obs} is the data that had been intended to collect. In the likelihood/Bayes situation, ignoring the missing-data mechanism means using

$$\int P(Y_{obs} = y_{obs}, Y_{mis}; \theta) dY_{mis}$$

as the likelihood function for θ . The conditions under which these approaches are valid differ. In the likelihood/Bayes case, ignoring the missing-data mechanism is valid when are distinct and the missing data are MAR. In the frequentist case, the stronger condition of MCAR is needed.

This definition of ignorability seems to implicitly assume that one is working within a likelihood-based or Bayesian context. The reason why the missing-data mechanism can be

ignored under this condition is that joint log-likelihood for δ and θ partitions as $l(\theta, \delta; y_{obs}, r) = l(\theta; y_{obs}) + l(\delta; r)$. Information about the complete-data population parameter is contained fully in the first term; inferences about θ are unaffected by R , and there is no need to model $P(R = r | y, \delta)$. However, if one is not working in likelihood-based or Bayesian frameworks, one may need to formally model R even when the missing data are MAR. Therefore, the appropriateness of not modeling the missing-data process is not a property of the mechanism alone, but a property of the mechanism and the method of analysis.

The precise meaning of ignorability and its implications have often been misunderstood and misapplied, because many statistical procedures in use today are actually a hybrid of likelihood and frequentist approaches. For example, the use of an expected information matrix is frequentist, because it takes an expectation over the distribution of all possible data values. Helpful discussion and clarification of this point is given by Kenward and Molenberghs (1998).

Nonignorable modeling

Any violation of MAR leads to a nonignorable missingness mechanism. No simplification of the joint distribution is possible, and inferences can only be made about marginal responses by making further assumptions about which the observed data alone carry no information (Little & Rubin, 2002; Little, 1995). Under MNAR, the missingness mechanism does not drop out of the likelihood; the missingness indicators provide information about the parameters of the complete-data population. In these situations, assuming MAR may lead to biased estimates of parameters of the complete-data population; joint modeling of longitudinal response and dropout mechanism is needed.

Completers only analysis

Omitting the subjects with missing observations tends to introduce bias, to the extent that the incompletely observed cases differ systematically from the complete cases. Completers may be unrepresentative of the

population for which the inference is usually intended: the population of all cases, rather than the population of cases with no missing data. In longitudinal studies with human or animal subjects, not all subjects complete the study and especially when completers and dropout seem to follow different trajectories, analyzing only the completers may be very misleading and inefficient.

Last observation carried forward (LOCF)

LOCF is often used in the analyses of clinical trials for FDA (Food and Drug Administration). It tends to understate differences in estimated time trends between treatment and control groups. Although LOCF is thought to be conservative, standard errors are biased downward as well, so it is not necessarily conservative. LOCF seems appealing only when between subject variation is high but responses within a subject is relatively stable over time. In this case, last observation may be a decent predictor for missing data points.

Mean imputation

Imputing the subject-mean seriously distorts trends over time and within-subject covariance structure. Imputing the occasion-mean distorts trends within subjects and between-subject variation. Both mean imputation methods introduce bias into longitudinal analyses and seriously impair standard errors and hypothesis tests.

Other single imputation techniques

Imputing from conditional means (e.g. through a regression prediction), from unconditional distributions (e.g. hot deck) or conditional distributions (through a predictive distribution) have been applied to longitudinal data, but the shortcomings of these methods have been well-documented (Little & Rubin, 2002; Schafer & Graham, 2002).

Single imputation strategies outlined above are designed to precisely predict the missing values. However, the goal of a missing-data procedure is to draw accurate inferences about the population quantities (e.g. mean change over time), not to accurately predict missing values. With imputation, the best way to achieve this goal is to preserve all aspects of the

data distribution (means, trends, within- and between-subject variation, etc.). Ad-hoc imputation methods inevitably preserve some aspect, but distort others.

Multiple imputation

Multiple imputation (MI) is a Monte Carlo technique (Rubin 1987, 1996) in which the missing values are replaced by a set of $m > 1$ simulated versions of them. These simulated values are drawn from a Bayesian posterior predictive distribution for the missing values given the observed values and the dropout times.

Carrying out MI requires two sets of assumptions. First, one must propose a model for the data distribution which should be plausible and should bear some relationship to the type of analysis to be performed. In the case of longitudinal analyses, the model should be capable of preserving the correlation structure and time trends within individuals. The second set of assumptions pertains to type of missingness mechanism. An assumption of MAR is commonly employed for MI. However, the theory of MI does not necessarily require MAR; MI may also be performed under nonignorable models.

The key idea of MI is that it treats missing data as an explicit source of random variability to be averaged over. The process of creating imputations, analyzing the imputed datasets, and combining the results is a Monte Carlo version of averaging the statistical results over the predictive distribution of the missing data,

$$\int P(\theta | Y) P(Y_{mis} | Y_{obs}) dY_{mis}.$$

In practice, a large number of multiple imputations is not required; sufficiently accurate results can often be obtained with $m \leq 10$. Once the imputations have been created, the m completed datasets may be analyzed without regard for dropout; all relevant information on nonresponse is now carried in the imputed values. Once the quantities have been estimated, the m versions of the estimates and their standard errors are combined by simple arithmetic as described by Rubin (1987). Let

$\hat{Q}^{(j)}$ and $\sqrt{U^{(j)}}$ denote the estimate and standard error for a scalar population quantity Q obtained from imputed dataset $j = 1, \dots, m$.

The overall estimate of Q is $\bar{Q} = m^{-1} \sum_j \hat{Q}^{(j)}$,

and the overall standard error is $\sqrt{T}$, $T = \bar{U} + (1 + m^{-1})B$, where $\bar{U} = m^{-1} \sum_j U^{(j)}$

and $B = (m-1)^{-1} \sum_j (\hat{Q}^{(j)} - \bar{Q})^2$. Interval

estimates and tests may be based on the approximation $(\bar{Q} - Q)T^{-1/2} \sim t_\gamma$, where

$\gamma = (m-1)(1+r^{-1})^2$, $r = (1+m^{-1})B/\bar{U}$, and the estimated rate of missing information is approximately $r/(1+r)$. Other rules for combining multidimensional estimates and test statistics are reviewed by Schafer (1997a Chap. 4).

MI may not be the best choice for every analysis, but it is a handy statistical tool and a valuable addition to a researcher's methodological toolkit. MI is attractive for a number of reasons. First, it allows researchers to use their favorite models and software; an imputed dataset can be analyzed by virtually any method that would be appropriate if the data were complete. Second, there are many classes of problems for which no direct ML procedure is available. For example, in longitudinal analyses, there is no direct ML method for incomplete covariates when occasions of measurement vary by individual. Third, MI singles out missing data as a source of random variation distinct from ordinary sampling variability. Finally, the separation of the imputation stage from the analysis stage provides flexibility to the entire modeling process.

Simple hypothesis testing and classical analysis of variance (ANOVA)

Let $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$ denote the responses for subject i , $i = 1, 2, \dots, m$ at a common sets of occasions $t = (t_1, t_2, \dots, t_p)$. If there are no missing values, it is said that the data are balanced in the sense that all subjects are measured at a common set of occasions. Simple t-tests based on change in scores (e.g.

$y_{ip} - y_{il}$) can be used to test the mean equality hypothesis. As a generalization, one may assume $y_i \sim N(\mu, \Sigma)$ where $\mu = (\mu_1, \mu_2, \dots, \mu_p)^T$. The classical ANOVA decomposition for repeated measures can be used to determine if means at each time point are equal. Let SS_A , SS_B and SS_{AB} denote sums of squares for subjects, occasions and subject-occasion interactions with degrees of freedom 1, $p-1$ and $(m-1)(p-1)$, respectively. Under the null hypothesis that all occasion-means are equal ($\mu_1 = \mu_2 = \dots = \mu_p$), the test statistics $F = SS_A(m-1) / SS_{AB}$ is distributed as $F_{(p-1), (m-1)(p-1)}$ provided that Σ satisfies the Huynh-Feldt circularity condition ($Var(y_{ij} - y_{ij'}) = 2\lambda$ for $j \neq j'$, for some $\lambda > 0$.) (Huynh & Feldt, 1970). One example of circularity is compound symmetry, which arises when $y_{ij} = \alpha_i + \mu_j + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ so that $V(y_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$ and $Corr(y_{ij}, y_{ik}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2)$. When the circularity assumption is violated, one can use more general multivariate regression models in which Σ is allowed to be unstructured (Seber, 1984).

When missing values destroy the balance, data analysts sometimes discard the subjects until balance is restored, or they impute missing values in such a way that the sums of squares are not distorted so that procedures requiring balanced data may be applied (Dodge, 1985). In agricultural experiments or laboratory settings, data are often balanced or nearly so. But in longitudinal studies with human or animal subjects, measurements at common sets of occasions are unlikely, so classical ANOVA is less common in these situations.

Linear mixed models

Linear mixed models (Laird & Ware, 1982) extend classical ANOVA to handle unbalanced data by relying on improved computational methods. That is, the inferential strategy is changed from exact distributional results to ML estimation. In linear mixed models, the variation in subjects' longitudinal

profiles arises at two levels: At the first level, the vector of repeated measurements for each subject is related to time and time-varying covariates by a relatively small number of estimated subject-specific regression coefficients.

At the second level, one relates these coefficients to additional time-varying and static covariates such as treatment, baseline characteristics, gender and so forth. The linear mixed-model paradigm combines these two stages into a single modeling procedure. These models—which are also known as multilevel models, linear mixed-effects models, random-effects models, random-coefficient models and hierarchical linear models—have been implemented in many software packages, including HLM (Bryk, Raudenbush & Congdon, 1996), MLwiN (Rasbash et al., 2000), the S-PLUS function `lme` (Pinheiro & Bates, 2000), SAS PROC MIXED (Littell et al., 1996) and Stata (Stata Corp., 1997).

Adopting the notation of Laird and Ware (1982), let $y_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^T$ denote the responses for subject i . The number of responses and the times of measurement may vary arbitrarily from one subject to another. The model is

$$y_i = X_i \beta + Z_i b_i + \varepsilon_i \quad (1)$$

where X_i ($n_i \times p$) and Z_i ($n_i \times q$) contain covariates, β are fixed effects, and b_i and ε_i are unobservable random errors distributed as

$$b_i \sim N_q(0, \psi) \quad (2)$$

$$\varepsilon_i \sim N_{n_i}(0, \sigma^2 V_i) \quad (3)$$

independently for $i = 1, \dots, m$. In this model, the vector of repeated measurements on each subject follows a linear regression model where some of the regression coefficients are common to population, whereas other coefficients vary by subject. Because the model does not assume any particular form for X_i and Z_i , it can handle time-varying covariates and unequally spaced

responses. The columns of Z_i usually span a subspace of the linear space spanned by X_i . Centering the distribution of b_i at zero causes β to become the population-averaged regression coefficients, and the random effects $b_1, \dots, b_m$ become perturbations due to inter-subject variation. When the number of measurements is small, the identity matrix (I_i) is typically used for V_i . Patterned correlation structures (auto-regressive, banded) are possible, in which case V_i contains some unknown parameters.

Averaging over the distribution of the latent random effects b_i , the marginal distribution of y_i is

$$y_i \sim N(X_i\beta, \Sigma_i), \quad (4)$$

where $\Sigma_i = Z_i\psi Z_i^T + \sigma^2 V_i$. Therefore, the elements of β represent the effects of covariates in X_i on the mean response, both for a single subject (i.e. given b_i) and on average for the population.

When the data entering the linear mixed model are unbalanced by design, ML estimation using a likelihood derived from (4) is entirely appropriate. If some responses for some subjects are missing, one may omit the missed occasions and apply ML to the reduced data; this is appropriate if the missing responses are MAR.

Nonlinear mixed models

Nonlinear mixed models generalize the linear mixed models to situations where the response is not necessarily normal. They are also known as generalized linear mixed models or generalized linear models with random effects. In these models, one supposes that y_{ij} belongs to an exponential family with $E(y_{ij}) = \mu_{ij}$ and $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$. The link function—the function that determines the relationship between expected mean and covariates—is $h(\mu_i) = X_i\beta + Z_i b_i$. If h is the identity function

and the responses are normal, then this reduces to a linear mixed model. More generally, the nonlinear mixed model can be applied to repeated observations of binary and count variables.

Except in special cases, the likelihood function for nonlinear mixed models

$$L = \prod_i \int P(y_i | b_i) P(b_i) db_i \quad (5)$$

cannot be computed analytically; it can only be approximated by numerical techniques such as Gauss-Hermite quadrature (Abramowitz & Stegun, 1964), adaptive quadrature (Kronrod, 1965) and Laplace expansions (Stroud, 1971). Algorithms for maximizing (5) are considerably more complicated than for the normal linear mixed model. Early programs used a technique called penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993), whereas later programs (HLM, MLWin, PROC NLMIXED) use true ML. True ML is better, because the resulting estimates tend to be less biased. Bayesian inference is also possible by Markov Chain Monte Carlo (MCMC) (e.g., Spiegelhalter et al., 1999).

In the linear mixed model, β is the effect of X_i on μ_i both for a single subject and on average for the population. In the nonlinear case, however, the distinction between subject-specific (SS) and population-averaged (PA) effects naturally emerges:

$E(y_{ij} | b_i) = h^{-1}(X_i\beta + Z_i b_i)$ is the SS mean response, whereas,

$$\mu_{ij} = E[E(y_{ij} | b_i)] = \int h^{-1}(X_i\beta + Z_i b_i) dP(b_i)$$

is the PA mean response. SS and PA effects have different interpretations and are appropriate in different circumstances (Zeger, Liang & Albert, 1988).

When missing data appear in nonlinear mixed models, as long as true ML or Bayesian techniques (not PQL) are used, the implications of missing responses are no different from normal linear mixed models; the procedures work as long as MAR is satisfied.

Semiparametric marginal models

Nonlinear mixed models are based on an SS formulation. Another way to formulate a model is to specify PA effects directly. Liang and Zeger (1986) proposed an estimation technique called generalized estimating equations (GEE) based on a multivariate version of quasi-likelihood (McCullagh & Nelder, 1989; Wedderburn, 1974). This formulation is semiparametric; rather than specifying a full distribution for the response, one only needs to specify its first two moments. That is, (a) the mean response as a function of covariates and (b) variances and covariances of the response as a function of the mean response are specified. In this approach, a broad class of non-Gaussian outcomes can be accommodated. Quasi-likelihood modeling is theoretically attractive, because it yields consistent and asymptotically normal estimates even when the covariance structure is misspecified. For this reason, GEE methodology has become quite popular for the analysis of longitudinal data.

The model is formulated as follows. Let $i = 1, \dots, m$ and n_i denote the subjects and the number of measurements for each subject, respectively. Let $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^T$ be the expectation of $y_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^T$ which is regarded as a function of covariates: $\mu_i = h^{-1}(X_i \beta)$ where β is a $p \times 1$ vector of unknown coefficients, X_i is an $n_i \times p$ covariate matrix, and h is the link function. The covariance matrix for y_i , denoted by V_i , is a function of μ_i (and hence β) and additional unknown parameters.

The estimate of β is obtained as the solution to the quasi-score equations

$$S(\beta) = \sum_{i=1}^m X_i^T \Delta_i A_i V_i^{-1} (y_i - \mu_i) = 0 \quad (6)$$

where $\Delta_i = \partial \beta / \partial \mu_i$. The covariance matrix for y_i is usually parameterized as $V_i = A_i^{1/2} M_i(\alpha) A_i^{1/2} / \Phi$, where A_i is $n_i \times n_i$ an

diagonal matrix with $g(\mu_{ij})$ as the j^{th} diagonal element; g is a hypothesized variance function; $M_i(\alpha)$ is a working correlation matrix and α is a vector that fully characterizes $M_i(\alpha)$; and Φ is a scale parameter. Therefore, the terms in equation (6) depend on β , α and Φ , but β is the parameter of interest whereas α and Φ are nuisance parameters. Solutions are obtained using iteratively reweighted least squares. At each iteration of the algorithm, one must plug in $\sqrt{m}$ -consistent estimates of α and Φ ; for details, see Liang and Zeger (1986). The solution to GEE, $\hat{\beta}$, is $\sqrt{m}$ -consistent, asymptotically normal, and efficient if the hypothesized covariance structure is correct (Zeger and Liang, 1986). But the popularity of the method stems from the fact that approximate unbiasedness and normality hold even if assumptions about second moments are wrong (Diggle et al., 2003). If the assumed covariance structure is correct, a consistent estimator of $\text{Cov}(\hat{\beta})$ is

$$(X^T \Delta A A^T \Delta^T X)^{-1} \quad (7)$$

where X_i is the matrix of stacked X_i 's, A is the stacked A_i 's and Δ is the stacked Δ_i 's. If $V_i \neq \text{Cov}(y_i)$, (7) can be biased. In that case, however, a consistent estimator of $\text{Cov}(\hat{\beta})$ can be obtained by the Huber-White information sandwich,

$$B \left[\sum_i X_i^T \Gamma_i (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^T \Gamma_i^T X_i \right] B \quad (8)$$

where $B = (X^T \Gamma \Gamma^T X)^{-1}$, $\Gamma = \Delta A$ and $\Gamma = \Delta_i A_i$. (Huber, 1967; White, 1980). In the literature (7) is often called a naive or model-based variance estimator, whereas (8) is called a robust or empirical variance estimator.

In practice, users of GEE typically select the variance function g based on the type of response variable. When y_{ij} is a frequency or count, for example, a natural choice is

$g(\mu_{ij}) = \mu_{ij}$. The working correlation matrix $M_i(\alpha)$ is chosen to reflect the hypothesized relationships among responses within subjects. Popular choices of $M_i(\alpha)$ include independence, exchangeable, one-dependent, auto-regressive or unstructured. In the independence model, $M_i(\alpha) = I$ and α is empty. Exchangeability means $M_i(\alpha) = (1 - \alpha)I + \alpha 11^T$. In the one-dependent case, the $(t, t+1)^{th}$ element of M is taken to be α_t . Auto-regressive correlations can be expressed as $Corr(y_{ij}, y_{ik}) = \alpha^{|t_{ij} - t_{ik}|}$, where t_{ij} and t_{ik} are the observation times associated with y_{ij} and y_{ik} , respectively. Under the unstructured model, $M_i(\alpha)$ is completely unspecified. In that case, the data must be able to support the estimation of all unknown correlation parameters, which requires measurements at a relatively small number of common time points.

The GEE and sandwich methods attempt to “robustify” inferences by relaxing assumptions on the data model, but in doing so, they impose stronger assumptions on dropout mechanisms. The impact of missing data in GEE is quite different from parametric modeling. When elements of y_i are missing, one can omit the missed occasions for certain covariance structures. Liang and Zeger (1986) noted that if the working covariance assumptions are correct, the GEE estimator and the model-based covariance matrix (7) are consistent under MAR, because GEE then becomes maximum likelihood (ML). If the covariance assumptions are wrong, consistency of the GEE estimation and the information sandwich generally requires the missing data to be MCAR, because the sandwich has no likelihood interpretation. Work on weighted estimating equations (WEE) attempts to resolve this problem.

Joint models for longitudinal response and dropout

In practice, the hypothesis of random dropout is essentially untestable; it cannot be

verified nor contradicted by examination of the observed data (Little & Rubin, 2002 Chap. 15). If this assumption is doubtful, alternative procedures should be developed, especially when the degree of departure from MAR is thought to be severe. When nonignorable missingness is suspected, it is necessary to make strong assumptions about the missingness mechanism and propose a specific model for it. That is, one needs to model the joint distribution of the longitudinal response and the dropout. From the likelihood point of view, there are two major ways to construct these models based on different factorizations of the joint distribution: selection models and pattern-mixture models.

Selection models

Selection models, which first appeared in the econometrics literature (Heckman, 1976; Amemiya, 1984), combine a model for the distribution of the complete data with a conditional model for the indicators of missingness given the data. In selection models (suppressing the parameters in the notation), the joint distribution of $f(y_i, r_i | x_i)$ is factored as $f(y_i | x_i)f(r_i | y_i, x_i)$. For example, one could assume that (a) a response variable follows a classical linear regression given a set of covariates, and (b) the probability that a response is observed is related to covariates and the response itself through a logit or probit regression function. These regression-type selection models have become a standard tool of econometricians (Maddala, 1983; Greene, 2000). The OSWALD software package (Smith et al., 1996) provides model-fitting routines for longitudinal data; this software is based on an extension of the work in Diggle and Kenward (1994).

Considering the responses and covariates to be the reasons for missingness, as a selection model does, can be intuitively appealing. Despite their conceptual appeal, the reputation of these models among statisticians is highly controversial. For example, Little and Rubin (2002, Chap. 15) argued that results from these models tend to be highly sensitive to departures from the assumptions about the shape of the complete-data population. In one example, Kenward (1998) demonstrated that a

slight perturbation to the population model—assuming a Student's *t*-distribution rather than a normal—caused drastic changes in parameter estimates. For these reasons, many statisticians tend to regard them as non-robust (see the discussion following the article by Diggle & Kenward, 1994).

Pattern-mixture models

Pattern-mixture models, a term coined by Little (1993), refers to the alternative strategy of first modeling the marginal distribution of the missingness indicators, and then the conditional distribution of the complete data given the pattern of missingness. The population of the complete data then becomes a mixture of distributions, weighted by the probabilities of the missingness patterns. Again, suppressing the parameters in the notation, $f(y_i, r_i | x_i)$ is factored as $f(r_i | x_i)f(y_i | r_i, x_i)$.

For example, consider a bivariate sample in which Y_1 is observed for all subjects but Y_2 is missing for some. A simple pattern-mixture model posits a Bernoulli distribution for R , a bivariate normal distribution for (Y_1, Y_2) given that $R=1$, and another bivariate normal distribution for (Y_1, Y_2) given that $R=0$. Because the conditional distribution of Y_2 given Y_1 is unobservable when $R=0$, unverifiable assumptions must be made about this distribution in order to estimate aspects of the distribution of Y_2 in the full population. The assumptions of pattern-mixture models are no less strong than those of selection models, but some consider them to be more honest, because one knows precisely which parameters in the model formulation cannot be estimated from the observed data. Results from fitting these pattern-specific models are then averaged to obtain parameter estimates for the overall population (e.g. Hedeker & Gibbons, 1997). Alternatively, this process of averaging can be performed through multiple imputation (Glynn, Laird & Rubin, 1993).

Little (1995) defined two types of pattern-mixture models for nonignorable dropout: those with outcome-dependent dropout

and those with random-effect-dependent dropout. In outcome-dependent models, subjects are grouped according to their dropout times and identifying restrictions are placed on the missing-value distributions for those groups (Little, 1993; Little & Wang, 1996; Molenberghs et al., 1998). In random-effect-dependent models, a random-coefficient model (1) is formulated with summaries of dropout time included as subject-level covariates (Wu & Bailey, 1989; Hedeker & Gibbons, 1997; Fitzmaurice et al., 2001). Little (1995) suggested that outcome-dependent models are appropriate when reasons for dropout seem closely related to the response variable itself, whereas random-effect-dependent models ascribe dropout to an underlying process (e.g. progression of a disease) which the outcome variable measures only imperfectly.

Weighted estimating equations

GEE may produce biased estimates if there are missing data, unless the data are MCAR. The method breaks down if the data are missing in a non-MCAR fashion, because the estimating equations on which they are based no longer have zero expectation. This problem suggests a method of modifying the estimating equations by applying weights which are proportional to the inverse-probabilities of response. Weighted estimating equations (WEE) that allow for non-MCAR missingness were first proposed by Robins, Rotnitzky and Zhao (1994, 1995). WEE are the semiparametric counterpart of joint modeling.

The price to be paid for incorporating weights is that a model must be specified for the missingness mechanism. Depending on the form of missingness model, WEE can handle MAR and MNAR mechanisms, but the parameters of an MNAR model are harder to estimate. Let W_i be an $n_i \times n_i$ matrix that contains the weights for subject i . W_i replaces the term $\Delta_i A_i V_i^{-1}$ in (6). So the information contained in Δ_i , A_i and V_i^{-1} about β and α is transferred to W_i . The weighted version of estimating equations becomes

$$S(\beta) = \sum_{i=1}^m X_i^T W_i (y_i - \mu_i) = 0 \quad (9)$$

The weight matrix W_i is, in most cases, an $n_i \times n_i$ matrix whose j^{th} diagonal element is an estimate of the reciprocal-probability that the j^{th} element of y_i is observed. In that case, it is easy to see how the weighting scheme leads to a set of unbiased estimating equations. Modifying the notation a bit, let y_{ij} , μ_{ij} , w_{ij} and r_{ij} be the observed response, expected response, weight and missingness indicator, respectively for subject i at occasion j , respectively. The estimating equations become

$$\begin{aligned} S^w(\beta_k) &= \sum_i \sum_j w_{ij} S_{ij}(\beta_k) \\ &= \sum_i \sum_j w_{ij} x_{ijk} (y_{ij} - \mu_{ij}) = 0. \\ w_{ij} &= 1 / P(r_{ij} = 1) = 1 / E(r_{ij}) \end{aligned}$$

implies

$$E_R \left[E_y(S(\beta_k)) \right] = 0.$$

(Carlin et al., 1999). In practice, the selection probabilities w_{ij}^{-1} are unknown and can, at best, be estimated by a logistic regression on similar-type of model for the r_{ij} 's. As shown by Robins et al. (1994, 1995), the asymptotic properties of the method are preserved if the inverse-weights w_{ij}^{-1} are $\sqrt{m}$ -consistent estimates of the actual response probabilities.

In WEE, one is simply discarding the subject-occasions that are difficult to use because of missing responses and/or covariates, and reweighting the rest to make them seem more representative of the population. Robins, Rotnitzky and Zhao (1994) discard subject-observations with missing covariates. Robins, Rotnitzky and Zhao (1995) discard subject-observations with missing responses. Rotnitzky and Robins (1997), Rotnitzky, Robins and Scharfstein (1998) and Scharfstein, Rotnitzky and Robins (1999) discard various sets of subject-occasions for which covariates and/or

responses are missing. The same idea is being applied in every case: estimating the inverse response probabilities using any information that seems to be related to missingness, including static covariates, time-varying covariates, baseline measures, pre-dropout responses or even post-dropout responses. With a post-dropout response, however, the influence on the response probability can only be guessed.

Application

Regarding the psychiatric dataset that was introduced before, Hedeker and Gibbons (1997) noted that the mean response profiles are approximately linear when plotted against the square root of week, and they express time on the square-root scale in their models. Adopting this convention, T (time) is defined to be the square root of week, and the time of last measurement R (which will be relevant in pattern-mixture models) is also expressed on the square-root scale. Furthermore, let G be an indicator for treatment group (0=placebo, 1=drug) and D an indicator of dropout status (0=completer, 1=dropout). The treatment effect is defined to be the difference in average slopes between the drug and placebo groups. In other words, the parameter of interest is the treatment by time interaction (drug effect over time) $G \times T$.

Two ad-hoc approaches (LOCF and completers only), model-based parametric approaches (selection and pattern-mixture models) and model-based semiparametric methods (unweighted and weighted generalized estimating equations) have been applied to this particular dataset and an estimate of treatment by time interaction and its standard error is obtained for each analysis method.

Model fitting procedures for selection models are implemented through OSWALD (Smith et al., 1996). It finds the most likely values of the data and dropout model parameters jointly by the simplex algorithm developed by Nelder and Mead (1965). It allows three components of variance: a random intercept between subjects (with variance ν^2), a measurement error realized independently between two responses (with variance τ^2) and a serial association component (with variance σ^2 and autocorrelation function

$\rho(u) = \exp(-\phi |u|)$. The marginal covariance matrix for y_i is $\sigma^2 H_i + \tau^2 I + \nu^2 J$, where $H_i = \rho(|t_{ij} - t_{ik}|)$, J is the matrix of ones and I is the identity. In linear mixed model notation it is equivalent to $Z_i \psi Z_i^T + \sigma^2 I$. Regression parameters for the data model part are interpreted in the same way as in linear mixed models. It is again assumed that $y_i = X_i \beta + Z_i b_i + \varepsilon_i$ where the columns of X_i are a constant (one), G , T , and $G \times T$. The columns of Z_i are a constant and T . The dropout (D) is assumed to depend on the time of the measurement (T), the treatment group (G) and some function of responses (see below) through a logit link.

Pattern-mixture models are implemented by incorporating summaries of R and their interactions with G and T into the fixed effects design matrix (X_i) in Equation (1). Then, it is proceeded by multiple imputation (MI) to obtain simulated values that are drawn from a Bayesian posterior predictive distribution for the missing values given the observed values and the dropout times. To create MI's for missing elements of y_i in a random-coefficient model, first a prior distribution for β and the covariance parameters in ψ , σ^2 and V_i must be specified. Then, a random value of these parameters is drawn from their joint posterior distribution given the observed elements of y_i . Finally, the missing elements of y_i are drawn from their conditional distribution given the observed elements derived from the marginal model $y_i \sim N(X_i \beta, Z_i \psi Z_i^T + \sigma^2 V_i)$, with β , ψ , σ^2 and V_i replaced by their simulated values. Repeating these steps m times produces m multiple imputations of the missing responses. Applications of MI to pattern-mixture models have been described by Verbeke and Molenberghs (2000) and Thijs et al. (2002). MI without large-sample approximations is possible by Markov chain Monte Carlo (MCMC), as described by Liu et al. (2000). SAS PROC MIXED provides an MCMC procedure for

simulating posterior draws of model parameters without large-sample approximations.

The PAN library for S-PLUS developed by Schafer (1997b) performs these computations rather quickly under conjugate priors for σ^2 and ψ (scaled inverted chi-square and inverted Wishart, respectively) and $V_i = I$. Important issues in using these techniques, including the choice of prior hyperparameters and monitoring convergence of the MCMC algorithm, are discussed in Schafer (2001). Once the imputations have been created, completed datasets are analyzed with a direct maximum likelihood approach under linear mixed effects model that includes G , T and $G \times T$. Finally estimates from $m=10$ imputations are combined by Rubin's (1987) rules. For a deeper discussion of these issues, see Demirtas and Schafer (2003).

Estimating equations-based approaches (GEE and WEE) are implemented through the software package YAGS (yet another GEE solver). An intercept, G , T and $G \times T$ are included in the model. In the unweighted version (GEE), correlation structure has chosen to be "independence" and "exchangeable". In WEE, weights are estimated based on the inverse probability of being observed for every subject-occasion in the dataset. Two ignorable mechanisms were assumed where weights are estimated by a logistic regression in which outcome variable is response/nonresponse indicator and covariates are T , G and some function of responses (see below).

In what follows, SM stands for selection model, PMM stands for pattern-mixture model; GEE and WEE are as defined before. Other details are described below:

LOCF: The last available measurement is carried forward to fill in unobserved cells.

COMP-ONLY: Only subjects having full set of measurements are considered for the analysis.

SM-1: D depends on G , T and the previous response; assumes ignorability.

SM-2: Same as SM-1 except that D depends on the average of available responses rather than the previous response.

- SM-3:** Same as SM-1 and SM-2 except that D depends on the current response rather than previous responses; assumes nonignorable dropout.
- SM-4:** Same as SM-3 except that D depends on the current and previous response.
- PMM-1** Pattern-mixture model with T , G , D , $G \times T$, $D \times T$, $G \times D$ and an intercept in the fixed effects part; random intercept and slope in the random part of the linear mixed model (1).
- PMM-2:** Same as PMM-1, except that a linear term is used for the time of last measurement (R) rather than D .
- PMM-3:** PMM that does the extrapolation within each pattern without borrowing any information from other patterns.
- PMM-4:** PMM that borrows information from completers for inestimable parameters.
- PMM-5:** Same as PMM-4 except that information is borrowed from the neighboring pattern rather than completers.
- PMM-6:** Same as PMM-5 except that information is borrowed from all available patterns (by a weighted average of estimable parameters from all other patterns) rather than the neighboring pattern.
- GEE-1:** Unweighted GEE with “independence” correlation structure.
- GEE-2:** Same as GEE-2 with “exchangeable” correlations.
- WEE-1:** Weighted version of GEE-1 where weights are assumed to depend on T , G and the average of observed responses for each subject.
- WEE-2:** Same as WEE-1 except that the previous response is used rather than the average of observed responses in weight calculations.

Conclusion

Estimated coefficients for drug effect over time ($G \times T$) and their standard errors under different analysis methods are tabulated in Table 1. Estimated coefficients are varying in a fairly wide range as well as their standard errors. Although one can safely conclude that there is a

drug effect over time, the true magnitude of this effect is disputable. True data model and dropout mechanism are rarely known in practice, therefore it is advisable that statisticians should attack the problem with the help of applied researchers/scientists to be more competent with discipline-specific issues. Subject-matter considerations are as important as the actual analysis method.

Another important issue is sensitivity. Models for incomplete data can be sensitive to untestable assumptions and/or inestimable parameters. Sensitivity analyses are universally acknowledged as crucial, because observed data cannot reveal the true missing-data mechanism. These analyses are usually conducted by applying a variety of models to one dataset to see how the estimated effects vary due to differing modeling assumptions. If our basic conclusions about effects of interest do not change drastically over this family, then the scientific validity of these conclusions is enhanced. Conversely, if the answers do exhibit great variation, drawing firm conclusions seems unwise. For examples of sensitivity analyses, see Little and Wang (1996) and Chapter 20 of Verbeke and Molenberghs (2000).

Robustness studies are less common than sensitivity analyses mentioned, but they are also extremely valuable. A robust method will perform well when applied to a variety of situations when its assumptions are not met. Considerations of robustness may allow us to prefer one model, $Model_1$, to another, $Model_2$, even when $Model_1$ and $Model_2$ achieve the same likelihood for the current data set. That is, if a variety of plausible joint population models is devised for response and dropout—different in nature but all tending to produce samples that resemble the observed data—and if, by simulation, it is discovered that $Model_1$ performs better than $Model_2$ across many of these populations, then there may be more of an inclination to trust $Model_1$ than $Model_2$.

Applying models to a variety of populations consistent with observed data is a useful tool to assess robustness of the models under consideration. These simulations can help us to answer important questions that are being

raised by potential users of nonignorable methods. When nonignorable dropout cannot be ruled out, robustness analyses are preferable to placing total faith in a single model. Although the truth is never known, a model that performs well under differing assumptions that yields simulated datasets which mimic the real data can be regarded as more trustworthy.

Although analyzing a real dataset using the proposed methodology is useful and insightful, simulations are needed to assess how well the method performs. Because there is no consensus among statisticians about which competing method is best, many advocate sensitivity analysis by trying a variety of method and then

sensitivity analysis is to simulate the performance of a method when its assumptions are wrong by proposing a variety of populations and dropout mechanisms capable of producing data like actually seen; then simulating behavior of various methods over repeated samples from each population; and identifying methods that seem to perform well for a variety of populations. Simulations driven by the latter approach are recommended to find arguably the best method that leads to accurate estimates and narrow, calibrated intervals under plausible population/dropout mechanisms.

Table 1: Estimated treatment effect with standard error.

<i>Method</i>	<i>Estimate</i>	<i>Standard Error</i>
LOCF	-0.61	0.11
COMP-ONLY	-0.36	0.08
SM-1	-0.74	0.13
SM-2	-0.69	0.12
SM-3	-0.81	-
SM-4	-0.77	-
PMM-1	-0.73	0.09
PMM-2	-0.75	0.09
PMM-3	-0.78	0.12
PMM-4	-0.99	0.15
PMM-5	-1.22	0.36
PMM-6	-0.95	0.17
GEE-1	-0.63	0.09
GEE-2	-0.66	0.10
WEE-1	-0.62	0.11
WEE-2	-0.69	0.11

seeing what happens, and/or identifying parameters that are nearly or truly inestimable and varying them over a plausible range.

This approach is certainly valuable, but limited. Methods that fit the data equally well may give different estimates and intervals for parameters of interest. But, that does not mean that the methods are equally robust to departures from the assumed model. Another approach to

References

- Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions*. National Bureau of Standards Applied Mathematics Series, Number 55. U.S. Government Printing Office: Washington.
- Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics*, 24, 3-61.

- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.
- Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with HLM/2L and HLM/3L programs*. Chicago: Scientific Software International, Inc.
- Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for nonignorable dropout. *Statistics in Medicine*, 22, 2553-2575.
- Diggle, P. J., Heagerty, P., Liang, K., & Zeger, S. L. (2003). *Analysis of longitudinal data* (Second Edition). Oxford: University Press.
- Diggle, P. J., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-94.
- Dodge, Y. (1985). *Analysis of experiments with missing data*. New York: Wiley and Sons.
- Fitzmaurice G. M., Laird, N. M., & Shneyer L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with nonignorable dropouts. *Statistics in Medicine*, 20, 1009-1021.
- Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88, 984-993.
- Greene, W. H. (2000). *Econometric analysis* (Fourth Edition). Upper Saddle River, NJ: Prentice-Hall.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5, 475-492.
- Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64-78.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Fifth Berkeley Symposium in Mathematical Statistics and Probability*, Berkeley: University of California Press, 221-133.
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, 65, 1582-1589.
- Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: an illustration of sensitivity. *Statistics in Medicine*, 17, 2723-2732.
- Kenward, M. G., & Molenberghs G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236-247.
- Kronrod, M. (1965). *Nodes and weights of quadrature formulas*. Consultants Bureau: New York. [English translation of Kronrod (1964)].
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Liang, K., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second Edition). New York: Wiley.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 84, 125-134.
- Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Little, R. J. A., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, 52, 98-111.
- Liu M., Taylor J. M. G., & Belin T. R. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics*, 56, 1157-1163.
- Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge: Cambridge University Press.

- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second Edition). London: Chapman & Hall.
- Molenberghs, G. M., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, 52, 153–161.
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308–313.
- Pinheiro J. C., & Bates D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Rasbash, J., Browne, W., Goldstein, H., & Yang M. (2000). *A user's guide to MLwiN* (Second Edition). London: Institute of Education.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.
- Rotnitzky, A., & Robins, J. M. (1997). Analysis of semi-parametric regression models with nonignorable nonresponse. *Statistics in Medicine*, 16, 81–102.
- Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93, 1321–1339.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–520.
- Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L. (1997b). *PAN: Multiple imputation for multivariate panel data, software library for S-PLUS*. University Park, PA: The Pennsylvania State University, Department of Statistics.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schafer, J. L. (2001). Multiple Imputation with PAN. In A. G. Sayer and L.M. Collins (Eds.) *New methods for the analysis of change* 355–377. Washington, DC: American Psychological Association.
- Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94, 1096–1146.
- Seber, G. A. F. (1984). *Multivariate observations*. New York: Wiley.
- Smith, D. M., Robertson, W. H., & Diggle, P. J. (1996). *Oswald: Object-oriented software for the analysis of longitudinal data in S*. Technical Report MA 96/192, Department of Mathematics and Statistics, University of Lancaster, United Kingdom.
- Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.
- Stata Corporation. (1997). *Stata reference manual*. College Station, TX: Stata Press.
- Stroud, A. H. (1971). *Approximate calculation of multiple integrals*. Prentice Hall: Eaglewood Cliffs.
- Thijs H., Molenberghs G., Michiels B., Verbeke G., & Curran D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3, 245–265.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439–447.
- White, H. (1980). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

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

Zeger, S. L., & Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.

Zeger, S. L., Liang, K., & Albert, S. A. (1988). Methods for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Monte Carlo Evaluation of Ordinal d with Improved Confidence Interval

Du Feng
Texas Tech University

Norman Cliff (Emeritus)
University of Southern California

This article reports a Monte Carlo evaluation of ordinal statistic d with modified confidence intervals (CI) for location comparison of two independent groups under various conditions. Type I error rate, power, and coverage of CI of d were compared to those of the Welch's t -test.

Key words: d statistic, Welch's t , computer simulation, Monte Carlo study

Introduction

One of the most commonly asked questions in social, behavioral, and biomedical research is concerned with whether scores from one group tend to be higher than those from the other (e.g., treatment effects). This type of location comparison questions (or two-sample problems) is usually answered by parametric tests such as Student's t test or Welch's t test, which requires interval level of measurement of the test variables. However, many behavioral and social variables have only ordinal justification (e.g., Likert-scaled data), thus, performing equivalence testing of two means can yield misleading results. Furthermore, Student's t test is known to be not robust when the normality and/or homogeneity of variance assumptions are violated (e.g., Wilcox, 1990, 1991), as they often are in empirical studies (Micceri, 1989; Wilcox, 1996, p.135). Although Welch's test was found to improve on t test under violations of these assumptions, ordinal methods are more appropriate, and can be more powerful, than the t tests for ordinal data.

Du Feng is an Associate Professor in the Department of Human Development and Family Studies, Box 41162, Lubbock, Texas 79409-1162. Email: du.feng@ttu.edu. Norman Cliff is a former president of the Psychometric Society, Society for Multivariate Experimental Psychology, and of Division 5 of the American Psychological Association. The authors wish to thank Dr. Rand Wilcox for his comments on an early draft of the manuscript.

Definition of d

Cliff (1993) introduced a dominance analysis summarized by the ordinal statistic d , which compares the proportion of times a score from one group or under one condition is higher than a score from the other, to the proportion of times when the reverse is true. The population analog of d is called δ (often written Δ). For random variables X_1 and X_2 , $\delta = \Pr\{x_1 > x_2\} - \Pr\{x_1 < x_2\}$. It is equivalent to the form of Kendall's τ called Somer's d (Somer, 1968) when one variable is dichotomous. This measure was introduced and discussed by nonparametric statistics books for years (e.g., Agresti, 1984; Hettmansperger, 1984; Randles & Wolfe, 1979), and its application was emphasized and extended by Cliff (1991, 1993, 1996).

Advantages of the ordinal statistics over the classical ones have been suggested repeatedly, including their robustness and power under departure from normality or equal variance assumptions, being invariant under monotonic transformation, suitability for much behavioral data which can only be given ordinal-scale status, and their descriptive superiority (Caruso & Cliff, 1997; Cliff, 1993; Long, Feng, & Cliff, 2003). From its definition, we can see that δ is the effect size itself. It is more directly related to the research question often asked: whether scores in one group or under one condition tend to be higher than those in another, than is through some kind of comparison of means or medians. The sample d as defined is an unbiased estimate of δ .

$$d = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_1 n_2}, \quad (1)$$

where n_1 and n_2 are the sample sizes for x_i and x_j , respectively.

It was noted that δ is a simple transformation of a measure, $p = \Pr\{x_1 > x_2\}$, proposed by Birnbaum (1956): $p = (\delta + 1) / 2$, which is estimated by a "common language effect size statistic" (McGraw & Wong, 1992), when there is no ties between random scores from the two groups (Long, Feng, & Cliff, 2003). However, δ has advantages over p because it takes into account ties in the data (Long, Feng, & Cliff, 2003). Similarly, Vargha and Delaney (2000) proposed a generalization of the "common language effect size statistic" (CL) suggested by McGraw and Wong (1992), in order to take into account ties between the two groups scores. They called the generalization "A measure of stochastic superiority," which was defined as $A = \Pr\{x_1 > x_2\} + .5 \Pr\{x_1 = x_2\}$. It was noted that A is simply a linear transformation of δ : $A = (\delta + 1)/2$ (Vargha & Delaney, 2000, p.104).

Inferences About δ

With traditional ordinal methods, for example, the Wilcoxon-Mann-Whitney (WMW) rank-sum test (Mann & Whitney, 1947; Wilcoxon, 1945), inferences are usually based on a randomization hypothesis which assumes that the two populations are identically distributed under the null hypothesis. The identical distribution assumption makes the test tend to be sensitive to differences in spread (also called "scale") and shape of the two distributions. However, this assumption is not necessary for making inferences about δ , because the sampling distribution of d is asymptotically normal and normal-based inferences can be made, with σ_d^2 being estimated from the sample. Several researchers (Birnbaum 1956; Cliff, 1991, 1993, 1996; Fligner & Policello, 1981; Hettmansperger, 1984; Mee, 1990; Siegel & Castellan, 1988; Zaremba, 1962) have suggested ways of making inferences about δ based on d with the sample

estimate of its variance, and described the calculation of the sample estimate of σ_d^2 .

Fligner and Policello (1981) introduced a robust version of the WMW test for comparing the medians of two independent continuous distributions, and tested behavior of d , using the sample estimate of its variance. Their results indicated that d behaved well in small samples in terms of Type I error rate and power over a variety of conditions of population distribution. Cliff (1993) suggested a modification of Fligner and Policello's (1981) procedure by deriving an unbiased sample estimate of the variance of d and setting a minimum allowable value for it in order to increase the efficiency of the estimate and to eliminate impossible values. Defining a dominance variable, which represents the direction of differences between scores, as: $d_{ij} = \text{sign}(x_{i1} - x_{j2})$, where x_{i1} represents any observation in the first group, x_{j2} in the second, Cliff (1993) showed that variance of d can be expressed as

$$\sigma_d^2 = \frac{(n_1 - 1)\sigma_{d_i}^2 + (n_2 - 1)\sigma_{d_j}^2 + \sigma_{d_{ij}}^2}{n_1 n_2}, \quad (2)$$

where d_i is

$$d_i = \frac{\#(x_i > x_j) - \#(x_i < x_j)}{n_1}, \quad (3)$$

and similarly for d_j .

The unbiased sample estimate of σ_d^2 was shown to be

$$s_d^2 = \frac{n_1^2 \Sigma(d_i - d)^2 + n_2^2 \Sigma(d_j - d)^2 - \Sigma \Sigma(d_{ij} - d)^2}{n_1 n_2 (n_1 - 1)(n_2 - 1)}. \quad (4)$$

To eliminate possible negative estimate of variance, $(1 - d^2)/(n_1 n_2 - 1)$ was introduced by Cliff (1993, 1996) as the minimum allowable value for s_d^2 . For detailed discussion on the σ_d^2 and its components, or the formulas presented above, see Cliff (1993, 1996).

Modification of CI for δ

The CI for δ is traditionally computed by $(d - z_{\alpha/2} s_d, d + z_{\alpha/2} s_d)$. However, this CI was

found to be unsatisfactory in Monte Carlo studies (Feng & Cliff, 1995; Vargha & Delaney, 2000). Delaney and Vargha (2002) used modifications of CI for δ that consisted of using Welch-like dfs . They adopted these dfs from Fligner and Policello (1981) procedure and Brunner-Munzel test (2000), and reported that these modifications improved performance of d (Delaney & Vargha, 2002).

These modifications, however, were used without paying attention to the specific situations in which d with traditional CI performed poorly. Long, Feng, and Cliff (2003) pointed out two reasons why d with traditional CI was unsatisfactory. One reason has to do with a zero estimated variance for d when $d = \pm 1$, in which case the conventional CI reduces to a point $\delta = \pm 1$. The other reason is that the traditional symmetric CI does not take into account the negative correlation between σ_d^2 and δ . They proposed using an asymmetric CI to account for boundary effects on the variance of d due to the negative correlation between σ_d^2 and δ . When $d \neq \pm 1$, using sample estimates of variance of d , the asymmetric CI for δ can be constructed based on the following equation:

$$\delta = \frac{d - d^3 \pm t_{\alpha/2} s_d (1 - 2d^2 + d^4 + t_{\alpha/2}^2 s_d^2)^{1/2}}{1 - d^2 + t_{\alpha/2}^2 s_d^2} \quad (5)$$

When $d = \pm 1$, a conservative approach, leading to relatively wide CI, is to assume the maximum possible variance for d , given δ . The maximum possible variance ($\sigma_{d_m}^2$) occurs when the scores in one group are bimodal with all the scores in the other group falling between the modes, leading to a variance of

$$\sigma_{d_m}^2 = (1 - \delta^2)/n_b, \quad (6)$$

where n_b is the sample size of the bimodal group.

This relation between σ_d^2 and δ^2 in the extreme case was used in constructing a CI for δ when $d = \pm 1$. The method is similar to the one used in constructing a CI for population proportion from a sample proportion (see Hayes, 1973, p.379). Assuming that $(d - \delta)/\sigma_{d_m} \sim$

$N(0,1)$, we have the CI with confidence level $1 - \alpha$: $Z_{\alpha/2} < (d - \delta)/\sigma_{d_m} < Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the critical z -score at the selected α level. The upper and lower limits of the CI for δ are the solutions of the equation

$$Z_{\alpha/2}^2 = \frac{(d - \delta)^2}{\sigma_{d_m}^2} \quad (7)$$

Inserting Equation (6) to the above for $\sigma_{d_m}^2$, when $d = 1$, the solution of Equation (7) gives

$$\delta = \frac{(n_b - Z_{\alpha/2}^2)}{(n_b + Z_{\alpha/2}^2)} \quad (8)$$

as the lower limit for the CI when $d = 1$ (in which case the upper limit is 1); and upper limit of the CI when $d = -1$ (in which case the lower limit is -1). With unequal groups, a conservative solution is to use the smaller sample size as n_b in Equation (8). This modification obviates the necessity of using a minimum allowable variance of d .

Methodology

A simulation study comparing rank t test, rank Welch test, Fligner-Policello test, and the d test found d to have inflated Type I error rate (Vargha & Delaney, 2000). However, the above mentioned modifications of CI was not used in this existing study. The primary purpose of the current study was to evaluate the performance of d with modifications of CI that were made based on theoretical and empirical concerns.

A Monte Carlo study was carried out in a variety of situations. To provide a basis for comparison for the behavior of d , the t -test with unpooled variance and Welch's adjustment of df (referred to as Welch's t , or t_w) was included in the analyses. Although this is known to be not completely robust (Wilcox, 1990), it is reasonably so for moderate variance heterogeneity, and it is clearly preferable to Student's t . It sacrifices a little power relative to the latter, but the sacrifice is realistic, especially in forming CI. It is now widely available in

statistical packages and is sometimes even the default statistic for mean comparisons.

Samples of small ($n = 10$) to moderate ($n = 30$) sizes were taken repeatedly from a large number of pairs of uncorrelated populations. In simulating the data, five factors were manipulated: form, mean, variance, skewness of the parent distributions, and sample size. Then, statistical inferences about δ were computed based on each selected pair of samples, and two-sided d and t_w tests at the .05 significance level were performed to compare the two independent groups. Subroutines of IMSL library were called by Fortran programs to generate the populations and samples. Another Fortran program was written to compute statistical inferences about δ for two independent groups and to perform d and t_w tests.

The intention of the present study was to investigate a variety of situations so that the results could be generalized to a wide spectrum of behavioral data. Behavioral variables are often strongly skewed (Miccerri, 1989; Wilcox, 1990, 1991), with concomitant kurtosis, whereas thick-tailed, but symmetric, distributions seem less common. Variables are often bounded by zero, and many are bounded at both ends. Furthermore, distributions differing in location can also differ in scale and/or skewness. Therefore, four families of distributions were selected for the Monte Carlo study: normal, skewed (defined below), chi-square, and beta-distributions. Chi-square and beta-distributions were employed to simulate one-side-bounded and two-side-bounded data with various degrees of skewness, respectively.

Within each family of distributions, certain combinations of means and variances were selected so that δ ranged from .3 to .8. The selection of effect sizes, in terms of δ , conforms to Cohen's (1988) guidelines for small, medium, and large effects for comparable location models.

Normal Distribution

The normal distributions selected had μ of 0, 1, 2, or 3, and σ^2 of 1, 4, or 9. While all pairs of groups with these means and variances were considered, only a subset of them, representing typical results, are reported here.

With symmetric distributions, the null hypothesis for the d analyses, $H_0 : \delta = 0$, is true when the null hypothesis for t_w , $H_0 : \mu_1 = \mu_2$, is true.

Skewed Distribution

Although there appears to be no satisfactory guidelines on what values of skewness are realistic, some studies found that estimated skewness of 2 was not uncommon (Miccerri, 1989; Wilcox, 1990). Thus, skewnesses of -2, 0, and 2 were used to examine the effect of unequal skewnesses. The logistic inverse transformation (Ramberg et al., 1979): $\pm \log(U - 1)^{-1}$, where U represents a uniform distribution on the interval zero to one ($0 \leq U \leq 1$), was used to generate skewed data. This yields distributions having skewnesses of 2 or -2. The transformed data were then re-scaled to have μ of -3, -1, 0, 1, 2, or 3, with σ^2 of 1, 4, or 9.

These skewed distributions also have heavier tails than the normal distribution; their kurtosis tends to be around 5. To avoid a possible effect of unequal kurtosis, and separate it from the effect of unequal skewness, the h -transformation: $Ze^{hZ^2/2}$ (Hoaglin, 1985), where Z is $N(0,1)$, was applied to generate symmetric populations with greater kurtosis. $h \approx .126$ results in kurtosis of around 5, which is comparable to kurtosis of the skewed distributions.

Given the levels of mean, variance, and skewness, there can be 54 different kinds of combinations for each group, and the number is squared when two groups are involved. However, only some representative combinations were selected, and a subset of these are reported here. Unlike in the normal case, for skewed data, the null hypothesis regarding δ and the null hypothesis regarding $(\mu_1 - \mu_2)$ are not necessarily both true or both false, although effects are quite small. Cases when both H_0 's are true or false, as well as one of them is true while the other is false, were included.

Chi-square Distribution

The one-side-bounded data were simulated using chi-square distributions with df ranging from 2 to 32. Certain combinations of the population groups were selected so that the

effect size, δ , fell into the low (.3) to high (.8) range. Several chi-square variates were rescaled by multiplying by constants in order to obtain the desired effect sizes.

Beta Distribution

The two-side-bounded data were generated, according to beta distributions with the first parameter (p) and the second parameter (q) ranging from 1 to 14. Again, certain population groups were selected for comparison, so that δ ranged from .3 to .8.

The null cases for bounded data were those when the two groups had identical chi-square or beta distributions. For the non-null cases, again, the populations compared could have equal or unequal variance, skewness, and kurtosis. For non-normal data, four non-null situations were considered: when two groups were (a) the same in shape (skewness and kurtosis) and scale (variance); (b) the same in shape but different in scale; (c) the same in scale but different in shape; and (d) different in shape and scale.

Sample size, particularly differences in sample size, can profoundly affect the behavior of location comparisons. For each population, observations were simulated for two independent groups using four combinations of the sample sizes $n_1 = 10, 30$, and $n_2 = 10, 30$. Both d and t_w tests were performed for the same data at the $\alpha = .05$ significance level. Two thousand simulation replications were employed under each distributional situation, so that for nominal $\alpha = .05$ and the 95% CI, a .01 difference is significant.

For example, empirical α 's that are higher than .06 are considered significantly higher than the nominal level .05; similarly, CI coverages that are lower than .94 are considered significantly lower than the nominal .95. With 2000 replications and $\alpha = .05$ for the proportions test, the power of the test to detect a departure of $\alpha \pm 1/2\alpha$, which was defined as the "liberal" tolerance criterion (Bradley, 1978) for robustness of Monte Carlo experiments, is .996; the power to detect a departure of $\alpha \pm 1/4\alpha$, the "intermediate" criterion (Robey & Barcikowski, 1992), is .7 (Cohen, 1988; Robey & Barcikowski, 1992).

The d and t_w tests were evaluated and compared in terms of three criteria: empirical Type I error rate, power, and CI coverage. The three criteria evaluate the tests from three different aspects. Coverage of CI has not been addressed as much as the other two by similar studies, though it is equally important and informative, and it is not necessarily implied by the others.

The proportion of the 2000 statistics that exceed the appropriate .05 critical values in the null case is the empirical Type I error rate. It is an estimate of the actual probability of a Type I error. Power is estimated by the proportion of rejection in the right direction at the .05 level in non-null cases. The CI coverage probability is estimated by the proportion of times that the CI constructed by each method covers the corresponding population parameter.

Results

Comparison of empirical α of d and t_w , revealed that with the adjusted CI, d gave rejection rates that were at or below .05 under all circumstances, tending to be conservative when at least one group was small ($n = 10$). On the other hand, use of the simple traditional CI led to liberal empirical α 's (greater than .06) when at least one group was small, particularly when the small n was paired with a larger variance. Welch's t gave several α 's above .06 when group sizes were unequal. It should be noted that none of these departures were above the liberal criterion, even though the range of conditions studied was wide.

The findings about the performance of d are similar to those of Fligner and Policello's (1981) in that d behaved well in terms of controlling the probability of Type I errors, but d appeared to be more conservative in this study with the adjustments on the CI for δ . Table 1 summarizes empirical Type I error rates of d and t_w .

Table 1. Empirical Type I Error Rate of d and t_w for $\alpha = .05$

			$n_1=10, n_2=10$		$n_1=30, n_2=30$		$n_1=10, n_2=30$		$n_1=30, n_2=10$	
	$\sigma_1:\sigma_2$	$\gamma_1-\gamma_2$	d	t_w	d	t_w	d	t_w	d	t_w
Normal	1:1	0	.021 ⁻	.035 ⁻	.048	.052	.041	.051	.037 ⁻	.047
	1:3	0	.031 ⁻	.056	.039 ⁻	.048	.048	.057	.032 ⁻	.048
Skewed	1:1	0	.038 ⁻	.044	.047	.048	.038 ⁻	.064 ⁺	.041	.064 ⁺
	1:3	4	.029 ⁻	--*	.049	--*	.050	--*	.032 ⁻	--*
Chi-square	1:1	0	.028 ⁻	.042	.050	.057	.039 ⁻	.060 ⁺	.049	.061 ⁺
Beta	1:1	0	.033 ⁻	.050	.048	.052	.037 ⁻	.053	.039 ⁻	.057

⁺ At least two standard deviation above .05, computed as if $\alpha = .05$.

⁻ At least two standard deviation below .05, computed as if $\alpha = .05$.

* No Type I error rate of t_w reported because this is a non-null case for means.

Power

Detailed results on the empirical power of the tests are summarized in Table 2. In general, t_w showed slightly higher power than d (when the adjusted CI was employed) in small samples. When both sample sizes were as large as 30, d and t_w had similar power. However, it should be noted that a direct power comparison between the two statistics is not always valid, because they usually had different actual α level and different CI coverage as well. It is also noted that many of the conditions under which t_w had the power advantage are those where its Type I error rate was too high in the null case, or the CI coverage was inadequate. Thus its advantages are largely spurious.

The power of both tests increased with sample size, and with effect size, in the expected ways. However, it appeared that the sample size had a stronger effect on d than on t_w , given that

with moderate samples ($n_1 = n_2 = 30$), the power advantage of t_w became less obvious or disappeared-- d sometimes had slightly higher power than t_w . Figure 1 shows an example of this condition with chi-square distributions.

Power of d with unadjusted CI was slightly higher compared to the reported power with the adjustments. However, as noted, this slight gain in power is associated with higher Type I error rate and poorer CI coverage.

Coverage of CI

With the aforementioned adjustments on the CI for δ (i.e., the adjustment when $d = \pm 1$, and the asymmetric adjustment), d performed well in general in terms of CI coverage, with a few exceptions. This coverage appeared to be a negative function of δ . It was at or above the nominal $1 - \alpha$ level independent of sample size, the form of the population distributions, and

Table 2. Empirical Power of d and t_w for $\alpha = .05$

δ	$\sigma_1:\sigma_2$	sk_1-sk_2	$n_1=10,n_2=10$		$n_1=30,n_2=30$		$n_1=10,n_2=30$		$n_1=30,n_2=10$	
			d	t_w	d	t_w	d	t_w	d	t_w
Normal Distribution										
.218	3:2	0.0	.093	.132	.269	.299	.120	.157	.177	.207
.363	1:1	0.0	.212	.275	.674	.709	.338	.398	.360	.425
.473	3:1	0.0	.280	.438	.848	.918	.301	.442	.773	.869
.520	1:1	0.0	.464	.553	.961	.972	.661	.719	.659	.705
.711	1:1	0.0	.803	.868	1.0	1.0	.947	.967	.943	.971
.820	2:1	0.0	.910	.967	1.0	1.0	.927	.984	1.0	1.0
Skewed Distribution										
.227	1:3	-2.0	.078	--*	.249	--*	.208	--*	.092	--*
.397	1:1	0.0	.254	.218	.743	.513	.158	.037	.311	.368
.472	1:1	4.0	.402	.566	.899	.997	.578	.852	.588	.844
.503	3:1	0.0	.422	.467	.940	.981	.216	.070	.731	.898
.781	1:3	-4.0	.905	.939	1.0	1.0	1.0	1.0	.971	.956
.816	3:1	0.0	.948	.907	1.0	1.0	.999	.941	.992	.992
Chi-square Distribution										
.242	14:1	.9	.089	.191	.289	.712	.096	.186	.259	.661
.346	1:1	-.5	.206	.218	.632	.524	.356	.225	.312	.381
.498	2:1	-.6	.405	.407	.942	.887	.642	.465	.615	.675
.662	1:1	-.1	.728	.794	.998	.998	.941	.942	.845	.915
.807	5:1	-.1	.939	.971	1.0	1.0	.976	.982	.998	.999
.835	1:1	-1.1	.945	.966	1.0	1.0	1.0	1.0	.962	.982
Beta Distribution										
.291	4.5:1	-.1	.130	.249	.463	.768	.143	.236	.334	.611
.327	2.5:1	-.4	.181	.250	.585	.723	.217	.241	.321	.481
.411	1:1	-.1	.276	.343	.802	.837	.440	.507	.454	.512
.553	7:1	-.5	.481	.612	.968	.994	.580	.650	.840	.940
.650	1:1	-.9	.700	.774	.998	.997	.941	.934	.818	.879
.814	12:1	-1.6	.904	.978	1.0	1.0	.937	.991	1.0	1.0

* No power of t_w reported because these are null cases for means.

Figure 1. Empirical power curve with chi-square distribution when $n_1 = n_2 = 30$.

Figure 1
Empirical Power Curve When $n_1 = n_2 = 30$
- Chi-Square Distribution

variance ratio, skewness, and boundedness of the populations compared, unless when δ was quite high (above .7). But it rarely dropped below .93 under all conditions considered. The adjustments provided improvement over the unadjusted CI—the coverage was lower without the adjustments when δ was above .7.

The Welch's t -test yielded good CI coverage for μ_d with normal data, regardless of variance ratio and sample size. However, it was not robust to skewness and nonnormality. The coverage was particularly poor when skewness was combined with heterogeneity of variance, or when high population variance ratio was combined with boundedness and/or small or unequal sample sizes. Table 3 shows results on the empirical CI coverage of d and t_w .

Conclusion

The ordinal method d does not involve excessive elaboration and complicated statistical analyses. Its concept can be easily understood by nonstatisticians. The aforementioned computer program for independent groups d analysis is easy to implement. Its output provides descriptive information, not only the null hypothesis is tested, but also a CI is provided. In addition, a dominance matrix that the program produces is a useful visual aid to the test.

It was a preliminary purpose of this study to evaluate the performance of d with comparison to the Welch's t . The performance of d was evaluated in terms of Type I error rate, power, and CI coverage using a variety of normal and nonnormal data, and was compared to that of Welch's t -test. The findings based on simulations generally show that d , with adjusted CI, has good control over α under all conditions considered. Welch's t controls α at its nominal level with normal data, but sometimes fails to do so under nonnormality. Theory indicates that unequal sample sizes and unequal skewnesses would affect the robustness of t_w (Wilcox, 1990), and the results support this conclusion.

The results on t_w is also consistent with previous researches which found the t_w to be robust when $n_1 = n_2$ (Tan, 1982; Wilcox, 1990), and which showed that t_w was not robust in terms of Type I errors when the two groups had

unequal variances, unequal sample sizes, and unequal skewnesses (Wilcox, 1990). Although, t_w behaved better here than Wilcox (1990) reported, probably because the levels of nonnormality examined in this study were not as high as in Wilcox (1990).

Adjustments of the CI for δ were proposed here, and it was examined whether and to what extent the adjustments improved the distributional behavior of d . The simulation results suggest that these adjustments improve the performance of d in term of Type I error rate and coverage, with a slight loss of power. However, the coverage is not completely satisfactory—it is adequate when δ is not too high, but can be low when the population δ is close to 1. Perhaps even further modification on the construction of the CI for δ is needed.

For both d and t_w , using normal or nonnormal data, under each selected effect size, the performance of the tests were better when the sample sizes were larger. This is accounted for by the central limit effect.

The findings of this study are partly consistent with those of Fligner and Policello's (1981) in that both studies suggest that the small sample behavior of d is good in terms of Type I error rate under normality, and it is robust when there is shift in scale. However, in our study, without the adjustments on the CI for δ , d sometimes appears to be more liberal in terms of actual α .

In this article, skewed, chi-square, and beta-distributions were selected for the purpose of assessment. More types of nonnormal distributions, such as heavy-tailed distributions, can be used in future simulation studies testing the behavior of the statistics. Further more, the distribution characteristics of the d statistic, its variance, and other components such as d_i and d_j should be investigated in further detail.

Several ad hoc analyses of d and s_d were carried out in an attempt to shed light on the reasons both for its good behavior and for the exceptions. No conclusions are possible so far, but some directions for investigation are suggested by these analyses. One aspect of the regular behavior of d may lie in the relative stability of s_d^2 as an estimate of σ_d^2 under most circumstances.

Table 3. Estimated Confidence Interval Coverage Probability of d and t_W for $\alpha = .05$

δ	$\sigma_1:\sigma_2$	sk_1-sk_2	$n_1=10,n_2=10$		$n_1=30,n_2=30$		$n_1=10,n_2=30$		$n_1=30,n_2=10$	
			d	t_W	d	t_W	d	t_W	d	t_W
Normal Distribution										
.218	3:2	0.0	.962 ⁺	.950	.955	.951	.959	.945	.957	.945
.363	1:1	0.0	.964 ⁺	.949	.952	.946	.953	.943	.954	.947
.473	3:1	0.0	.951	.946	.959	.950	.942	.946	.950	.945
.520	1:1	0.0	.951	.958	.960 ⁺	.945	.956	.950	.942	.949
.711	1:1	0.0	.933 ⁻	.945	.951	.956	.949	.953	.945	.942
.820	2:1	0.0	.921 ⁻	.950	.930 ⁻	.955	.862 ⁻	.941	.946	.955
Skewed Distribution										
.227	1:3	-2.0	.964 ⁺	.918 ⁻	.963 ⁺	.935 ⁻	.954	.941	.971 ⁺	.896 ⁻
.397	1:1	0.0	.962 ⁺	.965 ⁺	.960 ⁺	.950	.956	.913 ⁻	.942	.937 ⁻
.472	1:1	4.0	.958	.915 ⁻	.957	.951	.961 ⁺	.925 ⁻	.963 ⁺	.926 ⁻
.503	3:1	0.0	.956	.913 ⁻	.964 ⁺	.932 ⁻	.966 ⁺	.851 ⁻	.949	.948
.781	1:3	-4.0	.961 ⁺	.914 ⁻	.950	.924 ⁻	.953	.938 ⁻	.950	.910 ⁻
.816	3:1	0.0	.938 ⁻	.903 ⁻	.944	.934 ⁻	.951	.912 ⁻	.908 ⁻	.951
Chi-square Distribution										
.242	14:1	.9	.968 ⁺	.916 ⁻	.960 ⁺	.934 ⁻	.967 ⁺	.923 ⁻	.963 ⁺	.942
.346	1:1	-.5	.958	.956	.951	.947	.962 ⁺	.942	.950	.938 ⁻
.498	2:1	-.6	.956	.956	.963 ⁺	.951	.961 ⁺	.931 ⁻	.951	.953
.662	1:1	-.1	.947	.951	.947	.945	.951	.948	.928 ⁻	.945
.807	5:1	-.1	.938 ⁻	.951	.938 ⁻	.944	.922 ⁻	.939 ⁻	.931 ⁻	.957
.835	1:1	-1.1	.925 ⁻	.951	.924 ⁻	.954	.948	.956	.879 ⁻	.943
Beta Distribution										
.291	4.5:1	-.1	.959	.944	.957	.949	.962 ⁺	.925 ⁻	.956	.950
.327	2.5:1	-.4	.965 ⁺	.944	.953	.946	.962 ⁺	.933 ⁻	.958	.946
.411	1:1	-.1	.963 ⁺	.955	.949	.948	.957	.950	.950	.946
.553	7:1	-.5	.954	.939 ⁻	.955	.943	.946	.930 ⁻	.951	.945
.650	1:1	-.9	.946	.959	.949	.949	.957	.950	.930 ⁻	.935 ⁻
.814	12:1	-1.6	.932 ⁻	.934 ⁻	.931 ⁻	.945	.899 ⁻	.941	.946	.954

⁺ At least two standard deviation above .95, computed as if $\alpha = .05$.

⁻ At least two standard deviation below .95, computed as if $\alpha = .05$.

The cases of relatively poor behavior may result from two sources. One source is the correlation between s_d and d that becomes quite strong when δ is fairly high. The asymmetric CI, given by Equation (10), is one attempt at compensating for this effect, but it seems not to be strong enough when δ is very high, and may be too strong when it is low, at the expense of power. It may also be that there are a few circumstances where s_d^2 is less well behaved, although we do not clearly understand what these circumstances are.

Understanding the behavior of d may be facilitated by noting that it, too, is a mean difference. Let $(p_{11}, p_{12}, \dots, p_{1n1})$ be the values of a variable representing the proportion of x_{j2} scores that are less than each x_{i1} , respectively, and correspondingly for the second sample. That is, $p_{i1} = \frac{1}{2}(d_{i1} + 1)$, and $p_{j2} = \frac{1}{2}(d_{j2} + 1)$. Then d is the difference between the mean p_{i1} and the mean p_{j2} . Each p_{i1} reflects—although it does not equal—a corresponding value of a random variable P_1 . Given a distribution $F_1(X_1)$ and correspondingly $F_2(X_2)$, then for any x_{i1} , $p_{i1} = F_2(x_{i1})$, and vice versa, and each has a distribution, $G_1(P_1)$ and $G_2(P_2)$, respectively. Therefore, the behavior of d depends on the nature of these distributions in much the same way that the behavior of the sample mean difference depends on F_1 and F_2 . A difference is that p_{i1} is a binomial distribution of p_{i1} whose value depends on which x_{j2} happens to be in the sample. The two parts of the expression for the variance of d reflect these two aspects of the sampling process.

Not only does the variance of d depend on the variance of p_{ij} , but the other moments of its distribution depend on the other moments of their distributions. Thus, d is not distribution-free except in the limiting case where F_1 and F_2 coincide, but it depends on the distributions G_1 and G_2 rather than on F_1 and F_2 . The fact that it behaves more robustly than t_w simply reflects the fact that the distributions that determine its behavior tend to have better properties than the distributions of the variables themselves. However, we should not be surprised if situations can be found where the opposite is true. These issues can be investigated in future studies.

In sum, this article has shown that d behaves quite well in small and moderate samples in terms of Type I error rate, power, and coverage of the CI, but not perfectly. The adjustments to the CI improved matters in terms of Type I error rate and coverage. This ordinal statistic is robust to nonnormality, heterogeneity of variance, and unequal sample sizes. Yet, there are a few exceptions to the good behavior of d , and further modification may be needed when the population δ is very close to 1 or -1.

Welch's t -test performs well under normality, but is not robust to nonnormality. Its Type I error rate is inflated, power is lowered, and coverage is inadequate when the populations are skewed, and when nonnormality is combined with unequal variances and/or unequal sample sizes. It is particularly sensitive to skewness.

The d has attractive characteristics as a description of location difference. It is a direct numerical reflection of the tendency for scores in one group to lie generally above those of another. It is also invariant under monotonic scale transformations, so conclusions about location need less qualification. The additional fact that its sampling behavior has to be rated as very good seems to lead to a conclusion that it is the method of choice for location comparison in many situations.

References

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley.
- Best, D. J., & Rayner, J. S. W. (1987). Welch's approximate solution for the Behrens-Fisher problem. *Technometrics*, 29, 205-210.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, 57, 637-654.
- Cliff, N. (1991). Ordinal methods in the study of change. In Collins, L. M. & Horn, J. (Eds.) *Best methods for the analysis of change*. Washington, D.C.: American Psychological Association, 34-46.

- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.
- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. New Jersey: Lawrence Erlbaum.
- Cochran, W. G. (1951). Testing a linear relation among variances. *Biometrics*, 7, 17-32.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, (2nd ed). Hillsdale, NJ: Erlbaum.
- Davenport, J. M., & Webster, J. T. (1975). The Behrens-Fisher problem, an old solution revised. *Metrika*, 22, 47-54.
- Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, 7, 485-503.
- Fligner, M. A. (1979). A class of distribution-free tests for the two-sample scale problem. *Journal of the American Statistical Association*, 74, 889-893.
- Fligner, M. A., & Policello, G. E. II (1981). Robust rank procedure for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 76, 162-168.
- Hayes, W. L. (1973). *Statistics for the Social sciences*, (2nd Ed.). New York: Holt, Winehart, and Winston.
- Hettmansperger, T. P. (1984). *Statistical inferences based on ranks*. New York: Wiley.
- Hoaglin, D. C. (1985). Summarizing shape numerically: the g&h distributions. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.) *Exploring data tables, trends, and shapes*. New York: Wiley, 461-513.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.
- Micceri, T. (1989). The Unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- O'Brien, P. C. (1988). Comparing two samples: Extension of the t, rank-sum, and log-rank tests. *Journal of the American Statistical Association*, 83, 52-61.
- Ramberg, J. S., Tadikamalla, P.R., Dudewicz, E.J., & Mykytka, E.F., (1979). A probability distribution and its uses in fitting data. *Technometrics*, 21, 201-214.
- Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley.
- Robey, R. R., & Barcikowski, R. S. (1992). Type I Error and the number of iteration in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45, 283-288.
- Tan, W. Y. (1982). Sampling distribution and robustness of t, F and variance ratio of two samples and ANOVA models with respect to departure from normality. *Communications in Statistics: Theory and Computation*, 11, 2485-2511.
- Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Education and Behavioral Statistics*, 25, 101-132.
- Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Wilcox, R. R. (1990). Comparing the means of independent groups. *Biometric Journal*, 32, 771-780.
- Wilcox, R. R. (1991). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science*, 1, 101-105.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.
- Zaremba, S. K. (1962). A generalization of Wilcoxon's test. *Monatshefte für Mathematik*, 66, 359-370.

Multivariate Contrasts For Repeated Measures Designs Under Assumption Violations

Lisa M. Lix

Department of Community Health Sciences
University of Manitoba

Aynslie M. Hinds

Department of Community Health Sciences
University of Manitoba

Conventional and approximate degrees of freedom procedures for testing multivariate interaction contrasts in groups by trials repeated measures designs were compared under assumption violation conditions. Procedures were based on either least-squares or robust estimators. Power generally favored test procedures based on robust estimators for non-normal distributions, but was influenced by the degree of departure from non-normality, definition of power, and magnitude of the multivariate effect size.

Key words: à priori contrasts, robust estimators, covariance heterogeneity

Introduction

In a doubly multivariate repeated measures (RM) design, subjects provide data at K successive points in time or for each of K experimental conditions on p dependent variables. For example, measures of physical, social, psychological, and spiritual quality of life may be collected at multiple occasions during a course of treatment or therapy. A grouping factor (i.e., experimental vs. control group) is often included, resulting in a multivariate design in which both within-subjects main and interaction effects can be tested.

One approach for analyzing multivariate RM design is to follow statistically significant multivariate omnibus tests with multiple post hoc contrasts. Typically researchers will examine either strongly restricted contrasts, which are defined on the between- and/or

within-subjects factor levels for a single dependent variable, or moderately restricted contrasts, which are defined on between-subjects and/or within-subjects factor levels for two or more dependent variables (Elliot & Barcikowski, 1994). A multivariate simultaneous test procedure (STP) will control the familywise error rate (FWR), the probability of making at least one erroneous decision regarding the null hypothesis for the contrasts, to the nominal level of significance, α (Bird & Hadzi-Pavlovic, 1983; Elliot & Barcikowski, 1994). The FWRs for both types of contrasts tend to be well below the nominal level of significance, α . In addition, these contrasts may have low power to detect effects in multivariate designs.

An alternate approach is to bypass the omnibus test in favor of à priori multivariate contrasts which test focused hypotheses on the between-subjects or within-subjects factor levels for a linear combination of the dependent variables (Huberty, 1994; Huberty, Chou, & Benitez, 1994; Keselman et al., 1998; Krishnaiah & Reising, 1985). These multivariate contrasts enable researchers to draw conclusions about the localized source of an effect while taking account of the correlation across repeated measurements and dependent variables.

One issue in conducting these a priori contrasts in multivariate designs is controlling the FWR. Timm (2002) has recommended the

Lisa Lix is Assistant Professor in the Department of Community Health Sciences. Her research interests include multivariate methods and longitudinal data/repeated measurements. E-mail: lisa_lix@cpe.umanitoba.ca. Aynslie Hinds is a graduate student in the Department of Community Health Sciences at the University of Manitoba Email: umhinds0@cc.umanitoba.ca.

use of STPs based on the Bonferroni inequality or the studentized maximum modulus. If the development of confidence intervals for these multivariate contrasts is not of primary concern, a stepwise procedure may also be considered (Keselman, Lix, & Kowalchuk, 1998; Tamhane & Dunnett, 1999).

A second issue is the choice of a test statistic and its associated derivational assumptions. In multivariate between-subjects designs, it is known that conventional procedures for testing post hoc contrasts are sensitive to violations of the assumptions of normality and covariance heterogeneity, which underlie the usual multivariate analysis of variance (MANOVA) tests (Bird & Hadzi-Pavlovic, 1983; Sheehan-Holt, 1998). In multivariate RM designs, the two conventional approaches for testing effects are the multivariate mixed model (MMM) and doubly multivariate model (DMM) approaches (Thomas, 1983; Boik, 1988, 1991).

The MMM rests on the stringent assumption of multivariate sphericity (M-sphericity). M-sphericity is the assumption that all pairwise differences of the repeated measurements exhibit a common variance for all dependent variables. In addition, both the MMM and DMM approaches rest on the assumptions of homogeneity of the covariances across between-subjects factor levels and multivariate normality. Because M-sphericity is not likely to be satisfied in practice, the DMM approach has been recommended over the MMM approach. However DMM tests are sensitive to violations of the assumptions of covariance homogeneity and multivariate normality.

The purpose of this article is to compare the conventional DMM procedure to procedures that employ approximate degrees of freedom (ADF) multivariate test statistics that do not rest on the assumption of covariance homogeneity for testing multivariate contrasts in repeated measures designs. Recent research (Lix, Algina, & Keselman, 2003; Lix, Keselman, & Hinds, in press) has derived multivariate ADF tests using robust estimators instead of the usual least-squares estimators which are known to be sensitive to departures from multivariate normality. Thus, it should be possible to obtain a test for multivariate contrasts in RM designs that

is robust to both covariance heterogeneity and multivariate non-normality in multivariate RM designs, while controlling the FWR to α .

Definition of Test Statistics

Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y} = [Y_{ijkl}]$, and Y_{ijkl} is the score for the i th individual ($i = 1, \dots, n_j$; $\sum_{j=1}^J n_j = N$) in the j th group ($j = 1, \dots, J$), on the k th ($k = 1, \dots, K$) repeated measurement and l th dependent variable ($l = 1, \dots, p$). Then $\mathbf{X}$ is an $N \times J$ design matrix with $\text{rank}(\mathbf{X}) = J$, $\boldsymbol{\beta}$ is a $J \times L$ ($L = K \times p$) matrix of nonrandom parameters (i.e., population means), and $\boldsymbol{\varepsilon}$ is an $N \times L$ matrix of random error components. Each row of $\mathbf{Y}$ contains the L -dimensional response vector where the first K columns correspond to the repeated measurements obtained on the first dependent variable, the next K columns correspond to the repeated measurements obtained for the second dependent variable, and so on.

The null hypothesis for a multivariate contrast is

$$H_0 : \boldsymbol{\psi} = \mathbf{c}\boldsymbol{\beta}\mathbf{m} = \mathbf{0}, \quad (1)$$

where $\mathbf{c}$ is a vector that contains the contrast coefficients for the between-subjects effect and $\mathbf{m} = (\mathbf{I} \otimes \mathbf{u}')$, where $\mathbf{I} = \mathbf{I}_p$, the $p \times p$ identity matrix, $\otimes$ is the Kronecker product symbol, and $\mathbf{u}$ defines the contrast coefficients for the within-subjects effects. The best linear unbiased estimator for $\boldsymbol{\psi}$, which can be obtained by the least-squares method, is $\hat{\boldsymbol{\psi}} = \mathbf{c}\hat{\boldsymbol{\beta}}\mathbf{m}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Multivariate interaction contrasts are considered in this manuscript. There are a number of different types of interaction contrasts that may be defined for RM designs (Boik, 1993; Lix & Keselman, 1995). Tetrad contrasts, which are the simplest to define, test for differences between pairs of levels of two factors. For example, in a multivariate RM designs with $J = 3$ and $K = 4$, a tetrad contrast involving the first two levels of the between-subjects factor and the first and third levels of the within-subjects factor would require contrast

vectors of $\mathbf{c} = [1 \ -1 \ 0]$ and $\mathbf{u}' = [1 \ 0 \ -1 \ 0]$.

Under a DMM approach, H_0 is tested with one of several well-known multivariate tests that are functions of the eigenvalues of $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$, where

$$\mathbf{H} = \hat{\boldsymbol{\psi}} [\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}]^{-1} \hat{\boldsymbol{\psi}}', \quad (2)$$

and

$$\mathbf{E} = \mathbf{m}'\mathbf{Y}'[\mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}\mathbf{m}, \quad (3)$$

where $\mathbf{I}_N$ is an identity matrix of dimension N . The tests are the Lawley-Hotelling trace, Pillai-Bartlett trace, Roy's largest root, and Wilk's lambda (Timm, 2002). If the multivariate contrast is a single-degree of freedom contrast on the p dependent variables, then all of these procedures will reduce to Hotelling's (1931) T^2 .

When covariances are heterogeneous, Keselman and Lix (1997) demonstrated that DMM tests will produce inflated Type I error rates for omnibus tests of multivariate within-subjects effects, particularly when group sizes are unequal. Keselman and Lix (1997) showed that an ADF multivariate Welch-James (WJ) procedure due to Johansen (1980) can be used to test multivariate within-subjects main and interaction effects under covariance heterogeneity provided that sample sizes are sufficiently large. Moreover, Vallejo, Fidalgo, and Fernandez (2001) and Lix, Algina, and Keselman (2003) also demonstrated that a multivariate extension of the Brown and Forsythe (BF; Brown & Forsythe, 1974) procedure could be used to test within-subjects omnibus effects. The advantage of one procedure over the other depends on the omnibus effect of interest, total sample size, and the degree of covariance heterogeneity in the data.

Let $\mathbf{S}_j$ represent the sample covariance matrix for the j th group,

$$\mathbf{W}_j = \frac{\mathbf{m}'\mathbf{S}_j\mathbf{m}}{n_j}, \quad (4)$$

and $\mathbf{W} = \sum_{j=1}^J \mathbf{W}_j$. The WJ test statistic is

$$T_{WJ} = \hat{\boldsymbol{\psi}}\mathbf{W}^{-1}\hat{\boldsymbol{\psi}}'. \quad (5)$$

The statistic T_{WJ}/C , where $C = p + 2A - 6A(p + 2)$, is distributed as $F_{\alpha}[v_{WJ1}, v_{WJ2}]$, the $(1 - \alpha)$ percentile of the F distribution with $v_{WJ1} = p$, $v_{WJ2} = p(p + 2)/3A$, and

$$A = \sum_{j=1}^J \left[\text{tr}(\mathbf{I}_K - \mathbf{W}^{-1}\mathbf{W}_j)^2 \right] + \left[\text{tr}(\mathbf{I}_K - \mathbf{W}^{-1}\mathbf{W}_j) \right]^2 / 2(n_j - 1). \quad (6)$$

For the BF procedure, define

$$\mathbf{W}_j^* = \frac{(1 - n_j)}{N} \mathbf{m}'\mathbf{S}_j\mathbf{m}, \quad (7)$$

and $\mathbf{W}^* = \sum_{j=1}^J \mathbf{W}_j^*$. The test statistic is

$$T_{BF} = \left(\frac{v_h^*}{v_e^*} \right) \hat{\boldsymbol{\psi}}\mathbf{W}^{*-1}\hat{\boldsymbol{\psi}}', \quad (8)$$

which is distributed as $F_{\alpha}[v_{BF1}, v_{BF2}]$. The computations for v_{BF1} and v_{BF2} are lengthy, and the reader is referred to Vallejo et al. (2001) and Lix, Algina, and Keselman (2003) for the appropriate formulas.

Lix, Algina, and Keselman (2003) examined the WJ and BF procedures when least-squares estimators are replaced with robust estimators based on trimmed means. To define these procedures, let $Y_{(1)jkl} \leq Y_{(2)jkl} \leq \dots \leq Y_{(n_j)jkl}$, represent the ordered observations associated with the j th level of the between-subjects factor, the k th level of the repeated measures factor and the l th dependent variable. 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 is $h_j = n_j - 2g_j$. The trimmed mean is estimated by

$$\hat{\mu}_{yjl} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)jkl}.$$

(9)

Wilcox (1995a,b) has recommended that 20 percent trimming be adopted. It should be noted

that this is a univariate perspective on trimming, in which the most extreme scores for each column of $\mathbf{Y}$ are trimmed independently of the extreme scores in each of the other columns.

In order to obtain the sample Winsorized covariances, the sample Winsorized mean must first be computed and it is obtained by replacing the g_j smallest values with the γ percentile score, and the g_j largest values with the $(1 - \gamma)$ percentile score

$$\hat{\mu}_{wjkl} = \frac{1}{n_j} \sum_{i=1}^{n_j} Z_{ijkl}, \quad (10)$$

where

$$\begin{aligned} Z_{ijkl} &= Y_{(g_j+1)jkl} \text{ if } Y_{ijkl} \leq Y_{(g_j+1)jkl} \\ &= Y_{ijkl} \text{ if } Y_{(g_j+1)jkl} < Y_{ijkl} < Y_{(n_j-g_j)jkl} \\ &= Y_{(n_j-g_j)jkl} \text{ if } Y_{ijkl} \geq Y_{(n_j-g_j)jkl}. \end{aligned}$$

The sample Winsorized covariance is required to obtain a theoretically valid estimate of the standard error of a trimmed mean. The covariance matrix of the Winsorized sample, $\mathbf{S}_{wj} = [\hat{\sigma}_{wjqq'}]$, is

$$\hat{\sigma}_{wjqq'} = \frac{\sum_{i=1}^{n_j} (Z_{ijqq'} - \hat{\mu}_{wjqq'}) (Z_{ijqq'} - \hat{\mu}_{wjqq'})}{(n_j - 1)}, \quad (11)$$

for $q, q' = 1, \dots, L$.

To control the FWR for multiple tests, either a STP or a stepwise procedure may be adopted. For univariate RM designs under assumption violations, Lix and Keselman (1995) showed that the latter are more powerful, and recommended the use of either a step-up or step-down procedure based on the Bonferroni inequality, such as Hochberg's (1988) test. Under Hochberg's procedure, one begins by rank ordering the p -values corresponding to the statistics used for testing the hypotheses $H_{(1)}, \dots, H_{(B)}$, so that $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(B)}$ represent the ordered p -values. The decision rule is to reject

$H_{(m')}$ ($m' \leq m; m = B, \dots, 1$) if $p_{(m)} \leq \alpha/(B - m + 1)$. Testing begins with the hypothesis corresponding to the largest p -value, $p_{(B)}$. If $p_{(B)} \leq \alpha$, all B hypotheses are rejected; if not, $H_{(B)}$ is retained and testing moves to $H_{(B-1)}$. If $p_{(B-1)} \leq \alpha/2$, $H_{(B-1)}$ is rejected, as are all remaining hypotheses; if not $H_{(B-1)}$ is also retained, and $p_{(B-2)}$ is compared to $\alpha/3$, and so on. This continues, if all previous hypotheses have been retained, until $p_{(1)}$ is compared to α/B .

Methodology

A Monte Carlo study was used to evaluate the Type I error and power of the DMM, WJ and BF procedures for multivariate interaction contrasts. These three tests were investigated for a multivariate repeated measures design containing a single between-subjects factor with $J = 3$ levels and a single within-subjects factor with $K = 4$ levels.

The following variables were manipulated in the study. These were: (a) number of dependent variables, (b) total sample size, (c) equality/inequality of the group sizes, (d) the coefficient of variation of the group sizes for unbalanced designs, (e) degree of equality/inequality of the group covariance matrices, (f) nature of the pairing of group sizes and group covariance matrices, (g) multivariate normality/nonnormality, and (h) the non-null hypothesis for power comparisons. The degree of correlation between the dependent variables was set at $\rho = .80$. Keselman and Lix (1997) included both small and large p and ρ in their study; the former increased the total sample size required to obtain a robust solution for the WJ procedure, while the latter variable had little influence on the Type I error performance of the WJ procedure, which is consistent with previous research (Keselman & Lix, 1997). The pooled covariance of the repeated measurements had a non-spherical structure, with a value for ϵ , the index of non-sphericity, of $\epsilon = .57$. The pooled covariance matrix had an average variance of 1.0 and average covariance of 0.5.

The procedures were investigated for $p = 2$ and 4 dependent variables for total sample sizes ranging from 60 to 120. The WJ test is likely to perform less optimally for small to

moderate sample sizes, particularly for non-normal distributions (Keselman et al., 2000). Both balanced and unbalanced designs were included in the study. For unbalanced designs, the sample size conditions were selected based on previous research (Keselman & Lix, 1997; Vallejo et al., 2001; Lix, Algina, & Keselman, 2003). Table 1 contains the values of the total sample sizes that were examined, along with the values of the coefficient of variation of the group sizes, Δn_j , where

$$\Delta n_j = \frac{\sqrt{\sum_{j=1}^J (n_j - \bar{n})^2 / J}}{\bar{n}}. \quad (12)$$

Table 1. Group Sizes (n_j s) and Coefficient of Variation of Group Sizes (Δn_j) for Balanced and Unbalanced Designs.

N	n_j	Δn_j
60	20, 20, 20	0
	18, 20, 22	.08
90	30, 30, 30	0
	24, 30, 36	.16
	18, 30, 42	.33
120	40, 40, 40	0
	30, 40, 50	.20
	24, 40, 56	.33
	18, 40, 62	.45

This coefficient ranged in value from .08 to .45 when group sizes were unequal.

The procedures were investigated when the group covariance matrices were equal and unequal. For the latter case, the elements of the group covariance matrices were in a 1:3:5 ratio. These conditions are consistent with those selected by Keselman and Lix (1997) and Vallejo et al. (2001).

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

matrix with the smallest element values.

Type I error and power rates were obtained when the data were both multivariate normal and non-normal in form. With respect to the former condition, pseudorandom observation vectors $\mathbf{Y}_{ij}$ from a multivariate normal distribution with mean vector $\boldsymbol{\beta}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ were obtained using the SAS generator RANNOR (SAS Institute, 1999b). To obtain each $\mathbf{Y}_{ij}$, a row vector of L deviates in which each element has a standard normal distribution (i.e., $\mathbf{Z}_{ij}$), was transformed to a vector of multivariate observations via a triangular (Cholesky) decomposition, $\mathbf{Y}_{ij} = \boldsymbol{\beta}_j + \mathbf{L}\mathbf{Z}_{ij}^T$, where $\mathbf{L}$ is a upper triangular matrix satisfying the equality $\mathbf{L}^T\mathbf{L} = \boldsymbol{\Sigma}_j$. In this study, $\boldsymbol{\Sigma}_j$ was of the form $\boldsymbol{\Sigma}_j = (\boldsymbol{\Omega}_j \otimes \boldsymbol{\rho}_p)$ where $\boldsymbol{\rho}_p$ represent the p -dimension correlation matrix for the dependent variables and $\boldsymbol{\Omega}_j$ represents the K -dimension covariance matrix associated with a particular dependent variable for the j th group.

Two non-normal distributions were investigated: skewed and long-tailed. The skewed distribution had the same skewness (γ_1) and kurtosis (γ_2) values as a lognormal distribution, in which $\gamma_1 = 6.18$ and $\gamma_2 = 110.93$. The long-tailed distribution had skewness and kurtosis values equivalent to those of a double-exponential distribution, with $\gamma_1 = 0$ and $\gamma_2 = 3$.

These distributions and their associated measures of skewness and kurtosis are representative of those encountered in educational and psychological research (Micceri, 1989). The data were generated by the method developed by Fleishman (1978) and extended to the multivariate situation by Vale and Maurelli (1983).

For each distribution, a vector of constants, $\mathbf{w} = [a \ b \ c \ d]^T$ was obtained using Fleishman's method, to provide the desired degree of multivariate skewness and kurtosis. An intermediate covariance matrix (i.e., $\boldsymbol{\lambda}_j$) was computed so that $\mathbf{Y}_{ij}$ would have the desired $\boldsymbol{\Sigma}_j$. Elements of this intermediate matrix were computed using Vale and Maurelli's (1983) Equation 11 (p. 467), which involves finding the roots of a third-degree polynomial; these roots were computed using the SAS/IML POLYROOT function (SAS Institute, 1999a). The vector of univariate standard normal

deviates was transformed to a vector of multivariate normal deviates via the Cholesky decomposition, $\mathbf{Z}(\lambda)_{ij} = \beta_j + \mathbf{L}_\lambda \mathbf{Z}_{ij}^T$, where $\mathbf{Z}(\lambda)_{ij}$ is the vector of transformed variates, and $\mathbf{L}_\lambda$ is an upper triangular matrix of dimension L satisfying the equality $\mathbf{L}_\lambda^T \mathbf{L}_\lambda = \lambda_j$.

Next, each element of $\mathbf{Y}_{ij}$ was obtained by computing the zero through third powers of the corresponding elements of $\mathbf{Z}(\lambda)_{ij}$, so that $\mathbf{Z}(\lambda)_{ijkl} = [1 \ \mathbf{Z}(\lambda)_{ijkl} \ \mathbf{Z}(\lambda)_{ijkl}^2 \ \mathbf{Z}(\lambda)_{ijkl}^3]$ which represents the vector of powers of the k th components of $\mathbf{Z}(\lambda)_{ij}$. From this, $\mathbf{Y}_{ijkl} = \mathbf{Z}(\lambda)_{ijkl} \mathbf{w}$.

Three definitions of power were considered when non-null hypotheses were investigated. These were any-contrast power, that is, the power to detect at least one non-null hypothesis, all-contrast power, the power to detect all non-null contrasts, and average-per-contrast power, the average probability of detecting at least one non-null contrast. We examined the procedures when the effect size (f^2 ; Cohen, 1988) for the omnibus test of the within-subjects interaction was small and large for two patterns of non-null means.

For pattern 1, the first dependent variable had non-null means, while the second dependent variable had null means. For pattern 2, both dependent variables had the same non-null means. For patterns 1 and 2 respectively, the small effect size was equal to .16 and .08, respectively. The large effect size was 1.35 and .80 for patterns 1 and 2, respectively. The large effect size was selected to enable comparisons of all-contrast power across the investigated procedures; all-contrast power was zero for the small effect size.

The simulation program was written in the SAS/IML programming language (SAS Institute, 1999a). For the investigation of the FWR, the following factors were completely crossed: number of dependent variables (2), total sample size (3: small, moderate, large), relationship between group sizes and covariance matrices (4: equal group sizes/equal covariance matrices, equal group sizes/unequal covariance matrices, positive pairing of group sizes and

covariance matrices, negative pairing of group sizes and covariance matrices), and population distribution (3: normal, double exponential, lognormal). The degree of sample size inequality was nested within total sample size.

For the investigation of power, the following factors were completely crossed for $p = 2$: total sample size, relationship between group sizes and covariances, population distribution, effect size (2: small, large), pattern of non-null means (2: non-null means on one dependent variable, non-null means on both dependent variables). For $p = 2$, five thousand replications of each condition were performed using a .05 significance level. For $p = 4$, because of the size of the matrices and the computations required, only three thousand replications were conducted. For each replication, the conventional DMM, WJ and BF tests were computed using least-squares and robust estimators.

Results

Type I Error

Table 2 contains the empirical percentages of FWR for the conventional (i.e., DMM), BF and WJ procedures for both least-squares and robust estimators for $p = 2$. Bold values are not contained within the bounds for Bradley's (1978) liberal criterion of robustness, which, for the five percent level of significance that was adopted, is 2.5 to 7.5 percentage points.

The data reveal that when the data were multivariate normal and least-squares estimators were adopted, the conventional test for multivariate contrasts could control the FWR when sample size was small or moderate for all conditions with the exception of negative pairings of group sizes and covariance matrices, and the positive pairing when $\Delta n_j = .33$. When sample size was large, the FWR was outside the bounds of Bradley's (1978) criterion for almost all of the positive and negative pairing conditions. When the data were normal, both the BF and WJ ADF procedures based on least-squares estimators controlled the rate of Type I errors across all of the investigated conditions.

Table 2. Empirical Percentages of Familywise Type I Error for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$.

N	Test	Pairing	Δn_j	Normal		Double Exponential		Lognormal	
				LS	RE	LS	RE	LS	RE
60	DMM	$= n_j; = \Sigma_j$	0	3.99	2.34	3.80	2.22	1.66	2.40
	BF			2.55	0.89	2.29	0.78	0.44	0.78
	WJ			3.85	1.90	3.35	1.71	0.85	1.73
	DMM	$= n_j; \neq \Sigma_j$	0	7.11	4.72	6.72	4.25	2.76	4.56
	BF			3.24	1.25	2.85	1.10	0.66	0.93
	WJ			3.97	1.93	3.58	1.85	1.03	1.66
	DMM	+ pair	0.08	5.30	3.73	5.12	3.38	2.01	3.62
	BF			3.50	1.45	3.30	1.22	0.65	1.25
	WJ			3.78	2.10	3.44	1.92	0.94	1.78
	DMM	- pair	0.08	9.13	5.87	7.88	5.48	3.62	5.78
	BF			2.92	1.08	2.43	0.92	0.53	0.94
	WJ			3.92	1.97	3.17	1.81	1.02	1.75
90	DMM	$= n_j; = \Sigma_j$	0	4.09	2.75	3.71	2.50	1.93	2.69
	BF			3.02	1.46	2.72	1.31	0.75	1.34
	WJ			3.88	2.47	3.50	2.27	1.01	2.21
	DMM	$= n_j; \neq \Sigma_j$	0	6.95	5.20	6.44	4.93	3.07	5.59
	BF			3.75	1.88	3.40	1.81	0.78	1.85
	WJ			3.74	2.37	3.64	2.22	1.00	2.37
	DMM	+ pair	0.16	3.82	3.18	3.55	2.84	1.59	3.23
	BF			4.07	2.35	3.45	2.03	1.08	1.90
	WJ			3.76	2.63	3.27	2.26	0.91	2.28
	DMM	- pair	0.33	2.30	1.79	2.37	1.71	1.04	1.79
	BF			4.36	2.65	4.38	2.48	1.57	2.36
	WJ			3.70	2.39	3.58	2.36	1.12	2.37
	DMM	$= n_j; = \Sigma_j$	0.16	12.00	8.69	11.17	8.36	5.86	7.90
	BF			3.54	1.74	3.15	1.47	0.70	1.39
	WJ			3.80	2.52	3.53	2.19	1.06	2.14
	DMM	$= n_j; \neq \Sigma_j$	0.33	18.84	14.73	18.26	14.72	10.85	14.35
	BF			2.90	1.38	2.62	1.15	0.60	1.01
	WJ			3.80	2.44	3.59	2.04	1.16	1.88
120	DMM	$= n_j; = \Sigma_j$	0	4.10	3.11	3.85	2.88	2.15	3.15
	BF			3.21	1.89	3.10	1.76	1.00	1.74
	WJ			4.01	2.78	3.66	2.48	1.34	2.68
	DMM	$= n_j; \neq \Sigma_j$	0	7.06	5.65	6.34	5.12	3.30	5.58
	BF			4.12	2.44	3.47	2.10	1.19	2.21
	WJ			4.02	2.59	3.48	2.30	1.30	2.52
	DMM	+ pair	0.20	3.27	2.55	3.38	2.58	1.72	2.79
	BF			4.43	2.74	4.33	2.71	1.63	2.51
	WJ			4.09	2.62	3.71	2.60	1.17	2.73
	DMM	- pair	0.33	2.38	1.93	2.24	1.90	1.09	2.21
	BF			4.64	3.29	4.47	2.95	1.84	2.93
	WJ			3.98	2.72	3.63	2.64	1.27	2.73
	DMM	$= n_j; = \Sigma_j$	0.45	1.50	1.30	1.51	1.26	0.99	1.33
	BF			4.78	3.22	4.62	3.05	2.61	3.26
	WJ			3.74	2.62	3.43	2.48	1.35	2.64
	DMM	$= n_j; \neq \Sigma_j$	0.20	13.43	10.85	13.12	10.85	7.37	11.00
	BF			3.45	1.95	3.25	1.71	0.88	1.47
	WJ			3.90	2.28	3.36	2.15	1.05	2.07
	DMM	+ pair	0.33	19.41	15.10	18.49	15.67	11.59	14.54
	BF			3.54	1.70	2.90	1.45	0.62	1.29
	WJ			3.99	2.44	3.35	2.13	1.18	2.23
	DMM	- pair	0.45	26.89	23.04	26.11	23.11	17.83	20.58
	BF			3.07	1.45	2.37	1.15	0.52	1.12
	WJ			3.64	2.29	3.33	2.00	1.26	2.18

Note: + pair = positive pairing of group sizes and covariance matrices; - pair = negative pairing of group sizes and covariance matrices. Bold values are outside the range 2.5 - 7.5. LS = Least Squares estimators; RE = Robust estimators.

When the data were normal and robust estimators were adopted, the DMM test remained liberal for negative pairing conditions when sample size was moderate or large. The DMM, BF, and WJ procedures were frequently conservative, but this degree of conservatism decreased as the total sample size increased.

The results for symmetric and skewed distributions were substantially different. For the symmetric double exponential distribution, the FWR results for least-squares estimators was similar to those obtained for the normal distribution. That is, the DMM test was liberal for all negative pairing conditions and conservative for positive pairings when the degree of group size imbalance was large. The FWR for the ADF tests was well controlled. The same liberal tendencies of the DMM test were observed even when robust estimators were adopted, while the ADF tests were frequently conservative.

When the data were obtained from the skewed lognormal distribution, the error rates for the conventional and ADF procedures based on least-squares estimators were almost always conservative, except for negative pairings of group sizes and covariances when the DMM test could be liberal. When robust estimators were adopted, the FWRs for the DMM test could still be liberal. Those for the ADF tests tended to be less conservative than when least-squares estimators were adopted, and became even less so as total sample size increased.

The results for $p = 4$ (not reported) were similar to those provided in Table 2. However, the FWR for the conventional test were even more inflated than when $p = 2$. For example, when $N = 120$ and $\Delta n_j = .20$, the FWR was 18.14 and 12.32 percent for the double exponential distribution for least-squares and robust estimators, respectively.

Power

Table 3 contains the empirical percentages of any-contrast and average-per-contrast power for conventional and ADF procedures for the first mean pattern when the effect size was small. The data are averaged over all total sample size conditions. For the second mean pattern, any-contrast power attained its upper bound across most of the conditions;

therefore these data are not reported. To interpret these results, we describe mean power differences of less than ten percentage points as small, between ten and 20 percent as moderate, and those of greater than 20 percent as substantial.

When the data followed a multivariate normal distribution, procedures based on least-squares estimators were more powerful than those based on robust estimators. The differences in any-contrast power were moderate to large. For average-per-contrast power they were small to moderate. For positive and negative pairing conditions, the differences in any-contrast power for the BF and WJ procedures were small; for positive pairings the BF test was slightly more powerful than the WJ test.

When the data had a multivariate heavy-tailed distribution, any-contrast power and average-per-contrast power rates for the procedures based on least-squares estimators were larger than those based on robust estimators. However, the differences were generally small. The exception was for any-contrast power for the BF and WJ procedures for negative pairings of group sizes and covariances, where the differences were moderate.

However, when the data had a multivariate skewed distribution, the procedures based on robust estimators were consistent more powerful than those based on least-squares estimators. This held true for both any-contrast and per-contrast power. The power differences were small to moderate. The WJ procedure was more powerful than the BF test with robust estimators across all of the investigated conditions.

Table 4 provides all-contrast and average-per-contrast power when the effect size was large for both the first and second mean patterns. Again, when the data followed a multivariate normal distribution, procedures based on least-squares estimators were always more powerful than those based on robust estimators. The differences in all-contrast power between the procedures based on least-squares estimators and those based on robust estimators were moderate to large for the first mean pattern.

Table 3. Empirical Percentages of Power for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$; Small Effect Size, Mean Pattern 1.

Test	Pairing	Normal				Double Exponential				Lognormal			
		ANCP		PCP		ANCP		PCP		ANCP		PCP	
		LS	RE	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE
DMM	$= n_j; = \Sigma_j$	76.44	53.03	22.10	10.04	76.27	69.86	21.45	16.66	77.17	88.43	19.04	26.67
BF		72.17	44.28	20.15	7.81	71.66	61.12	19.31	13.44	68.60	81.57	14.71	21.39
WJ		74.92	48.72	20.96	8.77	74.95	65.66	20.45	14.82	79.24	85.06	19.55	24.07
DMM	$= n_j; \neq \Sigma_j$	78.49	57.99	22.01	10.93	78.64	73.32	21.53	17.01	79.90	89.14	20.02	26.00
BF		70.29	42.59	17.80	6.96	70.16	58.57	17.22	11.62	68.57	77.42	14.51	17.89
WJ		66.53	37.89	19.42	6.79	66.54	54.54	18.93	12.36	73.59	85.44	18.93	21.34
DMM	+ pair	80.61	62.00	20.03	11.00	80.70	78.25	19.59	17.18	81.56	93.17	17.98	25.24
BF		81.19	61.93	22.00	11.76	81.40	76.55	21.44	17.67	80.66	90.55	17.94	24.62
WJ		75.95	53.25	24.51	10.74	76.19	70.47	23.90	18.27	80.06	87.69	22.29	29.40
DMM	- pair	83.19	67.15	28.72	15.72	82.86	79.49	27.97	23.14	83.27	93.18	25.40	35.05
BF		62.12	31.13	14.10	4.48	60.66	45.59	13.65	7.73	61.42	69.90	12.51	14.37
WJ		63.23	34.63	16.99	5.59	63.13	50.91	16.63	10.26	72.30	75.12	18.02	19.41

Note: ANCP = Any-contrast power; PCP = average-per-contrast power; LS = Least-squares estimators; RE = robust estimators.

The greatest difference was for the WJ procedure for both the positive and negative pairing conditions, where the difference in power was 25.5 and 27.2 percent for the positive and negative pairing conditions, respectively. The smallest difference was for the BF procedure. For the second mean pattern, the differences between least-squares and robust estimators were small to moderate when the data were normally distributed. Again, the greatest differences were for the WJ procedure. For per-contrast power, the differences between least-squares and robust estimators were small to moderate. The largest difference for both mean patterns was for the BF procedure (13.1 percent) when group sizes and covariance matrices were negatively paired.

For normally distributed data with least-squares estimators, the differences among the procedures varied considerably depending on the relationship between the group sizes and

covariances. When the design was balanced and covariances were unequal, the WJ procedure was substantially more powerful than the BF procedure, and moderately more powerful than the DMM. The difference in power between the BF and WJ procedures was substantial for both the positive and negative pairing conditions for both mean patterns. This same pattern was evident when robust estimators were adopted.

For the double exponential distribution, the difference between procedures based on least-squares and robust estimators were small for both all-contrast and per-contrast power. The procedures based robust estimators were more power than those based on least-squares estimators for both types of power. Again, the differences between procedures based on least-squares and robust estimators were largest for the BF procedure.

Table 4. Empirical Percentages of Power for Robust and Least Squares Estimators for Multivariate Interaction Contrasts, $p = 2$; Large Effect Size.

		Normal				Double Exponential				Lognormal			
		ACP		PCP		ACP		PCP		ACP		PCP	
Test	Pairing	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE	LS	RE
Mean Pattern 1													
DMM	$= n_j; = \Sigma_j$	27.18	8.39	83.72	75.02	25.93	19.93	83.48	81.58	17.54	37.66	81.06	26.67
BF		23.23	5.35	81.71	71.53	21.56	14.16	81.32	78.27	10.59	26.94	76.95	21.39
WJ		25.55	6.96	82.98	73.37	24.32	17.22	82.77	80.10	18.28	33.45	80.87	24.07
DMM	$= n_j; \neq \Sigma_j$	27.34	7.43	83.20	73.92	25.81	19.73	82.90	80.96	18.00	38.88	80.77	26.00
BF		20.07	3.14	79.44	68.02	18.35	10.66	79.03	75.30	9.87	20.66	75.56	17.89
WJ		41.99	16.87	87.70	77.13	40.75	33.02	87.52	85.14	33.45	56.71	85.74	21.34
DMM	+ pair	15.49	2.68	79.19	71.21	13.99	10.46	78.78	77.78	8.23	23.46	76.28	25.24
BF		19.06	3.13	79.95	70.93	17.06	11.48	79.48	77.56	7.21	21.43	75.61	24.62
WJ		42.30	16.80	88.52	80.02	41.26	34.10	88.37	86.73	33.53	60.78	86.74	29.40
DMM	- pair	46.56	22.23	89.11	81.53	45.16	40.49	88.90	87.86	35.92	67.46	86.81	35.05
BF		23.12	2.75	80.27	67.21	21.17	9.62	79.81	75.12	14.64	24.70	76.58	14.37
WJ		49.24	22.06	89.22	78.33	47.90	42.27	88.97	87.55	38.12	69.81	86.65	19.41
Mean Pattern 2													
DMM	$= n_j; = \Sigma_j$	13.20	6.81	80.22	75.33	14.07	17.73	80.65	81.86	32.86	66.41	87.23	94.94
BF		9.13	3.71	78.17	72.05	9.73	11.32	78.49	78.66	23.09	54.93	83.72	92.52
WJ		11.95	5.62	79.47	73.83	12.94	15.13	79.97	80.50	34.36	61.33	87.50	93.86
DMM	$= n_j; \neq \Sigma_j$	11.89	5.58	79.37	74.29	12.24	16.60	79.70	81.27	33.98	69.45	87.45	95.40
BF		5.49	1.67	75.42	68.62	5.85	7.14	75.70	75.69	21.19	49.44	82.76	90.99
WJ		26.95	14.33	85.05	77.84	28.09	31.37	85.50	85.91	54.42	80.62	92.28	97.34
DMM	+ pair	3.71	1.67	74.72	71.32	4.08	7.37	75.15	77.50	19.22	59.08	82.89	93.86
BF		3.97	1.42	75.42	70.88	4.33	6.82	75.75	77.16	17.30	55.87	82.28	93.00
WJ		25.74	13.79	85.59	80.07	27.12	31.13	86.04	86.94	54.14	85.22	92.52	98.09
DMM	- pair	31.83	18.80	86.62	81.63	32.80	37.91	86.88	88.06	54.95	87.20	92.32	98.22
BF		6.10	1.35	76.14	67.78	6.90	5.99	76.57	75.39	28.14	57.49	84.40	93.19
WJ		34.33	17.84	87.18	79.21	35.76	40.28	87.57	88.47	60.65	89.10	93.43	98.56

Note: ACP = All-contrast power; PCP = average-per-contrast power; LS = Least-squares estimators; RE = robust estimators.

When the data were obtained from a skewed distribution, the results also favor the procedures based on robust estimators. For all-contrast power, the power differences were moderate to substantial. Moreover, the WJ procedure demonstrated substantially greater power than the BF procedure across most of the investigated conditions. It was also more powerful than the DMM test when the design was balanced but covariances were unequal, and for positive pairings of group sizes and covariances.

Conclusion

The purpose of this article was to examine procedures for conducting multivariate a priori contrasts in RM designs. Conventional tests for multivariate within-subjects effects are sensitive to violations of the assumptions of covariance homogeneity and multivariate normality. Approximate degrees of freedom procedures are an appealing alternative because they are robust to heterogeneous covariance matrices. Furthermore, these tests can be extended to the

case of non-normal data by substituting least-squares estimators with robust estimators which are insensitive to the presence of skewed distributions and/or extreme observations.

Consistent with results for omnibus tests of the interaction (Keselman & Lix, 1997), the data show that error rates of conventional tests of multivariate interaction contrasts can become inflated when the group with the smallest number of observations exhibits the greatest degree of heterogeneity. These tests can also become conservative when there is a positive relationship between group sizes and covariances. The liberal and conservative tendencies do not disappear as sample size increases, and they become exacerbated as the dimension of the data increases.

Approximate degrees of freedom procedures based on least-squares estimators will perform well under violations of covariance homogeneity. These procedures will never be liberal under departures from multivariate normality. They may lose a moderate amount of power compared to procedures based on least-squares estimators. For the moderate degree of kurtosis that characterized the double exponential distribution, the differences in power between the tests based on robust estimators and those based on least-squared estimators were negligible, but did not always favor robust estimators. This power difference depended on the magnitude of the effect, the nature of the non-null means, and the definition of power that was adopted by the researcher. When the data were obtained from skewed distributions, the procedures based on robust estimators demonstrated clear power advantages in terms of detecting all contrasts of interest. Average-per-contrast power and any-contrast contrast power also favored robust estimators.

Previous research suggests that the Welch-James procedure should be selected over the Brown-Forsythe test when covariances and group sizes are negatively paired (Vallejo et al., 2001; Lix, Algina, & Keselman, 2003), this recommendation does not hold for all of the conditions investigated in this simulation study.

The choice of a procedure for testing within-subjects effects in multivariate repeated measures designs is complex, and depends on a number of factors. In this article, we advocate

testing a set of hypotheses that enable the researcher to identify the localized source of multivariate interaction between a grouping factor and a repeated measures factor. If the data are in fact multivariate normal, then there is a modest gain in power to be obtained from adopting least-squares estimators. If the data are non-normal, there are power advantages by adopted a multivariate procedure that is robust to covariance heterogeneity and multivariate non-normality, particularly when the data are skewed. Which robust procedure to adopt is a function of the magnitude of the effect and the pattern of the non-null means. In closing, it should be noted that a SAS/IML (SAS Institute, 1999a) program to implement the Welch-James procedure with robust estimators for a variety of univariate and multivariate designs is available in Keselman, Wilcox, and Lix (2003).

References

- Bird, K. D., & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, 93, 167-178.
- Boik, R. J. (1988). The mixed model for multivariate repeated measures: Validity conditions and an approximate test. *Psychometrika*, 53, 469-486.
- Boik, R. J. (1991). Scheffe's mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communication in Statistics: Theory and Methods*, 20, 1233-1255.
- Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics*, 18, 1-40.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Brown, M. B., & Forsyth, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129-132.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

- Elliott, R. S., & Barcikowski, R. S. (1994). Simultaneous test procedures in exploratory multivariate analysis of variance. American Educational Research Association Conference (1994, New Orleans, Louisiana).
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- 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.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-802.
- Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Huberty, C. J. (1994). *Applied discriminant analysis*. New York: John Wiley.
- Huberty, C. J., Chou, T. F., & Benitez, E. B. (1994). The study of multivariate group contrasts. *Advances in Social Science Methodology*, 3, 123-138.
- Keselman, H. J., & Lix, L. M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, 50, 319-338.
- Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, 3, 123-141.
- Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, 53, 175-191.
- Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, 40, 586-596.
- Krishnaiah, P. R., & Reising, J. M. (1985). Multivariate multiple comparisons. In D. L. Banks, C. B. Read, & S. Kotz (Eds.), *Encyclopedia of Statistical Sciences*, 6, 88-95. New York: Wiley & Sons.
- Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research*, 38, 403-431.
- Lix, L. M., & Keselman, H. J. (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., & Hinds, A. (in press). A comparison of procedures for the multivariate Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine*.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 154-166.
- SAS Institute Inc. (1999a). *SAS/IML software, Usage and reference, version 8*. Cary, NC: Author.
- SAS Institute Inc. (1999b). *SAS language reference: Dictionary, version 8*. Cary, NC: Author.
- Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, 58, 861-881.
- Tamhane, A. C., & Dunnett, C. W. (1999). Stepwise multiple test procedures with biometric applications. *Journal of Statistical Planning and Inference*, 82, 55-68.
- Thomas, D. R. (1983). Univariate repeated measures techniques applied to multivariate data. *Psychometrika*, 48, 451-464.
- Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48, 451-464.
- Vallejo, G., Fidalgo, A., & Fernandez, P. (2001). Effects of covariance heterogeneity on three procedures for analyzing multivariate repeated measures designs. *Multivariate Behavioral Research*, 36, 1-27.
- Wilcox, R. R. (1995a). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, 65, 51-77.
- Wilcox, R. R. (1995b). Simulation results on solutions to the multivariate Behrens-Fisher problem via trimmed means. *The Statistician*, 44, 213-225.

On A Simple Method For Analyzing Multivariate Survival Data Using Sample Survey Methods

Pingfu Fu J. Sunil Rao
Department of Epidemiology and Biostatistics
Case Western Reserve University

A simple technique is illustrated for analyzing multivariate survival data. The data situation arises when an individual records multiple survival events, or when individuals recording single survival events are grouped into clusters. Past work has focused on developing new methods to handle such data. Here, we use a connection between Poisson regression and survival modeling and a cluster sampling approach to adjust the variance estimates. The approach requires parametric assumption for the marginal hazard function, but avoids specification of a joint multivariate survival distribution. A simulation study demonstrates the proposed approach is a competing method of recent developed marginal approaches in the literature.

Key words: sampling; design effect; survival analysis; clustered data

Introduction

Clustered survival events can occur in a number of ways. The form receiving considerable attention has been the scenario of when an individual is subject to experiencing repeat events (recurrent or multiple-type) over time. An illustration of this is the case where a child is diagnosed with chronic lung disease (CLD) for a period of time. The disease may or may not resolve. If resolution occurs, the child is susceptible to repeat occurrences of CLD over time (Norton, et. al., 2001). The time to the start of each CLD episode can be thought of a series of clustered events where the clustering unit is the child.

There have been a number of different methods proposed to handle inference in this situation. These include Andersen and Gill (AG) model (1982), Prentice, Williams and Peterson (PWP) model (1981), and Wei, Lin and Weissfeld (WLW) model (1989). In AG model, each subject is treated as a multi-event counting process with essentially independent increments; PWP model is a conditional approach; and WLW model is marginal method, in which one obtains the estimated coefficients, ignoring correlation, followed by fix of the variance of estimated coefficients.

More recently, Segal and Neuhaus (1993) showed how to use Poisson regression techniques to analyze such data. Their method made use of generalized estimating equation (GEE) machinery (Liang & Zeger, 1986) for doing point estimation. Robust inference was handled by using sandwich estimators for variance estimates of estimated regression parameters. In all of these applications, much of which has recently become widely available (Therneau & Grambsch, 2000) and can be fitted by major statistical software, such as SAS (SAS Institute, Cary, NC) and Splus (Insightful Corp., Seattle, Washington).

Survey sampling is another area where clustered events are quite common. The design effects approach, which is based on sample

Pingfu Fu is an Assistant Professor of Biostatistics. E-mail: pxf16@case.edu. J. Sunil Rao is Associate Professor of Biostatistics, Department of Epidemiology & Biostatistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106. E-mail: sunil@hal.cwru.edu.

survey techniques, has been used for analyzing such data. Design effect represents the estimated inflation in the variance of estimated coefficients due to correlated observations in each cluster (Rao and Scott, 1999). In order to account for the correlation among observations within each cluster, we can either transform the data by a design effect and apply standard methods afterwards assuming independence, or apply standard methods assuming independence, and then adjust the variances of the estimates by design effects. Work in non-survival setting includes that by Rao and Scott (1992, 1999) and Bieler and Williams (1995). In this paper, we are going to use the design effect approach under the survival analysis-Poisson regression and show how the design effects method can very simply handle clustered survival events, too.

Our method is similar to Segal and Neuhaus's approach in terms of the variance estimate - both use a sandwich estimator, but differ with respect with in the "filling". It's well-known that the Liang-Zeger's GEE application of quasi-likelihood on which Segal and Neuhaus's is based is essentially a special case of Binder's method (Binder, 1983) applied to with-replacement cluster sampling. Paik (1988) has shown that the GEE methods can lead to considerably biased parameter estimates in small sample settings. This is part of the motivation for the alternative approach we propose. Our method is parametric, and marginal, thus, it sacrifices the semi-parametric specification of AG, PWP and WLW. However, it provides another platform using only regular Poisson regression to analyze multivariate survival data.

Multivariate survival data and GEE

Assume that we have a sample of failure time data represented by

$$\left\{ \begin{array}{l} (T_{ijk}, \delta_{ijk}, x_{ijk}) : i \\ \quad \quad \quad = 1, \dots, G, j \\ \quad \quad \quad = 1, \dots, m_i, k = 1, \dots, n_{ij} \end{array} \right\}$$

where for observation k of individual j of treatment group i , T_{ijk} denotes a failure time, δ_{ijk}

is an event indicator taking the value 1 if T_{ijk} is uncensored and 0 otherwise, and x_{ijk} is a p -dimensional vector of covariates. There are assumed to be m_i individuals within treatment group i and G treatment groups in total. Let $S(t)$, $f(t)$ and $\lambda(t)$ be the survival distribution, density function and hazard function respectively for random variable T where $t \geq 0$ is a generic survival time.

Following Segal and Neuhaus (1993), we assume that the marginal hazard function for the k th observation of the j th individual in the i th treatment group involves covariates x_{ijk} through Cox's proportional hazards model

$$\lambda_{ijk}(t) = \lambda_0(t) \exp(\beta' x_{ijk}),$$

where β is a p -dimensional vector of regression parameters, and $\lambda_0(t)$ is the baseline hazard function. Thus

$$f_{ijk}(t) = \lambda_0(t) \exp\{\beta' x_{ijk} - \Lambda_0(t) e^{\beta' x_{ijk}}\}$$

and

$$S_{ijk}(t) = \exp\{-\Lambda_0(t) e^{\beta' x_{ijk}}\},$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function. As in Segal and Neuhaus, we depart from the standard Cox proportional hazards model which does not assign a parametric form for $\lambda_0(t)$.

Under the standard assumption of independent censoring, the likelihood for the k th observation of j th individual in the i th treatment group is

$$\begin{aligned} L_{ijk}(\alpha, \beta) &= f_{ijk}(t)^{\delta_{ijk}} (S_{ijk}(t))^{1-\delta_{ijk}} \\ &= [\lambda_0(t) \exp(\beta' x_{ijk})]^{\delta_{ijk}} \exp(-\Lambda_0(t) e^{\beta' x_{ijk}}) \\ &= (\mu_{ijk}^{\delta_{ijk}} e^{-\mu_{ijk}}) (\lambda_0(t) / \Lambda_0(t))^{\delta_{ijk}} \end{aligned} \quad (1)$$

where $\mu_{ijk}(t) = \Lambda_0(t) \exp(\beta' x_{ijk})$ and α are the parameters specifying the baseline survival distribution.

Because the δ_{ijk} takes on values of only 0 or 1, the first term in (1) can be thought of as a Poisson random variable with mean μ_{ijk} . A log-linear model for the hazard function implies a log-linear model for μ_{ijk} through

$$\log(\mu_{ijk}) = \log(\Lambda_0(t)) + \beta' x_{ijk}.$$

As mentioned earlier, we will give parametric form to $\lambda_0(t)$ or $\Lambda_0(t)$, say for example, by letting $\Lambda_0(t) = t$. Then $f(t)$ is simply an exponential density with mean $\exp(-\beta'x)$, and maximum likelihood estimates for the regression parameters β can be found by fitting a Poisson regression model where response is the censoring variable with an offset $\log t_{ijk}$.

By assuming the independence of responses within each cluster, Segal and Neuhaus (1993) handle the clustering by fitting a corresponding GEE model (Liang and Zeger, 1986) and use robust sandwich estimators for inference on the regression parameters. Obviously, using GEE machinery, we can also assume different variance and covariance structure built in to the procedure. The difficulty is the justification of the structure chosen. They also illustrate how to fit Weibull regression models and piecewise exponential models by changing the offset or augmenting with a time-dependent covariate respectively.

Adjusting inference by design effects

In randomized clinical trials, the usual primary research question is what is the treatment difference among all the treatments? Let's assume that correlated observations form a cluster which can be a patient, a family or a community, etc., and assume the observations between clusters are independent. The idea behind design effect approach we are using is first to derive Taylor linearization for implicitly defined parameter vectors, which was developed by Binder (1983) in generalized linear models, and then apply a between-cluster variance estimator for the linearized statistic, as described by Bieler and Williams (1995). The details of the design effect approach for our case are the

following. Let m_i be the number of clusters randomized to the treatment i , $i = 1, 2, \dots, G$; n_{ij} be the number of observations for cluster j in i th treatment, $j = 1, 2, \dots, m_i$; δ_{ijk} be the censoring indicator from the k th observation of j th cluster from i th treatment, $k = 1, 2, \dots, n_{ij}$;

$$\delta_{ij} = \sum_{k=1}^{n_{ij}} \delta_{ijk};$$

$x_{ijk} = (x_{1,ijk}, x_{2,ijk}, \dots, x_{p,ijk})'$ be the vector of covariates (i.e. treatment, sex, race, etc.) for the ijk th observation; $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ and $E(\delta_{ijk}) = \mu_{ijk}$. From earlier developments (section 2), treat δ_{ijk} as if it were Poisson, a log link function is used in the generalized linear model, i.e. $\log \mu_{ijk} = x'_{ijk} \beta$ and

$$p(\delta_{ijk} | x_{ijk}, \beta) = e^{-\mu_{ijk}} (\mu_{ijk})^{\delta_{ijk}} / \delta_{ijk}!$$

Thus,

$$\begin{aligned} \log(p(\delta_{ijk} | x_{ijk}, \beta)) &\propto -\mu_{ijk} + \delta_{ijk} \log(\mu_{ijk}) \\ &= -\Lambda_0(t_{ijk}) \exp(x'_{ijk} \beta) \\ &\quad + \delta_{ijk} (\log(\Lambda_0(t_{ijk})) + x'_{ijk} \beta). \end{aligned}$$

The log-likelihood equations are then

$$\begin{aligned} l(\beta) &= -\sum_i \sum_j \sum_k \Lambda_0(t_{ijk}) \exp(x'_{ijk} \beta) \\ &\quad + \sum_i \sum_j \sum_k \delta_{ijk} (\log(\Lambda_0(t_{ijk})) + x'_{ijk} \beta). \end{aligned}$$

The set of estimated Poisson regression coefficients, $\hat{\beta}$, that maximize $l(\beta)$ are found by solving the following score equations:

$$\begin{aligned} U(\beta) \equiv \frac{\partial l(\beta)}{\partial \beta} &= -\sum_i \sum_j \sum_k \Lambda_0(t_{ijk}) \exp(x'_{ijk} \beta) x_{ijk} \\ &\quad + \sum_i \sum_j \sum_k \delta_{ijk} x_{ijk} = 0. \end{aligned}$$

This can be done using the Newton-Raphson method. Then

$$\hat{\mu}_{ijk} = \Lambda_0(t_{ijk}) \exp(x'_{ijk} \hat{\beta}).$$

Since μ_{ijk} may contain other nuisance parameters, we have to estimate them from the likelihood function. For example, if we assume Weibull baseline hazard, $\lambda_0(t) = \nu t^{\nu-1}$, the shape parameter ν can be estimated by

$$\hat{\nu} = \frac{\sum \delta_{ijk}}{\sum \log(t_{ijk})(\hat{\mu}_{ijk} - \delta_{ijk})},$$

and an iterative procedure can be used to find the estimates of β and other nuisance parameters.

The associated sample information matrix for estimating β is

$$\begin{aligned} I &= -\frac{\partial^2 l(\hat{\beta})}{\partial \hat{\beta} \partial \hat{\beta}'} \\ &= -\sum_i \sum_j \sum_k \Lambda_0(t_{ijk}) \exp(x'_{ijk} \hat{\beta}) x_{ijk} x'_{ijk} + 0 \\ &= -\sum_i \sum_j \sum_k \hat{\mu}_{ijk} x_{ijk} x'_{ijk}. \end{aligned} \quad (2)$$

Under cluster sampling, the inverse of the information matrix is no longer a valid estimate of the variance $\hat{\beta}$ (Binder, 1983). To address this problem, Binder (1983) gave a general method for deriving the variance of parameter estimators under clustering in survey sampling, which satisfy estimating equations of the form:

$$U(\hat{\beta}) = \sum_l u_l(\hat{\beta}),$$

where the sum is over the observations. Thus, using Taylor series linearization:

$$\begin{aligned} U(\hat{\beta}) &= U(E(\hat{\beta})) \\ &+ \frac{\partial U(E(\hat{\beta}))}{\partial \hat{\beta}} (\hat{\beta} - E(\hat{\beta})) + o(\hat{\beta} - E(\hat{\beta})). \end{aligned}$$

By the delta method, the variance of $\hat{\beta}$ is then estimated by

$$\hat{V}(\hat{\beta}) = (I^{-1}) V_U (I^{-1})', \quad (3)$$

where

$$V_U = \hat{V}[U(\hat{\beta})].$$

Binder (1983) gave conditions under which (3) consistently estimates the asymptotic variance of $\hat{\beta}$. In order to obtain a cluster covariance matrix of $U(\hat{\beta})$, we first linearize $U(\hat{\beta})$, and then apply a between-cluster variance estimator for the linearized statistic. To this end, let

$$Z_{ijk} = x'_{ijk} \hat{r}_{ijk}$$

where $\hat{r}_{ijk} = \delta_{ijk} - \mu_{ijk}$ is the residual for the k th observation of the j th cluster from the i th treatment group. Accumulations of these linearized vectors are first formed at the cluster level, namely,

$$Z_{ij} = \sum_k Z_{ijk}. \quad (4)$$

The associated between-cluster within treatment group mean square matrix is

$$S_z = \sum_i m_i S_{zi},$$

where m_i denotes the number of clusters in treatment group i and

$$S_{zi} = \sum_j (Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z}_i)' / (m_i - 1),$$

depicts the $p \times p$ matrix of sample mean squares and cross products from treatment group i , with

$$\bar{Z}_i = \sum_j \frac{Z_{ij}}{m_i}.$$

Following (3), the estimated variance for β is given by

$$\hat{V}(\beta) = I^{-1} S_z (I^{-1})'.$$

The above estimate of the variance of $\hat{\beta}$ is called a “modified sandwich estimate” and converges to the true variance of $\hat{\beta}$ when the number of the clusters tends to infinity (Binder, 1983). If the total number of the clusters is small, then this estimate will be sharply biased towards zero, and some other estimate must be considered. Generally speaking, when the clusters are independent, the sum of the linearized vectors for each cluster, Z_{ij} in (4) can be unbiasedly estimated because $\hat{\beta}$ is usually a consistent estimate of β under usual regularity conditions without taking the correlation structure into account. Unlike the quasi-likelihood GEE approach of Liang and Zeger (1986), explicit specification of a correlation structure in the cluster is unnecessary, which is also mentioned in Bieler and Williams (1995).

Methodology

Generally speaking, there are two approaches for analyzing multivariate survival data. One is conditional model, and other is a marginal model. Conditional models induce dependence by including frailties (random effects) while marginal approach directly models fixed effects. We will employ a marginal-based approach when conducting simulations in order to evaluate the performance of the proposed design effect based approach. We specify a marginal survival distribution, and estimate the parameters characterizing the distribution. This approach however does not define the joint distributions for generating multivariate survival data, and thus the effect of dependence in repeating events over time cannot be studied. Hence we use a random effects approach as in Segal and Neuhaus (1993) where the joint distributions are forced to have proportional margins and a patterned covariance matrix.

We use positive stable mixing distributions (Hougaard, 1986) along with the random effects approach. Let T_{ijk} be the survival times of observation k of individual j with treatment group I conditional on an observed

covariate ζ_j . In this setup we assume that T_{ijk} 's in different clusters are independent.

Now assume ζ to be positive stable with index α . The Laplace transform for ζ is $E(\exp(-s\zeta)) = \exp(-s^\alpha)$. If we now define another random variable Y_{ijk} to be Weibull distributed with scale parameter $\exp(\beta'x_{ijk})$ and shape parameter α , then $T_{ijk} = Y_{ijk} \zeta_j^{-1/\alpha}$.

Thus, the T_{ijk} 's within a cluster are multivariate Weibull with Weibull margins having scale $\exp(\alpha\beta'x_{ij})$ and shape $\alpha\alpha$. The correlation between $\log(T_{ijk})$ and $\log(T_{ijl})$ is then just $1 - \alpha^2$ for $k \neq l$. The generation of positive stable variates ζ_j can be done using Splus which employs Chambers et al.'s (1976) algorithm.

In order to examine the performance of the newly proposed method for estimating regression parameters, we studied a number of scenarios. We first looked at varying the cluster size from $k = 5, 10$ and also the number of clusters $C = 20, 50$. The survival data was generated using the procedure just described with shape parameter $\alpha = 2$ and one covariate $\beta = 3$ for simplicity which are chosen arbitrarily.

The index of the positive stable distribution α was varied from 0.3 to 0.7 indicating decreasing levels of correlation between log survival times within a cluster. Survival times were censored at fixed times instead of random censoring to 10% and 20% censoring percentage. For each combination of experimental conditions, we conducted 200 simulations, and report biases of the regression parameter estimates from Poisson regression and GEE as well as mean variance of three types, i.e. naive, robust and new approach.

We fit Segal and Neuhaus's GEE-based method with independence correlation structures and compared the performance to the new method. The comparison will be made in terms of bias and variance. Since there is no explicit

formula available for the variance of $\hat{\beta}$ in this complex situation, so we don't know the true variance of $\hat{\beta}$. We use following approach to check the underestimation or overestimation of the estimate from each method in this finite sample situation. Let B be the number of simulations (in our case, we set $B = 200$, β_i , $i = 1, \dots, p$ be the true value of the coefficients, $\hat{\beta}_{ij}$ be the estimates of $\hat{\beta}_i$ in iteration j , where $i = 1, \dots, B$, and $\hat{\sigma}_{i,j}^2$ be the variance estimate of $\hat{\beta}_i$ in j th simulation after correction which accounts for the correlation of survival times within each cluster, then one way to check the biases of the variance estimate is the following efficiency quantity:

$$r_i = \frac{\hat{\sigma}_i^2}{(m(\tilde{\beta}_i))^2 + \text{var}(\tilde{\beta}_i)},$$

where

$$\hat{\sigma}_i^2 = \sum_j \hat{\sigma}_{ij}^2 / B,$$

$$\tilde{\beta}_i = (\hat{\beta}_{i,1} - \beta_i, \dots, \hat{\beta}_{i,B} - \beta_i)'$$

and $m(\tilde{\beta}_i)$ is the sample mean of $\tilde{\beta}_i$ and $\text{var}(\tilde{\beta}_i)$ is the sample variance of $\tilde{\beta}_i$. If $r_i > 1$, then the variance is empirically overestimated, if $r_i < 1$, then the variance is empirically underestimated.

The simulation was conducted in *S-plus*. Our approach can be implemented with minor programming, a call of glm function and several other lines of coding for matrix manipulation (the program is available upon request).

Tables 1-4 give the results of the simulation. Notice first, as number of clusters increases, the smaller the bias in estimating the scale parameter a , and the regression coefficient β for Segal and Neuhaus's approach and our approach.

This is because the estimates are consistent when the number of clusters gets large; and there is no systematic difference of the biases when the cluster size, percentage of

censoring, and value of index parameter α change.

Secondly, the variance estimate of $\hat{\beta}$ by the new method, the robust variance as well as naive variance estimates decrease when the number of cluster increases. Varying the cluster size does not change the variance, and there is no obvious evidence that a different percentage of censoring gives substantially different results. But increasing value of the index α , which changes the correlation of survival times in each cluster, does decrease the variance estimate in all three different types of estimates. This is because increasing α decreases correlation among the survival times within each cluster.

The naive variance estimates overestimate or underestimate the variance badly; the robust variance estimate and the new method usually underestimate the variance except in one case by our method with $r = 1.008$ ($C = 20$, $cen = 20\%$ and $\alpha = 0.4$). Overall, our method gives r values closer to 1 than the GEE approach, because correlation structure is not needed explicitly in calculating the variance of $\hat{\beta}$ as it is in GEE approach. The larger the number of clusters is the closer the r values are to 1.

A real data example (CGD)

The well-known Chronic Granulomatous Disease (CGD) dataset, which is described in the Appendix D of Fleming and Harrington (1991), has been analyzed by many authors. CGD is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections which usually present early in life and may lead to death in childhood. Phagocytes from CGD patients ingest microorganisms normally but fail to kill them, primarily due to the inability to generate a respiratory burst dependent on the production of superoxide and other toxic oxygen metabolites.

Thus, it is the failure to generate microbicidal oxygen metabolites within the phagocytes of CGD patients. There is evidence that gamma interferon is an important macrophage activating factor which could restore superoxide anion production and bacterial killing by phagocytes in CGD patients.

Table 1: Results for simulated multivariate Weibull distribution with number of clusters = 20 and 10% censoring. Mean bias and variance of regression parameter estimates over 200 simulations. In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and $cens$ is percentage of censoring.

$M_i = 20, cens = 10\%$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Bias of a , scale parameter					
$k = 5$	0.1052	0.1605	0.1236	0.1449	0.09872
$k = 10$	0.1429	0.1195	0.1180	0.1118	0.09502
bias of b , Poisson					
$k = 5$	0.09133	0.06826	0.12269	0.21668	0.1123
$k = 10$	0.15252	0.11607	0.06024	-0.01395	0.1347
bias of b , GEE					
$k = 5$	0.09138	0.06829	0.12272	0.21671	0.1123
$k = 10$	0.15256	0.11610	0.06027	-0.01392	0.1347
variance of b , mod. rob.					
$k = 5$	2.674	1.364	0.7706	0.4536	0.2924
$k = 10$	2.546	1.311	0.7452	0.4266	0.2410
variance of b , naive					
$k = 5$	3.517	1.6908	1.4923	0.8766	0.6618
$k = 10$	1.511	0.8239	0.7739	0.4706	0.3094
variance of b , robust					
$k = 5$	2.356	1.201	0.6781	0.3993	0.2575
$k = 10$	2.243	1.155	0.6558	0.3754	0.2121
efficiency (r), new app.					
$k = 5$	0.8457	0.9160	0.7519	0.6152	0.7793
$k = 10$	0.7825	0.7471	0.8183	0.8497	0.6449
efficiency (r), naive					
$k = 5$	1.1124	1.1352	1.4561	1.1891	1.7641
$k = 10$	0.4643	0.4694	0.8498	0.9373	0.8281
efficiency (r), robust					
$k = 5$	0.7451	0.8066	0.6617	0.5416	0.6863
$k = 10$	0.6893	0.6580	0.7202	0.7477	0.5676

In order to study the ability of gamma interferon to reduce the rate of serious infections, a double-blinded clinical trial was conducted in which patients were randomized to placebo vs. gamma interferon. The data used here, which is a little different from that was used by Fleming and Harrington in the example (on page 162), has 65 patients in the placebo

group, 63 in gamma interferon group, of 30 placebo patients who experienced at least one infection, 4 experienced 2, 4 experienced 3, 1 experienced 4, 1 experienced 5 and 1 experienced 7; of 14 treatment patients who experienced at least one infection, 4 experienced 2 and 1 experienced 3.

Table 2: Results for simulated multivariate Weibull distribution with number of clusters = 20 and 20% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.

In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and $cens$ is percentage of censoring.

$M_i = 20, cens = 20\%$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Bias of a , scale parameter					
$k = 5$	0.1057	0.1744	0.1304	0.1441	0.10299
$k = 10$	0.1357	0.1215	0.1276	0.1165	0.09683
bias of b , Poisson					
$k = 5$	0.1076	0.04897	0.10081	0.19289	0.08116
$k = 10$	0.1658	0.10761	0.02919	-0.05122	0.11866
bias of b , GEE					
$k = 5$	0.1078	0.04907	0.10092	0.19301	0.08126
$k = 10$	0.1659	0.10773	0.02929	-0.05112	0.11877
variance of b , mod. rob.					
$k = 5$	3.308	1.690	1.0056	0.6356	0.4219
$k = 10$	3.120	1.642	0.9672	0.5725	0.3428
variance of b , naive					
$k = 5$	4.064	2.159	1.8301	1.1675	0.919
$k = 10$	1.809	1.011	0.9245	0.6216	0.433
variance of b , robust					
$k = 5$	2.921	1.493	0.8879	0.5616	0.3725
$k = 10$	2.751	1.451	0.8540	0.5050	0.3028
efficiency (r), new app.					
$k = 5$	0.8984	1.008	0.9238	0.7869	0.9634
$k = 10$	0.8430	0.811	0.9650	0.9482	0.8272
efficiency (r), naive					
$k = 5$	1.1033	1.2876	1.6810	1.445	2.098
$k = 10$	0.4887	0.4993	0.9224	1.029	1.045
efficiency (r), robust					
$k = 5$	0.7931	0.8903	0.8156	0.6953	0.8506
$k = 10$	0.7432	0.7164	0.8520	0.8364	0.7306

Table 3: Results for simulated multivariate Weibull distribution with number of clusters = 50 and 10% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.

In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and $cens$ is percentage of censoring.

$M_i = 50, cens = 10\%$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Bias of a , scale parameter					
$k = 5$	0.04065	0.04009	0.06156	0.05327	0.04993
$k = 10$	0.05702	0.05781	0.07181	0.04742	0.02708
bias of b , Poisson					
$k = 5$	-0.01885	0.01177	0.03810	0.02688	0.04290
$k = 10$	0.10574	0.04004	0.09806	0.12305	0.02239
bias of b , GEE					
$k = 5$	-0.01884	0.01178	0.03810	0.02689	0.04291
$k = 10$	0.10575	0.04005	0.09806	0.12306	0.02240
variance of b , mod. rob.					
$k = 5$	0.9105	0.4661	0.2606	0.1546	0.09552
$k = 10$	0.8890	0.4462	0.2436	0.1422	0.08358
variance of b , naive					
$k = 5$	1.2961	0.7901	0.5114	0.4461	0.2702
$k = 10$	0.8169	0.5408	0.2750	0.2409	0.1648
variance of b , robust					
$k = 5$	0.8738	0.4476	0.2501	0.1484	0.09171
$k = 10$	0.8532	0.4285	0.2338	0.1365	0.08022
efficiency (r), new app.					
$k = 5$	0.860	0.9513	0.8769	0.6957	0.7103
$k = 10$	0.955	0.8215	0.7185	0.8391	0.6254
efficiency (r), naive					
$k = 5$	1.2242	1.6126	1.7212	2.008	2.009
$k = 10$	0.8775	0.9956	0.8111	1.422	1.233
efficiency (r), robust					
$k = 5$	0.8253	0.9135	0.8418	0.6678	0.6819
$k = 10$	0.9165	0.7889	0.6896	0.8056	0.6003

Table 4: Results for simulated multivariate Weibull distribution with number of clusters = 50 and 20% censoring. Mean bias and variance of regression parameter estimates over 200 simulations.

In the Table, a is scale parameter of Weibull distribution, b is regression parameter, α is index of positive stable distribution, k is the cluster size, m_i is number of clusters, and $cens$ is percentage of censoring.

$M_i = 50, cens = 20\%$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Bias of a , scale parameter					
$k = 5$	0.04621	0.04515	0.0594	0.04871	0.05098
$k = 10$	0.05784	0.06741	0.0750	0.05162	0.02529
bias of b , Poisson					
$k = 5$	-0.02937	-0.0003165	0.03197	0.02579	0.03369
$k = 10$	0.11790	0.0056062	0.08744	0.11434	0.01731
bias of b , GEE					
$k = 5$	-0.02928	-0.0002196	0.03208	0.02589	0.03379
$k = 10$	0.11803	0.00057196	0.08756	0.11443	0.01741
variance of b , mod. rob.					
$k = 5$	1.064	0.5632	0.3231	0.1986	0.1235
$k = 10$	1.040	0.5327	0.3021	0.1795	0.1089
variance of b , naive					
$k = 5$	1.5115	0.9028	0.6095	0.5428	0.3484
$k = 10$	0.9288	0.6249	0.3331	0.2876	0.2029
variance of b , robust					
$k = 5$	1.027	0.5440	0.3129	0.1926	0.1198
$k = 10$	1.005	0.5153	0.2930	0.1739	0.1057
efficiency (r), new app.					
$k = 5$	0.8992	0.9887	0.9812	0.8340	0.8235
$k = 10$	0.9821	0.8901	0.8226	0.9614	0.7481
efficiency (r), naive					
$k = 5$	1.2774	1.585	1.8508	2.28	2.323
$k = 10$	0.8765	1.044	0.9069	1.54	1.393
efficiency (r), robust					
$k = 5$	0.8676	0.9551	0.9502	0.8088	0.7987
$k = 10$	0.9490	0.8609	0.7976	0.9311	0.7259

In order to check how our method works in the real data situation, we fit the CGD using the newly proposed approach with single treatment indicator covariate without controlling other covariates. As we can see in Table 5, the coefficients from our method and Segal and Neuhaus's method with independent working correlation structure are the same, and the coefficients using Andersen-Gill's and Cox model are similar. In Cox model, only the first event was used. The former (our method and Segal and Neuhaus's) is different from the latter (Andersen-Gill's model and Cox model) because the models are different; the coefficients are proportional by a constant, which is the index parameter in the positive stable distribution. Currently, to obtain an estimate of this correlation parameter is problematic as mentioned in Segal and Neuhaus (1993). Nevertheless, the ratio of $\hat{\beta}/s.e(\hat{\beta})$ from our method is comparable with that from Andersen-Gill model. Thus, our method is effective to detect significance of the treatment effect (gamma interferon) though the coefficient is underestimated since the index from the positive stable distribution is between 0 and 1.

Table 5: Results of fitting the CGD (Chronic Granulomatous Disease) dataset of various methods under consideration.

	$\hat{\beta}$	$s.e(\hat{\beta})$	$ \hat{\beta} /s.e(\hat{\beta})$
New method	-0.856	0.2501	3.4226
Segal and Neuhaus	-0.856	0.2489	3.4389
Andersen-Gill	-1.2765	0.3774	3.3824
Cox model	-1.2062	0.4398	2.7426

Conclusion

It has been known that AG, WLW and PWP methods are extensions of survival models based on the Cox proportional hazards approach. They work well in one situation, but may not be appropriate in another (see Kelly and Lim, 2000, Therneau and Hamilton, 1997), since each method has different risk sets and risk intervals. Our new method was developed using a design effect approach from survey sampling and works well for the multivariate failure data. In addition, it's easy to implement. The strong assumption of the parametric form of the survival time can be relaxed by extending our method to the piecewise exponential case, which makes our method more flexible (Aitkin et. al., 1983). No covariance structure between the survival times in a cluster needs to be specified since it's implicitly built in our method.

As seen in our simulation study, the newly proposed method has slightly better finite sample performance than GEE based method. One limitation of our design effect method is that no time-dependent covariates are allowed. We also need to find a method to obtain an estimate of correlation parameter, as we saw it in Table 5; alternatively, a possible estimation strategy proposed by Segal, Neuhaus and James (1997) can be used for that. However, this limitation does not affect our ability to do inference about the regression parameters.

In our simulation, the censoring indicator is generated by fixed censoring time, a work on more general censoring mechanism, such as "independent censoring", is needed. In conclusion, the method of applying the cluster sampling techniques in the multiple failure data is a competing method of recent developed marginal approaches in the literature.

References

- Aalen O. O. (1988). Heterogeneity in survival analysis, *Statistics in Medicine* 7, 1121-1137.
- Aitkin M., & Clayton D. G. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics* 29, 156-163.

- Aitkin M., Laird N., & Francis B., (1983). Reanalysis of the Stanford heart transplant data. *Journal of the American Statistical Association* 78, 264-274 (1983).
- Andersen P. K., & Gill R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* 10, 1100-1120.
- Bieler G. S., & Williams R. L. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* 51, 764-776.
- Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51, 279-292.
- Chambers J. M., & Stuck B. W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association* 71, 340-344.
- Clayton D. G. (1983). Fitting a general family of failure-time distributions using GLIM. *Applied Statistics* 32, 102-109.
- Cochran, W. G. (1977). *Sampling Techniques*, Wiley, New York.
- Fleming, T. R., & Harrington D. P. (1991). *Counting Processes And Survival Analysis*, John Wiley & Sons, New York.
- Therneau, T. M., & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.
- Hougaard P. (1986). A class of multivariate failure time distributions. *Biometrika* 73, 671-678.
- Hougaard P. (1986). Survival models for heterogeneous population derived from stable distributions. *Biometrika* 73, 387-396.
- Huang X., Chen S., & Soong S. (1998). Piecewise exponential survival trees with time-dependent covariates. *Biometrics* 54, 1420-1433.
- Kelly P. J., & Lim L. L.-Y. (2000). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine* 19, 13-33.
- Laird N., & Oliver D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* 76, 3231-240.
- Liang K. Y., & Zeger S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.
- Lin D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine* 13, 2233-2247.
- McGilchrist C. A., & Aisbett C. W. (1991). Regression with frailty in survival analysis. *Biometrics* 47, 461-466.
- Norton K. I. et al., (2001). Chronic radiographic lung changes in children with vertically transmitted HIV-I infection. *American Journal of Roentgenology* 176, 1553-1558.
- Paik, M. C. (1988). Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics-Simulation* 17, 1155-1171.
- Prentice R. L., Williams B. J., & Peterson A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* 68, 373-379.
- Rao, J., & Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* 48, 577-585.
- Rao, J., & Scott, A. J. (1999). A simple method for analysing overdispersion in clustered Poisson data. *Statistics in Medicine* 48, 577-585.
- Segal M. R., & Neuhaus J. M. (1993). Robust inference for multivariate survival data. *Statistics in Medicine* 12, 1019-1031.
- Segal M. R., Neuhaus J. M., & James I. R. (1997). Dependence estimation for marginal models of multivariate survival data. *Lifetime Data Analysis* 3, 251-268.
- Therneau T. M., & Hamilton S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine* 16, 2029-2047.
- Wei L. J., Lin, D.Y., & Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* 84, 1065-1073.
- Woodruff R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 66, 411-414.

Variance Stabilizing Power Transformation for Time Series

Victor M. Guerrero Rafael Perera
Department of Statistics
Instituto Tecnológico Autónomo de México

A confidence interval was derived for the index of a power transformation that stabilizes the variance of a time-series. The process starts from a model-independent procedure that minimizes a coefficient of variation to yield a point estimate of the transformation index. The confidence coefficient of the interval is calibrated through a simulation.

Key words: Autocorrelation, coefficient of variation, confidence interval, model assumptions

Introduction

Applied model-based statistical analysis usually requires some assumptions to be satisfied by the data under study. When working with time-series, covariance-stationarity is often required to begin the modeling process. Therefore it is reasonable to look for a variance stabilizing transformation that will make the data get closer to fulfilling this assumption. Within the forecasting area, recall de Bruin and Franses' (1999) conclusion that data transformations should be considered prior to forecasting.

There are two approaches to search for the transformation. (i) Select the transformation before actually building a statistical model for the time series, or (ii) decide which transformation to use during the model building process. In the latter approach both model form and parameter estimation interact with the search for the transformation.

In the former, the scale where the analysis should be carried out is fixed before attempting to build a statistical model. This approach allows the analyst to select a transformation without conditioning on or interfering with a given model. Therefore it is called model-independent.

The focus in this article is on a model-independent method that is useful to select a power transformation that best stabilizes the variance of a time series variable $Z_t > 0$, for $t=1, \dots, N$. Such a method was proposed by Guerrero (1993) as a tool to be employed when the analyst wants to use the power transformation family: $T(Z_t) = Z_t^\lambda$ if $\lambda \neq 0$ and $T(Z_t) = \log(Z_t)$ if $\lambda = 0$ or when using its Box-Cox version: $Z_t^{(\lambda)} = (Z_t^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and $Z_t^{(\lambda)} = \log(Z_t)$ if $\lambda = 0$.

One of the most important works that proposed the second approach for choosing a transformation is the textbook by Box and Jenkins (1976). They suggested using the Box-Cox transformation in order to validate not only the constant variance assumption, but all the underlying assumptions of an Auto-Regressive Integrated Moving Average (ARIMA) model by estimating the transformation index (λ) together with the model parameters. Chen and Lee (1997) proposed a Bayesian method to choose the value of λ for a given model structure. Those works are supported by sound statistical theory, although in practice they present the problem that the model form may depend on the transformation selected. In fact, Gourieroux and Jasiak (2002) have shown that the

Victor M. Guerrero is Professor of Statistics at Instituto Tecnológico Autónomo de México. His research interests are in the analysis and forecasting of economic time series. He gratefully acknowledges financial support from Asociacion Mexicana de Cultura, A. C., in carrying out this work. Email him at: guerrero@itam.mx. Rafael Perera graduated from the University of Oxford and is employed at the Department of Primary Health Care at Oxford University. Email him at: rafael.perera@public-health.oxford.ac.uk.

autocorrelations (hence the ARIMA model structure) change as a function of the nonlinear transformation employed. Therefore, fixing the model form before selecting the transformation index could be inappropriate in some cases.

An advantage of the second approach for choosing the transformation is that a measure of variability, as well as a reference distribution, can be obtained for the estimated transformation index. Thus, it is possible to discriminate among different alternative values of λ based on a priori considerations. For instance, deciding whether the data should be analyzed in the original scale ($\lambda = 1$) or in logarithms ($\lambda = 0$), can be performed on the basis of the data at hand. This does not happen with the first approach because no model form and no reference distribution exist that will support the decision on an empirical basis. This fact can be considered a drawback of this approach. In this article, we consider this problem and work out a feasible solution by means of a confidence interval for the true λ value.

In the following section a summary of Guerrero's (1993) method is presented that produces a point estimate of the index λ by minimizing a coefficient of variation. Then, a confidence interval is derived for the true value of λ . Approximate expressions for some sample moments involved in the calculations are provided, and a reference distribution for the true coefficient of variation employed by the method is suggested. Some small sample simulations are used to calibrate the confidence coefficient of the interval and to get an insight into the performance of the procedure. Nominal confidence levels are related to realized levels and, useful empirical results are obtained. A section is devoted to illustrate the use of the method through some empirical applications. These examples help to understand how the method works in practice.

Selection of the Transformation

Guerrero (1993) proposed two methods for selecting the power transformation index λ . Underlying these methods is the theoretical result that states that the choice of the transformation index should be done in such a way that $[\text{var}(Z_t)]^{1/2} / [E(Z_t)]^{1-\lambda} = c$ holds valid

for all t and some constant $c > 0$. To use this result, it is necessary to estimate both the mean and the variance involved. In applied time series analysis there is usually only one observation at each time t , therefore $\text{var}(Z_t)$ cannot be estimated and that result cannot be applied directly. In order to operationalize the result, work with the observations grouped into $H \geq 2$ subseries. This enables the calculation of pairs of sample means and standard deviations, for example, $(\bar{Z}_h, S_h)$ for $h=1, \dots, H$, and then search for the λ value that produces

$$S_h / \bar{Z}_h^{1-\lambda} = c \text{ for } h=1, \dots, H \quad (1)$$

for some constant $c > 0$. The elements in this equation are given by $\bar{Z}_h = \sum_{r=1}^R Z_{h,r} / R$ and $S_h^2 = \sum_{r=1}^R (Z_{h,r} - \bar{Z}_h)^2 / (R-1)$, where $Z_{h,r}$ denotes the r th observation of subseries h . The subseries $\{Z_{h,1}, \dots, Z_{h,r}, \dots, Z_{h,R}\}$, for $h=1, \dots, H$, are formed by grouping R consecutive observations of the original series $\{Z_t : t=1, \dots, N\}$, trying to keep homogeneity between the subseries. For this to happen they must be equal-sized. Therefore, some number (n) of observations, with $0 \leq n < R$, will have to be left out of the calculations, leaving $R=(N-n)/H$. The subseries size must be chosen appropriately, and be equal to the length of the seasonality, if such an effect is present in the series.

The proposed methods stemmed from two empirical interpretations of equation (1). The first one led to minimizing the coefficient of variation of $S_h / \bar{Z}_h^{1-\lambda}$ as a function of λ . This method is not linked to a formal statistical model and therefore no assumptions need to be validated to be applied correctly in practice. The second empirical interpretation led to a method based on a simple linear regression in logarithms. The assumption of zero error autocorrelation that underlies this method needs careful attention as it is seldom valid when working with time series. Thus, the main method, because of its robustness against violation of assumptions, is the one that minimizes relative variation. We shall concentrate on that method.

A Confidence Interval for λ

To be able to make inferences about λ , estimated as the minimizer of the coefficient of variation, we require a reference statistical distribution. To get such a distribution we start by assuming that the random variables $W_h(\lambda) = S_h / \bar{Z}_h^{1-\lambda}$ for $h=1, \dots, H$, can be represented by a Moving Average model of order 1. That is $W_1(\lambda) - \mu = a_1$ and $W_h(\lambda) - \mu = a_h - \theta a_{h-1}$ for $h=2, \dots, H$, with $\{a_h\}$ a zero-mean white noise Gaussian process, $\mu > 0$ and θ a constant parameter such that $|\theta| < 1$ to ensure it is invertible. Thus, $E[W_h(\lambda)] = \mu$, $\text{var}[W_h(\lambda)] = \sigma^2$ and $\text{corr}[W_h(\lambda), W_{h'}(\lambda)] = \rho$ if $h' = h \pm 1$, and zero otherwise, with $\rho = -\theta / (1 + \theta^2) \in (-0.5, 0.5)$.

Such a model makes sense because λ is obtained in such a way that $W_1(\lambda), \dots, W_H(\lambda)$ are approximately constant, but a slight autocorrelation structure is expected in the process $\{W_h(\lambda)\}$ given that $\bar{Z}_h$ and S_h are calculated from time series observations. This assumption was validated by the simulations reported below as the expected behavior was observed. For the sake of simplicity, do not write $\mu(\lambda)$, $\sigma^2(\lambda)$ and $\rho(\lambda)$ even though these parameters are functions of λ .

The sample counterparts of μ and σ^2 will be denoted as $m = \sum_{h=1}^H W_h(\lambda) / H$ and $se^2 = \sum_{h=1}^H [W_h(\lambda) - m]^2 / (H-1)$ so that $CV(\lambda) = se/m$ is the sample coefficient of variation. In what follows we shall derive an approximate distribution for $CV(\lambda)$, from which a confidence interval for the true λ value can be obtained. Several proposals may be found in the literature to obtain the distribution, hence confidence intervals, for a Normal coefficient of variation (see Vangel, 1996, and the references therein), but none of them allows for autocorrelation in the observations.

We first apply the Theorem in Appendix 1 (known as the Delta Method) to the bivariate case, with $X_1 = se$, $X_2 = m$ and $g(X_1, X_2) = X_1 / X_2$, to get $E[CV(\lambda)] \approx E(se) / E(m)$ and

$$\text{var}[CV(\lambda)] \approx$$

$$\frac{E^2(se)}{E^2(m)} \left[\frac{\text{var}(se)}{E^2(se)} + \frac{\text{var}(m)}{E^2(m)} - \frac{2\text{cov}(se, m)}{E(se)E(m)} \right].$$

Then, evaluate each term in this expression as indicated in Appendix 2, so that $E(m) = \mu$,

$$\text{var}(m) = \sigma^2 [1 + 2\rho(H-1)/H] / H,$$

$$E(se) \approx \sigma \{1 - 2\rho / H - 1 / [2(H-1)]\}^{1/2},$$

$\text{var}(se) \approx \sigma^2 / [2(H-1)]$ and $\text{cov}(se, m) \approx 0$. Hence,

$$E[CV(\lambda)] \approx \frac{\sigma}{\mu} \left[1 - \frac{2\rho}{H} - \frac{1}{2(H-1)} \right]^{1/2}$$

and

$$\text{var}[CV(\lambda)] \approx \frac{\sigma^2}{2(H-1)\mu^2}$$

$$\left\{ 1 + \frac{2\sigma^2}{\mu^2} \left(\frac{H-1}{H} \right) \left[1 - \frac{2\rho}{H} - \frac{1}{2(H-1)} \right] \left[1 + \frac{2\rho(H-1)}{H} \right] \right\} \\ \approx \sigma^2 / [2(H-1)\mu^2]$$

where the last approximation follows from the fact that σ/μ must be close to zero, since λ is chosen to accomplish that goal. It is clear that $E[CV(\lambda)] \rightarrow \sigma/\mu$ as $H \rightarrow \infty$ and that it is a decreasing function of ρ . In fact, when $\rho \geq 0$ we observe that $E[CV(\lambda)] < \sigma/\mu$ for all H , and the opposite occurs when $\rho < 0$. Similarly, it is easy to see that $\text{var}[CV(\lambda)] \rightarrow 0$ as $H \rightarrow \infty$.

Because the variance of $CV(\lambda)$ is proportional to the square of its mean, the logarithm becomes an adequate variance-stabilizing power transformation (see Guerrero, 1993, eq. 4). In turn, assume that (roughly) $\log[CV(\lambda)] \sim N(\eta, \delta^2)$. From the Lognormal distribution, $E[CV(\lambda)] = \exp(\eta + \delta^2/2)$ and $\text{var}[CV(\lambda)] = [\exp(\delta^2) - 1] \exp(\delta^2 + 2\eta)$. Thus, solve for η and δ^2 , to get

$$\eta = \log\{E[CV(\lambda)]\} - \delta^2 / 2 \\ \approx \log(\sigma / \mu) + \log\{1 - 2\rho/H - 1/[2(H-1)]\} / 2 - \delta^2 / 2$$

and

$$\delta^2 = \log \left\{ \frac{\text{var}[\text{CV}(\lambda)]}{E^2[\text{CV}(\lambda)]} + 1 \right\}$$

$$\approx \log(1 - 2\rho/H) - \log\{1 - 2\rho/H - 1/[2(H-1)]\}$$

It is known that

$$1 - \alpha \approx \Pr \left\{ \frac{\log \text{CV}(\lambda) - \eta}{\delta} \geq -z_\alpha \right\}$$

$$= \Pr \left\{ \text{CV}(\lambda) \geq \frac{\sigma}{\mu} \frac{1 - 2\rho/H - 1/[2(H-1)]}{(1 - 2\rho/H)^{1/2}} \exp(-\delta z_\alpha) \right\}$$

with z_α the 100α upper percentile of the unit Normal distribution. The previous assertion leads us to an approximate $100(1-\alpha)\%$ confidence interval for the true coefficient of variation. Because, there is a one-to-one correspondence between coefficient of variation and λ value, it follows that an approximate $100(1-\alpha)\%$ confidence interval for λ is given by

$$CI_{1-\alpha}(\lambda) = \left\{ \lambda : \frac{\sigma}{\mu} \leq \frac{\text{CV}(\lambda)(1 - 2\rho/H)^{1/2}}{1 - 2\rho/H - 1/[2(H-1)]} \exp(\delta z_\alpha) \right\} \quad (2)$$

In order for this confidence interval to be useful in practice, estimate $\text{CV}(\lambda)$ as the minimum sample coefficient of variation, denoted as $\text{CV}(\hat{\lambda})$. Similarly, use the estimated first-order autocorrelation coefficient,

$$\hat{\rho} = \frac{\sum_{h=1}^{H-1} [W_h(\hat{\lambda}) - m] [W_{h+1}(\hat{\lambda}) - m]}{\sum_{h=1}^H [W_h(\hat{\lambda}) - m]^2}$$

and an estimate of δ , say $\hat{\delta}$, can be obtained by using $\hat{\rho}$ in place of ρ . Keep in mind that the interval (2) was derived from several approximations, in such a way that the actual confidence level may differ from the nominal level and calibration is required.

To appreciate numerically the effect that α , H and ρ have on the length of the confidence interval, some calculations are presented in Table 1 for selected values of those constants. This table shows values of the function

$$f(\alpha, H, \rho) \equiv \exp(\delta z_\alpha)(1 - 2\rho/H)^{1/2} / \{1 - 2\rho/H - 1/[2(H-1)]\}$$

which is the expanding factor of $\text{CV}(\lambda)$ that defines the length of $CI_{1-\alpha}(\lambda)$. It is clear that $f(\alpha, H, \rho)$ gets smaller as: (i) α gets larger, (ii) H gets larger (in fact, $f(\alpha, H, \rho) \rightarrow 1$ as $H \rightarrow \infty$) and/or (iii) ρ moves from positive to negative values. The first two of these conclusions have a clear interpretation in terms of confidence and sample size. The third has no clear explanation, but it should be borne in mind when trying to understand why two similar situations, differing only in the sign of ρ , will yield different results (especially when α and H are small). In practical applications, typically $H \geq 6$, so that ρ should not be expected to be the decisive factor in defining the size of the confidence interval, but we should be aware of its potential relevance.

In order to better understand how the method works, in Figure 1 the graph is presented of $\text{CV}(\lambda)$ against λ for the Sales Data that will be considered as an illustrative example below. Observe that the confidence interval is obtained by slicing the curve produced by the coefficient of variation of the variable $S_h / \bar{Z}_h^{1-\lambda}$, for $h=1, \dots, H$, as a function of λ . The minimum of this curve yields $\text{CV}(\hat{\lambda})$ and the required confidence interval is built by projecting on the horizontal axis the points where the curve reaches $\text{CV}(\hat{\lambda})f(\alpha, H, \hat{\rho})$, for a given α value.

Table 1: Expanding factor of $CV(\lambda)$ as a function of α , H and ρ .

α	$H \backslash \rho$	-0.45	-0.25	-0.05	0	0.05	0.25	0.45
0.01	2	5.75	7.86	12.1	13.9	16.2	39.7	544
	6	2.06	2.18	2.33	2.36	2.40	2.58	2.80
	50	1.26	1.27	1.28	1.28	1.28	1.29	1.29
0.05	2	3.69	4.83	6.99	7.87	8.98	19.4	189
	6	1.68	1.77	1.87	1.90	1.92	2.05	2.20
	50	1.18	1.19	1.19	1.19	1.19	1.20	1.21
0.1	2	2.92	3.73	5.22	5.82	6.56	13.3	108
	6	1.50	1.58	1.66	1.68	1.71	1.81	1.93
	50	1.14	1.14	1.15	1.15	1.15	1.16	1.16

Methodology

The confidence interval for λ was derived from several approximations that may cause the actual confidence level to differ from its nominal level. In order to calibrate the confidence coefficient, a small simulation study based on the following two model specifications was conducted.

Figure 1. 95% Confidence interval for λ built from $CV(\hat{\lambda})$ for the Sales Data.

$$1) Z_t = T_t + S_t + I_t,$$

where $T_t = t$, $S_t = \sum_{q=1}^{12} \delta_q D_{t,q}$ with $D_{t,q} = 1$ if $t = 12i + q$ for $i = 0, 1, \dots, H-1$ and $D_{t,q} = 0$ otherwise. $\{I_t\} \sim \text{independent } N(0, \sigma_t^2)$ with $\sigma_t^2 = E(Z_t)^{2(1-\lambda)}$ and $E(Z_t) = T_t + S_t$.

$$2) Z_t = (1 + \phi)Z_{t-1} - \phi Z_{t-2} + a_t - \theta a_{t-1}, \quad \text{where} \\ \{a_t\} \sim \text{independent } N(0, \sigma_t^2) \quad \text{with} \\ \sigma_t^2 = E(Z_t)^{2(1-\lambda)} \quad \text{and} \\ E(Z_t) = [(2 - \phi) - \phi^t] / (1 - \phi).$$

The first one is a seasonal model with seasonality length $R=12$. The parameter values for the seasonal effects were chosen as $\delta_1 = 2$, $\delta_2 = 4$, $\delta_3 = 5$, $\delta_4 = 0$, $\delta_5 = -1$, $\delta_6 = -2$, $\delta_7 = -3$, $\delta_8 = -3$, $\delta_9 = -2$, $\delta_{10} = 0$, $\delta_{11} = 1$, $\delta_{12} = -1$ so that $\sum_{q=1}^{12} \delta_q = 0$. The sample sizes were of the form $N=12H$, with $H=6, 12, 20, 30$. The second is an $ARIMA(1,1,1)$ model with initial values $Z_0 = 1$, and $Z_1 = 2$ with parameter values $\phi = 0.7$ and $\theta = 0.3$. In this case, the subseries size was taken as $R = 4$ and the sample sizes were $N=24, 48, 80, 120$, so that the values of H became again 6, 12, 20 and 30. Another exercise was carried out with the latter model and $R=3$, and sample sizes $N=18, 36, 60, 90$ to get the same values for H as before. For both models, $\lambda=0, 0.5, 1$ was employed; thus, when $\lambda \neq 1$ there is nonconstant variance, because it depends on the mean of the series.

Jennings' (1987) suggestion about the way that simulation studies should be reported was followed in order to provide information not only on coverage rates but also on bias. In Table 2, some results are presented from the simulations for the seasonal model. Similarly, Tables 3 and 4 show the corresponding results for the nonseasonal model, with $R=4$ and $R=3$, respectively.

Table 2: Observations below\above of a $100(1-\alpha)\%$ nominal confidence interval for λ and actual α , with Model 1 and $R=12$ (1000 samples).

H	α	z_{nom}	λ			Actual α average	z_{act}
			0	0.5	1		
6	0.005	2.575	5\5	9\13	4\10	0.015	2.170
	0.045	1.695	16\29	25\26	24\26	0.049	1.651
	0.050	1.641	18\32	26\26	26\26	0.051	1.635
	0.125	1.150	44\55	52\50	45\45	0.097	1.299
	0.130	1.128	44\58	56\53	49\47	0.102	1.270
12	0.040	1.750	3\6	11\1	3\4	0.009	2.365
	0.045	1.695	4\6	11\2	3\4	0.010	2.326
	0.195	0.860	19\25	35\23	28\19	0.050	1.645
	0.200	0.841	21\27	36\25	29\19	0.052	1.626
	0.295	0.539	40\48	61\50	59\40	0.099	1.289
	0.300	0.521	42\48	63\53	60\41	0.102	1.270
20	0.125	1.151	5\2	9\4	3\4	0.009	2.365
	0.130	1.128	5\3	9\4	5\4	0.01	2.326
	0.295	0.539	30\20	28\16	29\24	0.049	1.651
	0.300	0.521	32\20	31\16	29\26	0.051	1.635
	0.365	0.341	40\39	56\45	51\54	0.095	1.310
	0.370	0.330	43\42	57\47	57\59	0.102	1.270
30	0.190	0.879	5\1	12\2	8\2	0.010	2.326
	0.195	0.860	6\1	14\2	8\2	0.011	2.290
	0.335	0.421	18\17	37\15	30\24	0.047	1.679
	0.340	0.411	18\18	40\16	33\26	0.050	1.645
	0.395	0.251	37\35	76\25	60\58	0.097	1.300
	0.400	0.251	38\36	81\28	61\61	0.102	1.270

Table 3: Observations below\above of a $100(1-\alpha)\%$ nominal confidence interval for λ and actual α , with Model 2 and $R = 4$ (1000 samples).

H	α	z_{nom}	λ			Actual α average	z_{act}
			0	0.5	1		
6	0.005	2.575	8\9	7\12	9\13	0.019	2.075
	0.040	1.752	19\23	14\38	20\31	0.048	1.665
	0.045	1.695	21\23	14\46	23\36	0.054	1.607
	0.115	1.202	39\45	24\82	42\60	0.097	1.299
	0.120	1.175	39\46	24\87	43\66	0.102	1.270
12	0.035	1.812	1\3	0\10	7\6	0.009	2.365
	0.40	1.751	1\4	0\12	7\7	0.010	2.326
	0.165	0.974	10\19	11\61	20\28	0.050	1.645
	0.170	0.954	10\19	12\63	20\28	0.051	1.635
	0.245	0.693	19\41	20\113	37\65	0.098	1.353
	0.250	0.675	22\48	21\119	38\68	0.105	1.254
20	0.075	1.434	0\0	1\9	4\11	0.008	2.410
	0.080	1.405	1\0	1\10	4\13	0.010	2.326
	0.220	0.772	6\4	10\56	20\54	0.050	1.645
	0.225	0.755	6\5	10\58	20\57	0.052	1.626
	0.305	0.510	13\21	19\111	34\97	0.098	1.353
	0.310	0.496	14\21	20\112	35\101	0.101	1.275
30	0.105	1.254	0\0	2\17	1\9	0.010	2.326
	0.110	1.227	0\0	2\17	3\10	0.011	2.290
	0.270	0.611	2\3	11\66	16\53	0.050	1.645
	0.275	0.599	2\3	11\68	18\57	0.053	1.619
	0.340	0.411	7\10	24\117	37\94	0.096	1.308
	0.345	0.400	9\11	27\124	39\95	0.102	1.270

In these tables, the nominal confidence levels of the intervals were selected by trial and error. That is, we increased the confidence level by an amount of 0.005 units and looked for the levels that yield actual coverage rates of 99%, 95% and 90%, which are the most commonly used in practice. The *actual* α values were obtained by averaging over the different coverage rates obtained for $\lambda=0,0.5,1$. The group size $R=12$ was used for the monthly seasonal series because this is the usual practice. There is no commonly accepted value for nonseasonal time series. For instance, Guerrero's (1993) advice was to employ $R=2$ in order to minimize the loss of information by grouping. However, with this choice the estimation of variability required is very poor and perhaps a value $R>2$ could perform better. By looking at Table 2 it is reasonably clear that $H=6$ serves to obtain actual confidence levels similar to the nominal ones.

In Tables 3 and 4, the value of R was sought that makes the method work well also for $H=6$, when the series is nonseasonal. It was found that $R=4$ is preferable to $R=3$ in terms of having less bias and more comparable results for the different λ values. However, in Tables 2, 3 and 4, the value of the estimated autocorrelation coefficient was not considered, because it was not under our control. The simulations were carried out with the statistical package S-Plus 2000 (MathSoft, Inc.).

On the basis of these simulations, it was concluded that the nominal confidence level depends on the following factors: (i) the actual confidence level, (ii) the value of H , and (iii) the value of R . Thus, in order to calibrate the confidence intervals we estimated the following linear regression model (standard errors in parentheses) with $\bar{R}^2 = 0.9503$, $\hat{\sigma} = 0.1265$ and sample size=69

$$z_{\text{nom}} = 0.8845 - 0.0200R - 0.1426H + 0.0028H^2 + 0.9838z_{\text{act}} \\ (0.0932) \quad (0.0038) \quad (0.0091) \quad (0.0002) \quad (0.0361)$$

This result indicates that the Normal approximation derived previously requires a statistically significant numerical correction. With this equation, the appropriate z_{nom} may be calculated, given the values of R and H , as well as the desired z_{act} , corresponding to the actual

confidence level. Such a nominal value can then be introduced in expression (2) to obtain an appropriate confidence interval.

Illustrative Applications

The Sales dataset corresponds to the seasonal time series provided by Chatfield and Prothero (1973). The original series has $N=77$ observations on sales of an engineering firm. A time plot of the series without transformation appears in Figure 2(a) and power-transformed with $\lambda=0.254$ in Figure 2(b). This transformation index was obtained as minimizer of the coefficient of variation with $H=6$ subseries and $R=12$ observations per subseries (so that $n=5$ observations were left out of the calculations). In this case the autocorrelation required by the confidence interval was estimated as $\hat{\rho} = 0.2554$.

The following confidence intervals were obtained for the true λ value. 99%: (-0.0594,0.5646); 95%: (0.0216,0.4846); and 90%: (0.0616,0.4456). Figure 1 shows a graph of the coefficient of variation $CV(\lambda)$ for these data, together with a 95% confidence interval for λ . Thus, with a confidence level of 95%, it can be determined that $\lambda=0$ is not supported by the data as the index of a variance stabilizing power transformation. In other words, the logarithm is not a reasonable transformation to stabilize the variance of this time series. However, values such as $\lambda=0.25$ or $\lambda=0.34$, are reasonably adequate to represent the true value of λ , even with 90% confidence. This result is in agreement with the basic conclusion reached by previous authors (see Guerrero, 1993).

Now, for comparative purposes, assume that no autocorrelation exists in the series $W_h(\hat{\lambda}) = S_h / \bar{Z}_h^{1-\hat{\lambda}}$, for $h=1, \dots, H$, in such a way that Vangel's (1996) proposal can be used. In this situation, the $100(1 - \alpha) \%$ confidence interval is given by

$$\left\{ \lambda : \frac{\sigma}{\mu} \leq CV(\hat{\lambda}) \left[\left(\frac{\chi_{H-1,1-\alpha}^2 + 2}{H} - 1 \right) \right]^{-1/2} \right. \\ \left. \left[CV^2(\hat{\lambda}) + \frac{\chi_{H-1,1-\alpha}^2}{H-1} \right] \right\}$$

Table 4: Observations below\above of a $100(1-\alpha)\%$ nominal confidence interval for λ and actual α , with Model 2 and $R=3$ (1000 samples).

H	α	z_{nom}	λ			Actual α	
			0	0.5	1	average	z_{act}
6	0.005	2.575	1\5	2\12	7\11	0.013	2.229
	0.045	1.695	15\12	7\54	18\39	0.048	1.665
	0.050	1.641	15\13	7\55	21\40	0.050	1.645
	0.125	1.150	29\31	19\103	35\74	0.097	1.299
	0.130	1.128	31\31	21\105	37\78	0.101	1.279
12	0.040	1.750	1\0	0\17	3\8	0.010	2.326
	0.045	1.695	1\0	0\19	4\12	0.012	2.259
	0.150	1.032	5\5	12\61	16\44	0.048	1.669
	0.155	1.011	5\6	15\63	19\45	0.051	1.635
	0.250	0.671	9\21	23\118	42\85	0.099	1.289
	0.255	0.660	10\22	25\120	42\89	0.103	1.268
20	0.080	1.405	0\0	3\16	6\5	0.010	2.326
	0.085	1.370	0\0	3\17	6\7	0.011	2.290
	0.225	0.755	2\2	11\66	20\42	0.048	1.669
	0.2300	0.740	3\2	11\71	20\405	0.051	1.635
	0.315	0.480	13\10	25\120	39\91	0.099	1.289
	0.320	0.469	15\10	25\125	40\96	0.104	1.252
30	0.080	1.405	0\0	4\15	4\8	0.010	2.326
	0.085	1.370	0\0	4\17	4\11	0.012	2.259
	0.235	0.721	0\0	7\71	23\46	0.049	1.651
	0.240	0.709	0\0	7\74	23\48	0.051	1.635
	0.325	0.451	6\4	18\127	37\105	0.099	1.289
	0.330	0.440	7\5	19\132	38\113	0.105	1.254

Figure 2. Sales data. (a) Original and (b) power-transformed with $\lambda=0.254$.

Because $\hat{\lambda} = 0.254$, $CV(\hat{\lambda}) = 0.0838$, $H=6$ and $\alpha=0.05$, then $\chi^2_{5,0.95} = 1.15$. The confidence interval gets defined by the λ values satisfying the inequality $\sigma/\mu \leq 0.1708$, where it should be recalled that both σ and μ are functions of λ . Hence, (see Figure 1) the 95% confidence interval obtained is: $(-0.0797, 0.5717)$. The corresponding interval (2) on the assumption $\rho=0$, satisfies the inequality $\sigma/\mu \leq 0.1519$ and becomes $(-0.0204, 0.5266)$. Both intervals obtained on the no-autocorrelation assumption cover the value $\lambda=0$, but Vangel's interval is wider than ours. In this exercise, the autocorrelation coefficient changed from being a negative value to zero, leaving everything else constant. This change produced a larger expanding factor of $CV(\lambda)$, hence a wider interval.

Blowfly Data

Nicholson's blowfly data have been analyzed from several angles. Notably among these is the one that employs a nonlinear model for these data, in place of a power transformation (see Young, 2000). Nevertheless, because we are mainly concerned with the use of power transformations, we emphasize the analysis presented in the paper by Chen and Lee (1997). These authors used 82 observations of the original series (from 218 to 299) for comparison with previous works. They also mentioned that other authors used either a logarithmic or a square root transformation (i.e. $\lambda=0$ or $\lambda=0.5$). Then, they employed their method, conditioning on an autoregressive AR(1) model form, and made inferences on both λ and the parameters of that model (mean, autoregressive coefficient and error variance).

The point estimate of the transformation index was obtained as the posterior mean of a distribution obtained by Gibbs sampling with a uniform prior on the set $\{0.30, 0.31, \dots, 0.50\}$. The estimated value, $\hat{\lambda} = 0.39$ with standard error 0.001, clearly differs significantly from $\lambda=0$ and $\lambda=0.5$. However, we believe that Chen and Lee's method is misleading because it conditions on the model form, while the other methods against which they compared their results are model-independent. Moreover, it should be recalled that the model form may

change depending on the value of λ , as indicated by Gouriéroux and Jasiak (2002), thus the AR(1) specification might be in doubt.

We applied our procedure to the data employed by Chen and Lee, without conditioning on any given model structure. By so doing, $\hat{\lambda}=0.3997$, with $R=4$ and $H=20$; so that $n=2$ observations were not used. The point estimate of the transformation index took almost the same value as that obtained by Chen and Lee's method. The autocorrelation became in this case $\hat{\rho}=0.0215$ and the confidence intervals were 99%: $(-1.0448, 1.6272)$; 95%: $(-0.6048, 1.2892)$ and 90%: $(-0.3328, 1.0682)$. These intervals are inconclusive, because even with 90% confidence using the data in the original scale, in a square root scale or in logarithms, produces essentially the same results (in terms of variance stabilization). This result would have been expected just by looking at the graphs shown in Figure 3, where no relevant changes are observed in the time series behavior by changing the scale. We calculated again the interval proposed by Vangel (1996) on the assumption that $\rho=0$ (which may be deemed reasonable since $\hat{\rho}$ is indeed close to zero) with $\hat{\lambda}=0.3997$, $CV(\lambda)=0.54794$, $H=20$ and $\alpha=0.05$, so that $\chi^2_{19,0.95} = 10.12$. The corresponding 95% confidence interval was defined by the λ values satisfying the inequality $\sigma/\mu \leq 0.851204$, (see the graph of $CV(\lambda)$ in Figure 4) that is $(-1.7678, 2.1393)$. Thus, the previous conclusion holds valid even if the assumption $\rho=0$ were true.

Similarly, the graph of $CV(\lambda)$ shown in Figure 4 shows why the intervals are so wide: $CV(\lambda)$ is extremely flat for the range of usual λ values employed in practice. This is an example where the data are basically insensitive to the choice of a variance stabilizing transformation. To test this idea, we estimated the same AR(1) model for the data with the following choices of the transformation index: $\lambda=1, 0.39, 0$.

Figure 3. Blowfly data. (a) Original, (b) power-transformed with $\lambda=0.3997$, and (c) log-transformed. Figure 4. Confidence interval for λ with blowfly data.

The Maximum Likelihood estimation results appear in Table 5, where it may be observed that the estimated AR coefficients ($\hat{\phi}_1$) are almost the same in the three different scales. The Ljung-Box statistics $Q(24-1)$, when compared against a Chi-square distribution with 23 degrees of freedom, show no evidence of inadequacy.

The other two estimated parameters (mean $\hat{\phi}_0$ and residual standard error $\hat{\sigma}$) depend heavily on the scale of the analysis and do not allow a direct comparison. The t -statistics indicate that the estimated coefficients are significantly different from zero in the three cases and the residual graphs (not shown) are also very similar, showing no evidence of nonconstant variance by visual inspection. Thus,

it may be concluded that choosing one particular power transformation, within those indexed by $\lambda=1,0.39,0$, depends on some criterion different from variance stabilization. Perhaps, the forecasting ability of the model should be studied in the different scales, as Chen and Lee (1997) finally did, in order to select the λ value, but that task was outside the scope of this article.

Table 5. Estimation results of the AR(1) model for blowfly data with different choices of λ .

λ	1	0.39	0
$\hat{\phi}_0$	4249.83	63.448	8.311
t - stat	11.35	28.27	96.63
$\hat{\phi}_1$	0.735	0.726	0.712
t - stat	10.09	10.07	9.83
$\hat{\sigma}$	890.83	5.500	0.221
Q(24 - 1)	12.08	11.15	11.92

Conclusion

This article presents a procedure to calculate a confidence interval for the true index of a power transformation that best stabilizes the variance of a time series. This is useful as it enables a time series analyst to make statistical inferences about the transformation index, without relying on a model-dependent method. The procedure was derived from a study of the approximate mean and variance of the minimum coefficient of variation employed for choosing the transformation. Then, a small simulation study allowed us to calibrate the confidence coefficient. This calibration was justified because our analytical results were derived from several approximations that may yield inaccurate results in practical applications.

The coverage rates were found to be dependent on the nominal size of the confidence level, the subseries size R and the number H of subseries used. The simulations led to practical conclusions. For instance, the appropriate subseries size, when there is no seasonality in the time series, was found to be $R = 4$, while the length of the seasonal period is adequate for a seasonal time series (i.e. $R = 12$ for a monthly time series). A more extensive simulation study

would be required to consider negative λ values as well as some other time series models, in order to get more conclusive results.

The empirical illustrations provided evidence on the use the method may have in practical applications. The first example provided an empirical confirmation that our method can be trusted, because we obtained essentially the same results that were established previously by means of Maximum Likelihood. However, our method was applied with less effort, and we did not rely on knowledge of the model structure of the time series, as is required by the Maximum Likelihood method. The second illustration tested the recommendations derived from the simulation study. In fact, it was found that our method led to sensible results and it is relatively easy to apply it.

Finally, it is interesting to note that the confidence interval for the minimum coefficient of variation can also be used to construct confidence intervals for any coefficient of variation. Therefore, the results obtained here may lead to further research in the area of inference for a coefficient of variation in general.

References

- Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time series analysis. Forecasting and control*, 3rd ed. London: Prentice-Hall.
- Chatfield C., & Prothero, D. L. (1973). Box-Jenkins seasonal forecasting: problems in a case-study. *Journal of the Royal Statistical Society, A-136*, 295-315.
- Chen, C. W. S., & Lee, J. C. (1997). On selecting a power transformation in time-series analysis. *Journal of Forecasting*, 16, 343-354.
- de Bruin, P., & Franses, P. H. (1999). Forecasting power-transformed time series data. *Journal of Applied Statistics*, 26, 807-815.
- Gourieroux, Ch., & Jasiak, J. (2002). Nonlinear autocorrelograms: An application to inter-trade durations. *Journal of Time Series Analysis*, 23, 127-154.
- Guerrero, V. M. (1993). Time-series analysis supported by power transformations, *Journal of Forecasting*, 12, 37-48.

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

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics*, 1, 5th ed. London: Charles Griffin & Co. Ltd.

Vangel, M. G. (1996). Confidence intervals for a Normal coefficient of variation. *The American Statistician*, 15, 21-26.

Young, P. C. (2000). Stochastic, dynamic modelling and signal processing: Time variable and state dependent parameter estimation. In *Nonstationary and nonlinear signal processing*, ed. W. J. Fitzgerald, A. Walden, R. Smith, & P. C. Young, pp. 74-114. Cambridge: University Press.

Appendix

1. Approximate variances and covariances of functions of random variables

1) *Theorem.* Let $\mathbf{X} = (X_1, \dots, X_k)$ be a k -dimensional random vector, $g(\mathbf{X})$ be a real-valued function defined on $\mathbf{R}^k$ and $E(X_i) < \infty$ for $i = 1, \dots, k$. Assume that the partial derivatives $g'_i(\mathbf{X}) = \partial g(\mathbf{X}) / \partial X_i$ all exist and let $g'_i[E(\mathbf{X})]$ denote $g'_i(\mathbf{X})$ evaluated at $E(\mathbf{X})$. Then, the first-order Taylor expansion $g(\mathbf{X}) \approx g[E(\mathbf{X})] + \sum_{i=1}^k g'_i[E(\mathbf{X})][X_i - E(X_i)]$, so that $E[g(\mathbf{X})] \approx g[E(\mathbf{X})]$ and, if not all the $g'_i[E(\mathbf{X})]$ are zero, $\text{var}[g(\mathbf{X})] \approx \sum_{i=1}^k [g'_i[E(\mathbf{X})]]^2 \text{var}(X_i) + \sum_{i \neq j}^k \sum_{j=1}^k g'_i[E(\mathbf{X})]g'_j[E(\mathbf{X})]\text{Cov}(X_i, X_j)$

Similarly, for two functions $g_1(\mathbf{X})$ and $g_2(\mathbf{X})$, $\text{cov}[g_1(\mathbf{X}), g_2(\mathbf{X})] \approx \sum_{i=1}^k g'_{1i}[E(\mathbf{X})]g'_{2i}[E(\mathbf{X})]\text{var}(X_i) + \sum_{i \neq j}^k \sum_{j=1}^k g'_{1i}[E(\mathbf{X})]g'_{2j}[E(\mathbf{X})]\text{Cov}(X_i, X_j)$.

Proof. This result was established by Stuart and Ord (1987, Ch. 10).

2. Expected values, variances and covariance of m and se .

It is known that $E[W_h(\lambda)] = \mu$, $\text{var}[W_h(\lambda)] = \sigma^2$ and $\text{corr}[W_h(\lambda), W_{h'}(\lambda)] = \rho$ if $h' = h \pm 1$ and zero otherwise. Then,

$$\begin{aligned} 1) E(m) &= \mu, \\ 2) \text{var}(m) &= E \left\{ H^{-2} \sum_{h=1}^H \sum_{\ell=1}^H [W_h(\lambda) - \mu] [W_\ell(\lambda) - \mu] \right\} \\ &= H^{-2} \left(\sum_{h=1}^H E[W_h(\lambda) - \mu]^2 + 2 \sum_{h=1}^{H-1} E\{[W_h(\lambda) - \mu][W_{h+1}(\lambda) - \mu]\} \right) = \sigma^2 [1 + 2\rho(H-1)/H] / H, \\ 3) E[(H-1)se^2] &= E \left\{ \sum_{h=1}^H E[W_h(\lambda) - \mu]^2 - H(m - \mu)^2 \right\} \\ &= H\sigma^2 - H\sigma^2 [1 + 2\rho(H-1)/H] / H = (H-1)\sigma^2 (1 - 2\rho/H). \end{aligned}$$

Under Normal theory, with $\rho = 0$, the distribution of $(H-1)se^2 / \sigma^2$ is Chi-square with $H-1$ degrees of freedom. Since ρ cannot be far away from zero, it follows that $(H-1)se^2 / \sigma^2$ must have a distribution close to a χ_{H-1}^2 . The variance of such a distribution is derived by assuming approximately valid the following relationship that holds for a Chi-square distribution: *Variance* = 2 *Mean*, therefore

$\text{var}\left[(H-1)\text{se}^2/\sigma^2\right] \approx 2(H-1)(1-2\rho/H)$. From the Theorem in Appendix 1 with $k = 1$, $X = \text{se}^2$ and $g(X) = X^{1/2}$, $\text{var}(\text{se}) \approx \left\{g'\left[E(\text{se}^2)\right]\right\}^2 \text{var}(\text{se}^2) \approx \left[\frac{1}{2}E^{-1/2}(\text{se}^2)\right]^2 2\sigma^4(1-2\rho/H)/(H-1) = \frac{1}{2}\sigma^2/(H-1)$, hence,

$$4) E^2(\text{se}) = E(\text{se}^2) - \text{var}(\text{se}) \approx \sigma^2\{1-2\rho/H-1/[2(H-1)]\}.$$

Next, the Theorem in Appendix 1 applied with $k=2$, $\mathbf{X} = (\text{se}^2, m)$, $g_{11}(\mathbf{X}) = \text{se}$, $g_{12}(\mathbf{X}) = 0 = g_{21}(\mathbf{X})$ and $g_{22}(\mathbf{X}) = m$, allows

$$\text{cov}(\text{se}, m) \approx g'_{11}\left[E(\text{se}^2, m)\right]g'_{22}\left[E(\text{se}^2, m)\right]\text{cov}(\text{se}^2, m) = \frac{1}{2}E^{-1/2}(\text{se}^2)\text{cov}(\text{se}^2, m). \text{ Now,}$$

$$\begin{aligned} \text{Cov}(\text{se}^2, m) &= E\left[\frac{1}{H-1} \sum_{h=1}^H \{W_h(\lambda) - \mu\} - (m - \mu)\}^2 (m - \mu)\right] \\ &= \frac{1}{H-1} E\left\{\sum_{h=1}^H [W_h(\lambda) - \mu]^3 + \sum_{h \neq h'}^H \sum_{h'=1}^H [W_h(\lambda) - \mu]^2 [W_{h'}(\lambda) - \mu]\right\} - \frac{H}{H-1} E(m - \mu)^3 \\ &= \frac{1}{H-1} \left\{\sum_{h=1}^H E[W_h(\lambda) - \mu]^3 - HE(m - \mu)^3\right. \\ &\quad \left.+ \sum_{h=1}^H E(a_h^2 a_{h+1} - \theta a_{h-1} a_h a_{h+1} + \theta^2 a_{h-1}^2 a_{h+1} - \theta a_h^3 + \theta a_{h-1}^2 a_h^2 - \theta^3 a_{h-1}^2 a_h)\right\} \end{aligned}$$

then the normality assumption implies that the third central moments of a_h , $W_h(\lambda)$ and m are all zero. It follows that $\text{cov}(\text{se}^2, m) = 0$ and **5)** $\text{cov}(\text{se}, m) \approx 0$.

Size and Power of the RESET Test as Applied to Systems of Equations: A Bootstrap Approach

Ghazi Shukur

Departments of Economics and Statistics
Jönköping and Växjö Universities, Sweden

Panagiotis Mantalos

Department of Statistics
Lund University, Sweden

The size and power of various generalization of the RESET test for functional misspecification are investigated, using the “Bootstrap critical values”, in systems ranging from one to ten equations. The properties of 8 versions of the test are studied using Monte Carlo methods. The results are then compared with another study of Shukur and Edgerton (2002), in which they used the asymptotic critical values instead and found that in general only one version of the tests works well regarding size properties. In our study, when applying the bootstrap critical values, we find that all the tests exhibits correct size even in large systems. The power of the test is low, however, when the number of equations grows and the correlation between the omitted variables and the RESET proxies is small.

Key words: RESET, Systems of Equations, Bootstrap

Introduction

The RESET test proposed by Ramsey (1969) is a general misspecification test, which is designed to detect both omitted variables and inappropriate functional form. The RESET test is based on the Lagrange Multiplier principle and usually performed using the critical values of the F -distribution. While most authors (e.g., Ramsey and Gilbert (1972); Thursby and Schmidt (1977)) have studied the properties of the RESET tests in single equation situations, Shukur and Edgerton (2002), in what follows referred to as SE, examine the small sample properties of various generalization of the RESET test in an environment of equation systems.

The latter authors used Monte Carlo methods to study the properties of eight different versions of the RESET test in systems ranging from one to ten equations. By using the critical

values of the F -distribution, the authors find that the Rao's F -test exhibits the best performance as regards correct size, while, by using the critical values of the χ^2 -distribution, they find that the commonly used LRT (uncorrected for degrees-of-freedom), and LM and Wald tests (both corrected and uncorrected) behave badly even in a single equation situation. SE also find that the power of the test decreases when the number of equations grows and the correlations between the omitted variables and the RESET proxies are small.

Note that by using the critical values of the χ^2 -distribution, the LRT, LM and Wald tests are strictly valid only asymptotically. Therefore, making inferences on the basis of them can be a risky undertaking. Some authors, e.g., Kivit (1986), have used Monte Carlo methods to compare different LM, Wald and LR alternatives for single equation models. When testing for autocorrelation they have shown that the standard F -test, which is also only valid asymptotically, is in general more accurate as regards size properties.

However, an effective misspecification test should have correct significance levels under the null hypothesis, irrespective of the values of the regression parameters and other distributional parameters. It should also have

Ghazi Shukur is Professor of statistics and econometrics at two universities in Sweden (Jönköping University and Växjö University). Email: ghazi.shukur@ehv.vxu.se. Panagiotis Mantalos is an Associate Professor in statistics at Lund University, Sweden. Email address: panagiotis.mantalos@stat.lu.se.

reasonable power against the class of alternative specifications under investigation, but low power against other alternatives.

The purpose of this article is to improve the critical values of the test statistics by employing bootstrap technique, so that the size of the test approaches its nominal value. Horowitz (1994) and Mantalos and Shukur (1998) recommended this approach. Given the bootstrap critical values, analyzed here is the size and power of a different generalization of the systemwise RESET test, followed by a comparison with results found by SE.

Model Specification

The regression model investigated is the same model as in SE and consists of n linear stochastic equations given by

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{B} + \varepsilon_t, \quad (1)$$

where $\mathbf{Y}_t$ and ε_t are $(1 \times n)$ vectors of endogenous variables and disturbances respectively, $\mathbf{X}_t$ is a $(1 \times m)$ matrix of exogenous variables, $\mathbf{B}$ is a $(m \times n)$ matrix of parameters, and $t = 1, \dots, T$. The data matrices $\mathbf{Y}$ and $\mathbf{X}$ are $(T \times n)$ and $(T \times m)$ respectively. The null hypothesis of correct specification implies that the error terms will be independently and identically distributed conditional on the exogenous variables, and in many cases a normal distribution is also assumed,

$$\varepsilon_t | \mathbf{X}_t \sim N(0, \Sigma_{\varepsilon}). \quad (2)$$

The hypothesis of correct functional form is equivalent to assuming that the disturbances have zero conditional mean, $H_0: E(\varepsilon_t | \mathbf{X}_t) = 0$.

The class of alternative hypotheses to this null hypothesis is very general; omitted variables and incorrect functional form will obviously be members of the class, but so will endogeneity of the $\mathbf{X}$ variables.

The alternative hypothesis is specified through the following model:

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{B} + \mathbf{Z}_t \Gamma + \varepsilon_t. \quad (3)$$

$\mathbf{Z}$ is in general unknown, and the tests that we will investigate use a proxy $\hat{\mathbf{Z}}$. The following regression is estimated instead of (3),

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{B} + \hat{\mathbf{Z}}_t \Gamma^* + \delta_t. \quad (4)$$

If the null hypothesis is correct, then $\Gamma = \Gamma^* = 0$ whatever the choice of $\hat{\mathbf{Z}}$. If the hypothesis is incorrect, then the choice of $\hat{\mathbf{Z}}$ will obviously affect the power of any test based on (4). The greater the correlation between $\hat{\mathbf{Z}}$ and the non-linear part of the true conditional mean of $\mathbf{Y}$, then, in general, the greater the power will be. If we suspect certain variables to have been omitted, then using these variables as $\hat{\mathbf{Z}}$ will obviously be most appropriate.

Ramsey (1969) proposed approximating the unknown conditional expectation of $\mathbf{Y}$ by using a Taylor expansion around the conditional expectation *under the null hypothesis*, that is $\mathbf{X}\beta$ (Ramsey considered a single equation, and β was thus a vector). Because the parameters are unknown, this was in turn approximated using

$\hat{\mathbf{Y}} = \mathbf{X} \hat{\mathbf{B}}$, where $\hat{\mathbf{B}}$ was the OLS parameter estimate from the single equation version of (1). This is the RESET test procedure.

Define a systemwise version of the RESET test. Following common terminology of double regression tests, refer to equation (1) as the *primary regression*. The first stage of the RESET test is performed by calculating the least squares' predictions from the primary regression, i.e., $\hat{\mathbf{Y}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}$. These predictions are then used in the following *auxiliary regression*,

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{B} + \hat{\mathbf{Y}}_t^2 \Gamma_1^* + \hat{\mathbf{Y}}_t^3 \Gamma_2^* + \dots + \hat{\mathbf{Y}}_t^{G+1} \Gamma_G^* + \delta_t, \quad (5)$$

where the (t, i) :th elements of the power matrices are given by $[\hat{\mathbf{Y}}^j]_{ti} = \hat{y}_{ti}^j$. The RESET test is now performed by testing the hypothesis $\Gamma_1^* = \dots = \Gamma_G^* = 0$.

The practical implementation of the RESET test now depends on two factors. Firstly

it must be decided how many power matrices to include in the auxiliary regression (i.e., determine G). Secondly, it must be decided which test method to use. We concentrate on the second question, and set $G = 1$ throughout.

Denote by $\hat{\delta}_U$ the $(T \times n)$ matrix of estimated residuals from the *unrestricted* regression (5), and by $\hat{\delta}_R$ the equivalent matrix of residuals from the *restricted* regression with H'_0 imposed. The matrix of cross-products of these residuals will be defined as $\mathbf{S}_U = \hat{\delta}_U' \hat{\delta}_U$ and $\mathbf{S}_R = \hat{\delta}_R' \hat{\delta}_R$ respectively. Bewley (1986, Chapter 4) showed that the Wald, Likelihood Ratio and Lagrange Multiplier test statistics are given by

$$W = T(\text{tr } \mathbf{S}_U^{-1} \mathbf{S}_R - n), \quad (6)$$

$$LR = T \ln U, \text{ and} \quad (7)$$

$$LM = T(n - \text{tr } \mathbf{S}_R^{-1} \mathbf{S}_U), \quad (8)$$

where $U = \det \mathbf{S}_R / \det \mathbf{S}_U$. The above statistics are all asymptotically $\chi^2(p)$ distributed under the null hypothesis, where $p = Gn^2$ is the number of restrictions imposed by the null hypothesis. It is well known, however, that this asymptotic result becomes less and less accurate in small samples as the number of equations grows, see for example Laitinen (1978). A simple small sample correction is to replace T by $\Delta = T - (m + Gn)$, the degrees of freedom in the equations of the auxiliary regression (4). The *corrected* statistics are thus given by $WC = (\Delta/T)W$, $LRC = (\Delta/T)LR$ and $LMC = (\Delta/T)LM$, which have the same asymptotic distribution as given above.

Another more sophisticated approximation is that given by theorem 8.6.2 in Anderson (1958, p. 321). This uses an Edgeworth expansion, and if we choose the simplest form (which is accurate to the order T^{-2}) this corrected LR statistic is given by

$$LRE = \Delta_E \ln U, \quad (9)$$

where $\Delta_E = \Delta + \frac{1}{2}[n(G-1) - 1]$. Note that when $G = 1$, the difference between LRC and LRE is merely that the numerator in the correction is Δ in the first case and $\Delta - \frac{1}{2}$ in the second.

A final approximation is that given by Rao (1973, p. 556), namely

$$RAO = (q/p)(U^{1/s} - 1), \quad (10)$$

where p and Δ_E are defined above, $r = p/2 - 1$, $q = \Delta_E s - r$, and

$$s = \sqrt{\frac{p^2 - 4}{n^2(G^2 + 1) - 5}}. \quad (11)$$

RAO is approximately distributed as $F(p, q)$ under the null hypothesis, and reduces to the standard F statistic when $n = 1$.

Factors that Affect the Small Sample Properties of the RESET Test

A number of factors obviously can affect the size of the RESET tests, SE have investigated these factors systematically, and we therefore follow their line of investigation. The number of equations (n), the sample size (T), degrees of freedom (Δ) and the order of the restrictions (G) are four such factors. The power of the tests will also be affected by the size and form of $Z_t \Gamma$ in (3). In this paper we will also study the consequences of varying n and Δ , while T is chosen so as to give compatible values of Δ for different models ($T = \Delta + m + Gn$). We will also mainly concentrate on the case where $G = 1$.

A number of other factors can also affect the properties of the RESET tests, for example the distributions of $\mathbf{X}_t$, and $\mathbf{\varepsilon}_t$, and the values of $\mathbf{B}$. In the rest of this section we will consider these factors in some more detail. In this paper, we consider only stochastic exogenous variables $\mathbf{X}_t$ and although SE find that serial dependence in x has no practicable

effect on either the size or power of the RESET tests, we will allow autocorrelation in the exogenous variables in our study. The following generating processes are used,

$$x_{tj} = \alpha x_{t-1,j} + v_{tj}, \quad j = 1, \dots, m-1 \\ \text{and } t = 1, \dots, T \quad (12)$$

where and v_t is a multivariate normal white noise process with covariance matrix Σ_v . In our Monte Carlo study we have included a constant term among the exogenous variables, so that (12) has only been applied to the remaining $m - 1$ variables.

The power (but not the size) of the tests will also be affected by Z_t in (3). Intuitively, the power of the test ought to increase with an increase in the omitted portion of the regression. That is to say, an increase in the absolute value of Γ should imply an increase in the seriousness of the misspecification caused by using (1) instead of (3). Accordingly, we would expect the power of the RESET test to increase with Γ . The problem is to decide how large a value of Γ is needed to constitute "serious" misspecification.

SE found that a good measure of misspecification is given by the relative increase of goodness-of-fit, achieved by going from the incorrect model under the null (1) to the correct model under the alternative (3), i.e.,

$$R_D^2 = \frac{R_1^2 - R_0^2}{1 - R_0^2} \quad (13)$$

where R_0^2 and R_1^2 are the theoretical R^2 measures from the null and alternative models respectively. The reasoning behind this choice of misspecification measure, and the relationships that exist between goodness-of-fit and the other parameters of the model, are explained in the Appendix of their paper. An advantage of using R_D^2 as a measure of misspecification is that it is bounded between zero (no misspecification) and one (a perfect alternative).

The power of the test will also depend on the joint distribution of the included and omitted variables. If this distribution is joint

normal, then the regression of the omitted variables on the included variables is exactly linear, and no loss of fit will occur through the exclusion of the omitted variables.

The RESET test will have zero power in such circumstances, even though the parameter estimates will be biased, unless the omitted variable is also uncorrelated with the included variables. If the omitted variables are non-normal, then their conditional means can be non-linear in the included variables, and the RESET test can have power. The strength of the power, however, might depend on the correlation between the omitted variable (Z) and the proxy

variables ($\hat{Y}^j$) used in the auxiliary regression (4). In this paper, and as in SE, we concentrate on an omitted variable which is the square of one of the (normally distributed) included variables.

Bootstrap-hypothesis testing, critical values.

Two aspects are of primary importance when the properties of a test procedure are investigated. Firstly, determine if the *actual size* of the test (i.e., the probability of rejecting the null when true) is close to the *nominal size* (used to calculate the critical values). Given that actual size is a reasonable approximation to the nominal size, then investigate the *actual power* of the test (i.e., the probability of rejecting the null when false) for a number of different alternative hypotheses. When comparing different tests, therefore, those in which (a) actual size lies close to the nominal size and, given that (a) holds, (b) have greatest power are preferred. In most cases, however, the distributions of the test statistics used are known only asymptotically.

As a result, the tests do not have the correct size and inferential comparisons and judgments based on them might be misleading. However, by using bootstrap technique it is possible to improve the critical values so that the true size of the test approaches its nominal value.

In the regression model (1), the null hypothesis of correct specification implies that the error term ε_t will be independently and identically distributed, conditional on the exogenous variables. The most convenient way

to apply bootstrap, here, is to resample the ε_t . Since the errors are not observable and the usual solution is then to use the calculated Least Squares (OLS) residuals instead. A direct residual resampling gives:

$$\mathbf{Y}_t^* = \mathbf{X}_t \hat{\mathbf{B}}_{ols} + \varepsilon_t^*, \quad (1a)$$

where ε_t^* are i.i.d observations $\varepsilon_1^*, \dots, \varepsilon_T^*$, drawn from the empirical distribution ($\hat{F}_\varepsilon$) of the LS residuals. This method is called the *bootstrap based on residuals*, abbreviated as RB, proposed by Efron (1979). Note that, in what follows, all bootstrap statistics will marked by an asterisk (*). An important assumption for the RB is that ε_t are i.i.d, but even if this assumption holds, the empirical distribution $\hat{F}_\varepsilon$ is not based on exactly i.i.d data, namely observed residuals $\hat{\varepsilon}_t$. Therefore the following adjustments are necessary.

First, subtract the sample mean of the OLS residuals from the residuals: ($\hat{\varepsilon}_i - \bar{\varepsilon}$)

$$\text{where } \bar{\varepsilon} = T^{-1} \sum_{i=1}^T \hat{\varepsilon}_i \quad i = 1, \dots, T.$$

Thus, $E_*(\varepsilon_t^*) = 0$ for all t . And

$$E_*(\hat{\mathbf{B}}_{OLS}^*) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' E_*(\mathbf{Y}^*) = \hat{\mathbf{B}}_{OLS}, \quad \text{where}$$

$$\hat{\mathbf{B}}_{OLS}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}^*, \text{ and}$$

$$\text{Var}_* \hat{\mathbf{B}}_{OLS}^* = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1},$$

where

$$\hat{\sigma}^2 = \text{Var}_*(\varepsilon_t^*) =$$

$$T^{-1} \sum_{i=1}^T (\hat{\varepsilon}_i - \bar{\varepsilon})^2, \quad i = 1, 2, \dots, T.$$

This bootstrap procedure produces consistent variance but is downward biased (Efron, 1982). To remove this negative bias, Efron (1982) suggested the bootstrap data to be drawn from the empirical distribution $\hat{F}_\varepsilon$ putting mass $1/T$ to the adjusted OLS residuals

$(\hat{\varepsilon}_i - \bar{\varepsilon}) / \sqrt{1 - [m/T]}$, $i = 1, \dots, T$. This is called *the adjusted residual resampling ARR*.

The basic principle is to draw a number of bootstrap samples from the model under the null hypothesis. The bootstrap test statistic (T_s^*) can then be calculated by repeating this step N_b number of times. Then, take the $(1-\alpha)$:th quintile of the bootstrap distribution of T_s^* and get the α - level "bootstrap critical values" ($c_{1\alpha}^*$). Generally, the bootstrap procedure is summarized by the following steps:

(1) Estimate the test statistic as previously described, which is called (T_s).

(2) Use *the adjusted residual resampling ARR*, $(\hat{\varepsilon}_i - \bar{\varepsilon}) / \sqrt{1 - [m/T]}$ $i = 1, \dots, T$ to draw i.i.d. data $\varepsilon_1^*, \dots, \varepsilon_T^*$ and define: $\mathbf{Y}_t^* = \mathbf{X}_t \hat{\mathbf{B}}_{ols} + \varepsilon_t^*$.

Then, calculate the test statistic (T_s^*) as described, i.e., by applying the RESET test procedure to the (1a) model. Repeating this step N_b number of times and taking the $(1-\alpha)$:th quintile of the bootstrap distribution of T_s^* , we obtain the α - level "bootstrap critical values" ($c_{1\alpha}^*$), and finally, we then reject H_0 if $T_s \geq c_{1\alpha}^*$. This is our bootstrap test approach to investigate the size and power of the various generalization of the systemwise RESET test.

Monte Carlo Experiment

In a Monte Carlo study, the estimated size is estimated by simply observing how many times the null is rejected in repeated samples under conditions where the null is true. By varying factors such as described in the previous section, a succession of estimated sizes under different conditions is obtained. In general, the closer an estimated size is to the nominal size the better the test. Most of the factors discussed earlier either have very small effect, or have no effect at all on the estimated size of the tests. To show the effect of the remaining factors on the performances of the tests, the estimated sizes of the tests are displayed in the tables.

As regards the *estimated power functions* of the tests, these have mainly been compared graphically. This has proved to be quite adequate, since those tests that give reasonable results as regard size usually differed very little regarding power.

The Monte Carlo experiment was performed by generating data according to (1), (2) and (12), estimating the auxiliary regression (5) and then calculating the test statistics, T_s , defined above.

Because the number of regressors in the auxiliary regression (5) is $(m + n)$, we draw i.i.d. data $\varepsilon_1^*, \dots, \varepsilon_T^*$ from the empirical distribution $\hat{F}_\varepsilon$ putting mass $1/t$ to the adjusted (LS) residuals $(\hat{\varepsilon}_i - \bar{\varepsilon}) / \sqrt{1 - [(m + n) / T]}$, $i = 1, \dots, T$.

The bootstrap procedure described in the previous section is followed to obtain the α - level "bootstrap critical values" ($c_{t\alpha}^*$). The $\alpha = 0.05$ level, for example, is the $T_{sN_b, 96}^*$ of the order test statistic: $T_{sN_b, 1}^* \leq T_{sN_b, 2}^* \leq \dots \leq T_{sN_b, 100}^*$.

A final consideration is the significance levels to be used when judging the properties of the tests. Theoretically, it is possible to construct the empirical distributions of the test statistics, and to compare these with the theoretical asymptotic results. In this study, the tests of the null hypothesis were carried out using nominal significance levels (π_0) of 1%, 5%, 10% and 20%. Hence, for the 1%, 5%, 10%, and 20% levels, the "bootstrap critical values" $c_{t\alpha}^* = T_{sN_b, 99}^*$, $c_{t\alpha}^* = T_{sN_b, 95}^*$, $c_{t\alpha}^* = T_{sN_b, 90}^*$ and $c_{t\alpha}^* = T_{sN_b, 80}^*$ were chosen, respectively. Then, reject H_0 if $T_s \geq c_{t\alpha}^*$.

An approximate 95% confidence interval for the actual size (π) can be given as

$$\hat{\pi} \pm 2\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{N}}, \quad (13)$$

where $\hat{\pi}$ is the estimated size and N is the number of replications.

However, because the main interest is in the behavior of the distributions in the tails, only results using the conventional 5% significance level have been analyzed. A summary of the design can be found in Table 1 and 2, and in Table 3 approximate 95% confidence intervals for the actual size, based on (13) are presented. Letting the number of replications per model is 10,000, which by (13) seems to be sufficient when estimating size. Note again that SE's Monte Carlo design is followed, and a summary of the relationships between the various factors can be found in their article (in their Appendix).

Regarding the N_b number of the bootstrap samples used to estimate bootstrap critical value, Horowitz (1994) used the value of $N_b = 100$. However, it follows from Hall (1986) that the error in the size of a test using the "bootstrap critical values" is independent on the number of the bootstrap sample used to estimate $c_{t\alpha}^*$. $N_b = 500$ in the current study. Increasing the number of the bootstrap samples beyond 500 has little effect on the results of the experiment and takes longer time.

The primary interest lies in the analysis of systemwise tests, and thus the number of equations to be estimated is of central importance. As the number of equations grows, the computation time becomes longer. A system with ten equations was selected as the largest model when studying the size of the tests. This represents a fairly large consumption model of the type that is used in, for example, agricultural economics. Medium size models are represented by five- and seven-equation systems, and two- and three-equation systems are typical of the small models used when separability is imposed.

Another important factor that affects the performance of tests is the number of observations. The number of degrees of freedom, Δ , was held constant between models of different sizes, because this allows a fair comparison. If the number of observations, T , were held constant then tests in models with a large number of equations would automatically perform more poorly, simply due to the reduced degrees of freedom (a new predictor is included for each equation in the system).

Table 1. Values of Factors Held Constant for Different Models

Factor	Symbol	Design
Constant term		1 (size) or 0 (power)
Number of X variables (excl constant)	$n + 1$	number of equations + 1
Mean of X variables	μ_x	0
Parameters of X variables	B	E
Distribution of X variables		Normal
Covariance of X variables	Σ_x	$(1-\rho_x)\mathbf{I} + \rho_x\mathbf{E}$
Properties of X in repeated samples		Stochastic
Distribution of error terms		Normal
Covariance of error terms	Σ_ε	$\sigma^2\mathbf{I}$

Table 2a. Values of Factors that Vary for Different Models - Size Calculations

Factor	Symbol	Design			
Number of equations	n	1	2	3, 5, 7	10
Degrees of freedom	Δ	15, 25, 45, 75			
Nominal size	π_0	1%, 5%, 10%, 20%			
Goodness-of-fit in null	R_0^2	.1, .3, .5, .7, .9	.3, .5, .7	.3, .7	.3 .7
AR parameter for X	α	0, .5, .9	0, .5		
Correlation (X_i, X_j)	ρ_x	0, .5, .9	0, .5		

Table 2b. Values of Factors that Vary for Different Models - Power Calculations

Factor	Symbol	Design
Number of equations	n	1, 2, 3, 5, 7, 10
Degrees of freedom	Δ	15, 25, 45, 75
Nominal size	π_0	1%, 5%, 10%, 20%
Goodness-of-fit in null	R_0^2	.3, .5, .7
Relative difference in R^2	R_D^2	0, .1, .2, .3, .4, .5, .6, .7, .8, .9
AR parameter for X	α	0, .5
Correlation (η, z)	$\rho_{\eta z}$	.1, .3, .5, .7, .9

z is the omitted variable (the square of x_1) and η is the square of the conditional expected value of y .

Table 3. Approximate 95% Confidence Intervals for Actual Size

$\pi_0 \cdot N$	2000	10000
1%	± 0.44	± 0.20
5%	± 0.97	± 0.44
10%	± 1.34	± 0.60
20%	± 1.79	± 0.80

We have investigated samples typical for annual and quarterly consumption models, using degrees of freedom 15, 25, 45 and 75. This is equivalent to sample sizes of between 20 and 110 observations.

Various values of R_0^2 were chosen to represent different explanatory powers under the null with a greater variation in small models. The distribution of the exogenous variables was varied to account for a typical property of economic time series, i.e., that they are trended and/or autocorrelated. SE find that trending had no effect at all on the RESET tests, and it is therefore not considered here. The calculations were performed using GAUSS 3.2, and the results from different models were analysed using MSeExcel 4.0.

When calculating the power functions of the tests we used different values of R_D^2 to indicate different degrees of misspecification in the model. Different values of ρ_{nz} were used to illustrate different strengths in the relationship between the omitted variable and the proxy variable used in the auxiliary regression.

Analysis of the Size of the RESET Tests.

In this section, results are presented of the Monte Carlo experiment concerning the size of the RESET tests. When using the “bootstrap critical values”, our primary results reveal that the LM and Wald tests get results identical to their corrected correspondents (i.e., LMC, and WC).

All the LR tests (including the RAO) lead to identical results. Moreover, for a single equation, we find that all the eight test methods yield the same results. Noticeable effects on the estimated size were not found, however, by varying the number of equations, degrees of freedom, autocorrelation in the exogenous variables, the collinearity between the exogenous variables, or the goodness-of-fit under the null hypothesis. These results agree with the results obtained by SE regarding the Rao test only.

The results from the two articles are now compared to show the differences between our findings. Our results are shown in Table 4, where the same goodness-of-fit ($R_0^2 = 0.7$) was

used, multicollinearity ($\rho_X = 0.5$), and autocorrelation ($\alpha = 0.0$) in X as in Table SE 4. Note that changing the factors we have held constant in these tables (i.e., goodness-of-fit, multicollinearity and autocorrelation in X) would not change the conclusions in any way. Some important results regarding the different variants of the RESET test are presented in Table SE 4. They found that the number of equations in the system (n) and the degrees of freedom (Δ) have noticeable effect on the performances of the tests.

They also found that the RAO test was superior to all the other alternatives, with only one result (out of 30) lying slightly outside the 95% confidence interval, whereas the WALD and LRT tests performed extremely poorly.

When we use the “bootstrap critical values”, the results show that all tests perform well, i.e. the superiority of the Rao test to the other tests disappears. The WALD/Wald-C tests perform slightly badly in small samples and large systems. The Rao/LR and LM tests are shown to perform satisfactorily in all situations. Note that in our study, i.e. when we use the “bootstrap critical values”, all the tests have identical results for single equation models.

Analysis of the Power of the RESET tests

In this section, the most interesting results of our Monte Carlo experiment regarding the power of the various versions of the RESET test are discussed. The power of different versions of the RESET test was analyzed, using the “bootstrap critical values”, in systems ranging from one to ten equations. The power function was estimated by calculating the rejection frequencies in 2,000 replications using different values of the relative differences in goodness-of-fit, R_D^2 .

Even if a correctly given size is not sufficient to ensure the good performance of a test, it is a prerequisite. SE only present power results for the Rao test, since this test is shown to be superior in all situations. In our study, regarding the size, all tests perform well even in large systems of equations. To compare how the different test methods perform, consider the following power results:

Table 4. Estimated Size for the Alternative RESET Tests at 5% Nominal Size.

Δ	No. of Equations (n)					
	RAO = LRE = LRT = LRT-C					
	1	2	3	5	7	10
15	0.049	0.050	0.051	0.054	0.051	0.046
25	0.050	0.051	0.050	0.054	0.046	0.049
45	0.050	0.050	0.053	0.054	0.051	0.048
75	0.054	0.049	0.053	0.050	0.052	0.046

Wald = Wald-C						
Δ	1	2	3	5	7	10
15	0.049	0.049	0.049	0.054	0.048	0.044
25	0.050	0.054	0.050	0.053	0.046	0.048
45	0.050	0.050	0.053	0.052	0.050	0.048
75	0.054	0.049	0.053	0.050	0.052	0.047

LM = LM-C						
Δ	1	2	3	5	7	10
15	0.049	0.049	0.051	0.054	0.051	0.051
25	0.050	0.053	0.051	0.054	0.048	0.050
45	0.050	0.051	0.052	0.053	0.052	0.049
75	0.054	0.049	0.053	0.050	0.053	0.048

In this table $R_0^2 = 0.7$, $\rho_X = 0.5$ and $\alpha = 0.0$. The shading indicates bad performance as defined earlier in Table 3, i.e., when the results lie outside the approximate 95% confidence interval for actual size.

Table SE 4. Estimated Size for the Alternative RESET Tests at 5% Nominal Size.

	No. of Equations (n)						No. of Equations (n)					
	RAO						LRE					
Δ	1	2	3	5	7	10	1	2	3	5	7	10
15	.047	.048	.047	.050	.049	.058	.047	.048	.048	.062	.078	.182
25	.049	.047	.053	.048	.051	.048	.049	.047	.054	.051	.060	.078
45	.051	.051	.052	.049	.049	.048	.051	.051	.053	.050	.053	.055
75	.049	.050	.050	.054	.053	.054	.049	.050	.050	.054	.054	.056

LRT-C							Wald-C					
Δ	1	2	3	5	7	10	1	2	3	5	7	10
15	.051	.056	.058	.082	.110	.249	.069	.120	.193	.504	.841	.998
25	.052	.051	.060	.061	.075	.103	.062	.085	.132	.279	.559	.921
45	.052	.053	.057	.054	.059	.065	.057	.074	.096	.162	.291	.602
75	.050	.051	.052	.058	.058	.062	.054	.062	.074	.109	.167	.339

LRT							Wald					
Δ	1	2	3	5	7	10	1	2	3	5	7	10
15	.086	.164	.298	.756	.985	1.00	.101	.238	.457	.925	.999	1.00
25	.072	.107	.186	.468	.842	.999	.081	.150	.293	.708	.972	1.00
45	.062	.083	.116	.254	.500	.906	.067	.102	.165	.410	.760	.993
75	.058	.069	.086	.150	.284	.627	.060	.079	.113	.234	.469	.872

LM							LM-C					
Δ	1	2	3	5	7	10	1	2	3	5	7	10
15	.071	.087	.112	.285	.608	.970	.037	.012	.003	0.00	0.00	0.00
25	.063	.069	.090	.162	.353	.763	.042	.025	.011	.001	0.00	0.00
45	.058	.062	.073	.105	.187	.419	.047	.036	.026	.008	.001	0.00
75	.054	.058	.062	.083	.118	.241	.048	.042	.035	.021	.009	.002

Source : Shukur & Edgerton (2002, Table 4). In this table, $R_0^2 = 0.7$, $\rho_x = 0.5$ and $\alpha = 0.0$. The shading indicates bad performance, i.e., when the results lie outside the approximate 95% confidence interval for actual size.

The primary results reveal that, for single equation, the power functions for all the tests methods are identical. Moreover, in systems with more than one equation, we find that all the LR tests (uncorrected and corrected including the Rao's F -test) have identical results, and that the corrected and uncorrected Wald have identical results, and the same for the LM and corrected LM tests. This means that in single equation, the eight tests reduces to one and that we can present results from any one of them. In systems with more than one equation, the results differ between the three test groups (Wald, LR & LM).

The factors that affect the power of the RESET tests differ from those that affect the size. Although the number of equations (n), and degrees of freedom (Δ) had only a slight effect on the estimated size, they have a considerable effect on the power. As in the case of the size, changes in the autocorrelation between the exogenous variables (α), and the goodness-of-fit in the null (R_0^2) did not produce any noticeable effects on the power function of the tests, and will not be shown in the diagrams.

The power of the RESET test did, as expected, depend on the degree of misspecification (R_D^2) and the correlation between the proxy in the auxiliary regression and the omitted variable ($\rho_{\eta\epsilon}$). The greater the misspecification, and the better the RESET proxy mirrors the omitted variable, the greater the power of the tests.

In Figure 1, the power functions of the three test methods are shown at a nominal size of 5% for different degrees of freedom (Δ) and for systems with different numbers of equations (n). The autocorrelation in the exogenous variables ($\alpha = 0$) is fixed, the goodness-of-fit in the null ($R_0^2 = 0.7$) and the correlation between the included and omitted variables ($\rho_{\eta\epsilon} = 0.5$). The power functions have also been calculated at other values, but because the patterns obtained are essentially the same they are excluded to save space.

It can be seen from the diagrams in Figure 1 that the power functions satisfy the expected properties of increasing with Δ and R_D^2

(which is denoted $Rdif$ in the figure). The rate at which the power approaches one is heavily dependent on the values of Δ and n , however. It is quite clear that the Wald tests exhibits the best power among the others, especially in large samples (when $n = 10$). The LR tests (or the Rao test) is next best, while the LM test comes in third place. Note that in SE only results for the Rao test have been presented, which are very similar to our results for the LR tests groups, which we refer to as "Rao" in what follows.

A closer examination of the diagrams shows that in small samples the power functions decrease as n increases, while in large samples, i.e., when $\Delta = 75$, it can be seen that the power functions increase as n increases. The reason for this is that when n increases, the number of proxy variables that are included in the auxiliary regression also increases. Because each of these proxies is correlated with the omitted variable, their combined effect will tend to be greater when n increases (to hold this effect under control, the *multiple* correlation between the omitted variable and all of the proxy variables would have to be held constant) will obviously influence the power functions. Note also how, in small samples, the power functions become flatter as the number of equations increases, i.e., the tests become worse and worse, in particular the LM test. For large values of n and low degrees of freedom there is, little difference between the estimated size and estimated power.

Because SE only focus on the Rao test, and to facilitate comparison between the two papers, we will also present results for the Rao test. In Figure 2, the effect is shown of different values of $\rho_{\eta\epsilon}$ (rho in the figures) on the power function of the *RAO* test with 45 degrees of freedom, for systems with one, three, seven and ten equations. The power functions are shown at a nominal size of 5%, the autocorrelation in the exogenous variables ($\alpha = 0$) are fixed, and the goodness-of-fit in the null ($R_0^2 = 0.7$). The effect of the correlations between the proxies and the omitted variables is noticeable, and plays an important role on how quickly the power reaches the value of one. The effect of this factor is more dramatic in large systems, but again this is in part due to the usage of simple instead of multiple correlations.

Figure 1 : The Power Function of the Wald, Rao and LM Tests for Three and Ten equations, Using the Bootstrap Critical Values.

Wald test, 3 eq's :

Wald test, 10 eq's :

Rao test, 3 eq's :

Rao test, 10 eq's :

LM test, 3 eq's :

LM test, 10 eq's :

Figure 2 : The Power Function of Various Alternatives of the Rao Test with 45 df, Using the Bootstrap Critical Values.

Note also how the power functions become flatter for small $\rho_{\eta\zeta}$ as the number of equations increases. For high values of n and low $\rho_{\eta\zeta}$ there is very little difference between the estimated size and the estimated power. Note that, regarding the Rao test, our results are almost identical to those obtained by SE.

Conclusion

The *size and power* of systemwise generalisations of Ramsey's RESET test was examined for misspecification errors by using "bootstrap critical values" (c_{α}^*). Shukur and Edgerton (2002) (SE) studied the same properties of the test, but they used the asymptotic critical values instead. The purpose of this paper is to show the ability of the bootstrap technique to produce critical values that might be much more accurate than the asymptotic ones.

We followed the same principle as in SE to construct Wald, Lagrange Multiplier and Likelihood Ratio tests that are applicable to auxiliary regression systems. Various degrees-of-freedom corrections have been investigated, in particular the commonly used simple replacement of the number of observations (T) by the degrees-of-freedom (Δ) and, for the LR test, the Edgeworth correction developed by Anderson (1958). We also studied the properties of the systemwise F -test approximation proposed by Rao (1973).

The investigation has been carried out using Monte Carlo simulations. A large number of models were investigated, where the number of equations, degrees of freedom, error variance and stochastic properties of the exogenous variables have been varied. For each model, we performed 10,000 replications and studied four different nominal sizes. The power properties have been investigated using 2,000 replications per model, where in addition to the properties mentioned above the degree of misspecification (measured as the relative difference in the explanatory power between the null and true models) and the correlation between the omitted and included variables have also varied.

The analysis reveals that, in single equations, all test method are identical regarding the estimated size and power, while in systems with many equations the eight tests reduce to three groups, namely Wald, LR (or Rao), and LM. Although SE found that the Rao's F -test is the best and that the uncorrected LR test and both the corrected and uncorrected Wald and LM tests are shown to perform *extremely* badly in all situations, our analysis reveals that, in almost all

cases, the performance of all the tests are satisfactorily.

The factors that affect the power of the RESET tests differ from those that affect the size. While the number of equations (n), and degrees of freedom (Δ) had only a slight effect on the estimated size, they have a considerable effect on the power. As in the case of the size, changes in the autocorrelation between the exogenous variables (α), and the goodness-of-fit in the null (R_0^2) did not produce any noticeable effects on the power function of the tests. The power of the RESET test did, as expected, depend on the degree of misspecification (R_D^2) and the correlation between the proxy in the auxiliary regression and the omitted variable ($\rho_{\eta\pi}$). The greater the misspecification, and the better the RESET proxy mirrors the omitted variable, the greater the power of the tests.

As regards the power, the Wald test has been shown to perform somewhat better than the others especially in small samples and large systems, but the differences between the alternative RESET tests are minimal. The Rao test performs well in our study as well as in that of SE, i.e., when using the asymptotic critical values and the "bootstrap critical values", which reinforces our picture of good performance in both cases. Generally, the power functions become flatter for small $\rho_{\eta\pi}$ as the number of equations increases. For high values of n and low $\rho_{\eta\pi}$ there is indeed very little difference between the estimated size and the estimated power.

References

- Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, (2ed). New York, Wiley.
- Bewley, R. (1986). *Allocation Models: Specification Estimation and Applications*. Cambridge Mass. Ballinger.
- Edgerton, D. L., & Shukur, G. (1999). Testing Autocorrelation in a System Perspective. *Econometric Reviews*, 18, 343-386.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM: Philadel.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, 14, 1453-1462.

Horowitz, J. L. (1994). Bootstrap-based critical values for the information matrix test. *Journal of Econometric*, 61, 395-411.

Kiviet, J. (1986). On the Rigour of Some Specification Test for Modelling Dynamic Relationships. *Review of Economic Studies*, 53, 241-262.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21, 255-285.

Mantalos, Panagiotis, & Shukur, G. (1998). Size and Power of the Error Correction Model (ECM) of Cointegration Tests. A Bootstrap Approach. *Oxford Bulletin of Economics and Statistics*, 60, 249-255.

Ramsey, J. B. (1969). Test for Specification error in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B*, 31, 350-371.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, (2nd ed.). New York: Wiley.

Shukur, G., & Edgerton, D. L. (2002). The Small Sample Properties of the RESET Test as Applied to Systems of Equations. *Journal of Statistical Computation and Simulation*, 72 (12), 909-924, 2002.

Thursby, J. G., & Schmidt, P. (1977): Some Properties of Tests for Specification Error in a Linear Regression Model. *Journal of American Statistical Association*, 72, 635-641.

Interval Estimation For The Scale Parameter Of Burr Type X Distribution Based On Grouped Data

Amjad Al-Nasser Ayman Baklizi
Department of Statistics
Yarmouk University, Jordan

The application of some bootstrap type intervals for the scale parameter of the Burr type X distribution with grouped data is proposed. The general asymptotic confidence interval procedure (Chen & Mi, 2001) is studied. The performance of these intervals is investigated and compared. Some of the bootstrap intervals give better performance for situations of small sample size and heavy censoring.

Key words: Bootstrap, Burr type X distribution, grouped data, interval estimation

Introduction

In many applications, individual observations are very naturally categorized into mutually exclusive and exhaustive groups; such type of data is often called grouped data. Grouped data arise frequently in life testing experiments when inspecting the test units intermittently for failure, this procedure is frequently used because it requires less testing effort than continuous inspection. The data obtained from intermittent inspection consists only of the number of failures in each inspection interval. Other examples of natural occurrences of grouped data are given in Pettitt and Stephens (1977). In this article, a different computer intensive confidence interval is obtained based on grouped data for the scale parameter of Burr Type X Distribution “ $BurrX(\nu, \theta)$ ” whose distribution function is given by

$$F(x, \nu, \theta) = \left[1 - e^{-\left(\frac{x}{\theta}\right)^2} \right]^\nu, \quad x > 0, \theta > 0, \nu > 0 \quad (1)$$

By taking the first derivative on (1), the density function of the BurrX distribution can be obtained as:

$$f(x, \nu, \theta) = \frac{2\nu x}{\theta^2} \left(e^{-\left(\frac{x}{\theta}\right)^2} \right) \left(1 - e^{-\left(\frac{x}{\theta}\right)^2} \right)^{\nu-1}$$

This density function introduced by Burr (1942) providing new model of life time data. The applications of the Burr distribution may be found in the literature for the different twelve types of this distribution. Various authors, considered BurrX in different aspects (e. g., Mudholkar et al., 1995; Mudholkar & Hutson, 1996; Jaheen, 1995, 1996; Surles & Padgett, 1998, 2001; Ahmad et al., 1997). This distribution is a generalized Rayleigh distribution and also it is considered as a special case of exponentiated Weibull distribution that introduced by Mudholkar and Sirvastava (1993). The shape of this distribution depends on the parameter ν , by increasing its value the more symmetry of the distribution. Figure 1 represents the BurrX density function with $\nu = 0.4, 0.8, 1.2, 1.6$ and 2 with unity scale:

Amjad D. Al-Nasser, Department of Statistics, Faculty of Science. Yarmouk University, Irbid, Jordan. Email: amjadn@yu.edu.jo. Ayman Baklizi is Assistant Professor of Applied Statistics. His research interests are in accelerated life tests and censored data. Email: baklizi1@hotmail.com.

Figure 1. BurrX density with unity scale and different shaper parameters values.

In this article, bootstrap methods are used to construct confidence intervals for the scale parameter of BurrX distribution, θ . Many authors consider bootstrap methods in a vast range of domains (e. g., Davison et al., 2003; Noreen, 1989; Hall, 1988, 1992; Mooney & Duval, 1989). However, the widely used methods in constructing the confidence intervals consist of the second order accurate bootstrap confidence intervals, namely; “Bootstrap-t (BST)” which give good theoretical coverage probabilities but not reliable and the “Bias Corrected and Accelerated (BCa)” which is the second improved version of the Percentile intervals (PRC). These methods may provide good approximate confidence intervals and better than the usual standard intervals (Efron, 2003). We applied these methods, in addition to, the first improved version of PRC which is the Bias-corrected (BC). Other methods are used, Jackknife Intervals (JAC) and Intervals Based on the Bootstrap Standard Deviation (BSD) Intervals.

We review the approximated confidence interval proposed by Chen and Mi (2001). Next, different bootstrap confidence interval approaches are considered. Then, Monte Carlo evidence on the numerical performance of the bootstrap is presented. Finally, make recommendations.

Approximated Confidence Interval

This confidence interval was proposed by Chen and Mi (2001). They introduced an approximate confidence intervals of certain parameters for distribution function on $(0, \infty)$ which satisfy some monotonic conditions using grouped data. To explain their method, let $X_1, X_2, \dots, X_n$ be a random sample from a distribution $F(X, \theta)$ of a continuous type having probability density on $(0, \infty)$, where θ is unknown parameter. Assume that there are K (groups) inspection times $0 < t_1 < t_2 < \dots < t_k < \infty$, and the probability of an observation fail in the i th group is $p_i = P(t_{i-1} < X \leq t_i)$, $i = 1, 2, \dots, K+1$ where $t_0 = 0$ and $t_{k+1} = \infty$. Let $r_i, i = 1, 2, \dots, K+1$ is the number of observations fail in the i th interval $[t_{i-1}, t_i)$. Based on the upper endpoint approach they define the random variable

$$\zeta_n = \sum_{i=1}^k r_i t_i + r_{k+1} t_k \quad (2)$$

it follows that from the Central Limit Theorem

$$\frac{\zeta_n - ng(\theta)}{\sqrt{n}\sigma(\theta)} \rightarrow N(0,1) \text{ as } n \rightarrow \infty$$

where $g(\theta) = \sum_{i=1}^k t_i p_i + t_k p_{k+1}$,

and

$$\sigma^2(\theta) = \left(\sum_{i=1}^k t_i^2 p_i + t_k^2 p_{k+1} \right) - \left(\sum_{i=1}^k t_i p_i + t_k p_{k+1} \right)^2.$$

An approximate confidence interval for θ can be obtained based on the following estimates;

$$\hat{p}_i = \frac{r_i}{n}, \quad i = 1, 2, \dots, K+1$$

and

$$s_n^2 = \left(\sum_{i=1}^k t_i^2 \hat{p}_i + t_k^2 \hat{p}_{k+1} \right) - \left(\sum_{i=1}^k t_i \hat{p}_i + t_k \hat{p}_{k+1} \right)^2$$

for p_i and $\sigma^2(\theta)$, respectively. It follows that, asymptotically:

$$P\left(\frac{\zeta_n}{n} - z_{\alpha/2} \frac{s_n}{\sqrt{n}} < g(\theta) < \frac{\zeta_n}{n} + z_{\alpha/2} \frac{s_n}{\sqrt{n}}\right) = 1 - \alpha.$$

When the function $g(\theta)$ is monotone, an approximate $(1 - \alpha)\%$ confidence interval for θ , call it the CM interval, can be obtained as:

$$\left[g^{-1}\left(\frac{\zeta_n}{n} - z_{\alpha/2} \frac{s_n}{\sqrt{n}}\right), g^{-1}\left(\frac{\zeta_n}{n} + z_{\alpha/2} \frac{s_n}{\sqrt{n}}\right) \right] \quad (3)$$

However, the above interval possesses exact coverage probabilities and symmetry probabilities only for sufficiently large sample sizes.

Methodology

The bootstrap is a nonparametric technique introduced by Efron in 1979. In this study, we consider six bootstrap methods to construct confidence intervals for the scale parameter of BurrX distribution, θ . For each one of the methods described below, the random variable

ζ_n as defined in (2) calculated from the original data. We generate B bootstrap series $x^{*1}, x^{*2}, \dots, x^{*B}$ and then ζ_n^* be calculated from the bootstrap sample for each series. The standard normal cumulative distribution function is denoted by $\Phi(\cdot)$, and z_α is the α percentile of the standard normal distribution.

t Interval (BTS Intervals)

BST is very similar to confidence intervals based on the t-Student distribution. To construct this interval let z_α^* be the α percentile of the empirical distribution of $Z^* = \frac{(\zeta_n^* - \zeta_n)}{s_n^*}$, where s_n^* is estimated

standard error of ζ_n^* calculated from the bootstrap sample. Then BST interval for θ is given by

$$\left[g^{-1}\left(\zeta_n - z_{\alpha/2}^* s_n^*\right), g^{-1}\left(\zeta_n + z_{1-\alpha/2}^* s_n^*\right) \right] \quad (4)$$

The Percentile Interval (PRC Interval)

An alternative bootstrap interval has been proposed in Efron (1979), and is discussed in detail in Efron and Tibshirani (1993). Let $\hat{g}$ be the cumulative distribution function of ζ_n^* , then the $1 - \alpha$ PRC interval is given by

$$\left[\hat{g}^{-1}\left(\frac{\alpha}{2}\right), \hat{g}^{-1}\left(1 - \frac{\alpha}{2}\right) \right] \quad (5)$$

The Bias Corrected Interval (Interval)

The bias corrected interval (Efron, 1982) is calculated using the corrected percentiles of the bootstrap distribution of ζ_n^* . The determination of the appropriate percentiles depends on a number

$(\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\zeta_n^* < \zeta_n\}}{B}\right))$ which measure the median bias and called the bias correction. The $1 - \alpha$ BC interval is given by

$$\left[\hat{g}^{-1}(\alpha_1), \hat{g}^{-1}(\alpha_2) \right] \quad (6)$$

where

$$\alpha_1 = \Phi(2\hat{z}_0 + z_{\alpha/2}),$$

$$\alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2}).$$

Bias Corrected and Accelerated Interval (BCa Interval)

In this method, we calculate the bias correction $\hat{z}_0$ as before. We need to calculate the acceleration:

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

Thus in the same way we calculated the BC interval; the $1 - \alpha$ BCa interval is given by

$$[\hat{g}^{-1}(\alpha_1), \hat{g}^{-1}(\alpha_2)] \quad (7)$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right),$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right);$$

and $\zeta_n(i)$ is calculated using the original data excluding the i -th observation and

$$\zeta_n(\cdot) = \frac{\sum_{i=1}^n \zeta_n(i)}{n}.$$

Jackknife Intervals (JAC Intervals)

An interval based on the jackknife (Efron & Tibshirani, 1993) can be constructed as follows;

$$[\zeta_n(\cdot) - z_{\alpha/2} s\hat{e}, \zeta_n(\cdot) + z_{1-\alpha/2} s\hat{e}] \quad (8)$$

where

$$s\hat{e}^2 = \frac{n-1}{n} \sum_{i=1}^n (\zeta_n(\cdot) - \zeta_n(i))^2$$

is the jackknife estimate of the variance of ζ_n .

Bootstrap Standard Deviation (BSD Intervals)

An interval similar in form to the based on the jackknife can be constructed as follows;

$$[\zeta_n - z_{\alpha/2} s\tilde{e}, \zeta_n + z_{1-\alpha/2} s\tilde{e}] \quad (9)$$

where

$$s\tilde{e}^2 = \frac{1}{B-1} \sum_{i=1}^B (\zeta_{i,n}^* - \bar{\zeta}_n^*)^2,$$

$$\bar{\zeta}_n^* = \frac{1}{B} \sum_{i=1}^B \zeta_{i,n}^*$$

is the bootstrap estimate of the variance of ζ_n .

Results

A simulation study is conducted to investigate the performance of the interval methods. The 95% confidence intervals for θ was constructed using the seven methods proposed in (3)- (9). The criterion of attainment of lower and upper error probabilities (Jennings, 1987) which are both taken equal to 0.025 was used. In order to compare the performance of the bootstrap estimates, 2000 samples were generated from the BurrX distribution with $\theta = 1$ and $\nu = 0.4, 0.8, 1.2, 1.6$ and 2 ; with different sizes $n = 20, 30, 50$ and 100 .

For each combination, each sample was divided into $(K+1)$ groups where $K = 2, 4$ and 8 . The censoring proportion (cp) is taken as $0.2, 0.4$ and 0.6 . The empirical bootstrap distribution was constructed using $B = 2000$ replications. Then, the following quantities are simulated for each interval: Lower error rates (L): The fraction of intervals that fall entirely above the true parameter; Upper error rates (U): The fraction of intervals that fall entirely below the true parameter and Total error rates (T): The fraction of intervals that did not contain the true parameter value. The results are given in Tables 1-3.

Conclusion

We have compared the performance of several versions of bootstrap confidence intervals together and with the approximated (CM) confidence interval. Bootstrap confidence intervals outperform the CM interval in terms of

total error rates and symmetry in many cases with large sample sizes and appear to be better for small sample sizes.

It can be noted that for $k = 2$, small sample size ($n = 20, 30$) and with censoring proportion ($cp = 0.4$), the CM intervals tend to be anti-conservative. This is also true for JAC, BTS and BCa intervals. On the other hand, the BC, BSD and PRC intervals tend to attain the nominal sizes. As the censoring proportion is light to moderate with $cp = 0.8$, the JAC and the CM intervals tend to be equivalent and grossly anti-conservative while the BC and BCa intervals tend to be grossly conservative. For larger sample sizes ($n = 50, 100$) all intervals attain their nominal sizes except for the BC and BCa intervals where they remain anticonservative. In situations where $k = 2$ and small sample size, all intervals are asymmetric. As k increases, the intervals tend generally to be more symmetric.

The performance of the PRC, BC and BCa intervals improves considerably for larger values of k . Also their performance improves for higher values of r , that is, the more symmetric the parent BurrX distribution, the more symmetric the PRC, BC and BCa intervals tend to be. In conclusion, it appears that the intervals proposed by Chen and Mi (2001) have a good performance except for situations of small sample size and heavy censoring. In this case the BTS, JAC and especially BSD intervals provide better alternatives.

References

Chen, Z., & Mi, J. (2001). An Approximate confidence interval for the scale parameter of the gamma distribution based on grouped data. *Statistical Papers*, 42, 285-299.

Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science*, 18(2), 141-157.

Efron B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

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

Efron B. (2003). Second thoughts in the Bootstrap. *Statistical Science*, 18(2), 135-140.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. NY: Springer-Verlag.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, 16, 927-953.

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

Kulldorff, G. (1961). *Estimation from grouped and partially grouped samples*. New York: Wiley.

Mudholker, G. S., & Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42, 299-302.

Mooney, C. & Duval, R. (1993). *Bootstrapping a nonparametric approach to statistical inference*. London: Sage Publications.

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses*. New York: John Wiley and Sons.

Surles, J. G., & Padgett, W. J. (2001). Inference for reliability and stress-strength for a scaled Burr Type X Distribution. *Lifetime Data Analysis*, 7, 187-200.

Table.1 95% confidence interval for θ based on $k = 2$.

		n	20			30			50			100		
R	CP	Method	L	U	T	L	U	T	L	U	T	L	U	T
0.4	0.4	CM	0.000	0.023	0.023	0.020	0.019	0.039	0.012	0.027	0.039	0.015	0.027	0.042
		BTS	0.000	0.008	0.008	0.020	0.011	0.031	0.012	0.010	0.022	0.015	0.040	0.055
		PRC	0.000	0.026	0.026	0.000	0.045	0.045	0.000	0.096	0.096	0.002	0.087	0.088
		BC	0.000	0.070	0.070	0.000	0.048	0.048	0.012	0.066	0.078	0.052	0.035	0.087
		BCA	0.000	0.085	0.085	0.000	0.068	0.068	0.008	0.089	0.097	0.040	0.047	0.087
		JAC	0.000	0.023	0.023	0.020	0.019	0.039	0.012	0.027	0.039	0.042	0.027	0.069
		BSD	0.006	0.046	0.052	0.013	0.024	0.037	0.025	0.025	0.050	0.020	0.030	0.050
	0.6	CM	0.013	0.024	0.036	0.007	0.026	0.033	0.017	0.029	0.045	0.023	0.022	0.045
		BTS	0.013	0.009	0.022	0.007	0.020	0.027	0.017	0.005	0.022	0.022	0.020	0.041
		PRC	0.000	0.009	0.009	0.002	0.023	0.025	0.005	0.027	0.031	0.006	0.037	0.043
		BC	0.013	0.082	0.095	0.008	0.059	0.067	0.041	0.043	0.083	0.029	0.036	0.064
		BCA	0.000	0.083	0.083	0.007	0.064	0.071	0.041	0.046	0.087	0.023	0.042	0.065
		JAC	0.013	0.024	0.036	0.007	0.026	0.033	0.017	0.029	0.045	0.023	0.022	0.045
		BSD	0.018	0.037	0.054	0.016	0.033	0.049	0.022	0.021	0.043	0.021	0.027	0.048
	0.8	CM	0.029	0.049	0.078	0.018	0.024	0.042	0.015	0.027	0.042	0.024	0.034	0.058
		BTS	0.028	0.017	0.044	0.018	0.007	0.025	0.017	0.011	0.027	0.024	0.016	0.039
		PRC	0.006	0.017	0.022	0.005	0.023	0.028	0.006	0.015	0.021	0.019	0.016	0.034
		BC	0.006	0.091	0.097	0.008	0.101	0.109	0.013	0.091	0.104	0.037	0.056	0.093
		BCA	0.003	0.091	0.094	0.008	0.104	0.112	0.008	0.093	0.101	0.031	0.057	0.088
		JAC	0.029	0.049	0.078	0.018	0.024	0.042	0.015	0.027	0.042	0.024	0.034	0.058
		BSD	0.022	0.043	0.064	0.018	0.043	0.061	0.017	0.029	0.046	0.024	0.033	0.057
0.8	0.4	CM	0.000	0.020	0.020	0.024	0.021	0.045	0.014	0.027	0.041	0.019	0.023	0.042
		BTS	0.000	0.005	0.005	0.024	0.015	0.039	0.014	0.009	0.023	0.019	0.038	0.057
		PRC	0.000	0.022	0.022	0.000	0.050	0.050	0.000	0.090	0.090	0.001	0.086	0.087
		BC	0.000	0.060	0.060	0.001	0.052	0.053	0.014	0.068	0.082	0.059	0.032	0.091
		BCA	0.000	0.078	0.078	0.000	0.076	0.076	0.013	0.086	0.098	0.048	0.044	0.092
		JAC	0.000	0.020	0.020	0.024	0.021	0.045	0.014	0.027	0.041	0.041	0.023	0.064
		BSD	0.007	0.037	0.043	0.016	0.033	0.049	0.021	0.028	0.049	0.026	0.031	0.056
	0.6	CM	0.012	0.027	0.039	0.014	0.023	0.037	0.018	0.040	0.058	0.022	0.028	0.050
		BTS	0.012	0.011	0.022	0.014	0.015	0.029	0.018	0.011	0.029	0.022	0.025	0.046
		PRC	0.000	0.011	0.011	0.001	0.019	0.020	0.008	0.036	0.044	0.007	0.044	0.051
		BC	0.012	0.076	0.088	0.014	0.046	0.059	0.047	0.057	0.104	0.027	0.042	0.068
		BCA	0.000	0.077	0.077	0.014	0.050	0.063	0.047	0.062	0.108	0.022	0.050	0.072
		JAC	0.012	0.027	0.039	0.014	0.023	0.037	0.018	0.040	0.058	0.022	0.028	0.050
		BSD	0.017	0.035	0.052	0.023	0.037	0.060	0.024	0.041	0.065	0.025	0.035	0.060
	0.8	CM	0.029	0.072	0.100	0.018	0.032	0.050	0.020	0.035	0.055	0.023	0.030	0.053
		BTS	0.028	0.022	0.050	0.018	0.013	0.030	0.023	0.010	0.032	0.023	0.019	0.042
		PRC	0.007	0.022	0.029	0.003	0.031	0.034	0.011	0.018	0.028	0.015	0.019	0.033
		BC	0.008	0.107	0.115	0.005	0.097	0.102	0.018	0.099	0.117	0.035	0.050	0.085
		BCA	0.005	0.108	0.112	0.005	0.099	0.104	0.012	0.100	0.112	0.028	0.051	0.078
		JAC	0.029	0.072	0.100	0.018	0.032	0.050	0.020	0.035	0.055	0.023	0.030	0.053
		BSD	0.025	0.054	0.079	0.020	0.049	0.069	0.026	0.041	0.066	0.023	0.030	0.053

Table 1 Continued:

1.2	0.4	CM	0.000	0.022	0.022	0.026	0.019	0.045	0.014	0.030	0.043	0.022	0.022	0.043
		BTS	0.000	0.006	0.006	0.026	0.013	0.039	0.014	0.011	0.025	0.022	0.036	0.058
		PRC	0.000	0.026	0.026	0.000	0.047	0.047	0.000	0.088	0.088	0.000	0.076	0.076
		BC	0.000	0.066	0.066	0.016	0.054	0.069	0.014	0.062	0.076	0.042	0.036	0.078
		BCA	0.000	0.079	0.079	0.001	0.077	0.078	0.014	0.077	0.091	0.033	0.042	0.074
		JAC	0.000	0.022	0.022	0.026	0.019	0.045	0.014	0.030	0.043	0.044	0.022	0.066
		BSD	0.011	0.040	0.051	0.018	0.026	0.044	0.019	0.026	0.045	0.024	0.023	0.047
	0.6	CM	0.013	0.029	0.042	0.008	0.017	0.025	0.015	0.034	0.048	0.022	0.022	0.044
		BTS	0.013	0.014	0.026	0.008	0.015	0.023	0.015	0.008	0.023	0.022	0.016	0.038
		PRC	0.000	0.014	0.014	0.001	0.016	0.017	0.004	0.030	0.034	0.005	0.041	0.046
		BC	0.013	0.083	0.095	0.008	0.041	0.049	0.038	0.036	0.073	0.029	0.043	0.072
		BCA	0.006	0.083	0.088	0.008	0.041	0.049	0.035	0.045	0.079	0.026	0.049	0.074
		JAC	0.013	0.029	0.042	0.008	0.017	0.025	0.015	0.034	0.048	0.022	0.022	0.044
		BSD	0.016	0.038	0.054	0.021	0.029	0.050	0.020	0.030	0.050	0.021	0.029	0.050
	0.8	CM	0.020	0.062	0.081	0.012	0.032	0.043	0.018	0.027	0.045	0.020	0.037	0.056
		BTS	0.019	0.019	0.037	0.012	0.013	0.024	0.019	0.010	0.028	0.020	0.019	0.038
		PRC	0.004	0.019	0.022	0.003	0.032	0.035	0.007	0.017	0.023	0.014	0.019	0.032
		BC	0.004	0.139	0.143	0.004	0.145	0.148	0.018	0.095	0.113	0.022	0.054	0.076
		BCA	0.004	0.139	0.143	0.003	0.146	0.149	0.011	0.096	0.106	0.016	0.055	0.071
		JAC	0.020	0.062	0.081	0.012	0.032	0.043	0.018	0.027	0.045	0.020	0.037	0.056
		BSD	0.017	0.055	0.071	0.014	0.049	0.063	0.022	0.031	0.053	0.018	0.029	0.046
1.6	0.4	CM	0.000	0.024	0.024	0.029	0.016	0.045	0.014	0.033	0.046	0.017	0.020	0.036
		BTS	0.000	0.005	0.005	0.029	0.009	0.037	0.014	0.013	0.027	0.017	0.033	0.049
		PRC	0.000	0.028	0.028	0.000	0.052	0.052	0.000	0.094	0.094	0.001	0.070	0.071
		BC	0.000	0.065	0.065	0.001	0.057	0.057	0.014	0.063	0.077	0.057	0.028	0.085
		BCA	0.000	0.090	0.090	0.000	0.086	0.086	0.012	0.090	0.102	0.040	0.034	0.074
		JAC	0.000	0.024	0.024	0.029	0.016	0.045	0.014	0.033	0.046	0.037	0.020	0.057
		BSD	0.009	0.043	0.051	0.020	0.025	0.045	0.023	0.027	0.049	0.023	0.030	0.053
	0.6	CM	0.013	0.035	0.047	0.014	0.020	0.034	0.014	0.028	0.041	0.027	0.026	0.053
		BTS	0.013	0.016	0.028	0.014	0.014	0.028	0.014	0.011	0.025	0.026	0.021	0.047
		PRC	0.000	0.016	0.016	0.003	0.019	0.022	0.004	0.025	0.029	0.007	0.039	0.046
		BC	0.013	0.075	0.088	0.014	0.042	0.056	0.040	0.042	0.082	0.031	0.039	0.070
		BCA	0.000	0.075	0.075	0.014	0.045	0.059	0.040	0.046	0.086	0.027	0.045	0.072
		JAC	0.013	0.035	0.047	0.014	0.020	0.034	0.014	0.028	0.041	0.027	0.026	0.053
		BSD	0.020	0.041	0.061	0.018	0.027	0.045	0.016	0.032	0.048	0.019	0.030	0.049
	0.8	CM	0.024	0.066	0.090	0.018	0.028	0.046	0.017	0.029	0.045	0.024	0.036	0.060
		BTS	0.023	0.018	0.041	0.018	0.009	0.027	0.017	0.009	0.026	0.024	0.018	0.042
		PRC	0.004	0.018	0.022	0.004	0.027	0.031	0.010	0.017	0.026	0.019	0.018	0.037
		BC	0.005	0.104	0.109	0.006	0.097	0.103	0.016	0.094	0.109	0.033	0.060	0.092
		BCA	0.002	0.104	0.106	0.006	0.099	0.105	0.011	0.094	0.105	0.027	0.062	0.089
		JAC	0.024	0.066	0.090	0.018	0.028	0.046	0.017	0.029	0.045	0.024	0.036	0.060
		BSD	0.017	0.052	0.069	0.020	0.043	0.063	0.015	0.038	0.053	0.018	0.031	0.049

Table 1 Continued:

2	0.4	CM	0.000	0.024	0.024	0.024	0.016	0.040	0.013	0.032	0.044	0.019	0.024	0.043
		BTS	0.000	0.006	0.006	0.024	0.011	0.034	0.013	0.012	0.025	0.019	0.038	0.057
		PRC	0.000	0.028	0.028	0.000	0.049	0.049	0.000	0.085	0.085	0.000	0.090	0.090
		BC	0.000	0.074	0.074	0.005	0.055	0.060	0.013	0.047	0.060	0.032	0.036	0.067
		BCA	0.000	0.084	0.084	0.002	0.100	0.102	0.009	0.066	0.075	0.021	0.045	0.066
		JAC	0.000	0.024	0.024	0.024	0.016	0.040	0.013	0.032	0.044	0.043	0.024	0.067
		BSD	0.007	0.048	0.054	0.016	0.032	0.048	0.026	0.030	0.056	0.022	0.025	0.046
	0.6	CM	0.015	0.024	0.039	0.011	0.015	0.026	0.020	0.033	0.053	0.024	0.020	0.044
		BTS	0.015	0.011	0.026	0.011	0.011	0.022	0.020	0.011	0.030	0.024	0.019	0.042
		PRC	0.000	0.011	0.011	0.001	0.013	0.013	0.007	0.031	0.038	0.007	0.034	0.041
		BC	0.015	0.078	0.093	0.011	0.046	0.057	0.068	0.058	0.126	0.039	0.037	0.076
		BCA	0.000	0.079	0.079	0.011	0.052	0.063	0.056	0.062	0.118	0.037	0.039	0.076
		JAC	0.015	0.024	0.039	0.011	0.015	0.026	0.020	0.033	0.053	0.024	0.020	0.044
		BSD	0.018	0.034	0.051	0.018	0.027	0.045	0.022	0.032	0.054	0.022	0.028	0.050
	0.8	CM	0.023	0.067	0.090	0.024	0.032	0.056	0.015	0.037	0.052	0.020	0.034	0.054
		BTS	0.022	0.020	0.042	0.024	0.007	0.030	0.016	0.012	0.028	0.020	0.019	0.038
		PRC	0.005	0.020	0.025	0.005	0.032	0.037	0.007	0.022	0.029	0.017	0.019	0.035
		BC	0.008	0.127	0.135	0.005	0.133	0.138	0.017	0.094	0.111	0.023	0.049	0.072
		BCA	0.005	0.128	0.132	0.005	0.134	0.139	0.011	0.094	0.105	0.022	0.050	0.072
		JAC	0.023	0.067	0.090	0.024	0.032	0.056	0.015	0.037	0.052	0.020	0.034	0.054
		BSD	0.018	0.053	0.070	0.027	0.049	0.075	0.020	0.041	0.061	0.018	0.031	0.048

Table 2: 95% confidence interval for θ based on $k = 4$.

		n	20			30			50				100	
R	CP	Method	L	U	T	L	U	T	L	U	T	L	U	T
0.4	0.4	CM	0.000	0.036	0.036	0.005	0.026	0.031	0.009	0.025	0.034	0.014	0.023	0.037
		BTS	0.000	0.016	0.016	0.001	0.024	0.025	0.003	0.053	0.056	0.003	0.072	0.075
		PRC	0.000	0.067	0.067	0.000	0.078	0.078	0.000	0.097	0.097	0.001	0.104	0.104
		BC	0.014	0.035	0.049	0.032	0.019	0.051	0.048	0.013	0.061	0.069	0.010	0.078
		BCA	0.007	0.045	0.051	0.023	0.028	0.051	0.043	0.020	0.063	0.062	0.011	0.073
		JAC	0.000	0.039	0.039	0.005	0.029	0.033	0.011	0.028	0.039	0.015	0.024	0.039
		BSD	0.002	0.037	0.039	0.007	0.030	0.037	0.010	0.026	0.036	0.016	0.024	0.040
	0.6	CM	0.007	0.046	0.053	0.008	0.044	0.052	0.013	0.037	0.050	0.024	0.024	0.048
		BTS	0.002	0.021	0.022	0.004	0.030	0.033	0.005	0.038	0.043	0.010	0.042	0.052
		PRC	0.000	0.046	0.046	0.000	0.069	0.069	0.001	0.057	0.058	0.002	0.057	0.058
		BC	0.022	0.052	0.074	0.023	0.056	0.079	0.035	0.033	0.068	0.051	0.017	0.068
		BCA	0.012	0.055	0.067	0.017	0.063	0.079	0.029	0.038	0.067	0.046	0.019	0.065
		JAC	0.008	0.047	0.055	0.009	0.045	0.053	0.013	0.041	0.054	0.024	0.024	0.048
		BSD	0.006	0.047	0.053	0.009	0.045	0.053	0.013	0.041	0.054	0.022	0.028	0.050
	0.8	CM	0.008	0.057	0.065	0.012	0.042	0.054	0.014	0.039	0.053	0.019	0.031	0.050
		BTS	0.005	0.022	0.026	0.007	0.021	0.027	0.008	0.025	0.033	0.009	0.025	0.033
		PRC	0.001	0.027	0.028	0.002	0.030	0.031	0.003	0.028	0.031	0.006	0.028	0.034
		BC	0.014	0.092	0.106	0.014	0.072	0.086	0.021	0.050	0.071	0.026	0.040	0.065
		BCA	0.010	0.095	0.105	0.012	0.073	0.085	0.019	0.051	0.069	0.025	0.042	0.066
		JAC	0.011	0.057	0.068	0.012	0.043	0.054	0.015	0.040	0.055	0.020	0.031	0.051
		BSD	0.011	0.059	0.070	0.012	0.046	0.057	0.014	0.040	0.054	0.018	0.031	0.049

Table 2: Continued

0.8	0.4	CM	0.000	0.036	0.036	0.004	0.027	0.031	0.009	0.029	0.038	0.015	0.027	0.042
		BTS	0.000	0.016	0.016	0.002	0.027	0.029	0.002	0.052	0.054	0.003	0.089	0.091
		PRC	0.000	0.065	0.065	0.000	0.073	0.073	0.000	0.101	0.101	0.000	0.119	0.119
		BC	0.015	0.035	0.050	0.032	0.021	0.052	0.046	0.020	0.065	0.090	0.008	0.098
		BCA	0.006	0.043	0.048	0.025	0.028	0.053	0.043	0.024	0.067	0.082	0.013	0.095
		JAC	0.000	0.038	0.038	0.004	0.028	0.032	0.010	0.032	0.042	0.016	0.028	0.044
		BSD	0.003	0.036	0.039	0.009	0.027	0.036	0.012	0.030	0.041	0.017	0.026	0.043
	0.6	CM	0.005	0.039	0.044	0.009	0.034	0.043	0.013	0.027	0.040	0.015	0.030	0.045
		BTS	0.001	0.016	0.017	0.003	0.021	0.024	0.005	0.026	0.031	0.004	0.048	0.051
		PRC	0.000	0.041	0.041	0.000	0.068	0.068	0.001	0.044	0.045	0.000	0.065	0.065
		BC	0.019	0.042	0.061	0.018	0.053	0.071	0.035	0.025	0.059	0.049	0.021	0.070
		BCA	0.010	0.047	0.057	0.015	0.060	0.075	0.028	0.029	0.057	0.044	0.026	0.070
		JAC	0.009	0.039	0.048	0.010	0.037	0.046	0.013	0.029	0.042	0.015	0.030	0.045
		BSD	0.007	0.040	0.046	0.008	0.037	0.044	0.015	0.031	0.045	0.018	0.031	0.048
	0.8	CM	0.008	0.051	0.059	0.016	0.046	0.062	0.016	0.037	0.053	0.019	0.034	0.053
		BTS	0.004	0.014	0.018	0.008	0.021	0.029	0.010	0.023	0.032	0.009	0.028	0.036
		PRC	0.000	0.024	0.024	0.002	0.030	0.032	0.005	0.026	0.031	0.006	0.031	0.037
		BC	0.011	0.089	0.100	0.020	0.069	0.088	0.024	0.048	0.072	0.026	0.043	0.069
		BCA	0.010	0.089	0.098	0.018	0.069	0.087	0.021	0.048	0.069	0.024	0.044	0.068
		JAC	0.010	0.051	0.061	0.018	0.046	0.064	0.017	0.038	0.054	0.020	0.034	0.054
		BSD	0.009	0.056	0.065	0.019	0.045	0.064	0.016	0.039	0.054	0.015	0.037	0.051
1.2	0.4	CM	0.000	0.029	0.029	0.004	0.025	0.028	0.011	0.027	0.038	0.020	0.023	0.043
		BTS	0.000	0.012	0.012	0.001	0.027	0.028	0.002	0.049	0.051	0.004	0.088	0.092
		PRC	0.000	0.064	0.064	0.000	0.074	0.074	0.000	0.097	0.097	0.000	0.112	0.112
		BC	0.004	0.028	0.032	0.026	0.018	0.044	0.051	0.017	0.068	0.094	0.010	0.104
		BCA	0.000	0.038	0.038	0.011	0.022	0.033	0.041	0.022	0.062	0.087	0.012	0.099
		JAC	0.000	0.032	0.032	0.004	0.026	0.030	0.014	0.030	0.043	0.021	0.023	0.044
		BSD	0.002	0.034	0.036	0.004	0.029	0.033	0.013	0.033	0.046	0.019	0.024	0.043
	0.6	CM	0.007	0.049	0.056	0.011	0.036	0.047	0.012	0.027	0.039	0.018	0.034	0.052
		BTS	0.004	0.023	0.026	0.005	0.027	0.032	0.003	0.027	0.030	0.004	0.048	0.051
		PRC	0.000	0.054	0.054	0.000	0.064	0.064	0.000	0.047	0.047	0.001	0.059	0.060
		BC	0.013	0.060	0.073	0.019	0.048	0.066	0.029	0.022	0.051	0.047	0.021	0.068
		BCA	0.008	0.067	0.074	0.013	0.051	0.063	0.024	0.029	0.053	0.041	0.024	0.065
		JAC	0.009	0.050	0.058	0.011	0.039	0.050	0.012	0.027	0.039	0.018	0.034	0.052
		BSD	0.009	0.052	0.061	0.012	0.039	0.051	0.012	0.029	0.041	0.016	0.033	0.049
	0.8	CM	0.009	0.050	0.059	0.012	0.043	0.054	0.009	0.043	0.052	0.015	0.029	0.044
		BTS	0.006	0.016	0.022	0.004	0.019	0.022	0.004	0.024	0.027	0.007	0.023	0.030
		PRC	0.000	0.024	0.024	0.001	0.027	0.027	0.002	0.027	0.028	0.004	0.026	0.030
		BC	0.014	0.096	0.110	0.013	0.074	0.086	0.018	0.057	0.075	0.021	0.038	0.059
		BCA	0.010	0.095	0.104	0.010	0.075	0.085	0.015	0.060	0.074	0.021	0.040	0.061
		JAC	0.012	0.050	0.062	0.012	0.043	0.055	0.010	0.044	0.054	0.015	0.029	0.044
		BSD	0.012	0.053	0.065	0.007	0.046	0.053	0.009	0.042	0.051	0.015	0.033	0.048

Table 2: Continued

1.6	0.4	CM	0.000	0.033	0.033	0.009	0.029	0.038	0.011	0.029	0.039	0.018	0.025	0.043
		BTS	0.000	0.014	0.014	0.003	0.032	0.035	0.003	0.049	0.052	0.002	0.099	0.101
		PRC	0.000	0.060	0.060	0.000	0.082	0.082	0.000	0.088	0.088	0.000	0.135	0.135
		BC	0.011	0.033	0.044	0.040	0.021	0.061	0.054	0.020	0.074	0.080	0.011	0.090
		BCA	0.006	0.040	0.046	0.027	0.030	0.057	0.044	0.024	0.068	0.075	0.014	0.089
		JAC	0.000	0.034	0.034	0.009	0.033	0.042	0.012	0.031	0.043	0.019	0.025	0.044
		BSD	0.004	0.034	0.038	0.010	0.035	0.045	0.009	0.035	0.043	0.016	0.027	0.043
	0.6	CM	0.007	0.045	0.051	0.007	0.032	0.039	0.014	0.032	0.046	0.019	0.028	0.047
		BTS	0.003	0.019	0.022	0.003	0.022	0.025	0.003	0.030	0.033	0.005	0.046	0.051
		PRC	0.000	0.047	0.047	0.000	0.072	0.072	0.000	0.053	0.053	0.001	0.065	0.065
		BC	0.020	0.051	0.071	0.019	0.054	0.073	0.039	0.027	0.066	0.046	0.018	0.064
		BCA	0.011	0.059	0.070	0.014	0.059	0.073	0.033	0.032	0.065	0.044	0.023	0.067
		JAC	0.010	0.047	0.057	0.007	0.038	0.045	0.014	0.034	0.048	0.019	0.028	0.047
		BSD	0.010	0.046	0.056	0.006	0.037	0.042	0.014	0.030	0.044	0.015	0.032	0.047
	0.8	CM	0.008	0.056	0.063	0.011	0.045	0.056	0.014	0.034	0.048	0.024	0.032	0.056
		BTS	0.005	0.020	0.025	0.005	0.022	0.026	0.006	0.017	0.023	0.015	0.021	0.036
		PRC	0.001	0.026	0.027	0.002	0.031	0.032	0.002	0.018	0.020	0.010	0.025	0.034
		BC	0.010	0.093	0.103	0.014	0.077	0.090	0.021	0.045	0.065	0.032	0.038	0.070
		BCA	0.009	0.094	0.102	0.009	0.077	0.086	0.019	0.047	0.066	0.030	0.039	0.069
		JAC	0.012	0.056	0.068	0.012	0.046	0.057	0.016	0.034	0.050	0.024	0.033	0.057
		BSD	0.009	0.058	0.067	0.009	0.047	0.056	0.012	0.038	0.049	0.021	0.030	0.050
2	0.4	CM	0.000	0.031	0.031	0.004	0.026	0.030	0.011	0.021	0.032	0.018	0.023	0.041
		BTS	0.000	0.015	0.015	0.002	0.030	0.031	0.003	0.051	0.054	0.004	0.090	0.094
		PRC	0.000	0.063	0.063	0.000	0.075	0.075	0.000	0.094	0.094	0.000	0.115	0.115
		BC	0.005	0.031	0.036	0.033	0.023	0.056	0.055	0.011	0.066	0.079	0.010	0.089
		BCA	0.000	0.041	0.041	0.021	0.029	0.050	0.050	0.015	0.065	0.073	0.014	0.087
		JAC	0.000	0.035	0.035	0.004	0.031	0.035	0.014	0.023	0.037	0.020	0.024	0.044
		BSD	0.004	0.038	0.042	0.009	0.029	0.038	0.012	0.026	0.038	0.019	0.026	0.045
	0.6	CM	0.005	0.042	0.047	0.007	0.041	0.047	0.016	0.038	0.053	0.017	0.032	0.049
		BTS	0.002	0.014	0.016	0.003	0.028	0.031	0.003	0.037	0.040	0.007	0.053	0.060
		PRC	0.000	0.044	0.044	0.000	0.074	0.074	0.000	0.058	0.058	0.001	0.071	0.072
		BC	0.012	0.052	0.064	0.018	0.055	0.073	0.034	0.032	0.065	0.048	0.024	0.072
		BCA	0.006	0.057	0.063	0.012	0.068	0.079	0.025	0.038	0.063	0.043	0.027	0.069
		JAC	0.008	0.043	0.050	0.007	0.043	0.050	0.016	0.041	0.057	0.017	0.033	0.050
		BSD	0.007	0.046	0.053	0.009	0.043	0.052	0.013	0.036	0.049	0.016	0.035	0.051
	0.8	CM	0.008	0.053	0.061	0.013	0.043	0.056	0.010	0.038	0.047	0.020	0.030	0.050
		BTS	0.006	0.020	0.026	0.006	0.022	0.027	0.004	0.022	0.026	0.013	0.022	0.035
		PRC	0.001	0.025	0.025	0.001	0.030	0.030	0.001	0.024	0.025	0.009	0.025	0.034
		BC	0.012	0.088	0.099	0.015	0.064	0.079	0.016	0.053	0.069	0.027	0.044	0.070
		BCA	0.008	0.088	0.095	0.012	0.065	0.077	0.013	0.055	0.068	0.025	0.045	0.070
		JAC	0.010	0.053	0.063	0.014	0.043	0.056	0.010	0.039	0.049	0.020	0.031	0.051
		BSD	0.009	0.053	0.062	0.011	0.042	0.053	0.010	0.041	0.051	0.020	0.032	0.052

Table 3: 95% confidence interval for θ based on $k = 8$.

R	CP	n	20			30			50			100		
		Method	L	U	T	L	U	T	L	U	T	L	U	T
0.4	0.4	CM	0.000	0.033	0.033	0.003	0.027	0.030	0.011	0.029	0.040	0.013	0.028	0.041
		BTS	0.000	0.033	0.033	0.000	0.055	0.055	0.001	0.102	0.103	0.001	0.156	0.157
		PRC	0.000	0.153	0.153	0.000	0.143	0.143	0.000	0.164	0.164	0.000	0.190	0.190
		BC	0.012	0.030	0.042	0.040	0.015	0.054	0.073	0.013	0.086	0.092	0.011	0.102
		BCA	0.005	0.041	0.045	0.033	0.021	0.054	0.069	0.018	0.087	0.088	0.014	0.102
		JAC	0.000	0.040	0.040	0.004	0.030	0.034	0.011	0.030	0.041	0.013	0.029	0.042
		BSD	0.000	0.038	0.038	0.005	0.029	0.034	0.009	0.033	0.042	0.012	0.029	0.041
	0.6	CM	0.004	0.046	0.050	0.010	0.034	0.044	0.010	0.026	0.036	0.020	0.030	0.050
		BTS	0.001	0.035	0.036	0.004	0.037	0.041	0.004	0.046	0.050	0.006	0.073	0.078
		PRC	0.000	0.080	0.080	0.000	0.066	0.066	0.000	0.074	0.074	0.001	0.092	0.092
		BC	0.019	0.041	0.060	0.034	0.025	0.059	0.042	0.018	0.060	0.067	0.016	0.083
		BCA	0.013	0.046	0.059	0.027	0.028	0.054	0.035	0.020	0.055	0.063	0.019	0.082
		JAC	0.004	0.054	0.057	0.012	0.038	0.050	0.010	0.028	0.038	0.020	0.030	0.050
		BSD	0.003	0.049	0.052	0.011	0.036	0.047	0.010	0.029	0.038	0.021	0.029	0.050
	0.8	CM	0.005	0.060	0.065	0.005	0.057	0.062	0.013	0.044	0.057	0.016	0.035	0.050
		BTS	0.001	0.035	0.035	0.003	0.040	0.043	0.005	0.039	0.044	0.005	0.039	0.044
		PRC	0.000	0.043	0.043	0.000	0.051	0.051	0.003	0.047	0.050	0.002	0.043	0.045
		BC	0.012	0.076	0.088	0.015	0.065	0.080	0.028	0.047	0.075	0.036	0.030	0.065
		BCA	0.011	0.077	0.087	0.011	0.067	0.078	0.027	0.048	0.075	0.033	0.030	0.063
		JAC	0.006	0.068	0.074	0.005	0.061	0.066	0.013	0.047	0.060	0.016	0.035	0.051
		BSD	0.005	0.065	0.069	0.005	0.059	0.064	0.014	0.047	0.061	0.016	0.036	0.052
0.8	0.4	CM	0.000	0.035	0.035	0.003	0.031	0.034	0.007	0.024	0.031	0.013	0.024	0.037
		BTS	0.000	0.033	0.033	0.001	0.061	0.062	0.002	0.100	0.102	0.001	0.148	0.149
		PRC	0.000	0.146	0.146	0.000	0.144	0.144	0.000	0.162	0.162	0.000	0.191	0.191
		BC	0.014	0.028	0.042	0.038	0.018	0.056	0.070	0.008	0.078	0.108	0.009	0.116
		BCA	0.009	0.041	0.049	0.031	0.025	0.056	0.064	0.014	0.078	0.104	0.010	0.113
		JAC	0.000	0.037	0.037	0.005	0.034	0.038	0.007	0.027	0.034	0.013	0.025	0.038
		BSD	0.001	0.038	0.039	0.005	0.031	0.036	0.008	0.029	0.037	0.014	0.026	0.040
	0.6	CM	0.004	0.050	0.054	0.005	0.039	0.044	0.012	0.039	0.051	0.015	0.032	0.046
		BTS	0.001	0.033	0.034	0.001	0.044	0.045	0.001	0.065	0.066	0.002	0.073	0.075
		PRC	0.000	0.076	0.076	0.000	0.081	0.081	0.000	0.093	0.093	0.001	0.086	0.087
		BC	0.018	0.042	0.060	0.024	0.030	0.054	0.043	0.026	0.069	0.067	0.015	0.082
		BCA	0.013	0.048	0.061	0.021	0.035	0.056	0.040	0.029	0.069	0.064	0.017	0.081
		JAC	0.004	0.056	0.060	0.006	0.042	0.048	0.013	0.041	0.054	0.015	0.032	0.047
		BSD	0.004	0.054	0.058	0.005	0.040	0.045	0.013	0.041	0.054	0.016	0.033	0.048
	0.8	CM	0.003	0.066	0.069	0.009	0.056	0.065	0.013	0.043	0.056	0.019	0.032	0.051
		BTS	0.002	0.042	0.044	0.002	0.036	0.038	0.005	0.038	0.043	0.008	0.038	0.046
		PRC	0.001	0.050	0.051	0.000	0.048	0.048	0.002	0.043	0.044	0.005	0.042	0.046
		BC	0.011	0.083	0.093	0.019	0.066	0.085	0.026	0.043	0.069	0.039	0.029	0.067
		BCA	0.007	0.081	0.088	0.016	0.067	0.083	0.025	0.043	0.068	0.037	0.030	0.067
		JAC	0.004	0.072	0.076	0.011	0.059	0.070	0.013	0.044	0.057	0.019	0.033	0.052
		BSD	0.003	0.068	0.070	0.008	0.060	0.067	0.013	0.043	0.056	0.020	0.035	0.055

Table 3: Continued

1.2	0.4	CM	0.000	0.036	0.036	0.004	0.038	0.041	0.011	0.026	0.037	0.014	0.027	0.041
		BTS	0.000	0.036	0.036	0.001	0.059	0.059	0.001	0.098	0.099	0.001	0.147	0.148
		PRC	0.000	0.148	0.148	0.000	0.144	0.144	0.000	0.156	0.156	0.000	0.184	0.184
		BC	0.013	0.034	0.047	0.040	0.025	0.065	0.059	0.010	0.069	0.099	0.011	0.110
		BCA	0.005	0.042	0.047	0.033	0.030	0.063	0.057	0.015	0.071	0.096	0.016	0.111
		JAC	0.000	0.040	0.040	0.004	0.041	0.045	0.011	0.029	0.040	0.014	0.031	0.045
		BSD	0.001	0.042	0.043	0.005	0.039	0.044	0.011	0.029	0.039	0.015	0.028	0.043
	0.6	CM	0.007	0.052	0.059	0.008	0.050	0.058	0.012	0.038	0.050	0.016	0.031	0.046
		BTS	0.001	0.040	0.041	0.001	0.055	0.056	0.002	0.061	0.063	0.001	0.081	0.081
		PRC	0.000	0.070	0.070	0.000	0.086	0.086	0.000	0.084	0.084	0.000	0.099	0.099
		BC	0.019	0.047	0.066	0.032	0.038	0.069	0.045	0.026	0.071	0.069	0.016	0.084
		BCA	0.014	0.051	0.065	0.025	0.042	0.066	0.040	0.030	0.069	0.067	0.017	0.084
		JAC	0.007	0.056	0.063	0.008	0.054	0.061	0.012	0.040	0.052	0.016	0.032	0.047
		BSD	0.007	0.054	0.061	0.007	0.051	0.057	0.012	0.041	0.053	0.015	0.035	0.049
	0.8	CM	0.008	0.055	0.063	0.010	0.051	0.061	0.016	0.044	0.060	0.017	0.038	0.055
		BTS	0.004	0.032	0.036	0.002	0.038	0.040	0.008	0.038	0.045	0.006	0.041	0.047
		PRC	0.001	0.042	0.043	0.000	0.048	0.048	0.002	0.043	0.045	0.004	0.047	0.051
		BC	0.014	0.074	0.088	0.020	0.065	0.085	0.031	0.044	0.075	0.031	0.032	0.063
		BCA	0.011	0.073	0.084	0.016	0.064	0.080	0.029	0.044	0.073	0.029	0.033	0.062
		JAC	0.010	0.059	0.068	0.011	0.057	0.068	0.017	0.045	0.062	0.018	0.038	0.055
		BSD	0.009	0.058	0.066	0.010	0.054	0.064	0.016	0.046	0.062	0.019	0.037	0.055
1.6	0.4	CM	0.000	0.033	0.033	0.003	0.030	0.033	0.012	0.027	0.038	0.017	0.026	0.043
		BTS	0.000	0.030	0.030	0.000	0.055	0.055	0.002	0.093	0.095	0.002	0.160	0.162
		PRC	0.000	0.141	0.141	0.000	0.140	0.140	0.000	0.155	0.155	0.000	0.192	0.192
		BC	0.005	0.031	0.036	0.036	0.018	0.053	0.075	0.014	0.089	0.112	0.010	0.121
		BCA	0.002	0.042	0.044	0.030	0.024	0.054	0.071	0.017	0.087	0.109	0.012	0.121
		JAC	0.000	0.043	0.043	0.004	0.032	0.035	0.012	0.028	0.039	0.017	0.026	0.043
		BSD	0.000	0.036	0.036	0.005	0.033	0.037	0.012	0.028	0.040	0.017	0.026	0.043
	0.6	CM	0.004	0.045	0.049	0.008	0.038	0.045	0.010	0.038	0.048	0.019	0.030	0.049
		BTS	0.000	0.032	0.032	0.003	0.043	0.046	0.002	0.063	0.065	0.004	0.074	0.078
		PRC	0.000	0.067	0.067	0.000	0.076	0.076	0.000	0.081	0.081	0.000	0.091	0.091
		BC	0.014	0.038	0.052	0.030	0.029	0.059	0.039	0.027	0.066	0.066	0.017	0.082
		BCA	0.009	0.044	0.052	0.022	0.031	0.053	0.037	0.030	0.066	0.063	0.019	0.081
		JAC	0.004	0.048	0.052	0.009	0.042	0.051	0.010	0.040	0.050	0.020	0.031	0.050
		BSD	0.003	0.046	0.049	0.007	0.036	0.043	0.009	0.040	0.049	0.018	0.032	0.050
	0.8	CM	0.007	0.055	0.062	0.013	0.048	0.060	0.013	0.036	0.049	0.019	0.032	0.051
		BTS	0.002	0.026	0.028	0.005	0.026	0.031	0.004	0.033	0.037	0.006	0.039	0.045
		PRC	0.000	0.036	0.036	0.001	0.042	0.042	0.001	0.036	0.037	0.003	0.045	0.048
		BC	0.014	0.069	0.082	0.023	0.056	0.079	0.026	0.038	0.063	0.038	0.028	0.066
		BCA	0.011	0.068	0.079	0.021	0.055	0.076	0.023	0.037	0.060	0.037	0.029	0.066
		JAC	0.008	0.060	0.068	0.013	0.050	0.063	0.013	0.038	0.051	0.020	0.033	0.053
		BSD	0.007	0.056	0.063	0.012	0.048	0.060	0.013	0.038	0.051	0.017	0.033	0.050

Table 3: Continued

2	0.4	CM	0.000	0.031	0.031	0.007	0.031	0.038	0.006	0.029	0.034	0.017	0.034	0.051
		BTS	0.000	0.032	0.032	0.002	0.060	0.062	0.001	0.102	0.103	0.000	0.147	0.147
		PRC	0.000	0.133	0.133	0.000	0.143	0.143	0.000	0.158	0.158	0.000	0.178	0.178
		BC	0.006	0.031	0.037	0.042	0.021	0.063	0.064	0.013	0.077	0.113	0.011	0.123
		BCA	0.004	0.038	0.041	0.033	0.027	0.060	0.059	0.017	0.076	0.107	0.015	0.122
		JAC	0.000	0.039	0.039	0.007	0.035	0.041	0.006	0.029	0.035	0.017	0.035	0.052
		BSD	0.000	0.038	0.038	0.008	0.031	0.039	0.007	0.028	0.035	0.016	0.034	0.050
	0.6	CM	0.003	0.045	0.048	0.009	0.045	0.054	0.009	0.035	0.044	0.015	0.032	0.047
		BTS	0.001	0.031	0.032	0.003	0.049	0.052	0.002	0.058	0.060	0.002	0.092	0.093
		PRC	0.000	0.067	0.067	0.000	0.080	0.080	0.000	0.083	0.083	0.000	0.108	0.108
		BC	0.015	0.039	0.054	0.032	0.034	0.066	0.040	0.023	0.062	0.059	0.015	0.074
		BCA	0.009	0.046	0.055	0.029	0.040	0.068	0.035	0.025	0.060	0.055	0.016	0.071
		JAC	0.003	0.049	0.052	0.010	0.047	0.057	0.010	0.038	0.048	0.016	0.033	0.049
		BSD	0.002	0.046	0.048	0.009	0.046	0.055	0.009	0.039	0.048	0.016	0.033	0.049
	0.8	CM	0.004	0.057	0.061	0.009	0.046	0.055	0.014	0.040	0.054	0.018	0.031	0.049
		BTS	0.001	0.033	0.034	0.004	0.027	0.031	0.005	0.039	0.043	0.005	0.033	0.038
		PRC	0.000	0.043	0.043	0.001	0.042	0.042	0.001	0.043	0.044	0.003	0.039	0.041
		BC	0.011	0.071	0.081	0.017	0.059	0.075	0.034	0.042	0.076	0.035	0.026	0.061
		BCA	0.006	0.070	0.076	0.016	0.059	0.074	0.030	0.043	0.072	0.033	0.026	0.059
		JAC	0.004	0.062	0.066	0.010	0.050	0.060	0.015	0.042	0.057	0.019	0.032	0.050
		BSD	0.004	0.061	0.065	0.008	0.048	0.056	0.015	0.042	0.057	0.017	0.032	0.048

On Comparison of Hypothesis Tests in the Bayesian Framework without Loss Function

Vladimir Gercsik

International Institute of Earthquake
Prediction Theory & Mathematical Geophysics
Russian Academy of Science

Mark Kelbert

Department of Mathematics
University of Wales-Swansea
United Kingdom

The problem is how to compare the quality of different hypothesis tests in a Bayesian framework without introducing a loss function. Three different linear orders on the set of all possible hypothesis tests are studied. The most natural order estimates the Fisher information between indicators of event and decision.

Key words: Bayesian risk, Neyman-Pearson test, Fisher information, tests of independency

Introduction

It is well-known that no universal measure of a hypothesis test quality exists in statistics. In the Bayesian framework a linear ordering of tests is possible for a given loss function. It is said that a test is optimal if it minimizes the Bayesian risk.

However, the selection of a loss function is often somewhat arbitrary, and it is not always natural to measure the losses under different types of errors in the same units. For example, the cost of prevention measures in earthquake prediction is naturally expressed in money units. However, the losses from an earthquake including the psychological traumas, maiming and even the loss of human lives could be hardly expressed in money terms. Even if this expression is imposed, any estimation of these losses in money terms would depend a great deal on the variable economical and political situation. This loss function hardly looks as neutral and scientifically unbiased.

The subject of interest is in a situation when the loss function is unknown but the quality of any two hypothesis tests should be quantitatively compared. It turned out that all

hypothesis tests could be linearly ordered at least in three different ways:

Let p and $1-p$ be the Bayesian probabilities of random experiments $\omega \in \Omega$ with cumulative distribution functions F_1 and F_2 , respectively. A test is defined by a function $\Phi(\omega) = 1$ on the critical set B and 0 on B^c . If $\Phi(\omega) = 1$ then the alternative F_2 is accepted, and the hypothesis F_1 is accepted in the case $\Phi(\omega) = 0$. Clearly, the problem is symmetric with respect to interchange of hypothesis F_1 and F_2 and a simultaneous interchange of $\Phi(\omega)$ and $1 - \Phi(\omega)$.

The type I error is denoted (i.e., the probability to accept F_2 when F_1 is true) by α_1 , and the type II error (i.e., the probability to accept F_1 when F_2 is true) by α_2 . Considered are only *unbiased* tests (Barra, 1981), i.e., assume that $\alpha_1 + \alpha_2 \leq 1$. (If this condition is violated one could get an unbiased test by selecting B^c instead of B as a critical set.) Consider a random variable $X_1 = 0$ if F_1 is true, $X_1 = 1$ if F_2 is true, and call it *an indicator of events*. In a similar manner we define a random variable $X_2 = 0$ if a test accepts F_1 , and $X_2 = 1$ if the test accepts F_2 , called *an indicator of decision*. The joint distribution $\mathbf{P}(x_1, x_2)$ is defined by the relations $\alpha_1 = \mathbf{P}(X_2 = 1 \mid X_1 = 0)$, $\alpha_2 = \mathbf{P}(X_2 = 0 \mid X_1 = 1)$, in particular,

Vladimir Gercsik is a Senior Researcher in geophysics and earthquake prediction. Mark Kelbert is a Reader in statistics. Email him at M.Kelbert@swansea.ac.uk.

$$\begin{aligned} \mathbf{P}(X_2 = 0, X_1 = 0) &= p(1 - \alpha_1), \\ \mathbf{P}(X_2 = 1, X_1 = 0) &= p\alpha_1, \end{aligned} \quad (1)$$

$$\begin{aligned} \mathbf{P}(X_2 = 0, X_1 = 1) &= (1 - p)\alpha_2, \\ \mathbf{P}(X_2 = 1, X_1 = 1) &= (1 - p)(1 - \alpha_2). \end{aligned} \quad (2)$$

The marginal one-dimensional probabilities take the form:

$$\begin{aligned} \mathbf{P}(X_1 = 0) &= p, \quad \mathbf{P}(X_1 = 1) = (1 - p), \\ \mathbf{P}(X_2 = 0) &= p(1 - \alpha_1) + \alpha_2(1 - p), \end{aligned} \quad (3)$$

$$\mathbf{P}(X_2 = 1) = p\alpha_1 + (1 - p)(1 - \alpha_2). \quad (4)$$

Clearly, the worst possible unbiased test is determined by the condition that indicator of event X_1 and indicator of decision X_2 are independent. On the other extreme, an ideal test (normally, it does not exist) is one that provides the correct solution without any errors. Generically, the quality of a test is measured by some non-negative function of X_1 and $X_2 = \Phi(\omega)$ which takes the value 0 iff X_1, X_2 are independent, and the value 1 iff $X_1 = X_2$.

Measuring the quality of a test

Any of the following well-known Rachev (1991) functions used in tests of independency is acceptable as a measure of the quality of a test.

$$\beta_1(\Phi) = \sup_{x_1, x_2} |\mathbf{P}(x_1, x_2) - \mathbf{P}(x_1)\mathbf{P}(x_2)|, \quad (5)$$

$$\beta_2(\Phi) = \mathbf{E}_{x_1} \left[\sup_{x_2} |\mathbf{P}(x_2 | x_1) - \mathbf{P}(x_2)| \right] \quad (6)$$

($\mathbf{E}_{x_1}$ stands for the expectation with respect of distribution of X_1). It is easy to check that

$$\beta_1(\Phi) = \beta_2(\Phi) = p(1 - p)(1 - \alpha_1 - \alpha_2). \quad (7)$$

The quality of a test is measured by $\beta = (1 - \alpha_1 - \alpha_2)$. This is quite popular in practice as the Bayesian risk $\mathbf{R} = \mathbf{E}[w] = \alpha_1 + \alpha_2$ appears for the simplest loss function $w(x_1, x_2)$: $w(0,0) = w(1,1) = 0$, $w(0,1) = w(1,0) = 1$.

Another possibility to test the independence is to consider the maximal correlation coefficient

$$\rho(\Phi) = \sup \mathbf{E}[\phi_1(X_1)\phi_2(X_2)] \quad (8)$$

where sup is taken over the set of functions ϕ_1, ϕ_2 such that $\mathbf{E}[\phi_i(X_i)] = 0$, $\sigma^2[\phi_i(X_i)] = \mathbf{E}[\phi_i(X_i)^2] = 1$, $i=1,2$. Clearly, $\rho(\Phi) = 0$ iff X_1 and X_2 are independent. As function $\phi_i(x)$, $i=1,2$ could take only two values $\phi_i(0)$ and $\phi_i(1)$, and these values are defined in a unique way by the conditions imposed, the following relation holds

$$\rho(\Phi) = \mathbf{E} \left[\frac{(X_1 - \mathbf{E}(X_1))}{\sigma(X_1)} \frac{(X_2 - \mathbf{E}(X_2))}{\sigma(X_2)} \right] \quad (9)$$

A straightforward computation yields:

$$\rho(\Phi) = \beta \sqrt{\frac{p(1-p)}{P(1-P)}} \quad (10)$$

where

$$\mathbf{P} = \mathbf{P}(X_2 = 0) = \alpha_2(1 - p) + (1 - \alpha_1)p = \alpha_2 + \beta p.$$

The correlation coefficient $\rho(\Phi)$ is non-negative for any unbiased test; it equals to 1 when $X_1 = X_2$.

Perhaps, the most interesting way to measure the quality of a test is to consider an information $I(\Phi)$ in the indicator of solution $X_2 = \Phi$ about the indicator of event X_1 . To formalize this idea consider

$$\begin{aligned} I(\Phi) &= \\ &= \sum_{x_1=0,1} \sum_{x_2=0,1} \mathbf{P}(x_1, x_2) \log_2 \frac{\mathbf{P}(x_1, x_2)}{\mathbf{P}(x_1)\mathbf{P}(x_2)} \end{aligned} \quad (11)$$

which equals to $I(\Phi) = S(X_1) - \mathbf{E}[S(X_1 | X_2)]$.

$S(X_i)$ stands for the Fisher information of the prior distribution $\mathbf{P}(x_i)$, $S(X_i) = p \log_2 p + (1 - p) \log_2 (1 - p)$. Intuitively, it means that in a random trial with n outcomes where hypothesis F_1 and F_2 appear with probabilities p and $1 - p$, respectively, there are $\approx 2^{nS}$ quite

probable outcomes, and the rest could be neglected as $n \rightarrow \infty$. Next, $S(X_1 | X_2)$ is the *conditional entropy* of the conditional distribution $\mathbf{P}(x_1 | x_2)$ under condition that the decision x_2 is taken.

Therefore, the information $I(\Phi)$ equals to the mean reduction of uncertainty obtained by the using of the test Φ . Clearly, for an ideal test without any errors $I(\Phi)$ takes its maximal value $S(X_1)$, and $I(\Phi) = 0$ iff X_1 and X_2 are independent.

It is interesting to note that a sequence of random events $X_1^{(n)}$ and decisions $X_2^{(n)} = 1, 2, \dots$ can be treated as a message transmission over a channel without noise. In this interpretation the observation $x_1^{(n)}(\omega)$ could be treated as the coding of the random outcome F_i , $i = 1, 2$, and the decision $x_2^{(n)}(\omega)$ as its decoding. The maximization of $I(\Phi)$ means that the optimal decoding is applied.

Now following is proved:

Lemma 1. Fix the Type I error probability α_1 . Then, the Neyman-Pearson test Φ^* minimizing the Type II error probability α_2 , maximizes also the information $I(\Phi^*)$ among all unbiased tests.

Proof. A straightforward computation yields

$$\frac{\partial I}{\partial \alpha_2} = (1-p) \left[\log_2 \left(1 - p + \frac{\alpha_1 p}{1 - \alpha_2} \right) - \log_2 \left(1 - p + \frac{(1 - \alpha_1)p}{\alpha_2} \right) \right] \quad (13)$$

This derivative is non-positive for an unbiased test with $\alpha_1 + \alpha_2 \leq 1$. Hence, the information $I(\Phi)$ is maximal for a minimal possible value of α_2 , i.e. for the test Φ^* . A symmetric statement with interchanging of α_1 and α_2 is also true. •

The same property holds also for $\beta(\Phi)$ and $\rho(\Phi)$; this is immediate for $\beta(\Phi)$, and follows from the equality

$$\frac{\partial \rho}{\partial \alpha_2} = - \frac{\alpha_1 P + (1 - \alpha_1)(1 - P)}{P(1 - P)} \sqrt{\frac{p(1 - p)}{P(1 - P)}} \leq 0 \quad (14)$$

in the case of $\rho(\Phi)$.

Next, observe that the Fisher information $I(\Phi)$ is a convex function of α_1 and α_2 for all $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, $0 < p < 1$. To prove this, it suffices to compute

$$\frac{\partial^2 I}{\partial \alpha_2^2} = p \left[\frac{P(1 - P) - p\alpha_1(1 - \alpha_1)}{\alpha_1(1 - \alpha_1)P(1 - P)} \right], \quad (15a)$$

$$\frac{\partial^2 I}{\partial \alpha_1^2} = (1 - p) \left[\frac{P(1 - P) - p\alpha_2(1 - \alpha_2)}{\alpha_2(1 - \alpha_2)P(1 - P)} \right], \quad (15b)$$

$$\frac{\partial^2 I}{\partial \alpha_1 \partial \alpha_2} = \frac{p(1 - p)}{P(1 - P)}, \quad (15c)$$

$$\det \left\| \frac{\partial^2 I}{\partial \alpha_i \partial \alpha_j} \right\| = \frac{p(1 - p)(1 - \alpha_1 - \alpha_2)^2}{P(1 - P)\alpha_1(1 - \alpha_1)\alpha_2(1 - \alpha_2)} > 0. \quad (15d)$$

This convexity property implies that a randomized test $\bar{\Phi} = \int_{\Phi \in A} \Phi \mathbf{P}(d\Phi)$ where $\mathbf{P}(d\Phi)$ is a probability measure on a suitable set A , $\mathbf{P}(A) = 1$, can not be optimal in the sense of Fisher information. Indeed, Jensen's inequality yields

$$I(\bar{\Phi}) = I \left(\int_{\Phi \in A} \Phi \mathbf{P}(d\Phi) \right) \leq \int_{\Phi \in A} I(\Phi) \mathbf{P}(d\Phi) \leq \sup_{\Phi \in A} I(\Phi) \quad (16)$$

Hence, there always exists a non-randomized test Φ' such that $I(\bar{\Phi}) \leq I(\Phi')$. Clearly, a similar statement holds for $\beta(\Phi)$ as well. However, it is not true generically for $\rho(\Phi)$.

Fig.1 demonstrates this presenting the surface $\rho(\alpha_1, \alpha_2)$ for $p = 1/3$. Fig.2 presents the surface $I(\alpha_1, \alpha_2)$ for $p = 1/3$. Each of the characteristics $\beta(\Phi)$, $\rho(\Phi)$ and $I(\Phi)$ divides the set of all tests into equivalence classes with the same value of this characteristic inside the class, and defines a linear order between different equivalence classes.

Numerical Examples

Figure 3 presents the results of computation of optimal tests with respect of three criteria as above. As hypothesis F_1 and F_2 , select the normal distributions with pdf's

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+0.5)^2}{2}\right), \\ f_2(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-0.5)^2}{2}\right), \end{aligned} \quad (17)$$

respectively. The prior probability of appearing F_1 is $p = 1/3$. Fig.3 presents the type II error $\alpha_2 = g(\alpha_1)$ for Neyman-Pearson's problem in the cases $\sigma^2 = 0.2$, $\sigma^2 = 0.6$ and $\sigma^2 = 1$. Each of these curves represents all the tests of the form

$$\Phi(x) = \begin{cases} 1, & \text{if } f_2(x) \geq \lambda f_1(x) \\ 0, & \text{if } f_2(x) < \lambda f_1(x) \end{cases} \quad (18)$$

with different λ which are optimal for Neyman-Pearson problem, i.e., minimizes α_2 for a given value of α_1 . These curves of errors serve as a boundary of a convex domain of errors for all possible tests. The points are indicated (α_1, α_2) on the boundary where each of three characteristics of quality as above achieves its maximum. Clearly, the position of maximum is distinct in all three cases.

Checking Hypothesis of Uselessness of a Test

It is desirable to use empirical data for checking the uselessness of a test Φ . In the case of independence of the indicator of event and the indicator of decision any of three characteristics of quality as above equals 0. To reject the hypothesis of useless of test Φ with some level of confidence it is sufficient to demonstrate that $\alpha_1 + \alpha_2 < 1$ with this level of confidence.

Consider a series of n independent trials where the number M of appearance of

distribution F_2 equals m , the number L of selection of distribution F_2 by the test Φ equals l , and the number of correct choices K of F_2 by the test Φ equals k . The hypothesis of uselessness is formalized in the following form of H_0 : $\alpha_1 + \alpha_2 = 1$. If H_0 is true, then (with unknown probability p)

$$\begin{aligned} P(X_1 = 0) &= p, \\ P(X_2 = 0) &= \alpha_2(1-p) + (1-\alpha_1)p \equiv \alpha_2 \end{aligned} \quad (19)$$

and X_1 and X_2 are independent.

Probabilities $P(X_2=0, X_1=0)$, $P(X_2=1, X_1=0)$, $P(X_2=0, X_1=1)$ and $P(X_2=1, X_1=1)$ are estimated by the empirical frequencies $(n-m-1+k)/n$, $(1-k)/n$, $(m-k)/n$ and k/n , respectively. Hence, the estimates of conditional probabilities $\alpha_1 = P(X_2=1 | X_1=0)$ and $\alpha_2 = P(X_2=0 | X_1=1)$ looks like $(1-k)/(n-m)$ and $(m-k)/m$, respectively.

Clearly, the inequality $k > ml/n$ corresponds to the alternative H_1 : $\alpha_1 + \alpha_2 < 1$. It means that the critical sets have the form $\{K > K^*\}$ for given values $M = m$ and $L = l$.

If hypothesis H_0 is true then the independence of X_1 and X_2 and independence of different trials imply an explicit expression for test size

$$\begin{aligned} P(K > K^*, M = m, L = l) &= \\ &= \binom{n}{m} p^m (1-p)^{n-m} \times \\ &\quad \sum_{k=K^*}^{\min\{m, l\}} \binom{m}{k} q^k (1-q)^{m-k} \\ &\quad \binom{n-m}{l-k} q^{l-k} (1-q)^{n-m-l+k} \\ &= \binom{n}{m} p^m (1-p)^{n-m} q^l (1-q)^{n-l} \\ &\quad \sum_{k=K^*}^{\min\{m, l\}} \binom{n-m}{l-k} \binom{m}{k} \end{aligned} \quad (20)$$

Figure 1. Dependence of correlation coefficient on errors.

Figure 2. Dependence of Fisher information coefficient on errors.

Figure 3. Location of maximum points for different characteristics of test efficiency.

Figure 4. Dependence of maxima locations for different characteristics of test efficiency on variance.

Fig. 4 presents the dependence of locations of maximum points $\alpha_i(\beta)$, $\alpha_i(\rho)$ and $\alpha_i(I)$, $i = 1, 2$, for three characteristics of quality as above as functions of σ^2 . Observe that the Type II error tends to 0 as $\sigma^2 \rightarrow 0$, as the PDF f_2 tends to a delta-function located at the point $x = 0.5$.

Here $q = \alpha_2$. Using the equalities

$$\mathbf{P}(M = m) = \binom{n}{m} p^m (1-p)^{n-m}, \quad (21)$$

$$\mathbf{P}(L = l) = \binom{n}{l} q^l (1-q)^{n-l},$$

obtain

$$\mathbf{P}(K \geq K^* \mid M = m, L = l) = \sum_{k=K^*}^{\min\{m, l\}} \frac{\binom{m}{k} \binom{n-m}{l-k}}{\binom{n}{l}} \quad (22)$$

This conditional probability represents the level of confidence for the critical region $\{K > K^*\}$ for given values $M = m$ and $L = l$.

References

- Barra J. R. (1981). *Mathematical basis of statistics*. New York: Academic Press.
- Rachev S. (1991). *Probability metrics and the stability of stochastic models*. Chichester: John Wiley.

Early Scholars

Type I Error Rates For A One Factor Within-Subjects Design With Missing Values

Miguel A. Padilla James Algina
Educational Psychology
University of Florida

Missing data are a common problem in educational research. A promising technique, that can be implemented in SAS PROC MIXED and is therefore widely available, is to use maximum likelihood to estimate model parameters and base hypothesis tests on these estimates. However, it is not clear which test statistic in PROC MIXED performs better with missing data. The performance of the Hotelling-Lawley-McKeon and Kenward-Roger omnibus test statistics on the means for a single factor within-subject ANOVA are compared. The results indicate that the Kenward-Roger statistic performed better in terms of keeping the Type I error close to the nominal alpha level.

Key words: Type I error, Within-subjects, missing values, Kenward-Roger omnibus test, robustness

Introduction

A common problem in multivariate analysis is the missing data problem. Data values may be missing for a variety of reasons. For example, a subject may drop out of a longitudinal study because of death or illness, or refuse to respond to sensitive questions on a survey, or neglect to finish the survey because of its length, etc. These, of course, are just a few examples of processes that might cause the missing data.

There are several methods available for use when data are missing. The statistical properties of these procedures depend on the mechanism for the missing data. Rubin (1976, 1987) and Little and Rubin (1987) defined three types of missing data mechanisms. Two of these are missing completely at random (MCAR) and missing at random (MAR).

The third type consists of all other missing data mechanisms. Verbeke and Molenberghs (2000) advocate calling this third type missing not at random (MNAR). These types of missing data mechanisms will be described in the context of the design and analysis considered in this study. The design includes p repeated measurements made on a single group of participants. The purpose of the data analysis is to estimate parameters (i.e., the means, variances, and covariances of the repeated measurements) and to test the omnibus hypothesis that the p means are equal. To simplify the presentation the case of two repeated measurements (the simplest repeated measures design) will be used in the description.

Let X_1 and X_2 be two distinct variables. The missing data mechanism is MCAR when the pattern of missing data on X_1 and X_2 is completely independent of X_1 and X_2 . The missing data mechanism is MAR if the pattern of missing data on X_2 is dependent on observed values on X_1 but not on X_2 when X_1 is held constant and the pattern of missing data on X_1 is dependent on observed values on X_2 but not on X_1 when X_2 is held constant.

So what is a researcher to do if missing data are present in his or her study? A large number of methods have been proposed for analyzing incomplete data, but the most common solutions are probably listwise deletion

Miguel A. Padilla is a NIH Graduate Student Fellow in Educational Psychology. His research interests are in applied statistics. Email: mpadilla@ufl.edu. James Algina is Professor of Educational Psychology. His interests are in psychometric theory and applied statistics. Email: algina@ufl.edu.

and maximum likelihood ignoring the missing data mechanism. In listwise deletion all subjects with any missing data are excluded from the analysis. This is the procedure used in popular software packages (e.g., SAS and SPSS) for repeated measures ANOVA and MANOVA. Listwise deletion works reasonably well if the researcher has a large sample, a small percentage of missing data, and a MCAR missing data mechanism.

For example, if the researcher has a sample of 500 and 5% have missing data, the researcher will do the analysis with a sample of 475 and, if the data are MCAR, obtain unbiased estimates while still retaining power. However, if the researcher has a sample of 100 and 35% have missing data, doing the analysis with a sample of 65 could severely compromise power. Regardless of the sample size and amount of missing data, estimates will be biased and sampling distribution based inferences, such as hypothesis tests and confidence intervals, will be invalid if the missing data mechanism is MAR or MNAR.

As noted previously maximum likelihood ignoring the missing data mechanism is another procedure that can be used when data are missing. To understand the concept of ignoring the missing data mechanism, we must recognize that there are two types of data that can be taken into account in the analysis when there are missing data.

First, there are the independent and dependent variables that are the focus of the study and, second, there is a dichotomous indicator variable indicating whether or not a particular data point is missing. The missing data mechanism is a relationship of the indicator variable to the independent variables and the dependent variables and models the probability that data are missing as a function of the independent variables and dependent variables. The relationship might be modeled, for example, as a logistic regression function relating the presence or absence of the data points to the independent and dependent variables. Analyzing only the observed scores on the dependent variables is referred to as ignoring the missing data mechanism.

Rubin (1976) has shown that if the missing data mechanism is MCAR or MAR, ML

estimators of the parameters are consistent when the missing data mechanism is ignored. Thus, the MCAR or MAR missing data mechanisms are ignorable for purposes of ML estimation. If the data are MCAR both listwise deletion and ML ignoring the missing data mechanism will produce consistent estimators, but the ML estimators will be more accurate because they use all of the available data. Rubin (1976) has also shown that the MCAR missing data mechanism is ignorable for sampling distribution based inference procedures such as hypothesis tests and confidence intervals. So if the data are MCAR either listwise deletion or ML ignoring the missing data mechanism can be used for inference, but ML will result in more powerful tests and narrower confidence intervals because it does not delete the observed data for participants that have some missing data.

When ML estimation is used, whether the MAR missing data mechanism is ignorable for sampling distribution based inference depends on the how sampling variances and covariances are calculated. The MAR missing data mechanism is ignorable for sampling distribution based inferences on the means if the sampling covariance matrix is estimated from the observed information matrix for the means and the covariance parameter estimates but not if the matrix is estimated from the portion of the observed information matrix that pertains only to the means (Kenward & Molenberghs, 1998).

The MAR mechanism may not be ignorable for sampling distribution based inferences if the sampling covariance matrix is estimated from the expected information matrix. That is, for sampling distribution based inferences to be valid the expected value of the information matrix must be taken under the actual sampling process implied by the MAR mechanism (Kenward & Molenberghs, 1998). Kenward and Molenberghs refer to using this type of expected information matrix as the unconditional sampling framework whereas using the information matrix that ignores this sampling process is called the naïve sampling framework.

Additionally, the sampling covariance matrix for the means must be computed as the inverse of the unconditional information matrix for the means and the covariance parameters.

For a design with one-within subjects factor, as well as for more complicated multivariate designs, maximum likelihood ignoring the missing data mechanism can be implemented, by using PROC MIXED on SAS. However, it should be noted that many of the test statistic options in SAS use the expected information matrix under the naïve sampling framework.

Another method for analyzing incomplete data is multiple imputation (MI) (Little & Rubin, 1987; Rubin, 1976, 1987). In MI, multiple sets of plausible values are used to replace the missing values. This creates m data sets with plausible values replacing missing values. Each of the m data sets is analyzed to produce parameter estimates. The m estimates are then combined to create a single estimate and a standard error of the estimate.

One advantage of MI is that a single set of imputed data sets can be used for a variety of analyses. Second, inferences drawn from multiply imputed data are valid, provided that the missing data mechanism is MAR or MCAR, because MI accounts for missing data uncertainty (Schafer, 1997; Schafer & Olsen, 1998). MI is very efficient in that it only requires a small set of imputed data sets to conduct a valid analysis (Rubin, 1987; Schafer, 1997; Schafer & Olsen, 1998). However, MI can be cumbersome to use because of the need to analyze multiple data sets and combine the results to make one overall inference. This drawback has been overcome for some designs because software is available that combines the estimates automatically.

As noted previously, if the missing data mechanism is MNAR, the missing data mechanism is non-ignorable (NI) for purposes of ML estimation. Thus, if the missing data mechanism is not MAR or MCAR, the pattern of missing data must be taken into account in order to obtain consistent ML estimates. This can be accomplished by using a selection model that incorporates a model for the missing data indicator or by using a pattern mixture model, which stratifies the data on the basis of the pattern of missing data. See Little (1995) for additional details about these two approaches. For examples of these models the reader is referred to Diggle and Keward (1994), Troxel (1998), Kenward (1998), Albert and Follmann

(2000), and Fitzmaurice, Laird, and Shneyer (2001).

Sampling based inferences will also be valid under selection modeling that incorporates a model for the missing data and under a pattern mixture model. However, selection modeling incorporating the missing data mechanism and pattern mixture modeling are more difficult to implement than are analyses that ignore the missing data mechanism. For example, for the design considered in this study, the analysis ignoring the missing data mechanism can be implemented using PROC MIXED in SAS, but selection modeling incorporating the missing data mechanism cannot. Thus, it seems very likely that analyses that ignore the missing data mechanism will be widely used in the future. For this reason we focus on ML methods ignoring the missing data mechanism.

Let p denote the number of levels of the within-subjects factor, Σ the $p \times p$ population covariance matrix, S the $p \times p$ estimated covariance matrix, and Σ_i and S_i the $p_i \times p_i$ section of the population and sample covariance matrices, respectively, that pertain to the dependent variables on which subject i has observed scores. In addition let A_i denote a $p_i \times p$ indicator matrix obtained by eliminating the j^{th} row from the $p \times p$ identity matrix if the data for subject i is missing on x_j . Ignoring the missing data mechanism, the generalized least squares estimate of the mean vector is

$$\hat{\mu} = \left(\sum_i A_i' \Sigma_i^{-1} A_i \right)^{-1} \left(\sum_i A_i' \Sigma_i^{-1} x_i \right) \quad (1)$$

In practice Σ_i must be estimated and the estimated sample mean vector is

$$\bar{x} = \left(\sum_i A_i' S_i^{-1} A_i \right)^{-1} \left(\sum_i A_i' S_i^{-1} x_i \right).$$

If S is obtained by maximum likelihood or restricted maximum likelihood, $\bar{x}$ is the maximum likelihood estimate.

Let C be a $(p-1) \times p$ matrix of full row rank. Each row of C is a contrast vector. The

hypothesis that all p population means are equal is

$$H_0: \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$$

where the bold zero is a vector of length $(p-1)$ with all elements equal to zero. The default test statistic in PROC MIXED for testing the null hypothesis is

$$\frac{1}{(p-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}} \quad (2)$$

with critical value $F_{\alpha, p-1, n-1}$. An alternative is to use the test statistic

$$\frac{n-p+1}{(p-1)(n-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}} \quad (3)$$

with critical value $F_{\alpha, p-1, n-p+1}$. In SAS this is referred to as the Hotelling-Lawley-McKeon (HLM) test. If there are no missing data the test statistic simplifies to the usual F transformation of Hotellings T^2 for a repeated measures design with no between-subjects factors. According to Wolfinger and Chang (1995), when data are complete and the unstructured option for the covariance matrix is selected, the default test statistic tends to be liberal with small samples sizes and the HLM performs more satisfactorily.

In equations (1) and (2), the expression

$\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-}$ is the estimated sampling

covariance matrix of the mean vector $\bar{\mathbf{x}}$ and is based on the expected information matrix calculated under the naïve sampling framework. Even when data are MCAR or there are no

missing data, using $\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-}$ has two drawbacks

1. $\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-}$ is an estimate of

$\left(\sum_i \mathbf{A}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{A}_i \right)^{-}$, the sampling covariance matrix

of $\hat{\boldsymbol{\mu}}$ in equation (1). Results by Kackar and Harville (1984) show that, as a sampling

covariance matrix for $\bar{\mathbf{x}}$, $\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-}$ tends

to be too small because it fails to take into account the uncertainty in $\bar{\mathbf{x}}$ introduced by substituting $\mathbf{S}_i$ for $\boldsymbol{\Sigma}_i$.

2. Booth and Hobert (1998) and Prasad and

Rao (1990) show that $\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-}$ is biased

for $\left(\sum_i \mathbf{A}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{A}_i \right)^{-}$.

Harville and Jeske (1992) developed a better approximation, denoted by $\hat{m}^{\circledast}$, that can be used to estimate the sampling covariance matrix of $\bar{\mathbf{x}}$. Subsequently, Kenward and Roger (1997) developed an alternative estimator, denoted by $\hat{\Phi}_A$, that can also be used to estimate the sampling covariance matrix for $\bar{\mathbf{x}}$. Kenward and Roger also proposed a test statistic, which in the context of comparing p means is

$$\frac{\lambda}{p-1} \bar{\mathbf{x}}' \mathbf{C}' \left(\mathbf{C} \hat{\Phi}_A \mathbf{C}' \right)^{-1} \mathbf{C} \bar{\mathbf{x}}$$

with critical value $F_{\alpha, p-1, df}$ where λ and df are estimated from the data. The Kenward-Roger (KR) procedure is implemented in PROC MIXED. However, $\hat{m}^{\circledast}$ is used in place of $\hat{\Phi}_A$.

The Current Study

The purpose of this article is to compare Type I error rates for two procedures available in SAS: the HLM procedure and the Kenward-Roger (KR) procedure. Simulation methods were used to make the comparison. Data were generated under the MAR and MCAR missing data mechanisms because of the properties enjoyed by ML estimation under these mechanisms if the missing data mechanism is ignored. For comparison purposes data were also generated under a MNAR missing data mechanism. None of the procedures were expected to work well under this missing data mechanism.

Related literature

Fai and Cornelius (1996) developed and compared four alternative test procedures that can be used to test linear hypotheses on means in multivariate studies. The four test statistics, specialized to the context of this paper are shown in Table 1. For each of the four statistics Fai and Cornelius showed how to use the data to estimate the second degrees of freedom. The F_2 and F_4 statistics have a scale factor estimated from the data. The F_1 and F_2 statistics use

$\left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-1}$ to estimate the covariance

matrix of the mean vector whereas F_3 and F_4 use $\hat{m}^*$. The F_4 statistic is similar to the statistic obtained by using the Kenward-Rogers option in PROC MIXED, but the formula for the scale factors and the degrees of freedom are not identical to those used when the Kenward-Rogers option is employed in PROC MIXED. The test using F_1 is available in SAS when the Satterthwaite option is used in PROC MIXED.

Fai and Cornelius (1996) applied their tests to split-plot designs with a between-subjects factor with three levels and a within-subjects factor with four levels. The covariance structure was compound symmetric. The design was unbalanced in that the number of subjects varied across levels of the between-subjects factor and data were not generated for some combinations of subjects and the within-subjects treatment. Because the missing data were never generated, the missing data mechanism was effectively MCAR. Estimated Type I error rates

and power were reported for the main effect of the between-subjects factor. All four tests provided reasonable control of the Type I error rate. The performance of F_1 and F_3 , which do not include a scale factor was very similar. Type I error rates and power for F_4 were always larger than for F_3 .

Schaalje, McBride, and Fellingham (2001), reporting on a study conducted by McBride (2002), reported Type I error rates for F_1 and the test obtained using the Kenward-Roger option in PROC MIXED. McBride investigated performance of these tests in a split-plot design.

The following provides a social science example of the design investigated by McBride. Suppose three methods for structuring interactions among students in a mathematics classroom are to be compared; n schools are randomly assigned to each method, where n was three in half of the conditions studied by McBride and five in the other half. The methods will be implemented for three, six, or nine weeks. Each school contributes K classes. Each class is assigned a single interaction quality score. In half of the conditions studied by McBride, $K = 3$ and the design was balanced. In the other half, $K = 5$ so that within each school two classes would be assigned to two of the implementation periods and one class would be assigned to the remaining implementation period. In these conditions the design is unbalanced, but no data are missing.

McBride also investigated the effect of the covariance structure, including five structures: compound symmetric (equal correlations and equal variance for the repeated measures), heterogeneous compound symmetric (equal correlations, but unequal variances for the repeated measures), Toeplitz, heterogeneous first-order autoregressive (correlations conform to a first-order autoregressive pattern, but the variances for the repeated measures are unequal), and first-order ante-dependence (see Wolfinger, 1995, for examples of these covariance structures). The results indicated that employing the Kenward-Roger option provided better control than did employing the Satterthwaite option in PROC MIXED. Type I error rates were closer to the nominal level for balanced designs than for unbalanced designs.

For unbalanced designs, Type I error rates improved as n increased.

Kenward and Roger (1997) investigated how well the original Kenward-Roger procedure controlled Type I error rates in four situations: (a) a four-treatment, two-period cross-over design, (b) a row-column- α design, (c) a random

coefficients regression model for repeated measures data, and (d) a split-plot design. In (c) and (d) there were missing data. In (c) the missing data mechanism was MCAR. The missing data mechanism in (d) was not specified. In all situations, the Kenward-Roger test controlled the Type I error rate well.

Table 1. Test Statistics from Fai and Cornelius (1996).

Test Statistics	Critical values
$F_1 = \frac{1}{(p-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}}$	$F_{\alpha, (p-1), df_1}$
$F_2 = \frac{\lambda_2}{(p-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\sum_i \mathbf{A}_i' \mathbf{S}_i^{-1} \mathbf{A}_i \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}}$	$F_{\alpha, (p-1), df_2}$
$F_3 = \frac{1}{(p-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\hat{\mathbf{m}}^{\oplus} \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}}$	$F_{\alpha, (p-1), df_3}$
$F_4 = \frac{\lambda_4}{(p-1)} \bar{\mathbf{x}}' \mathbf{C}' \left[\mathbf{C} \left(\hat{\mathbf{m}}^{\oplus} \right)^{-} \mathbf{C}' \right]^{-1} \mathbf{C} \bar{\mathbf{x}}$	$F_{\alpha, (p-1), df_4}$

Methodology

The design of the simulation had three between-subject factors and three within-subjects factors. The between subjects-factors were number of variables (p), ratio of the number of subjects to number of variables (n/p), and correlation (ρ) for each pair of variables. The number of variables factor had three levels, $p = 2, 4$ and 6 . The ratio factor had two levels, $n/p = 5$ and 10 . The actual sample sizes are presented in Table 2.

The correlation factor had three levels, $\rho = .25, .50$, and $.75$ with all pairs of variables equally correlated (compound symmetric). The within-subjects factors were type of missing data mechanism (type), percent of missing data (percent), and test statistic (test). The type of missing data mechanism factor had three levels: MAR, MCAR, and MNAR. The percent of missing data factor had two levels: 10% and 20%. Finally, the test factor has two levels: HLM and KR. All factors in the design were crossed.

Table 2. Sample Size (n) According to Number of Variables and Sample Size Ratio (n/p).

Ratio	Variable		
	2	4	6
5	10	20	30
10	20	40	60

The model used to generate the data was

$$X_{ij} = \mu + e_{ij},$$

$i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p$. In matrix terms

$$\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{ip} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{bmatrix}$$

where $\mathbf{x}$ is a $p \times 1$ vector containing the random variables for the i^{th} subject on the p variables and $\boldsymbol{\mu}$ is a $p \times 1$ vector of means, with all elements equal. All of the means are equal because the study is concerned with Type I error rates. The common element was arbitrarily set to zero. The vector $\mathbf{e}$ is a $p \times 1$ vector of random errors with the following assumption, $\mathbf{e} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. In all conditions the diagonal elements of $\boldsymbol{\Sigma}$ were equal to one.

All data simulations were conducted using SAS version 9.0. For each combination of levels of the between-subjects factors, the following steps were used to simulate the data.

1. Simulate $\mathbf{Z}$, a $n \times p$ matrix of pseudorandom standard normal variables.
2. Calculate $\mathbf{T}$, the $p \times p$ upper triangular Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}$.
3. Calculate $\mathbf{E} = \mathbf{ZT}$, an $n \times p$ matrix of error scores.
4. Set $\mathbf{X} = \mathbf{E}$
5. Copy $\mathbf{X}$ five times, yielding six copies of the data set. The six copies were used to create data matrices with missing data for the six combinations of type of missing data mechanism and percentage of missing data.
6. Select data points for elimination. In all conditions there were no data missing on x_1 .
 - a. For the MCAR missing data mechanism, x_{ij} was eliminated from the matrix if $U_{ij} < \pi$ where π is the expected proportion of missing data on x_j .
 - b. For an MAR missing data mechanism, x_{ij} was missing if
$$U_{ij} < \Phi(kx_{i1} + c), j = 2, \dots, p$$

where Φ is the cumulative standard normal distribution function and k and c are parameters that control the dependence of the missing data on the x variables and the expected proportion of missing data.

- c. For the MNAR missing data mechanism, x_{ij} was deleted if

$$U_{ij} < \Phi(kx_{ij} + c).$$

That is, the probability that x_{ij} was missing depended on x_i . All conditions were replicated 5,000 times.

Setting k and c

The parameter k controls how dependent the missing data are on x in the MAR and MNAR conditions and was set equal to one. Let $R_{ij} = 1$ if X_{ij} is missing and zero otherwise. With $k = 1$, in the MAR conditions the biserial correlation between R_j and x_1 was .5 for $j = 2, \dots, p$; in an MNAR condition the biserial correlation between x_j and R_j was .5. Thus the missing data indicators depend fairly heavily on the x variables. With $k = 1$, the expected proportion of missing data on X_{ij} is dependent on c . In the procedure described in the preceding paragraphs the probability that $R_{ij} = 1$ is related to x_j or x_1 by a normal ogive (or probit model). Using well-known facts about the normal ogive model (see, for example, Lord & Novick, 1968, equations 16.9.3 and 16.9.4), it can be shown that

$$c = \sqrt{1+k^2} \{\Phi^{-1}(\pi)\}.$$

Thus, when $k = 1$,

$$c = \sqrt{2} \{\Phi^{-1}(\pi)\}.$$

For 10% and 20% missing data the expression becomes $c = -1.28\sqrt{2}$ and $c = -.84\sqrt{2}$, respectively.

Results

For each combination of the between-subjects factors (number of variables, correlation, and sample size) and the within-subjects factors (missing data mechanism, percent of missing data, and type of test) the Type I error rates for the HLM and KR tests were estimated as the proportion of the 5000 replications that resulted in a significant test statistic. This proportion variable was then analyzed by a 3 (number of variables) $\times$ 3 (correlation) $\times$ 2 (sample size ratio) $\times$ 3 (missing data mechanism) $\times$ 2 (percent of missing data) $\times$ 2 (test) ANOVA with missing data mechanism, percent of missing data, and test type as within-subjects factors. The main effect of test was significant with $F(1, 4) = 1066.70$, $p = .000$. The mean Type I error rates for the two tests were $M_{HLM} = .083$ and $M_{KR} = .065$. Inspection of the estimated Type I error rates indicated that, with the exception of four conditions, the estimated Type I error rate for the KR test was closer to the true

Type I error rate than was the Type I error rate for the HLM test. Consequently, results for the HLM test statistic were dropped from the model and Type I error rates for the KR test statistic were reanalyzed.

The new analysis showed no significant effects for correlation. The highest-order significant interaction was the interaction of missing data mechanism, percent missing data, and sample size ratio, $F(2, 8) = 15.58$, $p = .002$.

In addition the main effect of number of variables was significant, $F(2, 4) = 23.10$, $p = .006$. Because of this pattern of effects we present, in Table 3, the Type I error rates averaged over levels of the correlation factor. Bradley (1978) presented a conservative and liberal criterion for identifying conditions in which hypothesis testing procedures work adequately. His conservative criterion is $.9\alpha \leq \tau \leq 1.1\alpha$ ($.045 \leq \tau \leq .055$) and his liberal criterion is $.5\alpha \leq \tau \leq 1.5\alpha$ ($.025 \leq \tau \leq .075$). For this study, the liberal criterion was used to identify conditions in which the average Type I error rate was unacceptable. These are indicated in bold in Table 3.

Inspection of the results indicates that, as expected, Type I error rates for the KR test may be unacceptable when the missing data mechanism is MNAR. It appears that the error rate for the KR test is more likely to be unacceptable as the percent of missing data, sample size ratio, and number of variables increases. In regard to the effect of the number of variables, in our simulation the number of variables on which data were MNAR increased as the number of variables increased. Different results might have emerged if there had been missing data on only one of the variables, regardless of the number of variables.

When the data were MCAR or MAR, average Type I error rates were acceptable in all conditions. Inspection of the Type I error rates for individual cells in the design (i.e., not collapsing over correlation) indicated that when the data were MCAR or MAR, the Type I error rate was acceptable in all conditions. Reanalysis of the data, after dropping the results for MNAR conditions indicated that number of variables did not have a significant main effect and did not enter into any significant interactions.

Table 3. Mean Type I Error Rates for KR by Number of Variables, Sample Size Ratio, Percent of Missing Data, and Missing Data Mechanism

Number	Ratio	Percent	MCAR	MAR	MNAR
2	10	10	0.051	0.050	0.053
		20	0.061	0.052	0.068
	20	10	0.048	0.050	0.066
		20	0.049	0.057	0.098
4	10	10	0.053	0.049	0.063
		20	0.055	0.060	0.072
	20	10	0.052	0.054	0.072
		20	0.050	0.061	0.146
6	10	10	0.051	0.048	0.060
		20	0.058	0.059	0.096
	20	10	0.050	0.054	0.082
		20	0.052	0.062	0.184

Note. Each mean Type I error rate is an average of Type I error rates for three conditions. Unacceptable mean Type I error rates are in boldface.

The only significant effects were a two-way interaction of type of missing data and sample size ratio, $F(1,4) = 8.25$, $p = .045$, and a main effect of percent of missing data, $F(1,4) = 15.45$, $p = .017$. Average Type I error rates by type of missing data and sample size are presented in Table 4.

The results suggest that increasing the sample size ratio improves control of the Type I error rate when the data are MCAR, but not when the data are MAR. The means when 10% and 20% of the data were missing and the mechanism was MCAR or MAR were .051 and .056, respectively, suggesting that Type I error rate control declines as the percentage of missing data increases.

Table 4 Mean Type I Error Rates for KR by Sample Size Ratio and Missing Data Mechanism.

Ratio	MCAR	MAR
10	0.055	0.053
20	0.050	0.056

Conclusion

The aim of this study was to determine whether, when there are missing data and the sample size is small, using ML estimates of the means for a single factor repeated measures design in testing the omnibus hypothesis results in control of the Type I error rate. The specific methods used to test the hypothesis were the KR test and the HLM test as implemented in SAS. The results clearly showed that KR test provided better control of the Type I error rate than did the HLM test.

The results of this study support the conclusion that, in a single-factor repeated measures design, sampling distribution based inferences on the means using the KR test may not control the Type I error rate for the MNAR missing data mechanisms but do control the Type I error rate for the MCAR and MAR missing data mechanisms. However, sample size and percent of missing data may be key factors that affect ML based inferences for MCAR and MAR missing data conditions using the KR test.

For both MCAR and MAR data, the results suggest that increasing the percent of missing data tends to inflate the Type I error rates. The effect of increasing the sample size depended on the missing data mechanism, with a stabilizing effect when the data were MCAR, but not when the data were MAR.

Although the design investigated in this study was a simple one factor repeated measures design, the findings suggest further simulation work on using ML to directly estimate models with missing data with more complicated designs and with additional variation in the factors investigated in this study. One condition that can be introduced is a between-subjects factor. Designs with between-subjects factors and within-subjects factors, also known as split-plot designs, are even more common than the one investigated in this study. Split-plot designs are used in longitudinal studies with two or more treatment groups. In such designs, the number of time point at which observations are made may be larger than six, which is the largest number of measurements investigated in this study. Consequently, a repeated measures factor, with more levels than six, should be investigated in future work.

Although several correlation matrices were used in this study and the correlation matrix had little or no impact on the Type I error rate, in each correlation matrix the off-diagonal elements were the same (i.e., the matrices were compound symmetric). This type of matrix may occur in studies in which the levels of the within-subjects factor are treatments and the order of the treatment has been randomized. Nevertheless, the exclusive use of compound symmetric correlation matrices may have limited the generality of the results. And, in other repeated measures studies (e.g., longitudinal studies) the correlation matrix is not likely to be compound symmetric. Thus, another condition that can be fruitfully investigated in future work is correlation matrices that have varying off-diagonal elements.

The Type I error rates of the KR test were acceptable in both the MCAR and MAR conditions. However, the percent of missing data at which the KR test will begin to breakdown is still not clear, nor is it clear whether sample sizes larger than those studied in this research will improve the Type I error rate for the KR test applied to MAR data. Consequently, future work should increase both the sample size ratio and percent of missing data beyond what was used in this study.

Last, recall that in the MAR missing data mechanism the missing data pattern on one variable is related to or dependent on another variable in the model but not to the variable itself. Therefore, one question that can be asked is how does the KR test statistic perform with different degrees of dependence? So another condition that can be investigated in future work is different degrees of dependence for the MAR condition.

References

- Albert, P. S., & Follmann, D. A. (2000). Modeling repeated count data subject to information dropout. *Biometrics*, 56, 667-677.
- Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Diggle, P. D., & Keward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics*, 43, 49-93.

- Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65, 457-481.
- Fai, H. T., & Cornelius, P. L. (1996). Approximate F-tests for multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, 54, 363-378.
- Fitzmaurice, G. M., Laird, N. M., & Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine*, 20, 1009-1021.
- Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.
- Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, 81, 701-708.
- Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, 17, 2723-2732.
- Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236-247.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.
- 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: John Wiley & Sons.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.
- McBride, G. B. (2002). Statistical methods helping and hindering environmental science and management. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 300-305.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.
- Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 512-524.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33, 545-571.
- Troxel, A. B. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.
- Verbeke, G., & Molenbergh, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.
- Wolfinger, R. D., & Chang, M. (1995). *Comparing the SAS GLM and Mixed procedures for repeated measures*. Proceedings of the Twentieth Annual SAS Users Groups Conference.

Confidence Elicitation And Anchoring In The Respondent-Generated Intervals (RGI) Protocol

LiPing Chu
Department of Statistics
University of California, Riverside

S. James Press
Department of Statistics
University of California, Riverside

Judith M. Tanur
Department of Sociology
SUNY, Stony Brook

The Respondent-Generated Intervals protocol (RGI) has been used to have respondents recall the answer to a factual question by giving not only a point estimate but also bounds within which they feel it is almost certain that the true value of the quantity being reported upon falls. The RGI protocol is elaborated in this article with the goal of improving the accuracy of the estimators by introducing cueing mechanisms to direct confident (and thus presumably accurate) respondents to give shorter intervals and less confident (and thus presumably less accurate) respondents to give longer ones.

Key words: RGI, surveys, Bayes, hierarchical models, record-checks, anchoring

Introduction

The Respondent-Generated Intervals protocol (RGI) has been used to have respondents recall the answer to a factual question by giving not only a point estimate but also bounds within which they feel it is almost certain that the true value of the quantity being reported upon falls (Press, 2004). This paper reports on new thinking that aims to elaborate the RGI protocol with the goal of improving the accuracy of the estimators derived from the protocol.

There are two aspects to the new thinking. The first is a new analytical Bayesian procedure for estimating the population mean in an RGI survey; it is derived in the Appendix.

LiPing Chu is a doctoral candidate in applied statistics at the University of California, Riverside. Her dissertation centers on the robustness of Respondent Generated Intervals. Email: Liping13@netzero.com. Dr. S. James Press is a Distinguished Professor of Statistics, at the University of California, Riverside. Email: jpress@ucr.edu. Dr. Judith M. Tanur is a Distinguished Teaching Professor at the State University of New York, Stony Brook. Email: jtanur@notes.cc.sunysb.edu.

The second is a new type of anchoring questioning technique that cues and encourages confident (and presumably accurate) respondents to give short intervals and less confident (and presumably less accurate) respondents to give long intervals. The new analytical procedure is summarized briefly in the next section (and elaborated in the Appendix), followed by a section containing a discussion of a classroom survey experiment and how it incorporates the new questioning technique. The final section provides a discussion of the implications of these innovations.

Vague Prior Bayesian Point Estimator for the Population Mean

For a sample of n independent respondents in a survey, let y_i, a_i, b_i denote the basic usage quantity response, the lower bound response for where the true value to the question lies for that respondent, and the upper bound response for where the true value to the question lies for that respondent, respectively, of respondent $i, i = 1, \dots, n$. Suppose that the y_i 's are all independent and normally distributed. Suppose also that we adopt a vague prior distribution for the population mean, θ_0 , to represent knowing little, a priori, about the value of the population mean. It is shown in Press (2004) using a hierarchical Bayesian model, that

in such a situation, the posterior distribution of θ_0 is given by:

$$(\theta_0 | data) \sim N(\tilde{\theta}, \omega^2), \quad (2.1)$$

where the posterior mean, $\tilde{\theta}$, is expressible as a weighted average of the y_i 's, and the weights are dependent upon the intervals defined by the bounds, the smaller the interval the larger the weight. The posterior variance is denoted by ω^2 . The posterior mean is expressible as:

$$\tilde{\theta} = \sum_1^n \lambda_i y_i, \quad (2.2)$$

where the λ_i 's are non-negative weights that are given approximately by:

$$\lambda_i \doteq \frac{\left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(b_0 - a_0)^2}{k_2^2}} \right)}{\sum_{i=1}^n \left(\frac{1}{\frac{(b_i - a_i)^2}{k_1^2} + \frac{(b_0 - a_0)^2}{k_2^2}} \right)}, \quad \sum_1^n \lambda_i = 1, \quad (2.3)$$

where: $a_0 \equiv \min_{1 \leq i \leq n}(a_i)$; $b_0 \equiv \max_{1 \leq i \leq n}(b_i)$. The interval $(b_0 - a_0)$ represents the full range of opinions the n respondents have about the possible true values of their answers to the question, from the smallest lower bound to the largest upper bound. In equation (2.3), k_1 and k_2 denote pre-assigned multiples of standard deviations that correspond to how the bounds should be interpreted in terms of standard deviations from the mean. For example, for normally distributed data it is sometimes assumed that such lower and upper bounds can be associated with 2 standard deviations below, and above, the mean, respectively. With this

interpretation, take $k_1 = k_2 = 4$ to represent the length of the interval between the largest and smallest values the true value of the answer to the recall question might be for respondent i . If desired, take $k_1 = k_2 = k$, and then make a choice among reasonable values, such as: $k = 2, 4, 5, 6, 7, 8$, and study how the estimate of the population variance varies with k .

The new estimating procedure used here substitutes for $(b_0 - a_0)$:

$$4\tau \doteq (b_0 - a_0) + \frac{2}{\sqrt{n}}(s_a + s_b)$$

to form what will be called the extended range estimator, and

$$4\tau \doteq (\bar{b} - \bar{a}) + \frac{2}{\sqrt{n}}(s_a + s_b)$$

to form what will be called the extended average estimator (see Appendix). Here $\bar{b}$ and $\bar{a}$ are the means of the upper bounds and of the lower bounds given by the respondents, respectively; and s_a and s_b are the sample standard deviations of the lower bounds and upper bounds, respectively.

Methodology

The Classroom Survey: Confidence and Question Wording

Because point estimates of respondents who give short intervals are weighted more heavily in the Bayesian RGI estimator than are point estimates of respondents who give longer intervals (see 2.3), it is advantageous to encourage respondents who are more accurate to give shorter intervals and respondents who are less accurate to give longer ones. It is known from earlier uses of the RGI procedure that, among respondents who do not receive any special guidance about the length of their intervals, there is a substantial correlation between interval length and accuracy (with less accurate respondents giving longer intervals; Press & Tanur, 2003). There is also a

correlation between confidence and interval length (with less confident respondents giving longer intervals; Press & Tanur, 2002). The aim is to increase the correlation between accuracy and interval length, by working through respondents' confidence and cueing them appropriately. We have developed a questioning protocol that aims to increase that correlation.

First, the respondent is requested to give his/her best guess about the quantity being investigated, and then is asked how confident s/he is of that answer on a scale from 0 (least confident) to 10 (most confident). Figure 1 shows the form of this confidence scale for a question used in our experiment involving recall of the respondent's grade on a classroom exam. Respondents who represent themselves as highly confident (confidence ratings 7.5 or 10) are directed to a question that encourages them to give a narrow bounding interval. Less confident respondents (confidence ratings of 5 or less) are directed to a question encouraging a wide bounding interval.

The design for this experimental application of the new protocol used three versions of the bounding questions (and each version was completed by a different group of respondents). Version 1, referred to as unanchored, simply asks the respondent to give a narrow, or a broad, interval; this version was administered to Group 1. See Figure 2 for the wording of Version 1 for the question about the classroom exam. Version 2, administered to Group 2, which is referred to as the narrow-wide anchored condition, not only encourages respondents to give narrow or wide intervals, but it also tells them that the narrow interval should be no more than a specified width and that the wide interval should be at least a specified width. See Figure 3 for the wording of Version 2 as used for the question about the classroom exam. Version 3 (referred to as the wide-wide anchored condition and administered to Group 3) is the same as Version 2, except that the suggested width of the wide interval was considerably wider (see Figure 4).

Figure 1. New Form for RGI Protocol.

- 1) What is your best guess as to what your score was on your first exam in this class? (Please don't answer if you've missed the first exam)._____.
- 2) How confident are you about your answer to Question 1? Please answer on the following confidence scale. (Place a check in the first column next to the answer you prefer.)

Confidence Scale

Place a check somewhere in this column	Numerical Score	Interpretation of confidence rating	Which question should I answer next?
	0	I have absolutely no idea what my exam score was	Go to Question 3b
	2.5	I am uncertain what my exam score was	Go to Question 3b
	5.0	I might be right and I might be wrong about what my exam score was	Go to Question 3b
	7.5	I think that I know what my exam score was	Go to Question 3a
	10.0	I am absolutely certain what my exam score was	Go to Question 3a

Figure 2: Unanchored Bounds Condition.

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. Please fill in:

The smallest my exam score could have been is %_____.

The largest my exam score could have been is %_____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score

Please fill in:

The smallest my exam score could have been is %_____.

The largest my exam score could have been is %_____.

Figure 3: Narrow Wide Anchor-Type Bounds Question.

3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. For example, if your best guess about your exam score is 75%, give a narrow interval of no more than 4 points in length, such as (73%, 76%).

Please fill in:

The smallest my exam score could have been is %_____.

The largest my exam score could have been is %_____.

NOW GO TO QUESTION 4.

3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score. For example, if your best guess about your exam score is 75%, give a wide interval of at least 20 points in length, such as (65%, 85%).

Please fill in:

The smallest my exam score could have been is %_____.

The largest my exam score could have been is %_____.

Figure 4: Wide Wide Anchor-Type Bounds Question.

- 3a) If your answer to Question 2 is 7.5 or more, please give the **smallest possible interval** in which you believe that the exam score is included. For example, if your best guess about your exam score is 75%, give a narrow interval of no more than 4 points in length, such as (73%, 76%).

Please fill in:

The smallest my exam score could have been is _____,

The largest my exam score could have been is _____.

NOW GO TO QUESTION 4.

- 3b) If your answer to Question 2 is 5 or less, give a **sufficiently wide interval** so that the interval will most likely include the actual exam score. For example, if your best guess about your exam score is 75%, give a wide interval of at least 30 points in length, such as (60%, 90%).

Please fill in:

The smallest my exam score could have been is _____,

The largest my exam score could have been is _____.

Figure 5: Memory Evaluation Scale.

Does it ever happen that when you are sure you know something, it turns out that you are mistaken?
Please check one:

Never _____

} Good Memory

Seldom _____

Sometimes _____

} Poor Memory

Frequently _____

Ratings of Memory

Respondents were asked to evaluate their memory on the scale shown in Figure 5 (The designations “Good Memory” and “Bad Memory” as shown in Figure 5 did not appear in the questionnaire given to the respondents). If respondents are good judges of their own memory, then perhaps rather than asking confidence questions for each survey item we can use a procedure that simply classifies respondents into good memory and poor memory groups and encourage good memory respondents to give short intervals and poor memory respondents to give long ones. Such a procedure would impose considerably less respondent burden than does asking for confidence for each question.

The Survey

In the spring of 2003 we ran a small experimental record-check survey in an undergraduate, lower division, statistics class at the University of California at Riverside. In a randomized design three groups of students were each given a different version of the questionnaire and the students were asked to recall their midterm exam score, their score on their second homework assignment, and the amount they had paid at the beginning of the quarter as a registration fee. Because there were three versions of the questionnaire, and because participation was voluntary, sample sizes in the three groups were rather small, but sufficiently large for us to derive some preliminary results. (A similar experiment from a larger class was run several months later in the fall of 2003 – results will be available shortly.) With the students’ permission we were able to compare their reported grades with those recorded in the professor’s grade book; the registration fee was fixed by the university for all full-time students at \$239.

Results

The first finding was that the manipulation worked. Table 1 shows that the mean length of intervals generated by respondents who were

asked to give a wide interval were always wider than those from respondents asked to give a narrow interval. In every case in which a t-test was possible (that is, whenever both group sizes were greater than 1) this finding reached at least marginal statistical significance, in spite of the small sample sizes.

For both the homework question and the midterm question, the mean of the wide intervals for respondents given the wide-wide anchor was longer than the mean of the intervals for respondents given the narrow-wide anchor. This relationship did not hold for the question about registration fee, for which most respondents seem to have been very much lacking in knowledge about how much the actual fee was (which resulted in low confidence).

It is interesting to note that there seems to be a relationship between respondents’ confidence and the salience of the question. A large majority of respondents were quite confident that they remembered their midterm grade correctly, a large majority lacked such confidence for the registration fee, and for the homework grade the respondents split about half and half.

Table 2 further checks the manipulation, asking whether there was indeed a correlation between respondents’ confidence in the accuracy of their recall and their actual accuracy in reporting their usage quantities. The actual accuracy is measured as the absolute value of the differences between the reported usage quantity and recorded truth. Large values of these differences represent *inaccuracy*. If there is a relationship between accuracy and confidence, negative correlations would be expected, as indeed are indicated in Table 2. (We might have labeled the absolute value of the difference between truth and the usage quantity as *inaccuracy*, but calling it *accuracy* simplifies our discussion as long as the reader keeps in mind how the variable is measured and that we hope for negative correlations between it and interval length.)

Table 1: Manipulation Check.

Average Lengths of Intervals for Wide and Narrow Anchors.

		Narrow Interval	n	Wide Interval	n	p value** (1-tailed)
Midterm						
	Unanchored	6.7	19	14.6	8	0.006
	Narrow/Wide Anchor	9.0	18	16.7	3	0.069
	Wide/WideAnchor	8.5	19	25.0	3	0.026
RegFee						
	Unanchored	165.00	5	1280.40	19	0.004
	Narrow/Wide Anchor	0.00	1	763.90	13	*
	Wide/WideAnchor	20.00	1	608.33	18	*
Homework						
	Unanchored	2.9	12	6.6	10	0.033
	Narrow/Wide Anchor	2.8	10	6.4	10	0.030
	Wide/WideAnchor	2.7	9	7.6	8	0.020

*Narrow interval group n=1; no test of significance possible.

**p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers.

Table 2: Correlations between “*r*” Confidence and “Accuracy” (|usage-truth|).*

<i>All respondents</i>				
		<i>r</i>	<i>n</i>	p value** (1-tailed)
Midterm				
	Unanchored	-0.110	27	0.305
	Narrow/Wide Anchor	-0.518	21	0.008
	Wide/WideAnchor	-0.443	23	0.017
RegFee				
	Unanchored	-0.111	27	0.291
	Narrow/Wide Anchor	-0.049	14	0.434
	Wide/WideAnchor	-0.375	21	0.047
Homework				
	Unanchored	-0.385	20	0.047
	Narrow/Wide Anchor	-0.355	20	0.062
	Wide/WideAnchor	-0.184	17	0.289

*We expect high levels of confidence to go with greater accuracy (small error, the absolute difference between the usage quantity offered by the respondent and truth); thus increasing confidence should go with decreasing error. Hence, if our hypothesis is correct, the correlations should be negative. They are.

**p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. p-values in bold-face are significant at the 5% level or smaller.

Although these correlations are hardly enormous, there is a relationship between accuracy and confidence in all cases. Each group of respondents contributed at least one low correlation – the unanchored group showing a low correlation for both the midterm question and the registration fee question, the narrow-wide anchor group showing a low correlation for the registration fee question, and the wide-wide anchor group showing a low correlation for the homework question.

Hence, the low correlations cannot be attributed either to a particular group of respondents or to the difficulty of a particular question. It is suspected, however, that the correlations coming from the registration fee question are influenced by the fact that very few respondents were confident about their answers to this question – see the n 's in Table 1. We speculate that respondents knew more about the total fees they paid than about the specific registration fee, about which they knew almost nothing, so they guessed wildly. There is also some evidence from student comments that if their parents paid their fees or if they received financial aid, they have little knowledge about the amount of any fees.

Table 3 examines the relationship between interval length and accuracy (measured as explained above, that is, as inaccuracy). If, as hoped, respondents who are less accurate give longer intervals, positive correlations would be expected. The correlations in Table 3 are all positive. There are two panels for Table 3 – the top panel includes all respondents who gave the 4 pieces of data requested – confidence rating, usage quantity, lower bound, and upper bound – and whose usage quantity properly fell within the bounds.

The bottom panel includes only what is called obedient respondents – those who followed the directions given in the anchoring instructions and gave a wide interval at least as wide as prescribed, or a narrow interval at least as narrow as prescribed. Two comments are in order for this table. First, it seems to have been successful in increasing the correlation between interval length and accuracy from the level obtained from respondents without any special instructions regarding interval length. Most of these correlations are larger than those reported

in Press and Tanur (2002), where the median of 18 correlation coefficients (for 18 items) was 0.13; 6 of the 18 were negative; and the only correlations exceeding 0.40 were those relating to the frequencies of behaviors (a case where those who really had no occurrences of the requested behavior could easily remember that they had none, and could be quite confident about their recall).

Second, limiting ourselves to obedient respondents seems to be useful. (Note that, because the unanchored group was not given a suggested length of interval, the obedient vs. disobedient distinction does not pertain to this group and the data for this group in the lower panel of Table 3 simply repeat the data in the upper panel.) When we omit those respondents who were disobedient we find that the correlations never decrease substantially and two correlations that were originally small increase considerably.

Table 4 shows the results of the estimation process using the Bayesian estimators for the obedient respondents only. In Table 4 the estimator that is closest to the truth is presented in boldface. We see that although all the estimates were very close to one another, the extended average estimator is closest to truth for the midterm grades and for the registration fee. The sample mean seems to work best for the homework question, except for the unanchored condition where the extended range estimate is a tiny bit closer to truth.

Note that the median correlation between accuracy and interval length for the midterm question is 0.349; for the registration fee question it is 0.395; but for the homework question it is only 0.274. Hence we should not be surprised that the Bayesian estimator works better for the midterm and registration fee questions than it does for the homework question. The findings for the extended average estimator are also shown graphically in Figure 6. What is graphed is the absolute value of the difference between the RGI estimated value and average truth. G1 refers to the groups in the unanchored condition, G2 to groups in the narrow-wide anchor condition, and G3 refers to the groups in the wide-wide anchor condition.

Table 3: Manipulation Check

Correlations “r” between Interval Length and Accuracy ($|\text{usage-truth}|$).*All Respondents with Useable Data*

	r	n	p value* (1-tailed)
Midterm			
Unanchored	0.311	25	0.065
Narrow/Wide Anchor	0.149	19	0.272
Wide/WideAnchor	0.069	22	0.381
RegFee			
Unanchored	0.395	24	0.028
Narrow/Wide Anchor	0.671	12	0.008
Wide/WideAnchor	0.286	19	0.128
Homework			
Unanchored	0.011	20	0.482
Narrow/Wide Anchor	0.273	18	0.138
Wide/WideAnchor	0.320	17	0.105

Obedient Respondents Only

	r	n	p value* (1-tailed)
Midterm			
Unanchored	0.311	25	0.065
Narrow/Wide Anchor	0.624	11	0.020
Wide/WideAnchor	0.349	14	0.110
RegFee			
Unanchored	0.395	24	0.028
Narrow/Wide Anchor	0.638	9	0.032
Wide/WideAnchor	0.247	16	0.178
Homework			
Unanchored	0.011	20	0.482
Narrow/Wide Anchor	0.274	12	0.194
Wide/WideAnchor	0.305	11	0.181

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. p-values in bold-face are significant at the 5% level or smaller.

Table 4: Point Estimate Results Using Vague Prior and Extended Average and Extended Range Procedures.

<i>Obedient Respondents Only</i>						
		n	Average Truth	x-bar	Extended Average	Extended Range
Midterm						
	Unanchored	25	83.88	83.04	83.79	83.17
	Narrow/Wide Anchor	11	81.36	79.64	79.94	79.70
	Wide/WideAnchor	14	86.57	86.71	86.68	86.70
RegFee						
	Unanchored	24	\$239.00	\$1,366.46	\$1,202.25	\$1,328.28
	Narrow/Wide Anchor	9	\$239.00	\$1,090.78	\$974.77	\$1,047.50
	Wide/WideAnchor	16	\$239.00	\$1,190.88	\$1,122.65	\$1,176.97
Homework						
	Unanchored	20	16.91	18.00	18.06	17.99
	Narrow/Wide Anchor	12	16.72	17.92	18.04	18.01
	Wide/WideAnchor	11	16.69	18.63	18.81	18.65

Table 5: Average Confidence Scores by Respondents' Memory Rating.

<i>All Respondents</i>						
		Good Memory	n	Poor Memory	n	p value* (1-tailed)
Midterm						
	Unanchored	8.75	6	7.05	22	0.028
	Narrow/Wide Anchor	9.58	6	7.83	15	0.017
	Wide/WideAnchor	8.21	7	9.22	16	
RegFee						
	Unanchored	4.29	7	3.50	20	0.257
	Narrow/Wide Anchor	4.00	5	2.78	9	0.266
	Wide/WideAnchor	3.33	6	1.17	15	0.074
Homework						
	Unanchored	7.50	6	6.84	19	0.276
	Narrow/Wide Anchor	4.58	6	6.25	14	
	Wide/WideAnchor	4.50	5	6.35	13	

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. Pairs of numbers in bold-face are consistent with out hypothesis that good-memory respondents have higher confidence than poor-memory respondents. P-values in bold-face are significant at the 5% level or better.

Figure 6: Bias ($|\text{estimate} - \text{truth}|$) for Extended Average Bayesian Estimate Compared with ABS Bias (Absolute Error) of Sample Mean.

Table 6: Accuracy (|usage-truth|) by Respondents' Memory Rating.

All Respondents						
	Good Memory	n	Poor Memory	n	p value* (1-tailed)	
Midterm						
Unanchored	2.33	6	5.10	21	0.134	
Narrow/Wide Anchor	7.33	6	6.00	16		
Wide/WideAnchor	0.57	7	0.69	16	0.424	
RegFee						
Unanchored	961	7	1245	19	0.253	
Narrow/Wide Anchor	685	5	1519	10	0.112	
Wide/WideAnchor	1405	6	729	15		
Homework						
Unanchored	1.50	5	1.60	15	0.453	
Narrow/Wide Anchor	3.00	6	2.13	15		
Wide/WideAnchor	2.90	5	2.26	13		

*p-values are included for those readers for whom p-values are sensible. We recognize this will not be true for all readers. Pairs of numbers in bold-face are consistent with out hypothesis that good-memory respondents are more accurate than poor-memory respondents.

Table 5 relates respondents' ratings of their memory to their confidence as rated on the confidence scales for the questions. Those respondents who rated their memory good (those who claimed never or seldom to be mistaken when sure they knew something) in many cases give higher average confidence ratings than respondents who say their memory is less good (those who claimed sometimes or frequently to be mistaken when they were sure they knew something). This finding holds true for all questions for the unanchored-type condition, for the midterm and the registration fee questions for the narrow-wide anchored-type condition, and only for the registration fee for the wide-wide anchored-type condition.

Confidence ratings were higher on average for the good memory group than in the poor memory group in 6 of the 9 comparisons. Two of these 6 wins reached statistical significance at conventional levels and another was marginally significant.

Table 6 shows the accuracy achieved by respondents at different levels of self-rated memory. Note that accuracy is again measured by the absolute value of the difference between a respondent's reported usage quantity and truth. Thus large values represent inaccuracy, and smaller values are more accurate. We see that on the average respondents who rated themselves to have good memories were closer to the truth in 5 of the 9 possible comparisons (shown in boldface). None of these wins reached statistical significance at conventional levels.

Conclusion

There was some success with these new directions. We seem to have affirmed the need to ask confidence questions separately for each usage quantity, for while respondents' estimates of their own memory seem to be good predictors of that confidence, those memory estimates do not relate nearly as well to actual accuracy as do the confidence ratings themselves. It was hoped to minimize respondent burden by asking a single memory question, but it seems the burden of asking separate confidence questions is a necessary one.

We have established that respondents directed to give wide intervals give wider ones on the average than do respondents directed to give narrower ones. There does not seem to be much effect of the length of the anchoring-type interval, but the results of a considerably larger sample size experiment is necessary to see if that lack of effect is real. The correlation between accuracy and interval length was improved through the use of the confidence scale. It would be useful to increase that correlation even more, as it is the *sine qua non* for the successful application of the RGI protocol.

Other methods will be used to ask for respondents' confidence, but it will be limited by any imperfections in respondents' understanding of their own accuracy. Respondents who are honestly confident but nevertheless inaccurate, and respondents who honestly lack confidence but are nevertheless accurate, will continue to haunt us. Even in this test, however, it was apparent that the manipulation of respondents' interval length, based on their confidence, results in the RGI Bayesian estimator showing less bias than the sample mean in a majority of the cases examined.

References

- Press, S. J. (2004). Respondent-Generated Intervals (RGI) for Recall in Sample Surveys. *Journal of Modern Applied Statistical Methods*, 3, 104-116.
- Press, S. J., & Tanur, J. M. (2004). Relating RGI Questionnaire Design to Survey Accuracy and Response Rate. Special Issue on Questionnaire Development Evaluation. And Testing Methods. *Journal of Official Statistics*, (June, 20), 265-287.
- Press., S. J., & Tanur, J. M. (2003). The Relationship between Accuracy and Interval Length in the Respondent-Generated Interval Protocol, *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Washington, D. C.: American Statistical Association.
- Press, S. J., & Tanur, J. M. (2002). Cognitive and Econometric Aspects of Responses to Surveys as Decision-Making. *Technical Report #271*, Department of Statistics, University of California at Riverside, Riverside, CA 92521-0138.

APPENDIX

Each of n respondents in a sample survey provides a triple of data: (y_i, a_i, b_i) representing respondent i 's usage quantity (the term "usage" was introduced originally to reflect estimated frequency of a behavior), her/his lower bound (for true value of the usage), and his/her upper bound (for true value of the usage); $i = 1, \dots, n$. These quantities are jointly distributed. Suppose that marginally:

$$1) \quad (y_i | \theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2);$$

$$2) \quad (a_i | a_{i0}, \psi_{ai}^2) \sim N(a_{i0}, \psi_{ai}^2);$$

$$3) \quad (b_i | b_{i0}, \psi_{bi}^2) \sim N(b_{i0}, \psi_{bi}^2),$$

where θ_i denotes the true population value for the mean usage for respondent i ; a_{i0} , b_{i0} denote the true population values for respondent i 's lower and upper bounds, respectively; and $(\sigma_i^2, \psi_{ai}^2, \psi_{bi}^2)$ denote the corresponding population variances, respectively.

Note that although (y_i, a_i, b_i) are observed quantities, $(\theta_i, \sigma_i^2, a_{i0}, b_{i0}, \psi_{ai}^2, \psi_{bi}^2)$ are unknown and unobservable. Now, assume that the θ_i 's are exchangeable, and

$$4) \quad (\theta_i | \theta_0, \tau^2) \sim N(\theta_0, \tau^2).$$

Assuming $(\sigma_1, \dots, \sigma_n, \tau)$ are known, it has already been shown, adopting a vague prior on θ_0 , gives as the posterior distribution for θ_0 (see Press, 2004):

$$5) \quad (\theta_0 | y, \sigma_1, \dots, \sigma_n, \tau) \sim N\left(\sum_{i=1}^n \lambda_i y_i, \omega^2\right),$$

$$0 \leq \lambda_i \leq 1, \quad \sum_{i=1}^n \lambda_i = 1.$$

The λ_i 's and ω^2 are proportions of total precision. The development for a normal (rather than a vague) prior distribution on the population mean is simple and unchanged by the sequel.

Assessment of the variances

A) Assessment of the σ_i 's

Take $4\sigma_i \doteq b_i - a_i$, $i = 1, \dots, n$, as our assessment for the σ_i 's.

B) Assessment of τ .

Assume there are approximate bounds for all subjects in the population that are approximately 2 standard deviations on either side of the mean. Then, define:

$$a^* = \frac{1}{N} \sum_{i=1}^N a_{i0}; \quad b^* = \frac{1}{N} \sum_{i=1}^N b_{i0};$$

$$\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i; \quad \bar{b} = \frac{1}{n} \sum_{i=1}^n b_i,$$

where: a^*, b^* are averages of the *true* (unobserved) values of these bounds over the entire population; $\bar{a}, \bar{b}$ are the averages of the *observed* values of the bounds over the sample.

Assume:

$$\psi_{a1}^2 = \psi_{a2}^2 = \dots = \psi_a^2; \quad \psi_{b1}^2 = \psi_{b2}^2 = \dots = \psi_b^2.$$

Then,

$$6) \quad \bar{a} \sim N\left(a^*, \frac{\psi_a^2}{n}\right); \quad \bar{b} \sim N\left(b^*, \frac{\psi_b^2}{n}\right).$$

Next note that the true population mean value for respondent i must be between its bounds,

$$7) \quad a^* \leq \theta_0 \leq b^*.$$

Case 1: Extended Average Estimator

For 95% credibility on a^* :

$$8) \quad \bar{a} - 2 \frac{\psi_a}{\sqrt{n}} \leq a^* \leq \bar{a} + 2 \frac{\psi_a}{\sqrt{n}};$$

for 95% credibility on b^* :

$$9) \quad \bar{b} - 2 \frac{\psi_b}{\sqrt{n}} \leq b^* \leq \bar{b} + 2 \frac{\psi_b}{\sqrt{n}}.$$

From (7), (8) and (9):

$$\bar{a} - 2 \frac{\psi_a}{\sqrt{n}} \leq a^* \leq \theta_0 \leq b^* \leq \bar{b} + 2 \frac{\psi_b}{\sqrt{n}},$$

or:

$$10) \quad \bar{a} - 2 \frac{\psi_a}{\sqrt{n}} \leq \theta_0 \leq \bar{b} + 2 \frac{\psi_b}{\sqrt{n}}.$$

From (4) and 95% credibility,

$$11) \quad k\tau \equiv 4\tau = \left(\bar{b} + 2 \frac{\psi_b}{\sqrt{n}} \right) - \left(\bar{a} - 2 \frac{\psi_a}{\sqrt{n}} \right) \\ = (\bar{b} - \bar{a}) + \frac{2}{\sqrt{n}} (\psi_a + \psi_b).$$

But ψ_a and ψ_b are unknown. Estimate them by their sample quantities:

(12)

$$s_a^2 \equiv \hat{\psi}_a^2 \equiv \frac{1}{n} \sum_1^n (a_i - \bar{a})^2; \\ s_b^2 \equiv \hat{\psi}_b^2 \equiv \frac{1}{n} \sum_1^n (b_i - \bar{b})^2.$$

Then, the assessment procedure for τ becomes:

$$13) \quad 4\tau \doteq (\bar{b} - \bar{a}) + \frac{2}{\sqrt{n}} (s_a + s_b).$$

There is a Minitab 13 macro for computing the Bayesian RGI extended average estimator (see Remark c).

Case 2: Extended Range Estimator

From (10), since $a_0 < \bar{a}$, and $\bar{b} < b_0$, consider for an alternative assessment procedure,

$$10*) \quad a_0 - 2 \frac{\psi_a}{\sqrt{n}} \leq \theta_0 \leq b_0 + 2 \frac{\psi_b}{\sqrt{n}}$$

Then, (11) becomes:

$$11*) \quad k\tau \equiv 4\tau = \left(b_0 + 2 \frac{\psi_b}{\sqrt{n}} \right) - \left(a_0 - 2 \frac{\psi_a}{\sqrt{n}} \right) \\ = (b_0 - a_0) + \frac{2}{\sqrt{n}} (\psi_a + \psi_b).$$

Using (12) gives:

$$12*) \quad 4\tau \doteq (b_0 - a_0) + \frac{2}{\sqrt{n}} (s_a + s_b).$$

Remarks:

a) Note that these assessments give larger values of τ than our earlier assessments, $(\bar{b} - \bar{a})$, and $(b_0 - a_0)$ the assessments called average, and range, The credibility intervals for the population mean will accordingly be larger.

b) The second term in (13), and in (12*) disappear for large sample sizes, leaving just the average or range of the bounds, but for smaller sample sizes, the 2nd term can have a substantial effect.

c) Minitab 13 macros for computing the Bayesian RGI extended average and extended range estimators are available (for information about these macros, contact Diane Miller at diane.m.miller@nge.com).

A Note On Extending Scheffé's Modified Multiple-Comparison Procedure to Other Analysis Situations

Xinyue Zhou

Department of Psychology
Fudan University, China

Joel R. Levin

Department of Educational Psychology
University of Arizona

This article extends Scheffé's modified (sequential) multiple-comparison procedure in one-way analysis-of-variance to other analysis situations, including interaction comparisons in factorial ANOVA designs, tests of partial regression coefficients in multiple-regression analysis, and comparisons of means in one-factor multivariate analyses of variance. Researchers who are concerned with maintaining familywise Type I error rates while increasing statistical power relative to the original (simultaneous) Scheffé-based procedures are encouraged to consider these improved multiple-comparison methods.

Key words: controlled multiple-test procedures, modified Scheffé method, Type I error and power

Introduction

A two-step modification of the original Scheffé (1953) multiple-comparison procedure was proposed by Scheffé (1970) and recently brought to researchers' attention by Klockars and Hancock (2000). Specifically, the statistical power of the original Scheffé procedure can be improved by conducting an initial omnibus F -test with a Type I error probability of α before proceeding to investigate any contrasts of interest. If the omnibus test is not rejected, the process stops. On the other hand, if the omnibus test is rejected the degrees of freedom associated with both Scheffé's original multiplier (v_1 , or $K-1$ in the one-way analysis of variance [ANOVA]) and critical F -value may be decreased by one. That is, v_1-1 (or $K-2$) may be employed to test all contrasts (rather than v_1 used in the original Scheffé procedure).

In a recent study, Meyers and Beretvas (2003) compared the familywise Type I error rates and power of the original and modified Scheffé procedures. The modified Scheffé procedure maintained its familywise Type I error at the nominal but less conservative level and, as a direct result, demonstrated greater power. However, Meyers and Beretvas' investigation was restricted to the one-way ANOVA situation.

As with Scheffé's (1953) multiple-comparison procedure, the Roy-Bose (1953) procedure is congruent with an omnibus test in a multivariate analysis of variance (MANOVA) context. Because a similar correspondence exists between Roy's θ criterion and the Roy-Bose procedure, it should be possible to improve the Roy-Bose procedure by adding an initial omnibus MANOVA test of Roy's θ . That is, if the omnibus test is not rejected, no subsequent multiple comparisons are conducted. However, if the omnibus test is rejected, all subsequent contrasts may be tested against a modified (reduced) Roy-Bose critical value.

The advantage of the modified Scheffé and Roy-Bose procedures over the original procedures is evident: similar control over the familywise Type I error rates, similar versatility, and similar computational ease, but greater statistical power. The major disadvantage of the modified procedures is that they do not permit the construction of probability-based confidence

Xinyue Zhou is a faculty member in the Department of Sociology at Fudan University, Shanghai, China, with specializations in cognitive processes and applied statistical methods. Joel R. Levin (jrlevin@u.arizona.edu) is Professor of Educational Psychology at the University of Arizona. His areas of expertise are cognitive strategies, applied statistical methods and pedagogy, and research methodology.

intervals. Thus, if such intervals are of interest or importance to a researcher, then these techniques are not recommended. Even though the modified Roy-Bose procedure is based on the same sequential hypothesis-testing logic (Seaman, Levin, & Serlin, 1991; Shaffer, 1986) as the modified Scheffé procedure, it has not yet been subjected to empirical test. The present Monte Carlo simulation study examines the familywise Type I error rates associated with both modified Scheffé and Roy-Bose procedures, along with two other commonly used analysis approaches.

Methodology

The multiple-comparison procedures examined here are MS (modified Scheffé), MRB (modified Roy-Bose), LSD (Fisher's least significant difference procedure – see Kirk, 1995), and U (Uncontrolled, or multiple t tests each conducted at α). The first step in LSD is to perform an omnibus α -level F test involving all means. Given a rejection of the omnibus test, pairwise differences are then tested using a per-contrast α . The U approach tests each comparison at a separate α without attending to familywise Type I error rate protection.

The study includes three common multiple-comparison situations: (1) interactions in two-factor ANOVA; (2) tests of partial regression coefficients in multiple-regression analysis; and (3) mean comparisons in one-factor MANOVA.

The SAS/IML program was used to simulate various experimental conditions for all specified situations, with the selection of samples from normally distributed populations accomplished using PROC RANNOR. The selection of samples from multivariate normal distributions was simulated using the pseudorandom number generator provided by PROC VNORMAL.

Ten thousand replications were conducted for each design specification. Each test was conducted using the algorithm prescribed by the corresponding multiple-comparison procedure (MS, MRB, LSD, U) based on a familywise Type I error probability of .05. The number of replications producing at least one Type I error for a comparison set was

tallied to yield an estimate of the traditionally defined familywise Type I error rate (i.e., the probability of making at least one Type I error in the set of comparisons). Decision rules proposed by Serlin (2000) were applied. According to his .25 α acceptability rule, with $\alpha = .05$ any familywise errors of 625 or fewer in 10,000 runs (.0625) are considered reasonable.

Interaction Comparisons in Two-Factor Analyses of Variance

Interactions in both 2 x 4 and 3 x 3 factorial designs were examined with $n = 20$ and $n = 100$ participants per cell. The sample sizes were selected so that the omnibus test would be rejected virtually all the time in the large-sample case and not all the time in the small-sample case. Both the complete null situation (no interactions associated with any contrasts) and various partial null situations (interactions associated with one or more contrasts) were examined. The cell means consisted of 1s and 0s designed to reflect various complete null and partial null patterns. The population variance for each variable was set to 1.00 for each simulation. To keep the analyses manageable, only tetrad (four-cell difference-in-difference) interaction comparisons were considered (Marascuilo & Levin, 1970). Accordingly, in the 2 x 4 layout, there are six tetrad contrasts; and in the 3 x 3 layout, there are nine.

Tests of Partial Regression Coefficients in Multiple-Regression Analysis

In multiple regression, various patterns were examined with varying parameters: P (number of predictor variables) = 2, 3, 4; and N (total number of participants) = 20, 100. In this study, all parameters, including the population variance of each predictor and the covariance between predictors, varied so that a single nonzero population partial regression coefficient (beta weight) was equal to 0, 0.2, 0.5, 0.8, 1, or 2 and the rest of the coefficients were equal to 0. Both the complete null situation (no independent variables have any unique contributions to the dependent variable) and various patterns of a partial null situation (the dependent variable shares some variance with only one independent variable) were included in the analysis.

Multivariate Analysis-of-Variance Comparisons

In MANOVA, various patterns of mean differences were studied by varying several parameters: K (number of groups) = 2, 3, 4, 5; P (number of outcome variables) = 2, 3, 4; n (number of participants per group) = 20, 100; and ρ (the common within-group correlation between all outcome variables) = 0.2, 0.5, 0.8. Both the complete null situation (no mean group differences on any variables) and various partial null situations (group mean differences on one or more variables) were included. The mean vectors consisted of 1s, 0s, and -1s to represent different complete null and partial null patterns. The within-group population variance for each variable was set at 1.00 for each simulation.

Special case (one-variable-at-a-time comparisons).

For these MANOVA simulations, $K = 3$ and $P = 2, 3, 4$ situations were investigated. Only one-variable-at-a-time comparisons were included in these analyses. It was assumed that the original Scheffé procedure could be employed to examine all one-variable-at-a-time comparisons by splitting the familywise α by the number of dependent variables (P) using the Bonferroni inequality (e.g., Kirk, 1995).

If the omnibus test is rejected, that means at least one of the variables is statistically different across groups. Then it is possible to modify the original Scheffé procedure by dividing the familywise α by $P-1$ instead of P (see Table 1). This modification was investigated in the simulation to see how it preserves the familywise Type I error rate.

General case: $K = 2, P > 1$ (both one-variable-at-a-time and multiple-variable comparisons).

When $K = 2$, the approach used for modifying (improving) the Roy-Bose procedure involved reducing by one the hypothesis degrees of freedom associated with the critical value of Roy's θ following a rejection of the initial omnibus test. The specifications included in this part of the MANOVA simulations were $K = 2, P = 2, 3, 4, 5$. This situation is the two-group MANOVA equivalent of multiple-regression analysis.

General case: $K > 2, P > 1$ (both one-variable-at-a-time and multiple-variable comparisons).

With $K > 2$ and multiple-variable comparisons included, three situations were investigated: $K-1 > P$ ($K = 5, P = 3$); $K-1 = P$ ($K = 4, P = 3$); and $K-1 < P$ ($K = 4, P = 4$). Two MRB approaches (one reducing K and the other reducing P) were considered. Comparisons based on both one variable at a time and multiple variables were included. To keep the analyses manageable, only four-mean "groups-by-variables interaction" comparisons (specifically, 2 groups by 2 variables) were included as multiple-variable comparisons.

Results

Factorial ANOVA Interaction Comparisons

The MS method maintained the nominal familywise Type I error rate for both 2×4 and 3×3 designs in both the complete and partial null situations. The maximum familywise Type I error rate of the MS method was .048 and the average familywise Type I error rate was .033 (see Table 2). All replications yielded empirical familywise Type I error rates below the $\alpha = 0.05$ criterion.

The LSD procedure preserved the familywise Type I error rate under the complete null situation, with a maximum error rate of .052 and an average error rate of .051. However, in the partial null situation, familywise Type I error rates were seriously inflated with the LSD procedure (average = .146, maximum = .182).

The U approach completely failed to preserve the familywise Type I error rate as long as there was more than one true null comparison. The proportion of times that at least one Type I error was made was as high as .280 and averaged .197.

Tests of Partial Regression Coefficients in Multiple-Regression Analysis

Similar patterns were observed in the multiple-regression analyses (see Table 3). With an increase in the number of predictors, the Type I error rates increased accordingly for the LSD and U methods. However, the reverse pattern was true for the MS method.

Table 1. Critical Values for Original Scheffé-Based Methods and Modifications

Analysis situations	Original Scheffé-based methods	Modified Scheffé-based methods
Interaction contrasts in ANOVA	$S = \sqrt{(I-1)(J-1)F_{(I-1)(J-1),IJ(n-1)}^{(1-\alpha)}}$	$MS = \sqrt{[(I-1)(J-1)-1]F_{(I-1)(J-1)-1,IJ(n-1)}^{(1-\alpha)}}$
Tests of partial regression coefficients in multiple regression	$S = \sqrt{PF_{P,N-P-1}^{(1-\alpha)}}$	$MS = \sqrt{(P-1)F_{P-1,N-P-1}^{(1-\alpha)}}$
MANOVA special case: one-variable-at-a-time comparisons only	$S = \sqrt{(K-1)F_{K-1,N-K}^{[1-\frac{\alpha}{P}]}}$	$MS = \sqrt{(K-1)F_{K-1,N-K}^{[1-\frac{\alpha}{P-1}]}}$
MANOVA general case ($K=2$): one-variable-at-a-time and multiple-variable comparisons	$RB = \frac{P(N-K)}{N-P-1}F_{P,N-P-1}^{(1-\alpha)}$	$MRB = \frac{(P-1)(N-K)}{N-P-1}F_{P-1,N-P-1}^{(1-\alpha)}$
MANOVA general case ($K>2$): one-variable-at-a-time and multiple-variable comparisons	$RB = \sqrt{(\frac{\theta_{(s,m,n)}}{1-\theta_{(s,m,n)}})(N-K)}$ $s = \min(K-1, P);$ $m = \frac{abs[P-(K-1)]-1}{2}$ $n = \frac{(N-K-1)-P}{2}$	$MRB = \sqrt{(\frac{\theta_{(s,m,n)}}{1-\theta_{(s,m,n)}})(N-K)}$ based on reduced K or P

Table 2. Empirical Familywise Type I Error Rates of Three Methods at Nominal Alpha = .05 for Testing ANOVA Interaction Contrasts ($n = 20/n = 100$).

		LSD	U	MS
2 x 4 design	Complete null	.050/.052	.199/.203	.039/.048
	Partial null	.107/.128	.119/.128	.023/.038
3 x 3 design	Complete null	.051/.049	.275/.280	.034/.033
	Partial null	.167/.182	.187/.182	.024/.025

Table 3. Empirical Familywise Type I Error Rates of Three Methods at Nominal Alpha = .05 for Testing Partial Regression Coefficients in Multiple Regression ($N = 20/N = 100$)

$P = 2, B_2 = 0:$			
	LSD*	MS	U
$B_1 = 0.2$	.029/.040	.029/.040	.045/.047
$B_1 = 0.5$	.042/.053	.042/.053	.045/.053
$B_1 = 0.8$	.052/.053	.052/.053	.052/.054
$B_1 = 1$	.047/.050	.047/.050	.051/.050
$B_1 = 2$	.050/.051	.050/.051	.054/.052
$B_1 = 0$	.048/.046	.048/.046	.096/.097

$P = 3, B_2 = 0, B_3 = 0:$			
	LSD	MS	U
$B_1 = 0.2$	.046/.072	.027/.026	.094/.095
$B_1 = 0.5$	.076/.100	.028/.029	.091/.100
$B_1 = 0.8$	.093/.099	.030/.030	.093/.099
$B_1 = 1$	.081/.099	.030/.028	.096/.099
$B_1 = 2$	.076/.096	.029/.029	.089/.096
$B_1 = 0$	.046/.047	.035/.032	.133/.140

$P = 4, B_2 = 0, B_3 = 0, B_4 = 0:$			
	LSD	MS	U
$B_1 = 0.2$	.051/.092	.015/.015	.132/.135
$B_1 = 0.5$	.102/.135	.018/.015	.131/.135
$B_1 = 0.8$	.131/.143	.019/.015	.131/.143
$B_1 = 1$	.104/.145	.018/.017	.132/.145
$B_1 = 2$	.100/.134	.019/.015	.131/.134
$B_1 = 0$	.042/.042	.016/.014	.135/.144

* LSD in the $P = 2$ situation is the same procedure as MS.Table 4. Special Case: Empirical Familywise Type I Error Rates of Three Methods at Alpha = .05 for Testing One-Variable-At-a-Time Comparisons in MANOVA ($n = 20/n = 100$)

MS

	Complete null	Partial null
$K=3, P = 2, \rho = 0.2$	.038/.042	.054/.053
$K=3, P = 2, \rho = 0.5$	.037/.039	.055/.052
$K=3, P = 2, \rho = 0.8$	.030/.029	.053/.050
$K=3, P = 3, \rho = 0.2$	.033/.033	.049/.045
$K=3, P = 3, \rho = 0.5$	.027/.030	.050/.043
$K=3, P = 3, \rho = 0.8$	.022/.020	.042/.035
$K=3, P = 4, \rho = 0.2$	.025/.029	.040/.041
$K=3, P = 4, \rho = 0.5$	.024/.023	.045/.040
$K=3, P = 4, \rho = 0.8$	.016/.014	.036/.030

Table 4 Continued.

LSD

	Complete null	Partial null
$K=3, P=2, \rho=0.2$	.044/.051	.151/.165
$K=3, P=2, \rho=0.5$	.043/.047	.157/.158
$K=3, P=2, \rho=0.8$	.036/.038	.145/.148
$K=3, P=3, \rho=0.2$	.049/.048	.235/.260
$K=3, P=3, \rho=0.5$	.040/.046	.233/.235
$K=3, P=3, \rho=0.8$	.036/.035	.204/.196
$K=3, P=4, \rho=0.2$	.048/.052	.299/.351
$K=3, P=4, \rho=0.5$	.045/.046	.295/.316
$K=3, P=4, \rho=0.8$	.033/.033	.248/.244

U

	Complete null	Partial null
$K=3, P=2, \rho=0.2$	.227/.232	.160/.165
$K=3, P=2, \rho=0.5$	.211/.217	.161/.158
$K=3, P=2, \rho=0.8$	.180/.184	.145/.148
$K=3, P=3, \rho=0.2$	.313/.325	.263/.260
$K=3, P=3, \rho=0.5$	.290/.284	.243/.235
$K=3, P=3, \rho=0.8$	.230/.222	.205/.196
$K=3, P=4, \rho=0.2$	.400/.398	.354/.351
$K=3, P=4, \rho=0.5$	.356/.337	.309/.316
$K=3, P=4, \rho=0.8$	.259/.263	.248/.244

Table 5. Empirical Familywise Type I Error Rates of Three Methods at Alpha = .05 for Both One-Variable-At-a-Time and Multiple-Variable Comparisons in $K=2$ MANOVA ($n=20/n=100$)

MRB

	Complete null	Partial null
$P=2, \rho=0.2$	.040/.043	.048/.052
$P=2, \rho=0.5$	.044/.046	.046/.051
$P=2, \rho=0.8$	.040/.043	.043/.045
$P=3, \rho=0.2$	.039/.036	.042/.043
$P=3, \rho=0.5$	.035/.032	.040/.042
$P=3, \rho=0.8$	.031/.033	.040/.042
$P=4, \rho=0.2$	.026/.033	.023/.023
$P=4, \rho=0.5$	.026/.035	.022/.026
$P=4, \rho=0.8$	.030/.032	.022/.020
$P=5, \rho=0.2$	.019/.023	.014/.017
$P=5, \rho=0.5$	.016/.025	.012/.020
$P=5, \rho=0.8$	.018/.021	.013/.015

Table 5 Continued.

LSD

	Complete null	Partial null
$P = 2, \rho = 0.2$	.045/.054	.049/.056
$P = 2, \rho = 0.5$	.047/.051	.050/.053
$P = 2, \rho = 0.8$	.041/.049	.051/.052
$P = 3, \rho = 0.2$	.044/.045	.101/.098
$P = 3, \rho = 0.5$	.047/.047	.116/.101
$P = 3, \rho = 0.8$	.050/.049	.121/.093
$P = 4, \rho = 0.2$	.048/.051	.177/.197
$P = 4, \rho = 0.5$	.044/.052	.198/.204
$P = 4, \rho = 0.8$	.054/.048	.198/.188
$P = 5, \rho = 0.2$	.052/.053	.231/.285
$P = 5, \rho = 0.5$	.048/.050	.288/.289
$P = 5, \rho = 0.8$	.050/.049	.283/.274

U

	Complete null	Partial null
$P = 2, \rho = 0.2$	.108/.106	.060/.056
$P = 2, \rho = 0.5$	.095/.097	.051/.053
$P = 2, \rho = 0.8$	.084/.090	.057/.052
$P = 3, \rho = 0.2$	.185/.170	.101/.098
$P = 3, \rho = 0.5$	.164/.171	.099/.101
$P = 3, \rho = 0.8$	.169/.171	.088/.093
$P = 4, \rho = 0.2$	.275/.271	.202/.197
$P = 4, \rho = 0.5$	.293/.288	.207/.204
$P = 4, \rho = 0.8$	.283/.277	.199/.188
$P = 5, \rho = 0.2$	.362/.367	.279/.285
$P = 5, \rho = 0.5$	.375/.366	.288/.289
$P = 5, \rho = 0.8$	.370/.361	.283/.274

Note. Only the worst-case scenario (specifications with most null contrasts) is included in the table under the partial null situation.

Table 6. Empirical Familywise Type I Error Rates of Three Methods at Alpha = .05 for Multiple-Variable Comparisons in MANOVA ($n = 20/n = 100$).MRB (reducing P)

	Complete null	Partial null
$K = 4, P = 3, \rho = 0.2$ ($K-1 = P$)	.028/.025	.034/.028
$K = 4, P = 3, \rho = 0.5$	.030/.027	.034/.024
$K = 4, P = 3, \rho = 0.8$	.027/.028	.033/.023
$K = 4, P = 4, \rho = 0.2$ ($K-1 < P$)	.019/.014	.020/.013
$K = 4, P = 4, \rho = 0.5$	.019/.013	.019/.013
$K = 4, P = 4, \rho = 0.8$	.017/.016	.018/.012
$K = 5, P = 3, \rho = 0.2$ ($K-1 > P$)	.027/.023	.028/.016
$K = 5, P = 3, \rho = 0.5$	.021/.023	.028/.018
$K = 5, P = 3, \rho = 0.8$	.022/.018	.029/.016

Table 6 Continued.

MRB (reducing K)

	Complete null	Partial null
$K = 4, P = 3, \rho = 0.2$		
$(K-1 = P)$	.028/.025	.034/.028
$K = 4, P = 3, \rho = 0.5$	.030/.027	.034/.024
$K = 4, P = 3, \rho = 0.8$	.027/.028	.033/.023
$K = 4, P = 4, \rho = 0.2$		
$(K-1 < P)$	.020/.016	.024/.016
$K = 4, P = 4, \rho = 0.5$	.023/.015	.024/.015
$K = 4, P = 4, \rho = 0.8$	.018/.019	.022/.014
$K = 5, P = 3, \rho = 0.2$	.021/.019	.024/.012
$(K-1 > P)$		
$K = 5, P = 3, \rho = 0.5$	.017/.019	.022/.015
$K = 5, P = 3, \rho = 0.8$	.019/.017	.023/.013

LSD

	Complete null	Partial null
$K = 4, P = 3, \rho = 0.2$		
$(K-1 = P)$	.049/.048	.462/.522
$K = 4, P = 3, \rho = 0.5$	.051/.049	.513/.543
$K = 4, P = 3, \rho = 0.8$	.048/.049	.528/.526
$K = 4, P = 4, \rho = 0.2$		
$(K-1 < P)$	.048/.042	.558/.691
$K = 4, P = 4, \rho = 0.5$	.054/.047	.664/.702
$K = 4, P = 4, \rho = 0.8$	.052/.047	.678/.690
$K = 5, P = 3, \rho = 0.2$	.051/.045	.582/.688
$(K-1 > P)$		
$K = 5, P = 3, \rho = 0.5$	.052/.050	.680/.707
$K = 5, P = 3, \rho = 0.8$	.050/.049	.691/.704

U

	Complete null	Partial null
$K = 4, P = 3, \rho = 0.2$		
$(K-1 = P)$	.607/.613	.532/.522
$K = 4, P = 3, \rho = 0.5$	.626/.620	.539/.543
$K = 4, P = 3, \rho = 0.8$	.611/.621	.528/.526
$K = 4, P = 4, \rho = 0.2$		
$(K-1 < P)$	.747/.735	.690/.691
$K = 4, P = 4, \rho = 0.5$	.763/.751	.710/.702
$K = 4, P = 4, \rho = 0.8$	.743/.726	.678/.690
$K = 5, P = 3, \rho = 0.2$	.753/.756	.701/.688
$(K-1 > P)$		
$K = 5, P = 3, \rho = 0.5$	.782/.774	.693/.707
$K = 5, P = 3, \rho = 0.8$	.782/.769	.691/.704

The more predictors in the multiple-regression analysis, the less was the familywise Type I error rate. The MS procedure was successful in maintaining the nominal familywise Type I error rate, with a maximum of .053.

The LSD method exhibited control over the familywise Type I error rate when there were only two predictors (average = .047, maximum = .053), which is consistent with Levin et al. (1994). With two predictors, the LSD and MS procedures are equivalent and so both of them produced the same results. The LSD method also performed well under the complete null situation no matter how many predictors (maximum familywise Type I error rate = .048, average = .045). However, that method was not acceptable in partial null situations with more than two predictors (maximum familywise Type I error rate = .145, average = .099).

Not surprisingly, the U method maintained the familywise Type I error rate of .05 only when there was just one true null regression coefficient. In other situations, the familywise Type I error rate increased as the number of null regression coefficients increased. With one null coefficient, the average familywise Type I error rate was .050; with two null coefficients, the average familywise Type I error rate was .095; and with three null coefficients, the average familywise Type I error rate was .135.

MANOVA Comparisons

Special case ($K = 3$ one-variable-at-a-time comparisons only).

Univariate contrasts in $K = 3$, $P = 2, 3, 4$ designs were examined with $n = 20$ and $n = 100$ participants per group. The MS method maintained the nominal familywise Type I error rate for both complete and partial null situations within an acceptable level (see Table 4). In the $K = 3$, $P = 2$ situation, the maximum proportion of familywise Type I errors was .055 and the average familywise Type I error rate was .044. In the $K = 3$, $P = 3$ situation, the maximum familywise Type I error rate was .050 and the average familywise Type I error rate was .036. In the $K = 3$, $P = 4$ situation, the maximum familywise Type I error rate was .045 and the average was .030.

The LSD method preserved the familywise Type I error rate only under the complete null situation, with a maximum Type I error rate of .052. In the partial null situation, the familywise Type I error rate was seriously inflated (maximum = .351). The U method failed to protect familywise Type I error rate as long as there was more than one true null comparison. The proportion of times there was at least one Type I error was as high as .400.

General case for $K = 2$ (both one-variable-at-a-time and multiple-variable contrasts).

When both one-variable-at-a-time and multiple-variable contrasts were analyzed in the simulation, the MRB method preserved the familywise Type I error rate at the desired level, with a maximum error rate of .052 (see Table 5). The LSD method maintained the nominal familywise Type I error rate only under the complete null situation, with a maximum familywise Type I error rate of .054. In the partial null situation, the error rate inflated to as much as .289. The U method completely failed to preserve the familywise Type I error rate, with a maximum of .375.

General case for $P > 1$, $K > 2$ (both one-variable-at-a-time and multiple-variable contrasts).

When multiple-variable (2 groups by 2 variables) and one-variable-at-a-time comparisons were considered following a significant omnibus test, either P or K could be reduced by 1 in the MRB critical value. Reducing either one of these was found to be adequate for preserving the familywise Type I error rate (see Table 6). Reducing the minimum of $K-1$ and P produced lower critical values and, therefore, and greater power.

For example, when $K = 4$, $P = 4$ ($K-1 < P$), reducing K yielded an acceptable familywise error rate (maximum = .024) that was higher than that associated with reducing P . Conversely, when $K = 5$, $P = 3$ ($K-1 > P$), reducing P yielded an acceptable familywise error rate (maximum = .029) that was higher than that associated with reducing K . When $K-1 = P$ (e.g., $K = 4$, $P = 3$), reducing either $K-1$ or P produced the same critical values. Both of these

were acceptable for preserving the familywise Type I error rate (maximum = .034).

In all situations, the LSD method preserved the familywise Type I error rate only under the complete null situation, with a familywise Type I error rate of .054. However, in partial null situations, familywise Type I error rates were enormously inflated under the LSD method (maximum = .707). The U approach failed to preserve the familywise Type I error rate as long as there was more than one true null comparison. The proportion of times there was at least one Type I error was as high as .782.

Conclusion

The choice of multiple-test procedures should be based on considerations of Type I and Type II error characteristics instead of either tradition or increasing power at the risk of obtaining spurious statistical significance (Seaman et al., 1991). The results of this study offer insights for both educational statisticians and researchers. What we have shown is that there are valid reasons for replacing traditional Scheffé-based methods with an improved sequential version of the test when the researcher is not interested in constructing confidence intervals. These modified procedures are not only able to preserve the familywise Type I error rate, but they are also easy to perform, being based on the same test statistics as in the simultaneous case and merely requiring reducing the critical value at the second step. The modified Scheffé-based procedures are recommended if researchers need a more powerful and adequate alternative to the original procedures.

The choice of procedures depends on the parameter specifications. When v_1 is equal to 2, Fisher LSD-like methods are the most powerful while providing adequate familywise Type I error protection. Thus, for 3 x 2 designs in factorial ANOVA, when there are only 2 predictors in multiple regression, or for any situation in which v_1 equals 2 in MANOVA, LSD is recommended. However, beyond $v_1 = 2$ situations, the familywise Type I error rate is seriously inflated in partial null situations.

The results of this study demonstrate that indiscriminate use of LSD-like methods or U methods lead to inflated familywise Type I

error rates. Researchers are cautioned about applying these two classes of procedure in general multiple-regression analysis and MANOVA situations.

There should be an investigation of the power of the modified Scheffé-based methods relative to other commonly applied multiple-test procedures. This article was restricted to considering only ideal specifications with normal distributions and balanced designs. It is important to determine how robust modified Scheffé-based methods are in preserving the familywise Type I error rate under less than ideal distributional and design conditions.

References

- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.
- Klockars, A. J., & Hancock, G. R. (2000). Scheffé's more powerful *F*-protected post hoc procedure. *Journal of Educational and Behavioral Statistics*, 25, 13-19.
- Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal*, 7, 397-421.
- Meyers, J. L., & Beretvas, S. N. (2003, Apr.). *Type I error rates and power of variations of Scheffé's MCP*. Paper presented at the annual meeting of the American Educational Association, Chicago.
- Roy, S. N., & Bose, R. C. (1953). Simultaneous confidence interval estimation. *Annals of Mathematical Statistics*, 24, 513-536.
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, 87-104.
- Scheffé, H. (1970). Multiple testing versus multiple estimation: Improper confidence sets, Estimation of directions and ratios. *Annals of Mathematical Statistics*, 41, 1-29.
- Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, 110, 577-586.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, 5, 230-240.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81, 826-831.

Author Note

This study was conducted as part of the Xinyue Zhou's doctoral research in the Department of Educational Psychology at the University of Arizona. We are grateful to Jerome D'Agostino, Rosemary Rosser, Janice Streitmatter, and Rebecca Gomez for serving on the dissertation committee, and to Patricia Jones for her assistance with the simulations.

Multivariate And Multistrata Nonparametric Tests: The NonParametric Combination Method

Livio Corain Luigi Salmaso
Department of Management and Engineering
University of Padova

Researchers and practitioners in many scientific disciplines and industrial fields are often faced with complex problems when dealing with comparisons between two or more groups using classical parametric methods. The data arising from real problems rarely are in agreement with stringent parametric assumptions. The NonParametric Combination (NPC) methodology frees the researcher from stringent assumptions of parametric methods and allows a more flexible analysis, both in terms of specification of multivariate hypotheses and in terms of the nature of the variables involved in the analysis. An outline of NPC methodology is given, along with case studies.

Key words: Multivariate Inference, nonparametric combination, permutation test

Introduction

From a methodological point of view, when comparing NonParametric Combination (NPC) Test methodology to unconditional parametric testing it should be remembered that the latter suffers from the constraint that it is appropriate and applicable only when a set of conditions concerned with the likelihood model are all satisfied (Pesarin, 2002). Only if all conditions are jointly satisfied is the extension of inferential results to the population possible and appropriate. Otherwise when these conditions fail, especially if selection-bias procedures are used for data collection processes as in most real applications, most parametric inferential extensions are generally wrong or misleading.

Moreover, when all the above conditions are satisfied, in practice other assumptions regarding the validity of the parametric method, such as normality, are rarely

satisfied. Consequent inferences, when not improper, are necessarily approximated and their approximations are often difficult to assess. However, there are circumstances in which conditional testing procedures may be unavoidable as in the case of multivariate problems, when some variables are categorical and others are quantitative or when multivariate alternatives are subjected to order restrictions (for a detailed list of these circumstances see Pesarin, 2002). A short outline of the implementation of NPC methodology follows.

Brief overview of the NPC methodology

Without loss of generality, let us refer to a one-way MANOVA layout. The data structure is defined as follows. Denote by $\mathbf{X}$ an $(n \times k)$ data set:

$$\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_j, \dots, \mathbf{X}_C] = [\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_k],$$

where $\mathbf{X}_j, j=1, \dots, C, (C \geq 2)$ represents the j -th $n_j \times k$ group, $n_j \geq 2$ and $\sum_j n_j = n$, and $\mathbf{X}_i$ is the i -th univariate aspect of $\mathbf{X}, i=1, \dots, k (k \geq 1)$; moreover let $\mathbf{X}_{ji}$ represent the i -th univariate aspect of $\mathbf{X}_j$.

In the context of NonParametric Combination (NPC) of Dependent Permutation Tests a set of conditions should be jointly satisfied:

Livio Corain is a doctoral candidate at the University of Padova. The Email address is: livio.corain@unipd.it. Luigi Salmaso is an Associate Professor of statistics at the University of Padova. His research interests include multivariate analysis, design of experiments, nonlinear regression, and nonparametric statistics.

1) suppose that for $\mathbf{X}=[\mathbf{X}_1, \dots, \mathbf{X}_C]'$ an appropriate probabilistic k -dimensional distribution structure P exists, $P_j \in F, j=1, \dots, C$, belonging to a (possibly non-specified) family F of non-degenerate probability distributions.

2) the null hypothesis H_0 states the equality in distribution of the multivariate distribution of the k variables in all C groups:

$$H_0 : [P_1 = \dots = P_C] = \left[\mathbf{X}_1^d = \dots = \mathbf{X}_C^d \right].$$

Null hypothesis H_0 implies the exchangeability of the individual data vector with respect to the groups. Moreover H_0 is supposed to be properly decomposed into k sub-hypotheses $H_{0i}, i=1, \dots, k$, each appropriate for partial (univariate) aspects, thus H_0 (multivariate) is true if all the H_{0i} (univariate) are jointly true:

$$H_0 : \left[\bigcap_{i=1}^k X_{1i}^d = \dots = X_{Ci}^d \right] = \left[\bigcap_{i=1}^k H_{0i} \right].$$

H_0 is called the *global or overall null hypothesis*, and $H_{0i}, i=1, \dots, k$, are called the *partial null hypotheses*.

3) The alternative hypothesis H_1 is represented by the union of partial H_{1i} sub-alternatives:

$$H_1 : \left[\bigcup_{i=1}^k H_{1i} \right],$$

so H_1 is true if at least one of sub-alternatives is true. In this context, H_1 is called the *global or overall alternative*, and $H_{1i}, i=1, \dots, k$, are called the *partial alternatives*.

4) let $\mathbf{T}=\mathbf{T}(\mathbf{X})$ represent a k -dimensional vector of test statistics, $k \geq 1$, whose components $T_i=T_i(X_i), i=1, \dots, k$, represent the partial univariate and non-degenerate *partial test* appropriate for testing the sub-hypothesis H_{0i} against H_{1i} . Without loss of generality, all partial tests are assumed to be marginally unbiased, consistent and significant for large values (for more details, see Pesarin, 2001).

At this point, in order to test the global null hypothesis H_0 , the key idea comes from the partial (univariate) tests which are focused on k partial aspects, and then, combining them with an appropriate combining function, from a global (multivariate) test which is referred to as the global null hypothesis.

However, before introducing the combination methodology, we should observe that in most real problems, when the sample size is great enough, there is a clash over the problem of computational difficulties in calculating the conditional permutation space. This means it is not possible to calculate the exact p -value of observed statistic T_{i0} . This is overcome by using the CMCP (Conditional Monte Carlo Procedure).

The CMCP on the pooled data set $\mathbf{X}$ is a random simulation of all possible permutations of the same data under H_0 (for more details refer to Pesarin, 2001). Hence, in order to obtain an estimate of the permutation distribution under H_0 of all test statistics, a CMCP can be used. Every resampling without replacement $\mathbf{X}^*$ from the pooled data set $\mathbf{X}$ actually consists of a random attribution of individual data vectors to the C samples. In every $\mathbf{X}_r^*$ resampling, $r=1, \dots, B$, the k partial tests are calculated to obtain the set of values $[T_{ir}^*=T(\mathbf{X}_{ir}^*), i=1, \dots, k; r=1, \dots, B]$, the B independent random resamplings.

It should be emphasized that CMCP only considers permutations of individual data vectors, so that all underlying dependence relations that are present in the component variables are preserved. From this point of view, the CMCP is essentially a multivariate procedure.

The two-phases algorithm

Once the hypothesis system is defined and an appropriate set of k statistics $T_i=T_i(X_i), i=1, \dots, k$, the natural way to test the global null hypothesis consists of two sequential phases:

- 1) performing k partial tests;
- 2) combining them in a second-order global test.

It should be pointed out that this two-step procedure can be characterized by several

intermediate combinations if there is a more complex data configuration where the most interesting cases are given by testing in presence of stratification, closed-testing, multi aspect testing and repeated measures.

Assuming that the partial tests have real values and are marginally unbiased, consistent and significant for large values, then the first phase consists in:

1a. calculating the k -vector of observed values of test statistics $\mathbf{T}_0$:

$$\mathbf{T}_0 = \mathbf{T}(\mathbf{X}) = [T_{i0}(X_i), i=1, \dots, k];$$

1b. considering a data permutation of $\mathbf{X}$ by a random resampling $\mathbf{X}_r^*$, in order to randomly assign every individual data vector to a proper group and then calculate the vector statistics $\mathbf{T}_r^*$:

$$\mathbf{T}_r^* = \mathbf{T}_r^*(\mathbf{X}_r^*) = [T_{ir}^*(X_{ir}^*), i=1, \dots, k];$$

1c. carrying out B independent repetitions of step 1.b; the result is a set $\mathbf{T}^*$ of $B \times k$ CMC

$$\mathbf{T}^* = [\mathbf{T}_r^*, r=1, \dots, B] = [\mathbf{T}_1^*, \dots, \mathbf{T}_r^*, \dots, \mathbf{T}_B^*]'$$

is thus a random sampling from the permutation k -variate distribution of vector test statistics $\mathbf{T}$;

1d. the k -variate EDF (Empirical Distribution Function) $\hat{F}_B(\mathbf{z} | \mathbf{X})$

$$\hat{F}_B(\mathbf{z} | \mathbf{X}) = \left[1/2 + \sum_r \mathbf{I}(\mathbf{T}_r^* \leq \mathbf{z}) \right] / (B+1), \forall \mathbf{z} \in \mathbb{R}^k,$$

where $\mathbf{I}(\cdot)$ is the indicator function, and gives an estimate of the corresponding k -dimensional permutation distribution $F_B(\mathbf{z} | \mathbf{X})$ of $\mathbf{T}$. Moreover

$$\hat{L}_i(z | \mathbf{X}) = \left[1/2 + \sum_r \mathbf{I}(T_{ir}^* \geq z) \right] / (B+1), i=1, \dots, k,$$

gives an estimate $\forall z \in \mathbb{R}^1$ of the marginal permutation *significance level function*

$$L_i(z | \mathbf{X}) = \Pr \{ T_i^* \geq z | \mathbf{X} \};$$

thus

$$\hat{L}_i(T_{i0} | \mathbf{X}) = \hat{\lambda}_i$$

gives an estimate of the marginal p -value $\lambda_i = \Pr \{ T_i^* \geq T_{i0} | \mathbf{X} \}$ relative to test T_i , $i=1, \dots, k$. All these are unbiased and consistent estimates of corresponding true values;

1e. if $\hat{\lambda}_i < \alpha$, the null hypothesis H_{0i} relating to the i -th variable is rejected at the significance level α .

The second phase, based on a nonparametric combination of the dependent tests previously obtained, consists in the following steps:

2a. the combined observed value of the second-order test is evaluated through the same CMC results as the first phase, and is given by:

$$T_0'' = \psi(\hat{\lambda}_1, \dots, \hat{\lambda}_k);$$

2b. the r -th combined value of vector statistics (step 1.d) is then calculated by:

$$T_r'' = \psi(\hat{\lambda}_{1r}^*, \dots, \hat{\lambda}_{kr}^*),$$

where $\hat{\lambda}_{ir}^* = \hat{L}_i(T_{ir}^* | \mathbf{X})$, $i=1, \dots, k$, $r=1, \dots, B$;

2c. hence, the p -value of combined test T'' is estimated as:

$$\lambda_\psi'' = \sum_r \mathbf{I}(T_r'' \geq T_0'') / B;$$

2d. if $\lambda_\psi'' \leq \alpha$, the global null hypothesis H_0 is rejected at significant level α ; where ψ is an appropriate combining function.

Figure 1 summarizes graphically the complete framework of NPC solution. Remember that, in order to preserve the underlying dependence relations among variables, permutations must always be carried out on individual data vectors, so that all component variables and partial tests must be jointly analyzed.

[illegible]

binized test is unbiased and consistent; it also has interesting asymptotic properties.

a) it must be non-increasing in each argument:

$$\psi(..., \lambda_i, ...) \geq \psi(..., \lambda'_i, ...) \text{ if } \lambda_i < \lambda'_i, \\ i \in \{1, \dots, k\};$$

- b) it must attain its supreme value $\bar{\psi}$, possibly non finite, even when only one argument reaches zero:

$$\psi(..., \lambda_i, ...) \rightarrow \bar{\psi} \text{ if } \lambda_i \rightarrow 0, \\ i \in \{1, \dots, k\};$$

- c) $\forall \alpha > 0$, the critical value of every ψ is assumed to be finite and strictly smaller than the supreme value:

$$T''_{\alpha} < \bar{\psi}.$$

The above properties define the class C of combining functions. Some of the functions most often used to combine independent tests (e.g., Fisher, Lancaster, Liptak, Tippett, Mahalanobis) are included in this class. If in the overall analysis distinguishing the importance of partial tests by using appropriate weights opportunely fixed: $w_i \geq 0$, $i = 1, \dots, k$, with at least one strong inequality is considered more suitable, then the combined test using the Fisher combination is:

$$T'' = -\sum_i w_i \cdot \log(\lambda_i).$$

Nested combinations

Suppose that the k variables describing the testing problem can be classified into $m_1 < k$ classes according to some meaningful criteria. Moreover, the m_1 classes could themselves be put together in a further grouping, obtaining $m_2 < m_1$ classes and so on. After $T < k$ steps, this nested classification rule leads to only one final class which includes all variables. It is clear that in such a situation, before carrying out the global test by nonparametric combination of k partial tests, it is more appropriate to introduce T intermediate combination phases that reflect the meaningful classification rules.

This nested procedure can be represented by a graph (Figure 2) in which, from top to bottom, each node indicates a partial test (the corresponding p -value is displayed), and each arch indicates a nonparametric combination into a higher order test. Note that it is not necessary for all partial tests to be involved in

every phase. Some could be included after a given phase.

Features of Software *NPC Test 2.0*

NPC Test 2.0 (see details online at www.methodologica.it) implements completely NPC methodology offering both flexibility and a user-friendly interface. The available multivariate analyses are *Two or C Samples with Dependent Variables* (highlighting the dependence among responses) and *Two or C Samples with Repeated Measures*.

Data sets may be either created and manipulated inside the program on a normal spreadsheet or may be pasted or directly imported from the most utilized formats (see Figure 3).

The reader should be reminded that in NPC Test there are no limitations in the number of observations with respect to the number of variables, i.e., there are no problems regarding a possible lack of degrees of freedom. It is possible to consider one or more stratification factors in order to solve problems with extremely complex experimental designs.

All kinds of variables are dealt with (numeric or continuous, nominal, ordered categorical, or binary; see Figure 4) each provided with an appropriate set of test statistics should they also be suitable for an effective managing of missing values.

The testing procedure is easily performed by following a three step wizard where at first the user is requested to define the sample and the strata, then he has to specify the variables under testing and the test statistic (Figure 5) and finally he has to select a suitable Nonparametric Combination in to perform the global test. Four different functions for combining nonparametrically the partial tests are available: Fisher, Liptak, Tippet and Direct (Figure 6).

We highlight that every partial alternative hypothesis may be specified as being either one or two tailed. Moreover there is the possibility of testing both aspect X and X^2 of the same variable so multi-aspect testing (Pesarin, 2001) is also obtainable. Finally all performed tests are kept in an effective report that can easily be integrated and customised by means of an efficient text editor (Figure 7, 8).

Figure 2. Graphical representation of nested combinations.

Figure 3. NPC Test's interface for data management.

Figure 4. Type of variable definition.

Figure 5. Partial tests definition.

Figure 6. Nonparametric combination.

Figure 7. Performed tests in the report.

Figure 8. The report file editor.

Case studies: developing successful products and comparing two respiratory drugs

In order to better illustrate the NPC Test methodology let us develop two real case studies in the field of Management and Biostatistics. What does distinguish the best firms in the new product development (NPD) process? Over the past decade the New Product Development (NPD) process has been analysed in a number of works, both from an academic and a practitioner's point of view (Booz, Allen & Hamilton 1982; Madique & Zirger, 1984; Link, 1987; Cooper, 1990, 1993; Cooper & Kleinschidt, 1993; 1995; Pittiglio, Rabin, Todd & McGrath, 1995; Griffin, 1997, 1998). These works aimed at identifying NPD performance drivers, that is to say, all those practices, specific process configurations and internal business contexts which underlie the achievement of superior performances and company objectives.

However, these studies were carried out in different contexts and used both different measures of success and different methods of analysis. Griffin and Page (1993), in their literature review, identified 75 different measures previously used in papers on this topic, and classified them in the following groups: customer acceptance, financial performance, product level measures, firm based measures and program measures.

In general terms, in different industries and market types (i.e. B2C versus B2B) the relationship between drivers and performances and the appropriate set of measures of success to be considered may be different. For example, in a B2B marketplace a supplier involved in NP design, can be successful if the supplier is able to meet the specific needs of the client at a low cost and to carry out the task within an established time (Ragatz, Handfield & Scannell, 1997; Droge, Jayaram & Vickery, 2000). A company which produces industrial goods must consider the specific requirements of the customers and offer customized or semi-customized products. This can be done by using approaches and practices in NP development; making an effort to develop a partnership with customers (Hartley, Zirger & Kamath, 1997; Swink & Mabert, 2000; Tuten & Urban, 2001).

Recent studies have laid emphasis on the configuration of different drivers distinguishing between Best and Rest at a company level, considering the whole of the product the company developed in the last three or five years, i.e. the development program. Griffin (1997, 1998), for example, considered the NP program over a five year period and to do so, divided the sample on the basis of three sets of measures: market and financial success, relative success of the program in terms of meeting its objectives and, overall industry

success. Companies were classified as best when they were in the top third of their industry for NPD success and, also, were above the mean of the entire sample regarding the relative success of the program and market – financial success.

Context of the study, framework and key variables

This study aims at identifying the differences in driver configurations between successful and unsuccessful companies working in a B2B marketplace in two specific industries (Machinery Manufacturing, SIC35, and Electrical, Electronic Machinery, Equipment and Supplies, SIC36). We have considered all products developed and launched onto the market by each company in the last three years. Successful companies were defined as those above the median position for the global ranking of both the performances of the new product on the market and of the performances of the NPD process. This study has considered many different drivers: practices and processes, strategic guide and internal environment which support NP development.

The research considers companies which develop and produce industrial goods such as machinery, equipment and appliances to sell to other companies which use them in their production processes, or products, modules and components which will be incorporated into the client company's final products (in other words, these companies have other companies as clients, so their operations and businesses are conditioned by, for example: 1) the importance of the interaction between customer and supplier, so the NP department plays an important role in designing products based on the specific needs of the customer; 2) a limited number of customers with different requirements, 3) a short distribution channel and often direct sales; 4) a different and sometimes more critical role of marketing and promotion compared to a B2C environment; 5) customization or semi-customization of products; 6) a limited number of competitors (often companies that work in a niche or specialized market).

In this study, six categories of variables are considered, including performance and

driver use measures, referring to a three year NPD program:

PERFORMANCE

- NPD Operational Performances (IP, Internal Performances);
- Market, Products and Financial success (EP, External Performances).

DRIVER

- Product Architecture Approach;
- Organizational Mechanisms of NPD;
- Development Process of NPD;
- Strategic Capabilities.

Operational Performances (IP)

Operational Performances are those that depend on the NPD process, practices and environment support. Three types of performances are considered and are related to the time and quality dimensions of the development.

- Launch on Time;
- Time to Market Reduction;
- Product quality capability.

Market, Products and Financial success (EP)

The variables belonging to this category and considered in the present study are:

- Meet Profit Goals;
- Overall Product Success;
- Meet Revenue Goals.

Product Architecture Approach

The technical approach on product architecture.

- Standardization;
- Modularization;
- Platform.

Organizational Mechanisms

Organizational mechanisms refer to a set of techniques used during the various phases of the development process. Some of them concern technological aspects, others are concerned with organizational practices (PM, team, integration etc.).

- Project Manager Use;
- Customer Involvement (multi-item scale);
- Integration Design – Marketing;

- Integration Design – Manufacturing;
- Supplier Involvement (multi-item scale);
- Team Use.

Development Process

An NPD Process concerns the phases of the development itself and the overlapping level between these phases. The variables measure in how many cases during the development program each phase or approach has been used.

- Product Concept Development;
- Product Concept Test;
- Preliminary Design (multi-item scale);
- Late Engineering Changes (i.e. Early modifications);
- Overlapping Approach.

Strategic Capabilities

NP performances and success do not only depend on best practices and well defined process but also on the internal environment which supports NP development. This support can come from the management of the company (top management support, strategic guide) and from the capabilities of the employees.

- Up – Front Capabilities (VOC) (multi-item scale);
- Top Management Support;
- NP Strategic Guide (multi-item scale);
- Company Innovation Culture;
- Technological Capabilities (multi-item scale).

Distinguishing Best and Rest companies on the basis of high or low values in PI and PE, we obtain eight different groups (Figure 9):

- Best companies in PI (labeled BX);
- Rest companies in PI (labeled RX);
- Best companies in PE (labeled XB);
- Rest companies in PE (labeled XR);
- Best companies in both PI and PE (labeled BB);
- Rest companies in both PI and PE (labeled RR);
- Best companies in PI and Rest in PE (labeled BR);
- Rest companies in PI and Best in PE (labeled RB).

Among the set of all possible comparisons, after selecting only those more interesting from a research point of view (Figure 10), it is hypothesized that:

- H1: BX companies have higher level of drivers than RX companies;
- H2: XB companies have higher level of drivers than XR companies;
- H3: RB companies have higher level of drivers than RR companies;
- H3: BB companies have higher level of drivers than RR companies;
- H5: BB companies have higher level of drivers than BR companies.

In the empirical analysis conducted during the year 2000, we considered all NPs marketed from 1997 to 1999 by each company: this was defined as the NPD program. Market, product and financial measures of success refer to the results obtained as a result of the NPD program. For operational performances we considered the percentage of new products that have obtained high operational performances. As regards the drivers, in almost all cases we asked the company the percentage of projects which had adopted a certain driver. In other cases (i.e., capabilities and internal culture) we obtained the level of presence in the company as a whole, because it is practically impossible to discern the adoption percentage among projects for this type of variable.

Data and information were gathered through a questionnaire mailed to Italian manufacturing companies working in the B2B market in the mechanical and electronic sectors (SIC codes 35 & 36), with more than 100 and less than 1000 employees and a revenue of more than 20 billion Lire per year (approximately 10 million Euro). The addresses of the companies we mailed the questionnaire to were taken from Dun & Bradstreet's Business to Business database. The questionnaire was addressed to the new product development department manager. Phone assistance was provided to ensure that the information gathered was both complete and correct and some managers were interviewed. The sample was made up of 85 companies. Table 1 shows the composition of the sample used for the data analysis.

Figure 9. Graphical representation of Best and Rest definition.

Figure 10. Graphical representation of research hypotheses.

Table 1. Sample used for the data analysis.

Code	Description	N
SIC35	Machinery Manufacturing	60
SIC36	Electrical, Electronic Machinery, Equipment & Supplies	25
Total Sample Size		85

During the three year period considered (1997-1999), the firms launched a total of about 900 new products classified by the companies themselves as follows:

- 41% new products for new markets;
- 33% partially or totally substitute products;
- 26% products with significant improvements with respect to existing ones.

Best and Rest definition: the NonParametric Combination (NPC) of dependent rankings method

In many real situations we encounter the need to compare entities of a different nature (products, services, companies, behavior and so on) in order to obtain a ranking among the considered statistical units. If the comparison is based on only one feature the result is obtained in a trivial way but difficulties may arise when we are dealing with two or more informative variables jointly. We can build up as many rankings as the number of features we are dealing with. Apart from the case where units occupy the same position in every ranking, the need to summarize a set of rankings into one single global ranking arises.

The main purpose of the method (Pesarin, 2000) is to obtain a single ranking criterion for the statistical units under study, which summarizes many starting partial (univariate) criteria. This method is defined as nonparametric since it needs neither the knowledge of the underlying statistical distribution for the variables being studied, nor the dependence structure among variables, apart from the assumption that all dependences are monotonic regressions.

Methodology

Given a multivariate phenomenon $\mathbf{X}=[X_1, X_2, \dots, X_k]$, observed on N statistical units, and once we have calculated the k partial rankings $R_1, R_2, \dots, R_k$, starting from the variables $X_i, i=1, \dots, k$, each one being informative about a partial aspect of phenomenon $\mathbf{X}$, we want to build up a *global combined ranking* Y :

$$Y = \psi(X_1, X_2, \dots, X_k; w_1, w_2, \dots, w_k), \psi: \mathbb{R}^{2k} \rightarrow \mathbb{R}^1$$

where ψ is a real function allowing us to combine the partial dependent rankings and where $w_1, w_2, \dots, w_k$ is a set of weights, defined on the basis of technological, functional or economic considerations, which measure the relative degree of importance among the k aspects of $\mathbf{X}$.

In order to build up Y , a set of minimal reasonable conditions related to the variables $X_i, i=1, \dots, k$ are introduced:

1) For each of the k informative variables a partial ordering criterion is well established, in the sense that large is better; if it is not so, it is possible to recode the variables by means of any appropriate transformation φ :

a) if large is worse $\Rightarrow \varphi(X)=1/X$ or $\varphi(X)=-X$;

b) if δ is better (central target value) $\Rightarrow \varphi(X)=|X-\delta|$;

2) Regression relationships within the k informative variables are monotonic (increasing or decreasing)

3) The marginal distribution of each informative variable is non-degenerate.

Moreover, further assumptions need not be made, either on the statistical distribution of the informative variables, or on their dependence structure. Finally, notice that there is no need to assume the continuity of $X_i, i=1, \dots, k$, so that the probability of ex-equo can be different from zero.

Define the set of variables X_i as $\{Z_{ji}, i=1, \dots, k, j=1, \dots, N\}$, possibly after proper transformations. Without loss of generality, they are assumed to behave in accordance with the rule “large is better”. In this setting, we consider the rank transformations R_{ji} (*partial rankings*):

$$\{R_{ji} = R(Z_{ji}) = \#(Z_{ji} \geq Z_{hi}), i=1, \dots, k, j, h=1, \dots, N\}.$$

Associated with these ranks are the scores:

$$\left\{ \lambda_{ji} = \frac{R_{ji} + 0.5}{N + 1}, i = 1, \dots, k, j = 1, \dots, N \right\}.$$

Once a combining function ψ (for details of combining functions see paragraph 2.1 above) has been chosen, we compute the transformation

$$\psi: \{Y_j = \psi(\lambda_{j1}, \dots, \lambda_{jk}, w_1, \dots, w_k), j=1, \dots, N\},$$

and finally, applying the rank transformation, we obtain the global combined ranking Y :

$$\{Y_j = R(Y_j) = \#(Y_j \geq Y_h), j, h=1, \dots, N\}.$$

In the global ranking Y , each statistical units is ranked in a unique way, by taking into consideration the whole set of the k informative variables.

The method of nonparametric combination of dependent rankings has proved to be particularly useful for the problem of finding a meaningful classification criterion for the sample in groups, distinguishing companies which develop successful products from those which develop less successful products from the point of view of both Internal and External performances.

In fact, once the method is applied to the two sets of variables, the first measuring the *Market, Products and Financial* success (EP, External Performances) and the second measuring the *NPD Operational Performances* (IP, Internal Performances), obtain two global rankings of the companies, taking into account all success criteria.

Therefore, in these two global combined rankings the successful companies in External and Internal Performances were those in the upper positions while the worst companies were

those in the lower positions. As a discrimination rule, adopt the median positions. Those companies with a position above the median position in the global ranking were chosen as Best companies in EP and IP, and the remaining companies were labelled as Rest companies.

As a sensitivity analysis we performed an NPC testing procedure to verify whether the division was significant or not, that is to say whether Best companies in IP revealed a significantly higher level of operational variables and Best companies in EP revealed a significantly higher level of success variables.

As the associated p -values in Table 2 show, we can verify that at a 5% significance α -level the Best companies in IP are characterized by higher levels in all three operational variables and in the global test, taking into account the multivariate distribution of all three variables. In the same way the Best companies in EP are characterized by higher levels in all three success criteria and in the global test, taking into account the multivariate distribution of all three variables.

Internal Performances (IP)			Global
Launch on Time	Time to Market Red.	Quality Capability	
.000	.000	.015	.000

External Performances (EP)			Global
Meet Profit Goals	Overall Prod. Succ.	Meet Rev. Goals	
.000	.000	.000	.000

Table 2. Sensitivity analysis for testing the division in Best and Rest for both internal and external performances.

By simultaneously crossing the two rankings, the sample was divided into four classes, i.e. BB, BR, RB and RR (the first letter represents the internal performances), as shown in the Table 3. This final classification into four groups has been used to test the research questions.

Results

The NPC Test aims to identify the significant differences of the considered variables which characterize two specific groups. A p -value table is presented below for each of the five tested hypotheses (we use the graph representation only for the first hypothesis (Figure 11), where for sake of clarity a gray node means a significant p -value at a 5% α -level), reflecting the nested data set configuration in correspondence to the three variables classification: (1) multi-item scale variables, (2) variables belonging to the same driver group and (3) a final grouping which considers all driver groups together. As a result, the testing procedure is split up into the following phases:

- 1.1) is only for multi-item variables (if they are included in the driver group), performing the partial tests and
- 1.2) combines them into a single second order combined test;
- 2.) performs the other partial tests, in each group, for the remaining variables and
- 3.) combines them within the driver group, along with the combinations of step 1.2, obtaining a third order combined test of all variables within a driver group;
- 4.) finally, combines the four combined tests from step 2.2 (one for each driver group) in a global final test which is informative on the global null hypothesis.

In order to make the detecting of relevant differences easier, only significant p -values at 5% α -level have been printed in Table 4. The results suggest the Best companies on Market/Financial performances use the Architecture Approach more than the others. In particular, it is interesting to note that this group of variables does not discriminate between Best and Rest regarding the Operational Performances.

In other words it seems that an extensive use of product architecture related practices, such as the development of a product platform upon which to develop an entire new product line, the standardization of components to reduce production costs, modularity to offer a greater variety of products to the customer while

at the same time containing the internal variety the company has to deal with, allows the company to overcome any deficiencies in Operational Performances. This result is easier to understand if you consider the fact that some variables, which may influence external performances, have not been considered in the present study. These variables, such as for example the cost of the product on the market, are in turn influenced by company choices about the product architecture.

However, the main result is the great difference between the various groups in the Strategic Capability variables, and in particular the existence of a shared development strategy, well-defined development objectives and high technological capabilities. These variables represent the most noticeable difference between the various Best-Rest comparisons previously performed. In other words, strategic capabilities can help to achieve superior performance both on the operational and market/financial side. Perhaps these are the variables the companies have to act on in order to reach superior NPD performances, according to previous literature on this topic (see, for example, Griffin, 1998; and Cooper & Kleinschmidt, 1993).

A comparison between two different respiratory drugs

With the aim of comparing the features of two different respiratory drugs, a sample of 226 patients was recruited and then randomized into two distinct groups: group A, treated with a new drug labelled with A, and group B, treated with an old usual drug B. The purpose of the study is to establish, whether the new drug A is better than B, stressing the multivariate nature of the clinical comparison. In fact, we wish to make a decision on the basis of the three measured clinical end-points: D_MAT, D_SER and COMPL.

The first two clinical parameters are numeric variables which quantify the patient's health by means of a measure of respiratory airways expanding: D_MAT, is the difference, measured at noon, between the average of respiratory airways expanding evaluated two weeks before treatment (wash-out phase) and six weeks after treatment, and D_SER, the same difference, measured in the afternoon. The last

clinical parameter (COMPL) is a binary measure of therapy finishing: with this variable we can evaluate whether treatment A has a better degree of tolerance than B.

The null hypothesis states that A and B present no differences in their benefits, that is the equality in distribution of the multivariate distribution of the 3 responses in both groups:

$$H_0 : [P_A = P_B] = \left[\mathbf{X}_A \stackrel{d}{=} \mathbf{X}_B \right]$$

where $\mathbf{X}_A$ and $\mathbf{X}_B$ represent the multivariate random variables underlying group A and B, and P_A and P_B are the corresponding probability functions.

In the context of nonparametric combination, H_0 is supposed to be properly decomposed into

$$H_0 : \left[\begin{array}{l} D_MAT_A \stackrel{d}{=} D_MAT_B \\ \cap \left[D_SER_A \stackrel{d}{=} D_SER_B \right] \cap \left[COMPL_A \stackrel{d}{=} COMPL_B \right] \end{array} \right],$$

thus H_0 (multivariate) is true if all the H_{0i} (univariate) are jointly true.

The alternative hypothesis H_1 is represented by:

$$H_1 : \left[\begin{array}{l} D_MAT_A \stackrel{d}{>} D_MAT_B \\ \cup \left[D_SER_A \stackrel{d}{>} D_SER_B \right] \cup \left[COMPL_A \stackrel{d}{>} COMPL_B \right] \end{array} \right],$$

where $\stackrel{d}{>}$ means stochastic dominance. With 10000 CMC iterations results are shown in table 5.

It is concluded that from a multivariate point of view treatment A is better than B at 1% α -level. In order to take multiplicity into account, the FWE (Family Wise Error rate) must be considered to draw inferential conclusions, not only for the global test, but also for partial tests. At present, one of the best procedures is the *Closed Testing* (see e.g. Westfall et al., 1999).

For details on closed testing procedures with NPC, the reader should consider Finos et al. (2001). In this case-study, closed testing through NPC Test provided the result shown in Figure 12. Hence, after considering closed testing p-value corrections, the D_MAT and D_SER are both found to be significant, the first at 1% α -level and the second at 5% α -level.

The analysis can be extended by considering the same problem of the comparison between A and B treatments with the inclusion of a possible confounding factor, i. e., the patient's age. In order to do so we stratify the sample into Y, 4-8 year-old patients, and by O, 9-13 year-old patients.

In this way the hypothesis system is rewritten as:

$$H_0 : \bigcap_{j=Y,O} \left\{ \bigcap_{i=1}^k \left[X_{Ai} \stackrel{d}{=} X_{Bi} \right] \right\},$$

against the alternative

$$H_1 : \bigcup_{j=Y,O} \left\{ \bigcup_{i=1}^k \left[X_{Ai} \stackrel{d}{>} X_{Bi} \right] \right\}.$$

Notice that when we decide to aim our analysis at strata, we add a second step into our two-phase algorithm with the within-strata combination. Results are provided in Table 6.

The closed testing correction was also performed. Apart from a clinical interpretation of results which we do not consider here, it is worth noting that very complete information is provided by NPC Test analysis. Since the global test is significant at 1% α -level we are also able to identify:

- that only stratum O contributed to the overall significance;
- variables D_MAT and D_SER within stratum O contribute to the stratum significance.

Table 3. The four group definition.

EP External Perf.	IP (Internal Performances)		Tot.	Group	N.
	B	R			
B	25	13	38	BB	25
R	24	23	47	BR	24
Tot.	49	36	85	RB	13
				RR	23
				Tot.	85

Table 4. P-value table for each of the five tested hypotheses.

Driver / DRIVER GROUP	H1 BX>RX	H2 XB>XR	H3 RB > RR	H4 BB > RR	H5 BB > BR
Standardisation		.010			
Modularisation		.001	.001	.032	.019
Platform		.027	.039	.030	
PROD. ARCHITECTURE APPR.		.002	.006	.044	
Project Manager Use					
Customer Involvement 1					
Customer Involvement 2					
Customer Involvement					
Integration Design - Marketing					
Integration Design - Manufacturing					
Supplier Involvement 1					
Supplier Involvement 2					.022
Supplier Involvement					
Teame Use					
ORGANISATIONAL MECHANISMS					
Product Concept Development					
Product Concept Test					.024
Product Conceptpt					.041
Pre-Design 1					
Pre-Design 2					
Pre-Design					
Late Engineering Changes					
Overlapping Approach					
DEVELOPMENT PROCESS					
Up – Front Capabilities 1	.003			.000	.019
Up – Front Capabilities 2	.004			.004	
Up – Front Capabilities	.002			.000	.024
Top Management Support					.008
NP Strategic Guide 1	.010	.036		.009	
NP Strategic Guide 2	.010			.010	
NP Strategic Guide 3	.017				
NP Strategic Guide	.003			.002	
Company Innovation Culture		.004		.002	.003
Technological Capabilities 1	.023	.000		.000	.001
Technological Capabilities 2	.008	.034		.000	
Technological Capabilities	.008	.001		.000	.006
STRATEGICAL CAPABILITIES	.010	.003		.000	.012
GLOBAL	.037	.010	.022	.000	.044

Table 5.

A > B		D_MAT	D_SER	COMPL	GLOBAL
	<i>p-value</i>	0.0009	0.0014	0.0660	0.0003

Figure 11. Graphical representation of testing hypothesis BB>RR.

Figure 12: Closed testing procedure performed by NPC Test.

Table 6: P-values of partial, within-strata and global test considering stratification by patient’s age.

A > B	AGE	D_MAT	D_SER	COMPL	COMBINED
	Y: 4-8	0.0572	0.0520	0.1768	0.0520
	O: 9-13	0.0046	0.0353	0.2303	0.0068
	GLOBAL				0.0095

References

- Booz, Allen, & Hamilton. (1982). *New products management for the 1980s*. NY: Booz, Allen & Hamilton.
- Cooper, R. G. (1990, November - December). New products: what distinguishes the winners. *Research and Technology Management*, p. 27-31.
- Cooper, R. G. (1993). *Winning at new products: accelerating the process from idea to launch*. Reading, MA: Addison-Wesley.
- Cooper, R. G., & Kleinschmidt, E. J. (1993). New product success in the chemical industry. *Industrial Marketing Management*, 22 (2), 85-99.
- Droge, C., Jayaram, J., & Vickery, S. K. (2000). The ability to minimize the timing of new product development and introduction: an examination of antecedent factors in the north american automobile supplier industry. *Journal of Product Innovation Management*, 17 (1), 24-40.
- Finos, L., Pesarin, F., & Salmaso, L. (2001). *Test Combinati per il controllo della molteplicità mediante procedure di Closed Testing*. Dipartimento di Statistica, Università di Padova, Italy; Working Paper n.8/2001.
- Folks, J. L. (1984). *Combinations of independent tests*. In P.R. Krishnaiah & P. K. Sen (Eds.), *Handbook of Statistics*, 4, 113-121, Amsterdam: North-Holland.
- Griffin, A. (1997). PDMA research on new product development practices: Updating trends and benchmarking best practices. *Journal of Product Innovation Management*, 14 (6), 429-458.
- Griffin, A. (1998). *Drivers of new product development success: multivariate results from PDMA's best practices research*. In: EIASM - 5th International Product Development Management Conference, Como, I, p. 491-502.
- Griffin, A., & Page, A. L. (1993). An interim report on measuring product development success and factors. *Journal of Product Innovation Management*, 10 (3), 291-308.
- Hartley, J. L., Zirger, B. J., & Kamath, R. R. (1997). Managing the buyer - supplier interface for on-time performance in product development. *Journal of Operations Management*, 15 (1), 57-70.
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods*. (2nd Ed.). New York: Wiley.
- Joe, H. (1997) *Multivariate models and dependence concepts*. London: Chapman & Hall.
- Link, P. L. (1987). Keys to new product success and failure. *Industrial Marketing Management*, 16, 109-118.
- Madique, M. A., & Zirger, B. J. (1985). A study of success and failure in product innovation: The case of the u.s. electronics industry. *IEEE Transactions on Engineering Management*, 31 (4), 192-203.
- Methodologica Srl. (2001). *NPC Test© 2.0 - Statistical software for multivariate permutation tests*. Methodologica srl (<http://www.methodologica.it>), Treviso, Italy.
- Nelson, L. S. (1992). A randomization test for ordered alternatives. *Journal of Quality Technology*, 24, 51-53.
- Pesarin, F., & Lago, A. (2000). Nonparametric combination of dependent rankings with application to the quality assessment of industrial products. *Metron*, LVIII, 1-2.
- Pesarin, F. (2001). *Multivariate permutation tests with applications in biostatistics*. Chichester: Wiley.
- Pesarin, F. (2002, in press). Extending conditional inferences to unconditional ones. *Journal of the Italian Statistical Society*. 11, 161-173.
- Pittiglio, Rabin, Todd, & McGrath, (1995). *Product development leadership for technology - based companies: Measurement and management - a prelude to action.*, Weston, MA: Pittiglio Rabin Todd & McGrath.
- Ragatz, G. L., Handfield, R. B., & Scannell, T. V. (1997). Success factors for integration suppliers into new product development. *Journal of Product Innovation Management*, 14, 190-202.
- Swink, M. L., & Mabert, V. A. (2000). product development partnership: balancing the needs of OEMs and suppliers. *Business Horizons*, 43 (3), 59-68.
- Tuten, T. L., & Urban, D. J. (2001). An Expanded Model of Business-to-Business Partnership Formation and Success. *Industrial Marketing Management*, 30 (2), 149-164.
- Westfall, P., Tobias, R., Rom, D., Wolfinger, R., & Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests using the SAS®*, SAS Books by Users.

Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances

Xiaosheng Lei

Janet K. Holt

T. Mark Beasley

University of Alabama, Birmingham Northern Illinois University University of Alabama, Birmingham

Data simulation was used to investigate whether tests performed on aligned ranks (Beasley, 2002) could be used as robust alternatives to parametric methods for testing a split-plot interaction with non-normal data and heterogeneous covariance matrices. Results indicated the aligned rank method do not have any distinct advantage over parametric methods in this situation.

Key words: Nonparametrics, repeated measures, covariance heterogeneity, split-plot, interaction

Introduction

Repeated measures designs involving two or more independent groups are among the most common experimental designs (see Keselman & Algina, 1996). The parametric technique used to analyze a design in which a repeated measures (i.e., within-subjects) factor is crossed with a between-subjects (i.e., independent grouping or treatment variable) factor is the split-plot analysis of variance (ANOVA). It can be expressed with the following linear model:

$$Y_{ijk} = \mu_{**} + \beta_j + \pi_{i(j)} + \tau_k + \beta\tau_{jk} + \tau\pi_{ik(j)} + \zeta_{ijk}, \quad (1)$$

where j is referenced to the J groups of the between-subjects factor, i is referenced to the n_j subjects nested within the j^{th} group, k is referenced to the K levels of the within-subjects

factor, ζ_{ijk} is a random error vector, and $N = \sum n_j$ is the total number of subjects.

The interaction of the between-subjects and the repeated measures factors is often of most interest in many applications of the split-plot design (Boik, 1993). It is tested with an F -ratio, $F(Y)$, that is distributed approximately as $F_{[(J-1)(K-1), (N-J)(K-1)]}$ under the null hypothesis:

$$H_{0(J \times K)}: \beta\tau_{jk} = 0, \text{ for all } j \text{ and } k. \quad (2)$$

When the ANOVA model in (1) involves a within-subjects factor with $K > 2$, it requires the pooled within-group covariance matrix to be *spherical* (Huynh & Feldt, 1970). For the univariate $F(Y)$ from model (1), the sphericity assumption implies that the random error components, ζ_{ijk} , are NID(0, σ_ϵ^2) for each of the JK cells. Several procedures that correct $F(Y)$ by an ϵ factor have been developed to adjust the degrees of freedom so that $F_\epsilon(Y)$ will be a valid test of the interaction when there are departures from sphericity (e.g. Huynh, 1978).

Another suggested approach for dealing with non-spherical data is the use of multivariate tests because they do not require sphericity of the covariance matrix. Multivariate test statistics assume multivariate normality for the K repeated measures. Because repeated measures designs can be analyzed with multivariate tests applied to $(K-1)$ transformed variables (Marascuilo & Levin, 1983), the multivariate normality assumption applied to split-plot designs implies

Xiaosheng Lei is a teacher at International Industrial and Commercial College, South-Central University for Nationalities, China. This article is based on her Master's project in Biostatistics at the UAB. Janet K. Holt is Associate Professor in the Department of Educational Technology, Research, and Assessment. This article was conducted while she was a Visiting Professor in Section Statistical Genetics, Department of Biostatistics at UAB. T. Mark Beasley is Associate Professor in the Section Statistical Genetics, Department of Biostatistics. Email: mbeasley@uab.edu.

that the random error components are independent and multivariate normal with means of zero and a common covariance matrix (i.e., $\text{NID}[\mathbf{0}_{(K-1)}, \mathbf{C}_K \mathbf{\Sigma} \mathbf{C}_K']$, where $\mathbf{0}_{(K-1)}$ is a $(K-1)$ vector of zeros, $\mathbf{C}_K$ is a $(K-1) \times K$ normalized matrix of contrasts among the K repeated measures, and $\mathbf{\Sigma}$ is the $K \times K$ pooled within-group population covariance matrix. In order to pool these covariance matrices across the J groups, however, there is the implicit assumption that they are equal:

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \dots = \mathbf{\Sigma}_j \dots = \mathbf{\Sigma}_J. \quad (3)$$

If these covariance matrices are not equal, multivariate statistics are known to be invalid in terms of inflated Type I error rates, especially with unequal sample sizes (Olson, 1974).

In practice, it is likely that both the sphericity and normality assumptions are violated. However, multivariate tests are prone to inflate Type I error rates with violations of the multivariate normality assumption, especially with a small sample size to number of repeated measures (N/K) ratio (e.g., Blair, Higgins, Karniski, & Kromrey, 1994). By contrast, univariate tests are generally conservative with data sampled from heavy-tailed distributions (Wilcox, 1993). Thus, as compared to their multivariate extensions, univariate tests are noted to be more robust to non-normality. For example, simulation studies have indicated that $F_{\varepsilon}(Y)$ adequately corrects for non-sphericity (Huynh, 1978) and is reasonably robust to non-normality (Keselman, Algina, Kowalchuk, & Wolfinger, 1999). However, there are many skewed, heavy-tailed distributions that can affect the performance of both univariate (e.g., Wilcox, 1993; Zimmerman & Zumbo, 1993) and multivariate parametric tests (e.g., Blair et al., 1994; Keselman, Carriere, & Lix, 1993).

Beasley (2002) suggested an aligned rank procedure as a robust alternative to testing the interaction in split-plot designs when the normality assumption is violated. A univariate approach was detailed for situations in which the sphericity assumption holds and multivariate approach was also suggested for the more common case of non-spherical covariance structures. These procedures demonstrated more

statistical power than parametric procedures when error distributions were highly skewed; however, the issue of heterogeneous covariance matrices was not addressed.

Heterogeneity of variance is known to affect the Type I error rate of both univariate (Scheffé, 1957) and multivariate tests (Olson, 1974). Two approaches for testing interaction effects in repeated measures designs when the homogeneity of covariance assumption does not hold are the approximate degrees of freedom (df) multivariate Welch-James (WJ) statistic (Johansen, 1980; Keselman, Algina, Wilcox, & Kowalchuk, 2000) and the Huynh (1978) Improved General Approximation (IGA) tests. Simulation studies have shown these two approaches to be generally robust. However, under some conditions of departures from normality, sphericity and variance homogeneity, the WJ and IGA procedures have been found to yield inflated Type I error rates (Algina & Keselman, 1998; Keselman, Kowalchuk, & Boik, 2000). The purpose of this study was to investigate whether Beasley's (2002) aligned rank procedure could be used as a robust alternative to parametric procedures, when the normality and homogeneity of covariance assumptions were violated. Specifically, we investigated whether applying the WJ or IGA test to aligned ranks controlled Type I error rates when covariance matrices and sample sizes were unequal.

Rank-based competitors relax the normality assumptions by assuming that the random error components are *independent identically distributed* (IID) random variables from some continuous distribution, not necessarily the normal (i.e., NID). The rank transform concept is appealing because from a univariate perspective all data points (Y_{ijk}) are observations of one dependent variable measured under K different conditions or time points. Because the rank transform is monotonic, it is commonly believed that the null hypothesis for the parametric test of interaction (i.e., $F(Y)$) from model (1) is similar to the null hypothesis for similar tests performed on ranks (e.g., $F(R)$), except statistical inferences concern mean ranks. However, when test statistics for interactions used in parametric analyses of factorial designs

are applied to monotone transformations (e.g., rank transformation), the resulting tests lack an invariance property (Headrick & Sawilowsky, 2000). Specifically, the expected value of ranks for an observation in one cell will have a non-linear dependence on the original means of the other cells. Thus, interaction and main effect relationships are not expected to be maintained after rank transformations are performed (e.g., Blair, Sawilowsky, & Higgins, 1987).

Given these problems encountered by interaction tests based on the rank transform when other non-null effects are present (e.g., Blair et al., 1987; Toothaker & Newman, 1994), one solution is to treat other effects as nuisance parameters and remove them from the scores before ranking and analysis. McSweeney (1967) developed a chi-square approximate statistic for testing the interaction using aligned ranks in the two-way layout. Hettmansperger (1984) developed a linear model approach in which the nuisance effects are removed by obtaining the residuals from a regression model. However, both of these alignment procedures were developed for the two-way between-subjects factorial design and thus are not desirable because they do not remove the subjects' individual differences effect that is nested in the between-subjects factor, $\pi_i(j)$ from model (1). Higgins and Tashtoush (1994) proposed subtracting the subject effect and the repeated measures main effect and then ranking the aligned data from 1 to NK as follows:

$$R_{ijk} = \text{Rank}(Y_{ijk} - \bar{Y}_{ij*} - \bar{Y}_{*k} + \bar{Y}_{**}), \quad (4)$$

where $\bar{Y}_{*k}$ is the marginal mean of the k^{th} measure averaged over all N subjects, $\bar{Y}_{ij*}$ is the mean for the i^{th} subject averaged across the K measures, and $\bar{Y}_{**}$ is the grand mean of all NK observations. Following Hettmansperger (1984), this alignment could also be accomplished by obtaining the residuals from a linear model in which Y_{ijk} is regressed on a set of $(N-1)$ dummy codes that represent the subjects effect ($\pi_i(j)$) and a set of $(K-1)$ contrast codes that represent the repeated-measures main effect (τ_k) from model (1).

Univariate Approach

Consistent with Iman, Hora, and Conover (1984), Higgins and Tashtoush (1994) recommended applying the split-plot ANOVA from model (1) to the aligned ranks ($F(R)$), thus replacing Y_{ijk} with R_{ijk} . It should be noted, however, that many of the properties of the original data transmit to ranks, including heterogeneity of variance (Zimmerman & Zumbo, 1993) and non-sphericity (Harwell & Serlin, 1994). Thus, when performing the split-plot ANOVA F on aligned ranks, df -correction methods may be employed if the pooled covariance matrix is non-spherical (e.g., $F_{\varepsilon}(R)$) or if the between-subjects covariance matrices are heterogeneous (e.g., $IGA(R)$).

ε -adjusted F-test

With increasing departures from sphericity, the ANOVA F -ratio demonstrates a general lack of robustness, resulting in increasingly liberal tests. Huynh and Feldt (1976) developed an ε -adjusted test for split-plot models. Lecoutre (1991) corrected this formula so that in split-plot designs $\hat{\varepsilon}$ is replaced with $\tilde{\varepsilon}$:

$$\tilde{\varepsilon} = \frac{(N - J + 1)(K - 1)\hat{\varepsilon} - 2}{(K - 1)(N - J - (K - 1)\hat{\varepsilon})}, \quad (5)$$

where $\hat{\varepsilon}$ is a sphericity parameter estimated from the sample pooled within-group covariance matrix ($\mathbf{S}$), $\mathbf{S} = \sum[(n_j - 1)(N - J)]\mathbf{S}_j$. $\mathbf{S}_j$ is the sample covariance matrix for the j^{th} group with elements:

$$s_{kk'} = \sum (R_{ijk} - \bar{R}_{jk})(R_{ijk'} - \bar{R}_{jk'}) / (n_j - 1),$$

and

$$\hat{\varepsilon} = \frac{[\text{tr}(\mathbf{C}_K \mathbf{S} \mathbf{C}'_K)]^2}{(K - 1)[\text{tr}(\mathbf{C}_K \mathbf{S} \mathbf{C}'_K)]^2}. \quad (6)$$

The Lecoutre adjusted test for the interaction, $F_{\varepsilon}(Y)$, is distributed approximately as

$$F_{[\tilde{\varepsilon} (J-1)(K-1), \tilde{\varepsilon} (N-J)(K-1)]}.$$

Keselman et al. (1999) reported that $F_{\varepsilon}(Y)$ provided effective Type I error control for non-normal data with non-spherical covariance structures; however, it demonstrated low power under several conditions. We will examine the statistical properties of calculating the $\hat{\varepsilon}$ estimate and the ε -adjusted F -test from the aligned ranks ($F_{\varepsilon}(R)$).

Improved General Approximate

$F_{\varepsilon}(Y)$ was designed to correct for non-sphericity only. Jointly, the assumptions of sphericity and homoscedasticity in split-plot designs are referred to as *multi-sample sphericity* (Huynh, 1978). When covariance matrices are unequal across levels of the between-subjects factor and the design is unbalanced, the ε -adjusted F statistics as well as multivariate approaches are not robust for testing the interaction (Huynh, 1978; Keselman & Keselman, 1990).

In cases of arbitrary (i.e., non-spherical and/or heteroscedastic) covariance matrices, Huynh (1978) proposed the IGA procedure to estimate the dfs for the test statistics in the split-plot design. In order to adjust the tests for violations of multi-sample sphericity, the IGA procedure uses $\tilde{c}F_{[\alpha, \tilde{h}'', \tilde{h}]}$ as the critical value

for the interaction test. The statistics for these critical values are defined in terms of the separate covariance matrices for each of the J groups, $\mathbf{S}_j$. Let $\mathbf{S}^*$ denote a block diagonal matrix with $\mathbf{S}_j/n_j$ as the j th diagonal block. All off-diagonal blocks consist of a $(K \times K)$ matrix of zeros. Also let $\mathbf{D} = \{\mathbf{I} - (\mathbf{1}(\mathbf{1}')/K)\}$ where $\mathbf{I}$ is a K dimensional identity matrix and $\mathbf{1}$ is a $(K \times 1)$ vector of ones. Define $\mathbf{G}$ as a matrix constructed of $J^2(K \times K)$ blocks. The j th diagonal block of $\mathbf{G}$ is $n_j(1 - n_j/N)\mathbf{D}$ and the off-diagonal blocks are $(-n_j n_{j'}/N)\mathbf{D}$. For testing the split-plot interaction:

$$\tilde{c} = \frac{(N - J)\text{tr}(\mathbf{GS}^*)}{(J - 1)\sum_{j=1}^J (n_j - 1)\text{tr}(\mathbf{DS}_j)} \quad (7)$$

and

$$\hat{h}'' = \frac{[\text{tr}(\mathbf{GS}^*)]^2}{\text{tr}(\mathbf{GS}^*)^2} \quad (8)$$

Algina and Oshima (1994) applied the Lecoutre correction to the IGA so that

$$\tilde{h}'' = \frac{(J - 1)[(N - J + 1)\hat{h}'' - 2(J - 1)]}{(N - J)(J - 1) - \hat{h}''} \quad (9)$$

Let $\mathbf{A}_j = \text{tr}(\mathbf{C}_K \mathbf{S}_j \mathbf{C}_K')$, $\mathbf{B}_j = \text{tr}(\mathbf{DS}_j)^2$, and $\tilde{h} = \hat{\eta}/\hat{\delta}$, where

$$\hat{\eta} = \sum_{j=1}^J \frac{(n_j - 1)}{(n_j + 1)(n_j - 2)} (n_j A_j^2 - 2B_j) + \sum_{j=1}^J \sum_{j' \neq j}^J (n_j - 1)(n_{j'} - 1) A_j A_{j'} \quad (10)$$

and

$$\hat{\delta} = \sum_{j=1}^J \frac{(n_j - 1)}{(n_j + 1)(n_j - 2)} [(n_j - 1)B_j - A_j^2]. \quad (11)$$

We will examine the statistical properties of performing the Huynh's (1978) IGA test on the aligned ranks ($\text{IGA}(R)$).

Multivariate Approach

Another suggested approach for dealing with non-spherical data is the use of multivariate tests because they do not require sphericity of the covariance matrix. However, multivariate tests have strict sample size requirements based on the number of repeated measures. Furthermore, the degrees-of-freedom (dfs) for the error term of the univariate $F(Y)$ can be much larger than the error dfs (df_e) for the F approximate tests for the multivariate approach. Thus, the multivariate approach may have less statistical power in small sample situations (Keselman & Algina, 1996).

Agresti and Pendergast (1986) recommended a multivariate F -test based on Hotelling's (1931) T^2 for testing repeated measures effects in a single sample design. Their results showed that this multivariate test held the Type I error rate near the nominal alpha with departures from normality and sphericity.

Harwell and Serlin (1997) confirmed these results and also demonstrated that the Akritas and Arnold (1994) chi-square approximate test, which is functionally related to the Agresti-Pendergast test, inflated Type I error rates with total sample sizes of $N = 30$ or less. However, these findings are limited to the single sample repeated measures design.

To extend the Agresti and Pendergast (1986) approach for testing the interaction in a split-plot design, define $\mathbf{E}$ as a $K \times K$ pooled-sample cross-product error matrix for the aligned ranks (4) with elements:

$$e_{kk'} = \sum (R_{ijk} - \bar{R}_{jk})(R_{ijk'} - \bar{R}_{jk'}) . \quad (12)$$

Let $\mathbf{E}_p$ be a $JK \times JK$ block diagonal matrix where the j^{th} block of the main "diagonal" for $\mathbf{E}_p$ is defined as $\mathbf{E}/n_j$, and all other off-diagonal blocks are zero. That is, $\mathbf{E}_p$ is the Kronecker product of a diagonal matrix $\mathbf{n}^* = \text{diag}\{1/n_1, 1/n_2, \dots, 1/n_J\}$ and $\mathbf{E}$, $\mathbf{E}_p = \mathbf{n}^* \otimes \mathbf{E}$. Also, define $\mathbf{R}_{JK} = [\bar{R}_{11}, \bar{R}_{12}, \dots, \bar{R}_{1K}, \bar{R}_{21}, \dots, \bar{R}_{2K}, \dots, \bar{R}_{J1}, \dots, \bar{R}_{JK}]'$ as a JK -dimensional vector of mean ranks and $\mathbf{C}_{JK}$ as a $(J-1)(K-1) \times JK$ contrast matrix that represents the interaction. In general, $\mathbf{C}_{JK}$ can be defined as $\mathbf{C}_{JK} = \mathbf{C}_J \otimes \mathbf{C}_K$, where $\mathbf{C}_J$ is a $(J-1) \times J$ contrast matrix for the between-subjects effect and $\mathbf{C}_K$ is a $(K-1) \times K$ contrast matrix for the repeated measures effect.

Based on Agresti and Pendergast (1986), the distribution of the statistic,

$$H_{(R)} = (\mathbf{C}_{JK} \mathbf{R}_{JK})' (\mathbf{C}_{JK} \mathbf{E}_p \mathbf{C}_{JK}')^{-1} (\mathbf{C}_{JK} \mathbf{R}_{JK}) \quad (13)$$

multiplied by $(N-1)$, should approximate a χ^2 distribution with $df = (J-1)(K-1)$ asymptotically. It should be noted that $H_{(R)}$ is the Hotelling-Lawley trace for the interaction effect from a multivariate profile analysis performed on the Rank Transformed scores. Consistent with Agresti and Pendergast (1986), transforming H to an F -test may better control Type I error rates as opposed to comparing $(N-1)H_{(R)}$ to a chi-square distribution with $df = (J-1)(K-1)$, especially with smaller sample sizes (Harwell &

Serlin, 1997). Based on Hotelling (1951), $H_{(R)}$ (13) is transformed to an F approximation statistic by:

$$F_{H(R)} = [2(sn+1)/(s^2(2m+s+1))]H_{(R)} , \quad (14)$$

where $s = \min[(J-1), (K-1)]$, $m = [(K-J-1)/2]$, and $n = [(N-J-K)/2]$. This F approximation has numerator dfs of $df_h = [s(2m+s+1)] = [(J-1)(K-1)]$ and denominator dfs of $df_e = [2(sn+1)]$. Alternatively, a researcher could obtain a critical value for $H_{(R)}$ (13) from the sampling distribution of the Hotelling-Lawley trace using the s , m , and n parameters.

Keselman et al. (1993) suggested the use of the Welch-James test (Johansen, 1980) test for unbalanced within-subjects designs when covariance matrices were heterogeneous. The test statistic uses the same quadratic form as (13); however, separate covariance matrices are used:

$$WJ_{(R)} = (\mathbf{C}_{JK} \mathbf{R}_{JK})' (\mathbf{C}_{JK} \mathbf{S}^* \mathbf{C}_{JK}')^{-1} (\mathbf{C}_{JK} \mathbf{R}_{JK}) \quad (15)$$

where, $\mathbf{S}^*$ is a $JK \times JK$ block diagonal matrix where the j^{th} block of the main "diagonal" is defined as $\mathbf{S}_j/n_j$, and all other off-diagonal blocks are zero, $\mathbf{S}^* = \mathbf{n}^* \otimes \mathbf{S}$. The $WJ_{(R)/c}$ is distributed approximately as $F[f_1, f_2]$ with $f_1 = (J-1)(K-1)$, $f_2 = f_1(f_1+2)/3A$, $c = f_1 + 2A - 6A/(f_1+2)$ and

$$A = \frac{1}{2} \sum_{j=1}^J [\text{tr}\{\mathbf{S} \mathbf{C}_K' (\mathbf{C}_K \mathbf{S} \mathbf{C}_K') \mathbf{C}_K \mathbf{Q}_j\}^2 + \{\text{tr}(\mathbf{S} \mathbf{C}_K' (\mathbf{C}_K \mathbf{S} \mathbf{C}_K') \mathbf{C}_K \mathbf{Q}_j)^2\} / (n_j - 1) .$$

The $\mathbf{Q}_j$ matrix is a $JK \times JK$ block diagonal matrix corresponding to the j^{th} group. The $(s,t)^{\text{th}}$ block of $\mathbf{Q}_j$ is $\mathbf{I}_K$ if $s=t=j$ and $\mathbf{0}$ otherwise.

Olson (1974) showed that the Pillai-Bartlett trace (V) was more robust to violations to the normality and homogeneity of covariance assumptions. Applied to the aligned ranks it is computed as:

$$V_{(R)} = (\mathbf{C}_{JK} \mathbf{R}_{JK})' (\mathbf{C}_{JK} \mathbf{T} \mathbf{C}_{JK}')^{-1} (\mathbf{C}_{JK} \mathbf{R}_{JK}) \quad (16)$$

where, $\mathbf{T}$ is the Total sum of Squares matrix with elements defined as:

$$t_{kk'} = \sum \sum (R_{ijk} - \bar{R}_{*k})(R_{ijk} - \bar{R}_{*k'}) ,$$

and $\bar{R}_{*k}$ is the aligned rank mean for the k^{th} measure for all J groups combined. $V(R)$ (16) is transformed to an F approximation statistic by:

$$F_{V(R)} = [(2n+s+1)/(2m+s+1)][V/(s-V)] . \quad (17)$$

This F approximation has numerator dfs of $df_h = [s(2m+s+1)] = [(J-1)(K-1)]$ and denominator dfs of $df_e = [s(2n+s+1)]$. Again, a researcher could obtain a critical value for V (16) from the sampling distribution of the Pillai-Bartlett trace using the s , m , and n parameters.

For aligned ranks, the major purpose of the alignment process (4) is to remove the nuisance effects (i.e., main effects) so that test statistics will be sensitive to the effect of interest (i.e., interaction). The alignment process simply removes the mean values for the nuisance main effects, thus involving linear transformations of the data; however, the aligned ranks are a monotone transformation of the aligned data. Therefore, the aligned ranks (R_{ijk}) are placeholders for the percentiles of the original data (Y_{ijk}) with the nuisance location parameters removed. In either case, there is no guarantee that test statistics performed on R_{ijk} will reflect differences in location parameters without additional assumptions.

For the univariate test to be valid, under the null hypothesis in (2), not only are all J groups expected to have identical error distributions, but the error distributions for the K repeated measures are also expected to be identically distributed: $\text{NID}(0, \sigma_\epsilon^2)$ for all j and k . Similar to this sphericity assumption for univariate parametric tests, a rank-based version simply does not require *normal* error distributions. Thus, for rank-based tests, if the univariate assumption that all JK cells have *identically* shaped error distributions with a common variance (i.e., $\text{IID}[0, \sigma_\epsilon^2]$ for all j and k) is tenable, then statistically significant values for test statistics performed on the aligned ranks (4)

implies that the interaction is due to shifts in the location parameters (Lehmann, 1998). To illustrate the shift model for the univariate approach to the split-plot design, define the null hypothesis as:

$$\begin{aligned} H_0(J \times K): \mathbf{G}_1(\mathbf{Y}_1 - \mathbf{1}\Delta_1) &= \mathbf{G}_2(\mathbf{Y}_2 - \mathbf{1}\Delta_2) = \dots \\ &= \mathbf{G}_j(\mathbf{Y}_j - \mathbf{1}\Delta_j) = \dots = \mathbf{G}_J(\mathbf{Y}_J - \mathbf{1}\Delta_J) \end{aligned} \quad (18)$$

where $\mathbf{G}_j(\mathbf{Y}_j)$ is the K -dimensional distribution function of the original scores for the j^{th} group, $\mathbf{Y}_j$ is the $N \times K$ data matrix for the j^{th} group, $\Delta_j = [\delta_{j1} \ \delta_{j2} \ \dots \ \delta_{jk} \ \dots \ \delta_{jK}]$ is a $1 \times K$ vector of location parameters for the j^{th} group, and $\mathbf{1}$ is an $N \times 1$ vector of ones (Agresti & Pendergast, 1986, p. 1418). By requiring the univariate $\text{IID}[0, \sigma_\epsilon^2]$ assumption, if (18) is true then a statistically significant test statistic (i.e., $F(R)$) implies that the interaction is due to shifts in location parameters, a result conceptually similar to a rejection of the parametric null hypothesis in (2).

To illustrate the shift model for the multivariate approach to the split-plot design, define the null hypothesis as:

$$\begin{aligned} H_0(J \times K): \mathbf{G}_1(Y_{1k} - \delta_{1k}) &= \mathbf{G}_2(Y_{2k} - \delta_{2k}) = \dots \\ &= \mathbf{G}_j(Y_{jk} - \delta_{jk}) = \dots = \mathbf{G}_J(Y_{Jk} - \delta_{Jk}) , \end{aligned} \quad (19)$$

$$\text{for } k = 1, \dots, K .$$

$\mathbf{G}_j(Y_{jk})$ is the one-dimensional distribution function of the k^{th} repeated measure for the j^{th} group, Y_{jk} is the $N \times 1$ data matrix for the j^{th} group on the k^{th} measure and δ_{jk} is a scalar location parameter for the jk^{th} cell. This is similar to the $\text{NID}[\mathbf{0}_{(K-1)}, \mathbf{C}_K \Sigma \mathbf{C}_K']$ assumption for multivariate parametric tests except normal error distributions are not required. Under the multivariate model assumption that the random error vectors are $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{C}_K \Sigma \mathbf{C}_K']$ across the J groups, if (19) is true then a statistically significant multivariate test statistic performed on R_{ijk} implies that the interaction is due to shifts in location parameters. Again, this is a result conceptually similar to a rejection of the parametric null hypothesis in (2) and thus tests of shift parameter models (18 or 19) could be

used as robust alternatives to parametric procedures for testing interactions.

Note that the null hypotheses (18) and (19) are equivalent in terms of location parameters. If (18) is true so is (19); however, if (19) is true, it does not imply that (18) is true. Likewise, a false (18) does not imply a false (19). These distinctions are important because in order to test a null hypothesis of shifts in location parameters analogous to the null hypothesis in (2), the univariate null model for ranks (18) requires an assumption that the data for all JK cells are sampled from identically shaped distributions with a common variance. By contrast, the multivariate null model for ranks (19) only requires an assumption that the error distributions for each of the K repeated measures are identical for each of the J groups; however, there is no assumption that the error distributions for all K repeated measures are identically distributed. Thus, the relationship between the multivariate approach to analyzing aligned ranks and the F -ratio performed on aligned ranks is analogous to the relationship of the multivariate approach to repeated measures designs and the univariate approach that requires the sphericity assumption (Agresti & Pendergast, 1986).

Strictly speaking, if the assumption in (3) does not hold (i.e., the covariance matrices are heterogeneous), then neither the univariate (i.e., $\text{IID}[0, \sigma_{\epsilon}^2]$ for all j and k) nor multivariate $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{C}_K \Sigma \mathbf{C}_K']$ assumptions hold. The IGA test, Welch-James statistic, and the Pillai trace criterion have been shown to be generally robust to departures from homogeneous covariance assumption (3) for testing interaction among location parameters when normality holds. Thus, we investigated the use of the IGA, Welch-James, and Pillai tests applied to aligned ranks (4) as a robust alternative to testing interactions among location parameters (i.e., shift models 18 and 19) when assumptions of normality, sphericity, and homogeneous covariance matrices (3) do not hold.

Methodology

A 3 (sample size: $N = 30, 90$, and 150) $\times$ 3 (balanced, conservative unbalanced, and liberal unbalanced samples) $\times$ 2 (covariance structure:

spherical and non-spherical) $\times$ 3 (shape of error distribution: normal, double exponential, and exponential) factorial design was employed for this simulation study. For each of these conditions, 10,000 replications were generated using SAS/IML 8.2 (SAS Institute, 2001). Comparisons were made among procedures for testing the interaction effect in a $J=3 \times K=4$ split-plot design at the $\alpha=0.05$ significance level. For the aligned ranks (R_{ijk}), the following nine statistics were calculated: (a) the conventional F -test; (b) the Lecoutre (1991) ϵ -adjusted F ; (c) the $\text{IGA}_{(R)}$; (d) $H_{(R)}$ (13) using a critical value from the Hotelling-Lawley trace distribution, (e) the F approximate test for $H_{(R)}$ (14); (f) the $\text{WJ}_{(R)}$ test (15), (g) $V_{(R)}$ (16) using a critical value from the Pillai-Bartlett trace distribution, and (h) the F approximate test for $V_{(R)}$ (17).

For a $J=3 \times K=4$ split-plot design, the parameters for both the Hotelling-Lawley trace and Pillai-Bartlett trace distribution are $s = 2$, $m = 0$, $n = 11.5$ for $N = 30$, $n = 41.5$ for $N = 90$, and $n = 71.5$ for $N = 150$. The $\alpha=.05$ critical values for H are 0.587, 0.156, and 0.089 for $N = 30, 90$ and 150 , respectively. The $\alpha=.05$ critical values for V are 0.407, 0.139, 0.086 for $N = 30, 90$, and 150 , respectively.

The $N = 30$ condition was chosen because it has been used in other simulation studies (e.g., Agresti & Pendergast, 1986; Blair et al., 1987). Also, Harwell and Serlin (1997) reported that for a single sample, repeated measures design the multivariate F approximate test of rank transformed scores inflated Type I error rates with a total sample size of $N = 30$. For an unbalanced sample size, we used $\mathbf{n} = \{5, 10, 15\}$ for the "conservative" or positive pairing and the reverse for the "liberal" or negative pairing. For an unbalanced sample size with $N=90$ and $N=150$, we used $\mathbf{n} = \{15, 30, 45\}$ and $\mathbf{n} = \{25, 50, 75\}$, respectively, for the "conservative" or positive pairings and the reverse for the "liberal" or negative pairings.

The double exponential distribution was chosen as a condition where the errors were symmetric but heavy-tailed with skewness and kurtosis values of $\gamma_1=0$ and $\gamma_2=3$, respectively. The exponential distribution was selected as a condition where the errors were skewed ($\gamma_1=2$)

and extremely heavy-tailed ($\gamma_2=6$). Wilcox (1993) has noted that heavy-tailed distributions are common in practice and tend to inflate variances which in turn reduces power. In the case of empirical alpha rates, heavy-tailed distributions are likely to lead to Type I error rates that are below the nominal alpha. Micceri (1989) reported that 30.9% of the data from educational and psychological research had asymmetry as extreme as that of the exponential distribution. Furthermore, the exponential distribution condition is similar to the lognormal distribution ($\gamma_1=1.75$; $\gamma_2=5.90$) used in other simulation studies (e.g., Algina & Keselman, 1998; Algina & Oshima, 1994; Keselman et al., 1993). Moreover, it is representative of skewed, heavy-tailed distributions found in experimental psychology, most notably reaction time data (Zumbo & Coulombe, 1997).

Using the SAS/IML RANNOR function, a (n_j) by ($K=4$) matrix of normally distributed random variates with zero means and unit variances ($\mathbf{X}_j$) was generated for each of the $J=3$ groups. A covariance matrix Σ_j was subsequently imposed on the $\mathbf{X}_j$ scores by deriving a $K \times K$ matrix of principal component coefficients, $\mathbf{F}$, from the pre-specified covariance matrix (Σ_j) and pre-multiplying it by the transpose of $\mathbf{X}_j$ to create a data matrix $\mathbf{Y}_j$ that simulates Σ_j :

$$\mathbf{Y}_j' = \mathbf{F} \mathbf{X}_j' \quad (20)$$

(Beasley, 1994; Kaiser & Dickman, 1962).

In the first condition, all population correlations between measures (i.e., off-diagonal elements of Σ_j) were $\rho = 0.60$. This condition yielded results for a spherical covariance structure ($\varepsilon = 1$) in which case the univariate F -tests should not inflate Type I error rates with homogeneous covariance matrices. In the second condition, covariance structures with $\varepsilon = 0.64$ were imposed. The pairwise intercorrelations were ρ_{12} and $\rho_{34} = 0.70$ with all other population correlations equal to 0.30. These values were taken from Headrick and Sawilowsky (1999) and represent a realistic situation in which the sphericity assumption is

violated because a measure taken at time point $k=1$ is more correlated with a measure taken at time $k=2$ than it is with measures taken later in the experiment (i.e., time points $k=3$ and 4). Likewise, measures taken at time points $k=3$ and 4 were more correlated with each other than with previous measurements.

Two conditions of error non-normality were simulated: exponential and double exponential. To simulate the error distributions for both non-normal conditions, intermediate population correlation values were derived (see Headrick & Sawilowsky, 1999) for each of the three covariance structure conditions described above. First, the random normal variates ($\mathbf{X}_j$) were generated. Then, a matrix of principal component coefficients, $\mathbf{F}$, was derived from the intermediate values for the pre-specified correlation matrix. Subsequently, covariance structures with the intermediate values were imposed using (20). Then, data transformations using an extended Fleishman (1978) power method were performed (Headrick & Sawilowsky, 1999).

This process yielded data with zero means, unit variances, and the expected covariance structure (Σ_j) after the non-linear transformations were performed to make these values non-normal. Thus, these values were transformed so that the variances and shapes of each of the K error components were the same. This transformation process was also completed for each of the $J=3$ groups so that there were no between-group differences in variance or shape. Thus, under conditions in which the covariance matrices were homogeneous and spherical, the random error components (ζ_{ijk}) were IID($0, \sigma_\zeta^2$) for each of the JK cells, which permitted an investigation of the test statistics as robust alternative tests of interaction in terms of a univariate shift model for location parameters (18). Under conditions in which the covariance structures were homogeneous but not spherical, however, only the less restrictive multivariate assumption (IID[$\mathbf{0}_{(K-1)}, \mathbf{C}_K \Sigma \mathbf{C}_K'$]) was valid, thus creating a violation of the assumptions for the univariate parametric F -tests.

To impose heterogeneous variances, the second group ($j=2$) was multiplied by $\sqrt{3}$ and

the third group ($j=3$) was multiplied by $\sqrt{5}$, thus yielding a $\Sigma_1 = 3\Sigma_2 = 5\Sigma_3$ ratio. This variance ratio has been used in several other simulation studies (e.g., Keselman, et al., 2000). A repeated measures main effect pattern resulting in no interaction was imposed (Blair et al., 1987, p. 1143) after multiplication to increase variance was completed. Specifically for group 1, a vector of constants, $\mathbf{c}_1 = [0 \ 0 \ 1 \ 0]$, was added to each observation for the $K=4$ repeated measures. For group 2, $\mathbf{c}_2 = [-.5 \ -.5 \ .5 \ -.5]$. For group 3, $\mathbf{c}_3 = [-1 \ -1 \ 0 \ -1]$.

When covariance matrices were not homogeneous then both univariate and multivariate IID assumptions were violated, and thus, we investigated whether tests performed on aligned ranks (4) can be used as robust alternatives to testing interactions among location parameters under this extreme violation of the shift model assumptions.

Results

For all tables, $F_{(R)}$ refers to the univariate ANOVA F -test, $F_{\varepsilon(R)}$ refers to the Lecoutre (1991) ε -adjusted F , $IGA_{(R)}$ refers to the Improved General Approximate, $H_{(R)}$ refers to testing the Hotelling-Lawley trace (13) with a critical value from its referent distribution, $FH_{(R)}$ refers to the F approximation (14), $WJ_{(R)}$ refers to the Welch-James test (15), $V_{(R)}$ refers to testing the Pillai-Bartlett trace (13) with a critical value from its referent distribution, $FV_{(R)}$ refers to the F approximation (18). The subscript R indicates that the tests were performed on the aligned ranks (R_{ijk}). The results for the condition in which the $K=4$ repeated measures were equicorrelated and thus spherical are labeled as $\varepsilon = 1.00$ and $\varepsilon = 0.64$ refers to the non-spherical condition.

For this study, tests that demonstrated a Type I error rate lower than 0.05 were considered conservative but acceptable, while those with rates that were significantly above the nominal alpha were considered unacceptably liberal. Given $\alpha=0.05$ and 10,000 replications, a simulated estimate has a standard error of 0.0022. Thus, for empirical estimates of Type I error rates, any rejection rate two standard errors above 0.05 (i.e., 0.0544) was considered significantly liberal. This is consistent with Bradley's (1978) criterion of non-robustness in which the empirical Type I error rate should never exceed 1.1α . Likewise, any rejection rate below 0.0456 was considered significantly below the nominal alpha (i.e., conservative).

Tables 1, 2, and 3 show the rejection rates for the eight tests under conditions of heterogeneous covariance matrices. It is apparent that, for the conditions simulated in this study, none of the tests adequately controlled the Type I error rate when assumption (3) did not hold. As expected, most tests, with the exception of $IGA_{(R)}$ and $WJ_{(R)}$, produced rejection rates well above the nominal alpha with a liberal sample size-covariance pairing.

Also as expected, rejection rates for most tests were significantly below the nominal alpha with a conservative sample size-covariance pairing. The $IGA_{(R)}$ and $WJ_{(R)}$ were the best at controlling the Type I error rate. That is, these two procedures had rejection rates that were closest to the nominal alpha but were nevertheless unacceptably liberal under many conditions. Rejection rates for $IGA_{(R)}$ were similar for both sample sizes of $N=30$ and 90. By contrast, rejection rates for $WJ_{(R)}$ became less liberal with an increase in sample size from $N=30$ to 90. Therefore, $WJ_{(R)}$ was more sensitive to smaller sample sizes. A larger sample size of $N=150$ was used to investigate whether the $IGA_{(R)}$ and $WJ_{(R)}$ tests would eventually yield Type I error rates near the nominal alpha. Although these rejection rates reported in Table 3 are closer to $\alpha=0.05$, these values were consistently around 6 to 7.5% rejection.

Conclusion

One reason to use tests based on aligned ranks is that they have demonstrated superior power for detecting interactions in split-plot designs when error distributions are identically skewed with a common variance (Beasley, 2002). However, heterogeneous covariance matrices violate both the univariate (i.e., $\text{IID}[0, \sigma_\epsilon^2]$ for all j and k) and multivariate $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{C}_K \Sigma \mathbf{C}_K']$ assumptions. Results indicated that although the $WJ_{(R)}$ and $\text{IGA}_{(R)}$ produced relatively stable rejection rates across sample size – covariance pairing conditions, both tests yielded rejection rates significantly above the nominal alpha. However, $WJ_{(R)}$ required a much larger sample size ($N=150$) to produce rejection rates consistently around 6 to 7.5%. Perhaps, additional df correction could be applied, but it must be considered that the conditions imposed in this simulation study are rather extreme violations of the IID assumptions. Furthermore, for sample sizes this large the Type I error rates for the Welch-James test performed on the original non-normal ($WJ_{(Y)}$) are as close to the nominal alpha (Keselman et al., 2000) as the error rates for the Welch-James test performed on the aligned ranks ($WJ_{(R)}$; see Table 3). Moreover, for larger sample sizes the expected power advantage of $WJ_{(R)}$ over $WJ_{(Y)}$ is likely to be negligible, except for extremely small interaction effects. Thus, when covariance matrices are drastically unequal, it appears that aligned rank procedures cannot be used as robust alternatives to testing interaction among location parameters (i.e., shift models 18 and 19). Therefore, issues concerning the interpretation of rank-based tests are of concern.

Multivariate procedures performed on aligned ranks test a null hypothesis of distributional equivalence across the J groups for each of the K measures (Beasley, 2002).

However, situations where distributional equivalence does not hold but location parameters are identical only occur in symmetric distributions (Vargha & Delaney, 1998). Hence the null hypothesis being tested with asymmetric distributions and heterogeneous variances with rank data becomes one of location and variance differences. In other words, imposing the situation of unequal variances violates the restrictive assumption of the shift model (Lehmann, 1998) and explains the inflated Type I error rates that occur in the $F_{(R)}$ results. The effects of distributional nonequivalence are manifested in the Type I error rates of the other rank statistics tested in this study, including the Welch-James, the IGA, and the Pillai trace.

Therefore, WJ and IGA, as well as other tests, performed on the aligned ranks cannot be used as robust alternatives to testing the interaction in a split-plot design when assumption (3) does not hold. That is, when covariance matrices are heterogeneous, tests performed on the aligned ranks will detect between-group distributional differences to some extent, and thus, a statistically significant result cannot be attributed solely to differences among location parameters.

This is important because there are situations where the interaction null hypothesis in (19) would be rejected and the researcher might assume it was due to differences in location parameters when in actuality the rejection resulted from other between-group distributional (i.e., variance, shape) differences (Agresti & Pendergast, 1986; Beasley, 2002; Serlin & Harwell, 2001; Vargha & Delaney, 1998). For this reason, we do not recommend the Welch-James, the IGA, or the Pillai trace as tests of interaction among location parameters if covariance heterogeneity is suspected.

Table 2. Empirical Type I Error Rates ($\alpha=.05$) for the Interaction Tests in the Presence of a Repeated Measures Main Effect ($c = .50$) with a $\Sigma_1 = 3\Sigma_2 = 5\Sigma_3$ ratio and $N=90$.

n_1 n_2 n_3	Normal		Double Exponential		Exponential	
30 30 30 ^B	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.0748	.0981	.0734	.1015	.0797	.0982
$F_{\varepsilon(R)}$	.0743	.0744	.0732	.0746	.0786	.0737
	.0692	.0694	.0676	.0688	.0730	.0690
IGA _(R)						
$H_{(R)}$	.0766	.0736	.0724	.0731	.0761	.0751
$F_{H(R)}$	.0803	.0788	.0766	.0761	.0806	.0797
$WJ_{(R)}$	.0776	.0725	.0776	.0755	.0826	.0823
$V_{(R)}$	.0787	.0770	.0746	.0741	.0785	.0776
$F_{V(R)}$	.0766	.0744	.0724	.0710	.0767	.0751
15 30 45 ^C	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.0263	.0473	.0300	.0521	.0335	.0541
$F_{\varepsilon(R)}$	.0260	.0323	.0299	.0366	.0333	.0375
	.0601	.0602	.0644	.0649	.0647	.0623
IGA _(R)						
$H_{(R)}$	.0266	.0255	.0290	.0287	.0332	.0331
$F_{H(R)}$	.0298	.0273	.0314	.0310	.0345	.0355
$WJ_{(R)}$	.0713	.0727	.0783	.0823	.0875	.0842
$V_{(R)}$	.0275	.0253	.0291	.0295	.0335	.0338
$F_{V(R)}$	.0263	.0245	.0286	.0285	.0320	.0327
45 30 15 ^L	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.1441	.1518	.1374	.1460	.1259	.1483
$F_{\varepsilon(R)}$	.1433	.1204	.1373	.1138	.1245	.1154
	.0731	.0712	.0691	.0667	.0667	.0711
IGA _(R)						
$H_{(R)}$	.1382	.1370	.1366	.1307	.1197	.1270
$F_{H(R)}$	.1444	.1434	.1428	.1368	.1261	.1346
$WJ_{(R)}$	.0805	.0754	.0782	.0710	.0740	.0763
$V_{(R)}$	.1414	.1397	.1402	.1346	.1244	.1307
$F_{V(R)}$	.1389	.1359	.1360	.1314	.1221	.1284

Table 3. Empirical Type I Error Rates ($\alpha=.05$) for the Interaction Tests in the Presence of a Repeated Measures Main Effect ($c = .50$) with a $\Sigma_1 = 3\Sigma_2 = 5\Sigma_3$ ratio and $N=150$.

n_1 n_2 n_3	Normal		Double Exponential		Exponential	
50 50 50 ^B	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.0784	.0912	.0734	.0993	.0771	.0958
$F_{\varepsilon(R)}$	.0780	.0677	.0731	.0750	.0764	.0678
	.0719	.0651	.0687	.0707	.0716	.0645
IGA _(R)						
$H_{(R)}$	.0774	.0668	.0730	.0768	.0759	.0745
$F_{H(R)}$	.0794	.0684	.0743	.0789	.0775	.0767
$WJ_{(R)}$	.0798	.0706	.0740	.0778	.0791	.0802
$V_{(R)}$	.0784	.0663	.0724	.0779	.0759	.0747
$F_{V(R)}$	.0776	.0655	.0713	.0766	.0752	.0732
25 50 75 ^C	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.0266	.0477	.0274	.0481	.0387	.0573
$F_{\varepsilon(R)}$	.0264	.0338	.0274	.0344	.0384	.0421
	.0577	.0597	.0605	.0601	.0686	.0679
IGA _(R)						
$H_{(R)}$	.0265	.0271	.0281	.0295	.0381	.0336
$F_{H(R)}$	.0275	.0282	.0289	.0305	.0391	.0344
$WJ_{(R)}$	.0666	.0732	.0670	.0729	.0833	.0802
$V_{(R)}$	.0266	.0270	.0283	.0287	.0381	.0341
$F_{V(R)}$	.0262	.0263	.0275	.0284	.0372	.0337
75 50 25 ^L	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(R)}$	.1460	.1563	.1373	.1593	.1378	.1463
$F_{\varepsilon(R)}$	.1455	.1181	.1367	.1247	.1370	.1142
	.0697	.0668	.0720	.0711	.0767	.0721
IGA _(R)						
$H_{(R)}$	.1433	.1402	.1381	.1356	.1338	.1297
$F_{H(R)}$	.1464	.1431	.1404	.1387	.1366	.1324
$WJ_{(R)}$	.0765	.0755	.0792	.0744	.0787	.0751
$V_{(R)}$	.1437	.1403	.1373	.1366	.1347	.1305
$F_{V(R)}$	.1420	.1393	.1361	.1354	.1332	.1301

References

- Agresti, A., & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. *Communications in Statistics (A): Theory & Method*, 15, 1417-1433.
- Akritis, M.G., Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated-measures designs. *Journal of the American Statistical Association*, 89, 336-343.

- 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 & Behavioral Statistics*, 23, 152-169.

- Algina, J., Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. *British Journal of Mathematical & Statistical Psychology*, 47, 151-165.

- Beasley, T. M. (1994). CORRMTX: Generating correlated data matrices in SAS/IML. *Applied Psychological Measurement*, 18, 95.
- Beasley, T. M., (2002) Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*, 37, 197-226.
- Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research*, 29, 141-163.
- Blair, R. C., Sawilowsky, S. S., Higgins, J. J. (1987). Limitations of the rank transform statistic in test for interactions. *Communications in Statistics (B): Simulation & Computation*, 16, 1133-1145.
- Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics*, 18, 1-40.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, 31, 144-152.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532.
- Harwell, M. R., Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis*, 17, 35-49.
- Harwell, M. R., Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics (B): Simulation & Computation*, 26, 605-618.
- Headrick, T. C., Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal 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 (B): Simulation & Computation*, 29, 1059-1088.
- Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.
- Higgins, J. J., Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World*, 1, 201-211.
- Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Hotelling, H. (1951). A generalized T-test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics & Probability*, 2, 23-41.
- Huynh, H. (1978) Some approximate tests for repeated measurement designs. *Psychometrika*, 43, 161-175.
- Huynh, H., Feldt, L. S. (1970). Conditions under which mean squares ratios in repeated measurements designs have exact F distributions. *Journal of the American Statistical Association*, 65, 1582-1585.
- Huynh, H., Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1, 69-82.
- Iman, R. L., Hora, S. C., Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *Journal of the American Statistical Association*, 79, 674-685.
- Johansen, S. (1980). The Welch-James approximation of the distribution of the residual sum of squares in weighted linear regression. *Biometrika*, 67, 85-92.
- Kaiser, H. F., Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 27, 179-182.
- Keselman, H. J., Algina, J. (1996). The analysis of higher-order repeated measures designs. In Thompson, B. (Ed.), *Advances in social science methodology*, Vol. 4, pp. 45-70. JAI Press, Greenwich, CT.
- Keselman, H. J., Algina, J., Kowalchuk, R. K., Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical & Statistical Psychology*, 52, 63-78.

- Keselman, H. J., Algina, J., Wilcox, R. R., Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational & Psychology Measurement*, 60, 925-938.
- Keselman, H. J., Carriere, K. C., Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics*, 18, 305-319.
- Keselman, J. C., Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical & Statistical Psychology*, 43, 265-282.
- Keselman, H. J., Kowalchuk, R. K., Boik, R. J. (2000). An examination of the robustness of the Empirical Bayes and other approaches for testing main and interaction effects in repeated measures designs. *British Journal of Mathematical & Statistical Psychology*, 53, 51-67.
- Lecoutre, B. (1991). A correction for the e approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics*, 16, 371-372.
- Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks* (Revised 1st Ed.). Upper Saddle River, NJ: Prentice-Hall.
- Marascuilo, L. A., Levin, J. R. (1983). *Multivariate methods for the social science: A researcher's handbook*. Monterey, CA, Brooks/Cole.
- McSweeney, M. (1967). An empirical study of two proposed nonparametric test for main effects and interaction (Doctoral dissertation, University of California-Berkeley, 1968). *Dissertation Abstracts International*, 28(11), 4005.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, 69, 894-908.
- SAS Institute, 2001. *SAS/IML user's guide* (Release 8.2). Cary, NC.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Serlin, R. C., Harwell, M. R. (April, 2001). *A review of nonparametric test for complex experimental designs in educational research*. Paper presented at the American Educational Research Association. Seattle, WA.
- Toothaker, L. E., Newman, D. (1994). A. Nonparametric competitors to the two way ANOVA. *Journal of Educational & Behavioral Statistics*, 19, 237-273.
- Vargha, A., Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational & Behavioral Statistics*, 23, 170-192.
- Wilcox, R. (1993). Robustness in ANOVA. In Edwards, E. (Ed.), *Applied analysis of variance in the behavioral sciences*, pp. 345-374. Marcel Dekker, New York.
- Zimmerman, D., Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and the repeated-measures ANOVA on ranks. *Journal of Experimental Education*, 62, 75-86.
- Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology*, 51, 139-149.

Algorithms & Code The President's Problem

Jann-Huei Jinn
Department of Statistics
Grand Valley State University

A solution is offered in response to a complex combination problem challenged by Blom, Englund, and Sandell (1998). The problem is to determine the probability that a random permutation of the word BILLCLINTON has no equal neighbors.

Key words: Combinatorics, equal neighbors, random permutations, run

Introduction

The problem is to determine the probability that a random permutation of the word BILLCLINTON has no equal neighbors. Choose an initial order of the letters in the word Billclinton, for example, IINNLLBTCO. The problem is solved in three steps: Start with IINN, insert LLL, then insert B, T, C, and finally, insert O.

Methodology

Let X_1 be the number of equal neighbors in a random permutation of the four letters IINN. To obtain the solution, the probability function $P(X_1 = k)$ is needed. First, determine the probability function of the total number of runs (see references). Consider a random permutation of m 1's and n 0's. Denote by U the number of runs of 1's and by V the number of runs of 0's. The probability function $P(U = r, V = s)$ is needed. (Note that when $|r - s| > 1$, $P(U = r, V = s) = 0$.) The m 1's can be partitioned into r groups in $\binom{m-1}{r-1}$ ways.

Similarly, the n 0's can be partitioned into s groups in $\binom{n-1}{s-1}$ ways. It is known that the number of permutations with r 1-runs and s 0-runs is the product of these two binomial coefficients when $|r - s| = 1$, and twice that product when $r = s$. Since the total number of permutations is $\binom{m+n}{m}$, obtained is

$$P(U = r, V = s) = \frac{\binom{m-1}{r-1} \binom{n-1}{s-1}}{\binom{m+n}{m}} \quad (1.1)$$

for $r = 1, 2, \dots, m$ and $s = 1, 2, \dots, n$ such that $|r - s| = 1$. If $r = s$, then

$$P(U = r, V = r) = \frac{2 \binom{m-1}{r-1} \binom{n-1}{r-1}}{\binom{m+n}{m}} \quad (1.2)$$

Jann-Huei Jinn is a Professor in the Department of Statistics at Grand Valley State University. Allendale Michigan. Email: jinnj@gvsu.edu.

Step 1. Consider the permutation of IINN, we know that $m=n=2$ and Table 1 gives the probabilities $P(U = r, V = s)$ for this case.

Table 1: $m=n=2$ Probability Function
 $P(U=r, V=s)$.

$r \backslash s$	1	2
1	2/6	1/6
2	1/6	2/6

The probability function $P(U+V=k)$ of the total number of runs is obtained from the previous distribution (1.1) by summation. The result when $m=n=2$ is given in Table 2.

Table 2: $m=n=2$ Probability Function $P(U+V=k)$ of the Total Number of Runs

k	2	3	4
$P(U+V=k)$	2/6	2/6	2/6

Let W be the number of equal neighbors in the random permutation. The relation between runs and equal neighbors is $W = m+n-U-V$. Hence, when $m=n=2$, the probability of 2 equal neighbors is equal to the probability 2/6 of two runs, the probability of 1 equal neighbors is 2/6, and so on. Therefore, $X_1=4-(U+V)$. The probability function of X_1 required for the solution of the Statistician problem is given in Table 3.

Table 3: Probability Function of X_1 .

i	0	1	2
$P(X_1 = i)$	2/6	2/6	2/6

Step 2. Insert LLL. Let X_2 be the number of equal neighbors among the seven letters, IINNLLL, obtained. Since

$$P(X_2 = j) = \sum_i P(X_2 = j | X_1 = i) P(X_1 = i). \quad (1.3)$$

Consider $j=0,1,2,3,4$. Four letters B, T, C, and O can be inserted in the case of "IINNLLL" to get no equal neighbors. The conditional probabilities required are given in Table 4.

Table 4: Some Conditional Probabilities

$$P(X_2 = j | X_1 = i)$$

$i \backslash j$	0	1	2	3	4
0	60/210	120/210	30/210	0	0
1	36/210	72/210	78/210	24/210	0
2	18/210	48/210	78/210	48/210	18/210

All the probabilities given in Table 4 were calculated. Suppose that $X_1=0$, for example, ININ, the first L is inserted, the second L, and then the third L. It does not matter to insert the first L at any place, for example, if the first L is inserted to the right end of ININ, i.e., ININL, then !N!N!L! is obtained (where "!" represent the space to insert the second L). It is convenient to use a tree diagram to do illustration; see Figure 1.

Step 3. Insert B, T, C, and O (The order to insert).

Let X_3 be the number of equal neighbors among the eleven letters obtained.

$$P(X_3 = 0) = \sum_{j=0}^4 P(X_3 = 0 | X_2 = j) P(X_2 = j) \quad (1.4)$$

When $X_2=0$, for example, LILINLN, it does not matter where to insert B, T, C, O, there are no equal neighbors. Therefore, $P(X_3 = 0 | X_2 = 0) = 1$.

When $X_2=1$, for example, LIINLNL,

$$P(X_3 = 0 | X_2 = 1) = \frac{1}{8} + \frac{7}{8} \left(\frac{1}{9} + \frac{8}{9} \times \frac{1}{10} + \frac{8}{9} \times \frac{9}{10} \times \frac{1}{11} \right) = \frac{2880}{7920}$$

(see Figure 2 - 4).

Conclusion

Thus, the solution to the Presidents problem, the probability that a random permutation of the word BILLCLINTON has no equal neighbors, is 39/110.

Figure 1: Tree diagram for calculation of the conditional probability $p(X_2 = i | X_1 = 0)$.

Therefore, $P(X_2 = 2 | X_1 = 0) = \frac{2}{6} \times \frac{3}{7} = \frac{6}{42} = \frac{30}{210}$,

$P(X_2 = 1 | X_1 = 0) = \frac{2}{6} \times \frac{4}{7} + \frac{4}{6} \times \frac{4}{7} = \frac{24}{42} = \frac{120}{210}$, and $P(X_2 = 0 | X_1 = 0) = \frac{4}{6} \times \frac{3}{7} = \frac{12}{42} = \frac{60}{210}$.

Suppose that $X_1=1$, an example where this occurs is INNI . We obtain $X_2=0$ by separating the pair NN with the first L , the second L or the third L inserted, but not both or all three. It is convenient to use a tree diagram; see Figure 2. We obtain

$$P(X_2 = 0 | X_1 = 1) = \frac{1}{5} \times \frac{4}{6} \times \frac{3}{7} + \frac{4}{5} \times \frac{3}{6} \times \frac{1}{7} + \frac{4}{5} \times \frac{1}{6} \times \frac{3}{7} = \frac{36}{210},$$

$$P(X_2 = 1 | X_1 = 1) = \frac{1}{5} \times \frac{4}{6} \times \frac{4}{7} + \frac{1}{5} \times \frac{2}{6} \times \frac{4}{7} + \frac{4}{5} \times \frac{2}{6} \times \frac{1}{7} + \frac{4}{5} \times \frac{3}{6} \times \frac{2}{7} + \frac{4}{5} \times \frac{1}{6} \times \frac{4}{7} = \frac{72}{210},$$

$$P(X_2 = 2 | X_1 = 1) = \frac{1}{5} \times \frac{2}{6} \times \frac{3}{7} + \frac{4}{5} \times \frac{2}{6} \times \frac{3}{7} + \frac{4}{5} \times \frac{3}{6} \times \frac{4}{7} = \frac{78}{210}, \text{ and}$$

$$P(X_2 = 3 | X_1 = 1) = \frac{4}{5} \times \frac{2}{6} \times \frac{3}{7} = \frac{24}{210}.$$

Figure 2: Tree diagram for calculation of the conditional probability $P(X_2 = i | X_1 = 1)$.

Suppose that $X_1 = 2$, an example is II NN. Refer to Figure 3.

Figure 3: Tree diagram for calculation of the conditional probability $P(X_2 = i | X_1 = 2)$

$$P(X_2 = 0|X_1 = 2) = \frac{3}{5} \times \frac{2}{6} \times \frac{1}{7} + \frac{2}{5} \times \frac{3}{6} \times \frac{1}{7} + \frac{2}{5} \times \frac{1}{6} \times \frac{3}{7} = \frac{18}{210},$$

$$P(X_2 = 1|X_1 = 2) = \frac{3}{5} \times \frac{2}{6} \times \frac{2}{7} + \frac{3}{5} \times \frac{2}{6} \times \frac{2}{7} + \frac{2}{5} \times \frac{2}{6} \times \frac{1}{7} + \frac{2}{5} \times \frac{3}{6} \times \frac{2}{7} + \frac{2}{5} \times \frac{1}{6} \times \frac{4}{7} = \frac{48}{210},$$

$$P(X_2 = 2|X_1 = 2) = \frac{3}{5} \left[\frac{2}{6} \left(\frac{2}{7} + \frac{1}{7} + \frac{4}{7} \right) \right] + \frac{2}{5} \left(\frac{2}{6} \times \frac{3}{7} + \frac{3}{6} \times \frac{4}{7} \right) = \frac{78}{210},$$

$$P(X_2 = 3|X_1 = 2) = \frac{3}{5} \left(\frac{2}{6} \times \frac{2}{7} + \frac{2}{6} \times \frac{4}{7} \right) + \frac{2}{5} \times \frac{2}{6} \times \frac{3}{7} = \frac{48}{210}, \text{ and } P(X_2 = 4|X_1 = 2) = \frac{3}{5} \times \frac{2}{6} \times \frac{3}{7} = \frac{18}{210}.$$

Based on (1.3) and combining Table 3 and Table 4, we obtain

$$P(X_2 = 0) = \sum_{i=0}^2 P(X_2 = 0|X_1 = i) P(X_1 = i) = \frac{2}{6} \times \left(\frac{60}{210} + \frac{36}{210} + \frac{18}{210} \right) = \frac{228}{1260}. \text{ Similarly,}$$

$$P(X_2 = 1) = \frac{2}{6} \times \left(\frac{120 + 72 + 48}{210} \right) = \frac{480}{1260}, \quad P(X_2 = 2) = \frac{2}{6} \times \left(\frac{30 + 78 + 78}{210} \right) = \frac{372}{1260},$$

$$P(X_2 = 3) = \frac{2}{6} \times \left(\frac{24 + 48}{210} \right) = \frac{144}{1260}, \text{ and } P(X_2 = 4) = \frac{2}{6} \times \frac{18}{210} = \frac{36}{1260}.$$

Table 5: The Probability Function of X_2

j	0	1	2	3	4
$P(X_2 = j)$	228/1260	480/1260	372/1260	144/1260	36/1260

Figure 4: Tree diagram for calculation of the conditional probability $P(X_3 = 0|X_2 = 1)$.

When $X_2 = 2$, for example, NII LLNL , $P(X_3 = 0|X_2 = 2) = \frac{2}{8} \times \left(\frac{1}{9} + \frac{8}{9} \times \frac{1}{10} + \frac{8}{9} \times \frac{9}{10} \times \frac{1}{11} \right) + \frac{6}{8} \times \left(\frac{2}{9} \times \frac{1}{10} + \frac{2}{9} \times \frac{9}{10} \times \frac{1}{11} \right) + \frac{6}{8} \times \frac{7}{9} \times \frac{2}{10} \times \frac{1}{11} = \frac{864}{7920}$ (see Figure 5).

Figure 5: Tree diagram for calculation of the conditional probability $P(X_3 = 0 | X_2 = 2)$.

When $X_2 = 3$, for example, $\underline{\text{NN}} \underline{\text{LL}} \underline{\text{IIL}}$,

$$P(X_3 = 0 | X_2 = 3) = \frac{3}{8} \times \left(\frac{2}{9} \times \frac{1}{10} + \frac{2}{9} \times \frac{9}{10} \times \frac{1}{11} + \frac{7}{9} \times \frac{2}{10} \times \frac{1}{11} \right) + \frac{5}{8} \times \frac{3}{9} \times \frac{2}{10} \times \frac{1}{11} = \frac{192}{7920}$$

(see Figure 6).

Figure 6: Tree diagram for calculation of the conditional probability $P(X_3 = 0|X_2 = 3)$.

When $X_2 = 4$, for example, NN II LLL, $P(X_3 = 0|X_2 = 4) = \frac{4}{8} \times \frac{3}{9} \times \frac{2}{10} \times \frac{1}{11} = \frac{24}{7920}$
(see Figure 7).

Figure 7: Tree diagram for calculation of the conditional probability $P(X_3 = 0|X_2 = 4)$.

Therefore,

$$\begin{aligned}
 P(X_3 = 0) &= P(X_3 = 0|X_2 = 0)P(X_2 = 0) + P(X_3 = 0|X_2 = 1)P(X_2 = 1) + \\
 &P(X_3 = 0|X_2 = 2)P(X_2 = 2) + P(X_3 = 0|X_2 = 3)P(X_2 = 3) + P(X_3 = 0|X_2 = 4)P(X_2 = 4) \\
 &= 1 \times \frac{228}{1260} + \frac{2880}{7920} \times \frac{480}{1260} + \frac{864}{7920} \times \frac{372}{1260} + \frac{192}{7920} \times \frac{144}{1260} + \frac{24}{7920} \times \frac{36}{1260} = \frac{39}{110}.
 \end{aligned}$$

This is the answer to the Bill Clinton (President) problem.

References

- Blom, G., Englund, J., & Sandell, D. (1998). The Mississippi Problem. *The American Statistician*, 52, 49-50.
- Blom, G., Holst, L., & Sandell, D. (1994). *Problems and Snapshots from the World of Probability*, New York: Springer-Verlag.
- Constantine, G. M. (1987). *Combinatorial Theory and Statistical Design*. New York: Willey.
- David, F. N., & Barton, D. E. (1962). *Combinatorial Chance*, London: Griffin.
- Mood, A. M. (1940). The Distribution Theory of Runs, *The Annals of Mathematical Statistics*, 11, 367-392.

Pseudo-Random Number Generation In R For Commonly Used Multivariate Distributions

Hakan Demirtas
School of Public Health
University of Illinois at Chicago

An increasing number of practitioners and applied statisticians have started using the R programming system in recent years for their computing and data analysis needs. As far as pseudo-random number generation is concerned, the built-in generator in R does not contain multivariate distributions. In this article, R routines for widely used multivariate distributions are presented.

Key words: Simulation; computation; pseudo-random numbers

Introduction

Monte Carlo simulation has become one of the key tools in all fields of science. Simulation methodology relies on a good source of numbers that appear to be random. Methods for producing pseudo-random numbers and transforming those numbers to simulate samples from various distributions are among the most important issues in computational statistics.

For doing Monte Carlo studies, it is generally better to use a software system with a compilable programming language, such as Fortran or C. In addition to more flexibility and control, the programs built in the compiler languages execute faster. Libraries that are available in both Fortran and C contain a large number of pseudo-random number generation routines. However, using these libraries efficiently may be a daunting task for practitioners largely due to the fact that operational characteristics depend on the type of

compiler, computing platform, operating system, linker and debugger, which in turn may lead to implementation difficulties.

A fundamental shift has been witnessed in recent years among statistically oriented researchers towards an extensive usage of Splus. Splus is both a data analysis system and an object-oriented programming language. Unlike Fortran or C, Splus is an interpreted (not a compiled) language. A publicly available package, called R, provides essentially the same functionality as Splus. The R programming system can be downloaded and installed at www.r-project.org.

The built-in pseudo-random number generator in R does not have routines for multivariate distributions, therefore built-in codes are not available. The purpose of this paper is to provide complementary R routines for generating pseudo-random numbers from some important multivariate distributions. In the next section, eleven R functions are presented. The quality of the resulting variates has not been tested in the computer science sense (in terms of independence, d-variate uniformity, measures based on lattice structure, etc.). However, the first two moments for random vectors and the first moment for random matrices were rigorously tested. For the purposes of most applications, fulfillment of this criterion should be a reasonable approximation to reality.

Hakan Demirtas is an Assistant Professor of Biostatistics at the University of Illinois at Chicago. His research interests are the analysis of incomplete longitudinal data, multiple imputation and Bayesian computing. E-mail: demirtas@uic.edu.

Functions for Random Number Generation

The following abbreviations are used: *PDF* stands for the probability density function; *PMF* stands for the probability mass function; *CDF* stands for the cumulative distribution function; *GA* stands for the generation algorithm; *nrep* stands for the number of identically and independently distributed random variates; *d* is the dimension. The formal arguments other than *nrep* and *d* reflect the parameters in *PDF* or *PMF*. Auxiliary functions are included as needed.

Multivariate normal distribution

$$PDF : f(x | \mu, \Sigma) = c \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

for $-\infty < x < \infty$ and $c = (2\pi)^{-d/2} |\Sigma|^{-1/2}$, Σ is symmetric and positive definite, where μ and Σ are the mean vector and the variance-covariance matrix, respectively. *GA*: Using the Cholesky decomposition and a vector of univariate normal draws (see Code 1).

Code 1. Multivariate normal distribution:

```
draw.d.variate.normal<-function(nrep,d,mean.vec,cov.mat){
  if (nrep<1)|(floor(nrep)!=nrep)){ stop("Number of replicates must be
  an integer whose value is at least 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(length(mean.vec)!=d){
    stop("Mean vector is misspecified, dimension is wrong!\n")}
  if((ncol(cov.mat)!=d)|(nrow(cov.mat)!=d)){
    stop("Variance-covariance matrix is misspecified, dimension is
    wrong!\n")}
  if(min(eigen(cov.mat)$values)<0)
  {stop("Variance-covariance matrix must be symmetric and positive
  definite!\n")}
  z<-matrix(rnorm(nrep*d),nrep,d)
  x<- z*%chol(cov.mat)+t(matrix (rep(mean.vec,nrep),nrow=d))
  x}
```

Multivariate *t* distribution

$$PDF : f(x | \mu, \Sigma, \nu) = c \left(1 + \frac{1}{\nu}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)^{-(\nu+d)/2}$$

for $-\infty < x < \infty$ and

$$c = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}} |\Sigma|^{-1/2},$$

Σ is symmetric and positive definite and $\nu > 0$, where μ , Σ and ν are the mean vector, the variance-covariance matrix and the degrees of freedom, respectively. *GA*: Using the Cholesky decomposition, a vector of univariate normal and χ^2 draws (see Code 2).

Multivariate uniform distribution

The function in Code 3 generates a *d*-variate $U_d(0,1)$ distribution with specified covariance matrix Σ . *GA*: An approximate method of Falk (1999) based on *CDF* of multivariate normal deviates.

Code 2. Multivariate t distribution:

```

draw.d.variate.t<-function (df,nrep,d,mean.vec,cov.mat){
if((nrep<1)|(floor(nrep)!=nrep)){
stop("Number of replicates must be an integer whose value is at least
1!\n")}
if((d<2)|(floor(d)!=d)){
stop("Dimension must be an integer whose value is at least 2!\n")}
if(length(mean.vec)!=d){
stop("Mean vector is misspecified, dimension is wrong!\n")}
if((ncol(cov.mat)!=d)|(nrow(cov.mat)!=d)){
stop("Variance-covariance matrix is misspecified, dimension is
wrong!\n")}
if(min(eigen(cov.mat)$values)<0)
{stop("Variance-covariance matrix must be symmetric and positive
definite!\n")}
if (df<=1){
stop("Degrees of freedom must be greater than 1!\n")}
z<-matrix(rnorm(nrep*d),nrep,d)
x<-z%*%chol(cov.mat)
xt<-sqrt(df/rchisq(1,df))*x+t(matrix(rep(mean.vec,nrep),nrow=d))
xt}

```

Code 3. Multivariate uniform distribution:

```

draw.d.variate.uniform<-function(nrep,d,cov.mat){
if((nrep<1)|(floor(nrep)!=nrep)){
stop("Number of subjects must be an integer whose value is at least
1!\n")}
if(d<2){
stop("Number of variables must be at least 2!\n")}
if(length(mean.vec)!=d){
stop("Mean vector is misspecified, dimension is wrong!\n")}
if((ncol(cov.mat)!=d)|(nrow(cov.mat)!=d)){
stop("Variance-covariance matrix is misspecified, dimension is
wrong!\n")}
if(sum(cov.mat!=t(cov.mat))+min(eigen(cov.mat)$values<=0)){
stop("Variance-covariance matrix must be symmetric and positive
definite!\n")}
draw<-draw.d.variate.normal(nrep,d,mean.vec,cov.mat)
x<-pnorm(draw)
x}

```

Multivariate Bernoulli distribution (correlated binary data)

The function in Code 4 generates correlated binary variates using an algorithm developed by Park, Park and Shin (1996) based on sums of Poisson random variables in which the sums have some common terms. In Code 4, *mean.vec* corresponds to the expectations for each variable and *corr.mat* is the correlation matrix.

Multivariate hypergeometric distribution

PMF for the univariate hypergeometric distribution: $f(x | M, L, N) = \frac{\binom{M}{x} \binom{L-M}{N-x}}{\binom{L}{N}}$ for $x = \max(0, N-L+M), \dots, \min(N, M)$, where N is the number of items to be sampled, independently with equal probability and without replacement, from a lot of L items of which M are special; the realization of x is the number of special items in the random sample. In the multivariate case are more than two outcomes. *GA*: Sequential generation of succeeding conditionals which are univariate hypergeometric. In Code 5, *mean.vec* stands for the number of items in each category and k is the number of items to be sampled.

Code 4. Multivariate Bernoulli distribution (correlated binary data):

```
draw.correlated.binary<-function(nrep,d,mean.vec,corr.mat){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at
    least 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if((max(mean.vec)>=1)|(min(mean.vec)<=0)){
    stop("Expectations should be greater than 0 and less than 1!\n")}
  if(length(mean.vec)!=d){
    stop("Mean vector is misspecified, dimension is wrong!\n")}
  if((ncol(corr.mat)!=d)|(nrow(corr.mat)!=d)){
    stop("Correlation matrix is misspecified, dimension is wrong!\n")}
  if(sum(corr.mat!=t(corr.mat))>0){
    stop("Correlation matrix is not symmetric!\n")}
  if(sum(diag(corr.mat)!=rep(1,d))>0){
    stop("Not all diagonal elements of correlation matrix are 1!\n")}
  if((max(corr.mat)>1)|(min(corr.mat)<0)){
    stop("Correlations should be greater than or equal to 0 and less
    than or equal to 1!\n")}
  alpha<-matrix(0,d,d) ; cor.limit<-matrix(0,d,d)
  for (i in 1:d){
    for (j in 1:d){
      cor.limit[i,j]<-min(sqrt((mean.vec[j]*(1-mean.vec[i]))/
      (mean.vec[i]*(1-mean.vec[j]))),sqrt((mean.vec[i]*(1-mean.vec[j]))/
      (mean.vec[j]*(1-mean.vec[i]))))}
```

Code 4 Continued

```

if(sum(corr.limit>=corr.mat)<d^2){
stop("Correlations are beyond their upper limits imposed by
expectations")}
for (i in 1:d){for (j in 1:d){
alpha[i,j]<-log(1+corr.mat[i,j]*sqrt((1-mean.vec[i])*
(1-mean.vec[j]))/(mean.vec[i]*mean.vec[j])))}
beta<-matrix(0,d,d*d)
summ<-1 ; counter<-0 ; while (summ>0){
counter<-counter+1 ; minloc<-min.loc.finder(alpha); w<-matrix(1,d,d)
my.min<-apply(matrix(alpha[,-minloc],d,
d-length(unique(minloc))),2,min)
if (length(my.min)==1){w[,-minloc][my.min==0]<-0
w[-minloc,][my.min==0]<-0}; if (length(my.min)>1){
w[,-minloc][,my.min==0]<-0 ; w[-minloc,][my.min==0,]<-0
w[alpha==0]<-0}
for (i in 1:d){
beta[i,counter]<-
alpha[minloc[1],minloc[2]]*1*((minloc[1]==i)|(minloc[2]==i)|
(sum(w[,i])==d))}
alpha<-alpha-alpha[minloc[1],minloc[2]]*w
summ<-sum(alpha)} ; tbeta<-t(beta) ; w<-(tbeta!=0)
x<-matrix(0,nrep,d); y<-matrix(0,nrep,d)
pois<-numeric(nrow(tbeta)); sump<-numeric(d)
for (k in 1:nrep){for (j in 1:nrow(tbeta)){
pois[j]<-rpois(1,max(tbeta[j,]))} ; for (i in 1:d){
sump[i]<-sum(pois*w[,i])}
x[k,]<-sump}; y[x==0]<-1 ; y[x!=0]<-0
y}

min.loc.finder<-function(my.mat){
w<-is.matrix(my.mat)
if (w==F){stop("This is not a matrix!\n")}
if (nrow(my.mat)!=ncol(my.mat)){
stop("This is not a square matrix!\n")}
n<-nrow(my.mat) ; my.vec<-as.vector(t(my.mat))
my.vec[my.vec==0]<-999
my.index<-min((1:length(my.vec))[my.vec==min(my.vec)])
row.index<-floor((my.index-1)/n)+1
col.index<-my.index-d*floor((my.index-1)/n)
c(row.index,col.index)}

```

Code 5. Multivariate hypergeometric distribution:

```

draw.multivariate.hypergeometric<-function(nrep,d,mean.vec,k){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least
    1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(length(mean.vec)!=d){
    stop("Number of items are misspecified, dimension is wrong!\n")}
  if(min(mean.vec)<=0){
    stop("Number of items vector cannot contain non-positive numbers!\n")}
  if(sum(floor(mean.vec)!=mean.vec)>0){
    stop("Number of items vector cannot contain non-integer numbers!\n")}
  if((k<=0)|(floor(k)!=k)){
    stop("Number of items to be sampled must be a positive integer!\n")}
  if(k>sum(mean.vec)){
    stop("Number of items to be sampled cannot be greater than the total
    items!\n")}
  x<-matrix(0,nrep,d) ; tot.m<-sum(mean.vec) ; myk<-k
  for (i in 1:nrep){
    summ<-tot.m ; k<-myk
    for (j in 1:(d-1)){
      x[i,j]<-rhyper(1,mean.vec[j],summ-mean.vec[j],k)
      k<-k-x[i,j] ; summ<-summ-mean.vec[j]} ; x[i,d]<-k
    x}

```

Multivariate beta (Dirichlet) distribution

$$PDF: f(x|\alpha_1, \dots, \alpha_d) = \frac{\Gamma(\sum \alpha_j)}{\prod \Gamma(\alpha_j)} \prod_{j=1}^d x_j^{\alpha_j-1} \text{ for}$$

$\alpha_j > 0$, $x_j \geq 0$ and $\sum_{j=1}^d x_j = 1$, where

$\alpha = (\alpha_1, \dots, \alpha_d)$ is the shape vector. GA: Using the ratios of gamma variates with common scale parameter (β), (see Code 6).

Multinomial distribution

$$PDF: f(x|\theta_1, \dots, \theta_d) = \frac{N!}{\prod x_j!} \prod_{j=1}^d \theta_j^{x_j}$$

for $0 < \theta_j < 1$, $x_j \geq 0$ and $\sum_{j=1}^d x_j = N$,

where $\theta = (\theta_1, \dots, \theta_d)$ is the vector of cell probabilities and N is the size. GA: Sequential generation of marginals which are binomials (see Code 7).

Code 6. Multivariate beta (Dirichlet) distribution:

```

draw.dirichlet<-function(nrep,d,alpha,beta){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(length(alpha)!=d){stop("Shape vector is misspecified, dimension is wrong!\n")}
  if(min(alpha)<=0){
    stop("Shape vector cannot contain non-positive numbers!\n")}
  if(beta<=0){stop("Common scale parameter must be positive!\n")}
  mygamma<-matrix(rgamma(nrep*d,alpha,beta),nrep,d,byrow=T)
  mybeta<-matrix(0,nrep,d); for (i in 1:nrep){
    mybeta[i,]<-mygamma[i,]/sum(mygamma[i,])}
  mybeta}

```

Code 7. Multinomial distribution:

```

draw.multinomial<-function(nrep,d,theta,N){
  if((nrep<1)|floor(nrep)!=nrep){
    stop("Number of replicate samples must be integer whose value is at least 1!\n")}
  if((d<1)|floor(d)!=d){
    stop("Dimension must be integer whose value is at least 2!\n")}
  if (length(theta)!=d){
    stop("Length of the parameter vector does not match the dimension!\n")}
  if (min(theta)<0){
    stop("Parameter vector contains negative values!\n")}
  if (sum(theta)!=1){
    stop("Sum of probabilities must be 1!\n")}
  if((N<2)|floor(N)!=N){
    stop("Size must be an integer whose value is at least 2!\n")}
  mult<-matrix(0,nrep,d) ; mytheta<-theta; for (r in 1:nrep){
    theta<-mytheta ; size<-N ; mult[r,1]<-rbinom(1,size,theta[1])
    for (j in 2:(d-1)){
      size<-N-sum(mult[r,1:(j-1)])
      theta[j]<-theta[j]/sum(theta[j:d])
      mult[r,j]<-rbinom(1,size,theta[j])}
    mult[r,d]<-N-sum(mult[r,1:(d-1)])}
  mult}

```

Dirichlet-Multinomial distribution

This is a mixture distribution that is a multinomial with parameter θ that is a realization of a random variable having a Dirichlet distribution with shape vector α . As before, N is the size, β is the common scale parameter of the gamma variates that are used for generating Dirichlet variates. GA: An appropriate Dirichlet is generated which, in turn, is employed to generate the multinomial conditionally (see Code 8).

Multivariate Laplace (double exponential) distribution

PDF : $f(x | \mu, \Sigma, \gamma) = c \exp\left(-((x - \mu)^T \Sigma^{-1} (x - \mu))^{\gamma/2}\right)$

for $-\infty < x < \infty$ and

$$c = \frac{\gamma \Gamma(d/2)}{2\pi^{d/2} \Gamma(d/\gamma)} |\Sigma|^{-1/2}, \Sigma \text{ is}$$

symmetric and positive definite, where μ , Σ and γ are the mean vector, the variance-covariance matrix and the shape parameter, respectively. GA: Involves in generation of a point s on the d -dimensional sphere (see the auxiliary function below for $d=2,3,4$ and Marsaglia, 1972) and a generalized univariate gamma variate (Ernst, 1998) y from the density

$$f(y | \alpha, \gamma) = \frac{\gamma}{\Gamma(\alpha)} y^{\alpha\gamma-1} e^{-y^\gamma}$$

with $\alpha = d$. Finally, $yT^T s + \mu$ delivers variates from multivariate Laplace distribution, where $T^T T = \Sigma$ (see Code 9).

Code 8. Dirichlet-Multinomial distribution:

```
draw.dirichlet.multinomial<-function(nrep,d,alpha,beta,N){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(length(alpha)!=d){
    stop("Shape vector is misspecified, dimension is wrong!\n")}
  if(min(alpha)<=0){stop("Shape vector cannot contain non-positive numbers!\n")}
  if(beta<=0){stop("Common scale parameter must be positive!\n")}
  if((N<2)|floor(N)!=N){
    stop("Size must be an integer whose value is at least 2!\n")}
  dirichlet<-apply(draw.dirichlet(nrep,d,alpha,beta),2,mean)
  if(sum(dirichlet)!=1){dirichlet[d]<-1-sum(dirichlet[1:d-1])}
  draws<-draw.multinomial(nrep,d,dirichlet,N)
  draws}
```

Code 9. Multivariate Laplace (double exponential) distribution:

```

generate.point.in.sphere<-function(nrep,d){
  if ((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  my.mat<-matrix(0,nrep,d) ; if ((d==2)|(d>4)){
    for (i in 1:nrep){index<-0
      while (index<1){u.mat<-runif(d)*sample(c(-1,1), d, replace = TRUE)
        summ<-sum(u.mat^2) ; my.mat[i,]<-u.mat/sqrt(summ)
        index<-1*(summ<=1)}}}
    if (d==3){for (i in 1: nrep){
      index<-0 ; while (index<1){u1<-runif(1,-1,1) ; u2<-runif(1,-1,1)
        s1<-u1^2+u2^2 ; w<-(s1<=1)
        index<-1*w ; my.mat[i,1][w]<-2*u1[w]*sqrt(1-s1[w])
        my.mat[i,2][w]<-2*u2[w]*sqrt(1-s1[w]) ; my.mat[i,3][w]<-1-2*s1[w]}}}
    if (d==4){for (i in 1: nrep){index<-0 ; while (index<1){
      u1<-runif(1,-1,1) ; u2<-runif(1,-1,1) ; u3<-runif(1,-1,1)
      u4<-runif(1,-1,1) ; s1<-u1^2+u2^2 ; s2<-u3^2+u4^2
      w1<-(s1<=1) ; w2<-(s2<=1) ; index<-1*(w1&w2)
      my.mat[i,1][w1&w2]<-u1[w1&w2] ; my.mat[i,2][w1&w2]<-u2[w1&w2]
      my.mat[i,3][w1&w2]<-u3[w1&w2]*sqrt((1-s1[w1&w2])/s2[w1&w2])
      my.mat[i,4][w1&w2]<-u4[w1&w2]*sqrt((1-s1[w1&w2])/s2[w1&w2])}}}
    my.mat}

draw.multivariate.laplace<-function(nrep,d,gamma,mu,Sigma){
  if ((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(gamma<=0){stop("Shape parameter must be positive!\n")}
  if((nrep<2)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least
    2!\n")}
  if(d<2){stop("Dimension must be at least 2!\n")}
  if(length(mu)!=d){stop("Mean vector is misspecified, dimension is
  wrong!\n")}
  if((ncol(Sigma)!=d)|(nrow(Sigma)!=d)){
    stop("Variance-covariance matrix is misspecified, dimension is
    wrong!\n")}
  if(sum(Sigma!=t(Sigma))+min(eigen(Sigma)$values<=0))
  {stop("Variance-covariance matrix must be symmetric and positive
  definite!\n")}
  mul.laplace<-matrix(0,nrep,d)
  for (i in 1: nrep){s<-generate.point.in.sphere(1,d)
    mul.laplace[i,]<-(rgamma(1,d,1)^(1/gamma))*t(chol(Sigma))%*%t(s)+mu}
  mul.laplace}

```

Wishart distribution

$$PDF : f(x | \nu, S) = (2^{\nu d/2} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma((\nu+1-i)/2))^{-1} |S|^{-\nu/2} |x|^{(\nu-d-1)/2} \exp(-\frac{1}{2} \text{tr}(S^{-1}x)),$$

x is positive definite, $\nu \geq d$ and S is symmetric and positive definite, where μ and S are the degrees of freedom and the scale matrix, respectively. GA: Using a simple function of the variates that follow d-variate normal distribution (see Code 10).

Inverted Wishart distribution

$$PDF : f(x | \nu, S) = (2^{\nu d/2} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma((\nu+1-i)/2))^{-1} |S|^{\nu/2} |x|^{-(\nu+d+1)/2} \exp(-\frac{1}{2} \text{tr}(Sx^{-1})),$$

x is positive definite, $\nu \geq d$ and S is symmetric and positive definite, where μ and S^{-1} are the degrees of freedom and the inverse scale matrix, respectively. GA: Using Wishart variates (see Code 11).

Code 10. Wishart distribution:

```
draw.wishart<-function(nrep,d,nu,sigma){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least 1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(nu<d){
    stop("Distribution is not proper !\n")
    stop("Degrees of freedom should be greater than or equal to the dimension!\n")}
  if(floor(nu)!=nu){
    stop("Degrees of freedom should be an integer!\n")}
  if((ncol(sigma)!=d)|(nrow(sigma)!=d)){
    stop("Scale matrix is misspecified, dimension is wrong!\n")}
  if(min(eigen(sigma)$values)<0)
  {stop("Scale matrix must be symmetric and positive definite!\n")}
  wishart<-matrix(0,nrep,d^2)
  for (i in 1:nrep){
    alpha.i<-draw.d.variate.normal(nu,d,rep(0,d),sigma)
    wishart[i,]<-t(alpha.i)%*%alpha.i }
  # This function generates Wishart deviates in the form of rows.
  # To obtain the Wishart matrix, convert each row to a matrix where
  # rows are filled first.
  wishart}
```

Code 11 Inverted Wishart distribution:

```

draw.inv.wishart<-function(nrep,d,nu,inv.sigma){
  if((nrep<1)|(floor(nrep)!=nrep)){
    stop("Number of replicates must be an integer whose value is at least
    1!\n")}
  if((d<2)|(floor(d)!=d)){
    stop("Dimension must be an integer whose value is at least 2!\n")}
  if(nu<d){
    stop("Distribution is degenerate!\n")
    stop("Degrees of freedom should be greater than or equal to the
    dimension!\n")}
  if(nu==d+1){
    warning("Expectation does not exist!\n")}
  if(floor(nu)!=nu){
    stop("Degrees of freedom should be an integer!\n")}
  if((ncol(inv.sigma)!=d)|(nrow(inv.sigma)!=d)){
    stop("Inverse scale matrix is misspecified, dimension is wrong!\n")}
  if(min(eigen(inv.sigma)$values)<0)
  {stop("Inverse scale matrix must be symmetric and positive definite!\n")}
  inv.wishart<-draw.wishart(nrep,d,nu,solve(inv.sigma))
  # This function generates Wishart deviates in the form of rows.
  # To obtain the Inverted-Wishart matrix, convert each row to a matrix
  # where rows are filled first.
  inv.wishart}

```

Results

For each distribution, the parameters can take infinitely many values and first two moments virtually fluctuate on the entire real line. Although the quality of random variates was tested by a broad range of simulations to see any potential aberrances and abnormalities in some subset of the parameter domains and to avoid any selection biases, it is constructive to report the empirical and distributional moments for arbitrarily chosen parameter values.

Table 1 tabulates the theoretical and empirical means for each distribution for arbitrary values. Throughout the table, the number of replications (*nrep*) and the dimension (*d*) are chosen to be 10,000 and 3, respectively. A similar comparison is made for the variance-

covariance matrices, as shown in Table 2. In both tables, the deviations from the expected moments are found to be negligible, suggesting that random number generation routines presented are accurate.

Conclusion

The reader is invited to be cautious about the following issues: 1) It is not postulated that algorithms presented are the most efficient. Furthermore, implementation of a given algorithm may not be optimal. Given sufficient time and resources, one can write more efficient routines. 2) Quality of every random number generation process depends on the uniform number generator. McCullough (1999) raised some questions about the quality of the Splus.

Table 1: Comparison of theoretical and empirical means for arbitrarily chosen parameter values.

<i>Distribution</i>	<i>Theoretical mean</i>	<i>Empirical mean</i>
Normal	(0, 0, 0)	(0.00139, 0.00384, 0.00377)
t	(0, 0, 0)	(0.00109, 0.00300, 0.00295)
Uniform	(0.5, 0.5, 0.5)	(0.50027, 0.49985, 0.50029)
Bernoulli	(0.9, 0.8, 0.7)	(0.89925, 0.80004, 0.69890)
Hypergeometric	(1.66667, 3.33333, 5)	(1.66911, 3.33117, 4.99972)
Dirichlet	(0.16667, 0.33333, 0.5)	(0.16670, 0.33337, 0.49992)
Multinomial	(8, 6, 6)	(7.98714, 6.01304, 5.99982)
Dirichlet-Multinomial	(30, 40, 30)	(29.98379, 40.00643, 30.00978)
Laplace	(0, 0, 0)	(0.00292, -0.00213, -0.00288)
Wishart-Inverted Wishart	5.0 1.0 1.5	5.00752 1.00138 1.50141
	1.0 5.0 1.0	1.00138 5.00305 1.00462
	1.5 1.0 5.0	1.50141 1.00462 4.99576

Table 2: Comparison of theoretical and empirical variance-covariance matrices for arbitrarily chosen parameter values.

<i>Distribution</i>	<i>Theoretical variance-covariance</i>	<i>Empirical variance-covariance</i>
Normal	1.0 0.2 0.3	0.99149 0.19815 0.29178
	0.2 1.0 0.2	0.19815 0.99946 0.20227
	0.3 0.2 1.0	0.29178 0.20227 1.00232
t	1.0 0.2 0.3	1.01178 0.20221 0.29773
	0.2 1.0 0.2	0.20221 1.01989 0.20640
	0.3 0.2 1.0	0.29773 0.20640 1.02281
Uniform	1.0 0.2 0.3	0.99958 0.19164 0.28748
	0.2 1.0 0.2	0.19164 0.99988 0.19065
	0.3 0.2 1.0	0.28748 0.19065 1.00039
Bernoulli	0.09000 0.01200 0.06874	0.09060 0.01295 0.06920
	0.01200 0.16000 0.09165	0.01295 0.15998 0.09228
	0.06874 0.09165 0.21000	0.06920 0.09228 0.21044
Hypergeometric	1.17702 -0.47978 -0.70043	1.17180 -0.47571 -0.69609
	-0.47978 1.88332 -1.42106	-0.47571 1.89074 -1.41503
	-0.70043 -1.42106 2.11864	-0.69609 -1.41503 2.11112
Dirichlet	0.01068 -0.00427 -0.00641	0.01067 -0.00424 -0.00643
	-0.00427 0.01709 -0.01282	-0.00424 0.01699 -0.01275
	-0.00641 -0.01282 0.01923	-0.00643 -0.01275 0.01917
Multinomial	4.8 -2.4 -2.4	4.79330 -2.39451 -2.39880
	-2.4 4.2 -1.8	-2.39451 4.19743 -1.80293
	-2.4 -1.8 4.2	-2.39880 -1.80293 4.20172
Dirichlet-Multinomial	21 -12 -9	20.96299 -11.98345 -8.97954
	-12 24 -12	-11.98345 24.00701 -12.02356
	-9 -12 21	-8.97954 -12.02356 21.00310
Laplace	1.0 0.2 0.3	0.99830 0.19518 0.29604
	0.2 1.0 0.2	0.19518 0.99736 0.19502
	0.3 0.2 1.0	0.29604 0.19502 0.99869

At the time of this writing, a source that tested the R generator is unknown to the author. In addition, the differences between empirical and distributional moments have merely been examined for each distribution. More comprehensive and computer science-minded tests are needed possibly using DIEHARD suite (Marsaglia, 1995) or other well-regarded test suites.

In a nutshell, the R routines provided may be useful for applied scientists for simulation and computation purposes. Acknowledging the fact that dependence of the random number generation libraries on specific linkers, debuggers, compilers, operating systems and computing platforms may create problems in practice, these routines could be a handy addition to a practitioner's set of tools given the growing interest in R.

References

- Ernst, M. D. (1998). A multivariate generalized Laplace distribution. *Computational Statistics*, 13, 227-232.
- Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Communications in Statistics. Simulation and Computation*, 28, 85-791.
- Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *Annals of Mathematical Statistics*, 43, 645-646.
- Marsaglia, G. (1995). *The Marsaglia Random Number CDROM*, including the DIEHARD Battery of Tests of Randomness. Department of Statistics, Florida State University, Tallahassee, FL.
- McCullough, B. D. (1999). Assessing the reliability of statistical software: Part 2. *The American Statistician*, 53, 149-159.
- Park, C. G., Park, T., & Shin D. W. (1996). A simple method for generating correlated binary variates. *The American Statistician*, 50, 306-310.

An Algorithm And Code For Computing Exact Critical Values For The Kruskal-Wallis Nonparametric One-Way ANOVA

Sikha Bagui Subhash Bagui
University of West Florida, Pensacola

In this article, an algorithm and code to compute exact critical values (or percentiles) for Kruskal-Wallis test on k independent treatment populations with equal or unequal sample sizes using Visual Basic (VB.NET) is provided. This program has the ability to calculate critical values for any k , sample sizes (n_i), and significance level (α). An exact critical value table for $k = 4$ is also developed. The table will be useful to practitioners since it is not available in standard nonparametric statistics texts. The program can also be used to compute any other critical values.

Key words: Kruskal-Wallis test, ANOVA, visual basic

Introduction

Headrick (2003) wrote an article for generating exact critical values for the Kruskal-Wallis (K-W) one-way ANOVA using Fortran 77. In this article we present Visual Basic (VB.NET) Code for generating exact critical values for K-W tests using the Visual Basic Programming Language. VB.NET is more user friendly and more accessible than Fortran 77. While Fortran 77 may not be available to all, the proposed VB.NET program can be a simpler alternative to Fortran 77.

When one or more treatment populations violate normality assumption or the homogeneity of treatment population variances, it is customary to use Kruskal-Wallis (1952) rank-based nonparametric test as an alternative to the conventional F test for one-way analysis

of variance (ANOVA) for k independent treatment populations.

In order to find the critical values of K-W tests, one needs to find the null distribution of the K-W statistics. In one-way ANOVA, the null hypothesis is that the effect of all treatment populations are the same. Thus, it is reasonable to use such a type of null distribution of the K-W statistics which are derived under the assumption that all observations for treatment populations $n_1, n_2, \dots, n_k$ are from the same population to calculate the critical values of K-W tests, where n_i is the sample size of the i th treatment population. The K-W statistic depends on the rank-sums of each treatment population that are obtained from the combined ranks of $N = n_1 + n_2 + \dots + n_k$ observations.

It is known that the large sample null distribution of K-W statistic is approximately a chi-square (χ^2) distribution with $(k-1)$ degrees of freedom (d.f.). Conover (1999) suggested that whenever $k \geq 4$ and $n_i > 5$, for each treatment population, a chi-square critical with $(k-1)$ d.f. ($\chi^2_{\alpha, k-1}$) be used to test the null hypothesis. But for small samples, say $n_i \leq 5$, the null-distribution of K-W statistic is not known and a chi-square approximation will not be a good approximation. The common nonparametric text books such as

Sikha Bagui is an Assistant Professor in the Department of Computer Science. Her areas of research are database and database design, data mining, pattern recognition, and statistical computing. Email: bagui@uwf.edu. Subhash Bagui is a Professor in the Department of Mathematics and Statistics. His areas of research are statistical classification and pattern recognition, bio-statistics, construction of designs, tolerance regions, statistical computing and reliability. Email: sbagui@uwf.edu.

Conover (1999), Gibbons (1992); Siegel and Castellan (1989) provided exact critical values for the K-W test for $k=3$ and $n_i \leq 5$ observations per treatment population. Major statistical software such as MINITAB, SPSS provide only the asymptotic P -value of the K-W statistics. In view of all these, in this article, we provide a VB.NET program as an alternative to Fortran 77 to compute exact critical points for the K-W tests, and also report a table for exact critical values for the K-W test for $k=4$ treatment populations and $n_i \leq 5$ observations per treatment population. Even though the number of ways N ranks can be divided into groups of $n_1, n_2, \dots, n_k$ grows, our VB.NET program works well with reasonable values of k and n_i .

Methodology

To calculate the K-W statistics, first we need to generate N uniform pseudo-random numbers from the interval $(0,1)$. We assume that the probability of a tie is zero. Then the random variates are ranked to form permutation of numbers from 1 to N . The program then sequentially divides the permutation of ranks into k classes according to the users specific sample sizes of $n_1, n_2, \dots, n_k$. The program then calculates rank sums of each treatment population, R_j , and next computes the value of K-W statistic

$$H = \frac{12}{N(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(N+1).$$

This process is replicated a sufficient number of times until the null distribution of H is modeled adequately. Then the program selects a critical value that is associated with a percentile values of 0.90, 0.95, 0.975 or 0.99 (or equivalently a alpha level of 0.10, 0.05, 0.025 or 0.01). In some cases, returned values may coincide with two different alpha values, since returned values are true for a range of P -values.

For example, given $\alpha = 0.05$, $k = 3$, and $n_i = 5$, our VB.NET program will return a critical value of 5.659997 with a replication of 100,000 runs which is same as the value reported by Headrick (2003) for $\alpha = 0.05$, $k = 3$, and $n_i = 5$. Also, with adequate number of runs, our VB.NET program yields the same values reported by Conover (1999) in Table A8. Table A8 is for $k = 3$, $n_i \leq 5$ and $\alpha = 0.1, 0.05$, and 0.01. In Table 1 critical values are provided for K-W statistic for $k = 4$, $n_i \leq 5$ and $\alpha = 0.1, 0.05, 0.025$, and 0.01. At the bottom of the table, the asymptotic chi-square critical values of H from a chi-square critical value table are also provided. The notation K_α is a $(\alpha)100\%$ percentile for the K-W statistics which is equivalent to $(1-\alpha)$ level critical value of the K-W statistic. This table will be very useful to the practitioners because it is not available in standard nonparametric text books. The critical values in Table 1 are generated using 1 million replications in each case.

Conclusion

In case of large N , the program needs large number of replications in order to adequately model the null distribution of K-W statistic H . So the replication number should be in increasing order such as 10,000, 50,000, 100,000, 500,000, and 1,000,000, etc. and stop the process once two consecutive values are almost the same. If there are k independent treatment populations, then at least $N!/(n_1!)(n_2!)\dots(n_k!)$ replications are necessary for a near fit of H . For a good fit of H , one needs much more replications than $N!/(n_1!)(n_2!)\dots(n_k!)$. The VB.NET code is given in the Appendix. The VB.NET program is very user friendly. The VB.NET program allows the user to provide the values of replication numbers, total number of observations, percentile fractions, and separate class sizes based on which the program will return a critical value.

Table 1. Critical values for K-W statistics for small sample sizes.

Sample sizes	$K_{0.90}$	$K_{0.95}$	$K_{0.975}$	$K_{0.99}$
2, 2, 2, 1	5.0357	5.3571	5.6786	5.6786
2, 2, 2, 2	5.5000	6.0000	6.1667	6.1667
3, 2, 2, 1	5.3889	5.8056	6.0556	6.5000
3, 2, 2, 2	5.5778	6.2444	6.6444	7.0000
3, 3, 2, 1	5.6222	6.1556	6.5111	7.0444
3, 3, 2, 2	5.7273	6.4727	7.0000	7.5636
3, 3, 3, 1	5.5818	6.5273	6.8909	7.3273
3, 3, 3, 2	5.8182	6.6818	7.4697	7.9545
3, 3, 3, 3	5.9744	6.8462	7.6154	8.4359
4, 2, 1, 1	5.2083	5.4583	6.0833	6.0833
4, 2, 2, 1	5.5000	6.0000	6.5000	6.8000
4, 2, 2, 2	5.6727	6.4364	6.9818	7.3091
4, 3, 2, 1	5.5636	6.3000	6.9091	7.3636
4, 3, 2, 2	5.7121	6.6136	7.3409	7.8712
4, 3, 3, 1	5.6667	6.5379	7.2727	7.7500
4, 3, 3, 2	5.8590	6.7820	7.5256	8.3205
4, 3, 3, 3	5.9780	6.9670	7.7582	8.6538
4, 4, 2, 1	5.5682	6.3864	7.1364	7.8864
4, 4, 2, 2	5.7692	6.6923	7.5000	8.3077
4, 4, 3, 1	5.6603	6.6154	7.4808	8.2179
4, 4, 3, 2	5.8901	6.8626	7.7527	8.6044
4, 4, 3, 3	6.0048	7.0333	7.8905	8.8619
4, 4, 4, 1	5.6374	6.7088	7.6319	8.5714
4, 4, 4, 2	5.9000	6.9429	7.8857	8.8571
4, 4, 4, 3	6.0292	7.1292	8.0542	9.0625
4, 4, 4, 4	6.0441	7.1691	8.1618	9.2426
5, 2, 2, 1	5.5309	6.0327	6.5782	7.2000
5, 2, 2, 2	5.6091	6.4818	7.1545	7.6636
5, 3, 2, 1	5.5030	6.3303	7.0939	7.7455
5, 3, 2, 2	5.7538	6.6564	7.4641	8.1949
5, 3, 3, 1	5.6564	6.6000	7.4205	8.1179
5, 3, 3, 2	5.8571	6.8220	7.6505	8.5824
5, 3, 3, 3	6.0210	7.0114	7.8267	8.8400
5, 4, 2, 1	5.5615	6.4077	7.2115	8.1346
5, 4, 2, 2	5.7725	6.7220	7.6154	8.4725
5, 4, 3, 1	5.6396	6.6813	7.5253	8.3989
5, 4, 3, 2	5.8933	6.9143	7.7933	8.8143
5, 4, 3, 3	6.0292	7.0892	7.9892	9.0292
5, 4, 4, 1	5.6686	6.7714	7.6857	8.7286
5, 4, 4, 2	5.9350	6.9775	7.9600	9.0000
5, 4, 4, 3	6.0346	7.1640	8.1412	9.2294
5, 4, 4, 4	6.0569	7.2431	8.2569	9.3824
5, 5, 2, 1	5.5648	6.5341	7.2725	8.3077
5, 5, 2, 2	5.7771	6.7714	7.6400	8.6342
5, 5, 3, 1	5.6476	6.7371	7.6286	8.5886
5, 5, 3, 2	5.9067	6.9417	7.8667	8.9417
5, 5, 3, 3	6.0000	7.1176	8.0588	9.1706
5, 5, 4, 1	5.6625	6.7800	7.7500	8.8425
5, 5, 4, 2	5.9338	7.0257	8.0096	9.1434
5, 5, 4, 3	6.0523	7.2176	8.2092	9.3562
5, 5, 4, 4	6.0614	7.2684	8.3281	9.5053
5, 5, 5, 1	5.6224	6.8294	7.8176	9.0529
5, 5, 5, 2	5.9412	7.0745	8.1059	9.2863
5, 5, 5, 3	6.0433	7.2456	8.2889	9.4772
5, 5, 5, 4	6.0774	7.3200	8.4253	9.6521
5, 5, 5, 5	6.0971	7.3314	8.4629	9.7886
$\chi^2_{1-\alpha; k-1} \rightarrow$	$\chi^2_{0.10; 3} = 6.251$	$\chi^2_{0.05; 3} = 7.815$	$\chi^2_{0.025; 3} = 9.348$	$\chi^2_{0.01; 3} = 11.345$

References

- Conover, W.J. (1999). *Practical nonparametric statistics*. (3rd ed.). New York: Wiley.
- Gibbons, J.D. (1992). *Nonparametric statistical inference*. (3rd ed.). New York: Marcel Dekker.
- Headrick, T.C. (2003). An algorithm for generating exact critical values for the Kruskal-Wallis One-way ANOVA. *Journal of Modern Applied Statistical Methods*, 2, 268-271.
- Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion analysis of variance. *Journal of American Statistical Association*, 47, 583-621.
- Minitab (2000). *Minitab for Windows*, release 13.3, Minitab Inc., State College, PA.
- Siegel, S., & Castellan, N.J. (1989). *Nonparametric statistics for the behavioral sciences*, (2nd ed.). New York: McGraw-Hill.
- SPSS(2002). *SPSS for Windows*, version 11.0, SPSS, Inc., Chicago, IL.

Appendix

Public Class Form1

Inherits System.Windows.Forms.Form
'Kruskal-Wallis One-Way Anova

Dim a, m, n, i, j, k, v, x, y, z, sumy, row, count As Integer
Dim output, prompt1, prompt2, prompt_value, group_value As String
Dim output2 As String
Dim sums, sumz, H, percentile As Single

Dim file1 As System.IO.StreamWriter

Private Sub Button1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Button1.Click
 m = Val(TextBox2.Text)
 n = Val(TextBox1.Text)
 percentile = Val(TextBox3.Text)

 file1 = System.IO.File.CreateText("C:\Documents and Settings\sbagai\Desktop\file1.txt")

 prompt1 = "Please enter number of groups"
 prompt_value = InputBox(prompt1)

x = Val(prompt_value)

 Dim array1(n) As Single
 Dim array2(n) As Single
 Dim array3(n) As Integer
 Dim array4(m) As Single
 Dim array5(n) As Integer

 For i = 1 To (x - 1)
 prompt2 = "Please enter number in group" & i
 group_value = InputBox(prompt2)
 array5(i) = Val(group_value)
 Next

 For k = 1 To m
 For i = 1 To n
 array1(i) = Rnd()
 Next

```

For i = 1 To array1.GetUpperBound(0)
    array2(i) = array1(i)
Next

Array.Sort(array2)

'ranking row
For row = 1 To array1.GetUpperBound(0)
    For i = 1 To array2.GetUpperBound(0)
        If array1(row) = array2(i) Then
            array3(row) = i
        End If
    Next
Next

For i = 1 To array3.GetUpperBound(0)
    output &= array3(i) & " "
Next

y = 0
z = 0
sumy = 0
sumz = 0
For i = 1 To (x - 1)
    y = array5(i)
    For j = (1 + z) To (y + z)
        sumy += array3(j)
    Next
    a = y + z
    z = a
    sums = (sumy * sumy) / y
    sumz += sums
    sumy = 0
    sums = 0
Next

count = 0
sumy = 0
For i = (z + 1) To array3.GetUpperBound(0)
    sumy += array3(i)
    count = count + 1
Next
sums = (sumy * sumy) / count
sumz += sums

H = 12 / (n * (n + 1)) * sumz - 3 * (n + 1)
array4(k) = H

H = 0
Next

Array.Sort(array4)      'array4 - sorted F values

output = " "
For i = 1 To array4.GetUpperBound(0)
    output &= array4(i) & " "
Next

count = 0
For i = 1 To array4.GetUpperBound(0)
    count += 1
Next

```

```
        v = percentile * count
        output = " "
        output = array4(v)
        MessageBox.Show(output, percentile & " percentile value")
        output2 = percentile & " = " & output
        file1.WriteLine(output2)
        file1.Close()
    End Sub

    Private Sub Form1_Load(ByVal sender As System.Object, ByVal e As
System.EventArgs) Handles MyBase.Load
        Randomize()          'Calling the random number generator
    End Sub
End Class
```

A Modification Of The EM Algorithm To Estimate An Andersen-Gill Gamma Frailty Model For Multivariate Failure Time Data

Maria Antònia Barceló Marc Saez
Research Group on Statistics, Applied Economics and Health (GRECS)
University of Girona, Spain

A modification of the Andersen-Gill gamma shared frailty model is presented. The variance of the frailty is directly modeled by means of a generalized linear model, the EM algorithm is modified in order to simultaneously estimate a semiparametric model for the failure times and a model for the variance of the frailty. A simulation study is conducted to evaluate the performance of the proposed algorithm (EMB algorithm) and compared with other methods, a marginal model, and a conditional model. Multivariate data from a nosocomial infection study is used to illustrate the methods. The EMB fit turned out to be better than the fit obtained from a marginal model or from a conditional model. The EMB provided the best fit (being the least over-dispersed and having the highest AIC and the highest pseudo-R square) and estimated the parameters most efficiently. The proposed method is able to capture and to take into account unobservable random effects in semiparametric models.

Key words: Frailty, marginal and conditional models, generalized linear models, EM, nosocomial infections

Introduction

Patients admitted to intensive care units (ICU) run a high risk of contracting a nosocomial infection due not only to the susceptibility associated with the severity of their conditions, but also to medical procedures that the ICUs use.

The authors thank Josep Maria Sirvent and Alfons Bonet from the ICU of the Hospital Universitari, and Dr. Josep Trueta, Girona, Spain, for comments and suggestions. Acknowledgments are also due to Maria Teresa Butiñá, Blanca Martínez (pharmacological service of the hospital) and Jordi Batlle and Montse Motjé (clinical analysis service of the hospital) for providing the data. This project was funded in part by a grant from the Agència d'Avaluació de Tecnologia i Recerca Mèdica, Servei Català de la Salut, Generalitat de Catalunya, AATRM 07/50/98. Correspondence regarding this article should be address to: marc.saez@udg.es.

Barceló and Saez (2001) analysed the factors that determine the occurrence of nosocomial infections in the ICU of a tertiary-level hospital in Girona, Spain, during the second quarter of 1999 (March-June, 1999). The authors tried to determine which factors, those associated to patients (such as their immunodeficiency) or those related to ICU (such as invasive medical procedures or the inappropriate use of antimicrobial agents), were the most relevant in the explanation of the occurrence of nosocomial infections in the ICU. They were interested in analysing the factors that determine both the occurrence of an infection and also the time leading up to the onset of the infection.

In that context, the standard approach to obtain adjusted risk (hazard) factors for the infection would be the Cox model (Cox, 1972). The problem was that a patient could have several episodes of infection during her/his ICU stay. As a consequence the data set had multiple events per subject, i.e. recurrent events. As is well known the main problem of the Cox model with multivariate data is that the observations are not independent, implying, among other things, the violation of the proportionality hypothesis. It is also known that standard Cox

models ignore such dependence, leading to estimates that are inefficient and biased.

The Andersen-Gill approximation (AG) to the Cox model (Andersen & Gill, 1982; Andersen et al., 1993) overcomes, in part, this problem. The AG model is a counting process approach in which each patient is represented as a set of rows with time intervals of (entry time, first infection], (first infection, second infection], ... , (nth infection, last follow-up]. Each row is treated as a different patient and, therefore, risk proportionality is not violated. However, the underlying hypothesis in AG, called the hypothesis of independent increments, is very restrictive and may be untenable. Under this hypothesis the multiple observations of an individual are independent, although conditioned on the explanatory variables. Therefore, a suitable alternative is needed.

In choosing a model for the time to recurrent infection one needs to consider the biological process of disease. It was very likely that after experiencing the first infection, the risk (hazard) of subsequent infections would increase. This could happen if each infection permanently compromised the ability of the immune system to combat subsequent infection. If this were the case one would use a model containing separate strata for each episode of infection (Therneau & Hamilton, 1997). In this sense, the first choice was the Prentice, Williams and Peterson (PWP) model (Prentice, Williams & Peterson, 1981). The PWP is a marginal model with respect to the estimation of the parameters, which treats the dependence between event times as a nuisance to control for, without explicitly specifying models for this dependence.

Conditional methods, in contrast, explicitly model the dependence between recurrences. Amongst them, frailty models (Clayton & Cuzick, 1985) have become the most popular for analysing multivariate survival data. In those models the dependence between the events is accounted for by the introduction of frailties or unobservable random effects into the marginal hazards (Klein, 1992). The frailties are shared among recurrences from the same individual. Maximum likelihood estimation in the AG shared frailty model (with gamma-distributed frailties) is usually performed using

the EM algorithm as suggested by Gill (1985) and further discussed by Nielsen et al. (1992) and Klein (1992).

In particular, the estimation of the model using the EM algorithm is carried out by fixing through the variance of the frailty until its convergence. Then, the algorithm iterates between the E and the M steps. In the E step the frailties are replaced in the complete data log likelihood by their conditional expectation. The M step consists of computing the Nelson-Aalen estimator as if the frailties had been observed. This procedure is repeated for other arbitrary values of the variance obtaining in each case the log incomplete data profile likelihood as a function of the variance. Finally, the estimate of the variance is computed either numerically or graphically. The EM algorithm, however, could converge arbitrarily and slowly and, furthermore, the final estimate of the variance obviously depends on the initial choices for that parameter.

Here, the directly modelling of the variance is proposed. In this sense, Clayton (1988) and Lindsey (1999) were followed. The former proposes the possibility of extending the EM algorithm by simultaneously estimating the variance of the frailty. Lindsey (1999) pointed out that "dispersion varying with the explanatory variables is surprisingly common" (Lindsey, 1999, p. 2230) and suggests estimating a separate regression equation for the dispersion parameter. Besides the extension of the model, a modification of the EM algorithm is also proposed, which is called EMB, to simultaneously estimate such a two-equation model.

An alternative to the frailty models can be found in the penalised likelihood models (Behrman et al., 1991; Therneau & Grambsch, 1998). The idea is to use a penalty function for a constrained solution, equal to the log gamma density. The penalty function captures the local variability underlying the joint density of data. The problem is that such variability is in fact approaching two different things, frailty (false contagion) and serial correlation or dependence (true contagion). In addition, the choice of the shrinkage parameter used in the penalisation is a controversial question in survival analysis.

Methodology

Suppose that there is a random sample of I individuals from an underlying group population and that each individual can have J observations. For this framework the most straightforward mathematical notation derives from the theory of counting processes (Fleming & Harrington, 1991; Andersen et al., 1993).

Let i ($I = 1, \dots, I$) denote individual and (i, j) denote the j th observation in the i th individual. For each observation (i, j) , where $i = 1, \dots, I$ and $j = 1, \dots, J$, let $N_{ij}(t)$ be an observed multivariate counting process. $N_{ij}(t)$ is the cumulative number of events observed for the i -th subject. A process $Y_{ij}(t)$ is further observed, indicating whether individual i is observed to be at risk for experiencing an j th event at time t . Finally covariates $X_{ij}(t)$ (possibly time-dependent) are observed (Andersen, 1992). The multivariate counting process $N_{ij}(t)$ has an intensity process given by,

$$\lambda_{ij}(t) = Y_{ij}(t) \alpha_{0j}(t) e^{\beta' X_{ij}(t)} = \lambda_{0j}(t) e^{\beta' X_{ij}(t)} \quad [1]$$

where $\lambda_{0j}(t)$ denotes an unknown baseline intensity; and β is a vector of unknown parameters.

The Prentice, Williams and Peterson (PWP) model

Although the PWP is a marginal model, it is conditional in relation to the construction of the risk set. In this sense, the model allows the baseline risk to vary between recurrences, i.e. λ_{0j} , $j = 1, \dots, J$,

$$\lambda_{ij}(t) = \lambda_{0j}(t) e^{\beta' X_{ij}(t)} \quad (2)$$

Thus, it is actually an AG model with time dependent strata. That is to say, the risk set for the recurrence j , for instance, only contains those individuals who experienced $j-1$ recurrences. Such a strategy makes it possible to control

dependence between the recurrences, stratifying according to them.

Although estimates obtained from PWP models are consistent (Prentice, Williams & Peterson, 1981), the dependence between observations remains in fact uncontrolled. As a consequence, standard errors are biased (usually overstated). For this reason we propose to robustly estimate the standard errors of the parameters. In particular we chose a grouping jackknife estimate (Therneau & Hamilton, 1997). The idea is to compute the i change in the estimates of the parameters with all the observations for the i -th subject removed from data set. This will result in a matrix D , each row i of which will be an estimate of the leverage, i.e., average change, of the i -th subject. The matrix $D'D$ will approximate the grouped jackknife estimate of variance and will be an asymptotically unbiased estimate of the variance of the robust parameter estimates (Therneau & Hamilton, 1997; Lin & Wei, 1989).

The AG Gamma frailty model.

Following Nielsen et al. (1992), we formulate now an intensity process λ satisfying,

$$\lambda_{ij}(t | \varpi_i) = \varpi_i Y_{ij}(t) \alpha_{0j}(t) e^{\beta' X_{ij}(t)} \quad (3)$$

where ϖ_i denotes subject specific frailties independently drawn from a gamma(ν, η) distribution. Note that here we also allow the baseline hazards to vary between recurrences, i.e. α_{0j} .

The gamma density of the frailties is

$$f(\varpi; \nu, \eta) = \frac{\varpi^{\nu-1} e^{-\eta \varpi} \eta^\nu}{\Gamma(\nu)}$$

where η is the scale parameter and ν the shape parameter.

Due to identification problems it is usually assumed that $\nu = \eta$, i.e. that the distribution of the frailties has unitary mean and variance equal to $1/\eta$, say θ (Clayton, 1978;

Vaupel et al., 1979; Nielsen et al., 1992; Klein, 1992).

Let Z be a $n \times q$ design matrix that describes how the frailties applied to individuals subjects, $\varpi_i = e^{Z_i \omega}$. It is also assumed that the frailty consists of independent clusters of observations, i.e. $Z_{ij} = 1$ iff recurrence j belongs to individual i . Let us define

$$D_i = \sum_{j=1}^J Z_{ij} \delta_j$$

$$E_i^* = \sum_{j=1}^J Z_{ij} \Lambda_0(t_j) e^{\beta' X_j}$$

where δ_i is an indicator equal to one in a failure time case and zero otherwise; Λ_0 is the cumulative baseline hazard.

D_i is the number of events in the i -th individual and $E_i = E_i^* e^{\omega_i}$ is the expected number of events in the individual based on the covariates and the model.

The full log likelihood, when ω is observed, is then,

$$L_f = \sum_{i=1}^I \left[\left(\frac{1}{\theta} \omega_i - \frac{1}{\theta} e^{\omega_i} \right) + \frac{1}{\theta} \log \left(\frac{1}{\theta} \right) - \log \Gamma \left(\frac{1}{\theta} \right) \right]$$

$$+ \sum_{i=1}^I [D_i \omega_i - E_i^* e^{\omega_i}] + \sum_{j=1}^J \delta_j [\log(\lambda_{0j}(t_j)) + X_j \beta] \quad (4)$$

It can be shown (Therneau & Grambsch, 1998) that as a function of any single ω_i , [4] is proportional to,

$$\left(\frac{1}{\theta} + D_i \right) \omega_i - \left(\frac{1}{\theta} + E_i^* \right) e^{\omega_i}$$

therefore, conditional on the data, the ϖ_i are distributed as gamma variates with shape

$\frac{1}{\theta} + D_i$ and scale $\frac{1}{\theta} + E_i^*$. In this sense, we can write,

$$E \left(e^{\omega_i} \mid \tilde{\beta}, \frac{1}{\tilde{\theta}}, y \right) = \frac{D_i + \frac{1}{\tilde{\theta}}}{E_i^* + \frac{1}{\tilde{\theta}}} \equiv e^{\hat{\omega}_i} \quad (5)$$

where y denotes the observed data and the tilde denotes either provisional or definitive estimates.

The maximisation of the log likelihood (4) can be done using the EM algorithm. Therneau and Grambsch (1998) suggested how to use only the quantities returned by an ordinary Cox model program. Starting with the case of a fixed variance, the quantities D_i and $M_i = D_i - E_i$ can be obtained by summing over the input data and the returned martingale residuals, respectively. E_i^* is obtained from E_i and the current estimates of $\hat{\omega}_i$ (E-step). The next estimates of $\hat{\omega}_i$ are obtained from equation (5) and, finally, $Z_{ij} \hat{\omega}_i$ is used as a prior in the next invocation of the Cox model (M-step).

One problem with the EM algorithm is that variance estimates for the estimated parameters are not immediately provided (Louis, 1982). It was suggested by Gill (1989) and further discussed by Nielsen et al. (1992) and, above all, Andersen et al. (1997), that a non parametric information calculation was likely to provide consistent variance estimators. A simpler possibility lies in using the robust estimate of the covariance matrix of the estimated parameters described above (Therneau & Hamilton, 1997).

The Penalised Cox model

As mentioned above, Behrman et al. (1991) proposed to alternatively use the penalised log likelihood,

$$\log L_\alpha(f) = \sum_i \log(f(x_i)) - \alpha R(f)$$

where $\alpha R(f)$ is a penalty term that takes account of the roughness or local variability in the joint

density of the data. The smoothing parameter α , which controls the balance between smoothness and goodness of fit, must be typically chosen by cross-validation.

Therneau and Grambsch (1998) suggested using the log gamma density as the penalty function for the constrained solution,

$$-\alpha R(f) = \frac{1}{\theta}(\omega - e^\omega) + \frac{1}{\theta} \log\left(\frac{1}{\theta}\right) - \log \Gamma\left(\frac{1}{\theta}\right)$$

Therefore, in our case,

$$PLL = Cox PL + \sum_{i=1}^I \frac{1}{\theta} (\omega_i - e^{\omega_i}) + \frac{1}{\theta} \log\left(\frac{1}{\theta}\right) - \log \Gamma\left(\frac{1}{\theta}\right) \quad (6)$$

where PLL denotes the penalised log likelihood and by Cox PL we mean the numerical value returned as the partial likelihood by a standard Cox model program for the given values of β and ω , ω having been entered as an offset term.

Therneau and Grambsch (1998) pointed out that for any fixed value of the variance of the frailty the EM algorithm and the constrained minimisation of the penalised likelihood have the same solution.

A modification of the EM algorithm for the estimation of the AG gamma frailty model: The EMB algorithm.

In both the AG gamma frailty and the penalised Cox models, frailty is assumed to be constant between individuals and also within each individual, i.e. between recurrences. As an alternative, and following Wassell and Moeschberger (1993), we propose to directly model the variance of the frailty,

$$\theta_{ij} = e^{c + \gamma' X_{ij}} \quad (7)$$

where γ is a vector of parameters, X_{ij} denotes the covariates for individual i .

To model the variance we propose to use a generalised linear model, GLM (Nelder &

Pregibon, 1987; Nelder & Lee, 1991, 1996; Nelder, 1998). In particular,

$$\text{Link function:} \quad \log(\theta_{ij}) = c + \gamma' X_{ij} \quad (8)$$

$$\text{Variance function:} \quad \text{Var}(\theta_{ij}) = 2\theta_{ij}^2$$

The link function is simply a transformation of (7). For the variance function we have chosen the deviance transformation. This is because it is close to the optimal normalising transform for the GLM distributions irrespective of the distribution chosen for the link (Pierce & Schafer, 1986).

Note that we allow the frailty to differ between different individuals. It is also possible that the frailty may vary through the recurrences. In this sense we have introduced flexibility into the gamma frailty model.

In the estimation of the model we propose a modification of the EM algorithm, which we called EMB. In particular a new step (step 1) is introduced in the algorithm,

- 0.- From the provisional value of $\tilde{\theta} = 1$, estimate a standard AG model and compute $\tilde{\theta}_i$.
- 1.- Estimate a model for the variance and obtain the fitted values of $\tilde{\theta}_i$.
- 2.- Use the values of the variance computed in step 1 to fit the AG gamma frailty model using the standard EM algorithm.
- 3.- Compute $\tilde{\theta}_i$ and return to step 1.

The EMB algorithm is iterated until convergence. The complete EMB algorithm is shown in the appendix.

An obvious starting value for $\tilde{\theta}$ would be 0, i.e. no frailty, the problem is that in this case the frailty remains fixed at zero in the update formula. For this reason, we have preferred here $\tilde{\theta} = 1$.

Results

A simulation study was conducted to evaluate the performance of the proposed EMB algorithm and to compare it with other methods, the PWP model and the penalized Cox model in particular.

Multivariate failure times were generated from an AG gamma frailty model with the following hazard function, where i denoted individuals and j denoted repeated measures, i.e. recurrences, within the same individual. In particular, we considered $I = 100$ individuals and $J = 2$ recurrences.

$$\lambda_{ij}(t) = \lambda_{0j}(t) e^{X_{ij}\beta + Z_{ij}\omega}$$

We simulated two Weibull baseline hazards,

$$1) \lambda_{0j}(t) = 0.5 (e^{0.1X1_i + 0.3X2_{ij}})^{-0.5}$$

$$2) \lambda_{0j}(t) = 0.5 (e^{0.1X1_i + 0.3X2_{ij} + 0.2X3_{ij}})^{-0.5}$$

Two fixed variables (X_1 , X_2) and one time-varying explanatory variable (X_3) were first simulated, although they were maintained fixed throughout the simulation. In particular, $X1 \sim \text{binomial}(1, p = 0.6)$, $X2 \sim \text{normal}(32.639, 12.967)$ and $X3 \sim \text{binomial}(4, p = 0.25)$.

Let $\omega = \exp(\omega)$ follow a gamma distribution with parameters ν and η . Without loss of generality, we will assume that $\nu = \eta$, i.e. that the subject-specific ω_i has mean one and variance $(1/\nu) = \theta_i$, where θ dictates the heterogeneity across individuals. Summing up,

$$\omega_i \sim \Gamma\left(\frac{1}{\theta_i}, \frac{1}{\theta_i}\right).$$
 Furthermore, we will assume

that the random effect consists of independent clusters of observations, i.e. $Z_{ij} = 1$ if recurrence j belongs to individual i , and zero otherwise.

From (8), we simulated two cases: Case A. $\text{Var}(\omega_i)$ fixed over time ($\theta_{ij} = 1$); Case B.

$\text{Var}(\omega_i)$ time-dependent ($\theta_{ij} = 0.5 + 0.1ID + 0.25X3_{ij}$), where

$ID = 1, 2, \dots, 100$. Finally, we compute $t_{ij} = \lambda_{ij} / \omega_i$.

Simulated failure times for units (i, j) were independently censored by three uniform variables across all datasets to achieve overall censoring levels of 95%, 80% and 40%. Summing up, 500 datasets were simulated for twelve possible designs, 1A, 2A, 1B and 2B with 95%, 80% and 40% censoring for each one.

Three methods were used to fit the simulated data sets, PWP, penalised Cox (PC) and our proposed modification (EMB). In all cases baseline hazards were allowed to vary between recurrences. Only in the penalised Cox case the variances of the parameters were not robustly estimated. Furthermore, in this latter case, the design matrix for the frailty was set equal to a diagonal matrix, each element of the diagonal corresponding to a different individual.

In Table 1, we show the results of the simulation. It was expected that failure to model existing frailty would result in biased estimates of parameters and reduced efficiency (Wassell & Moeschberger, 1993). In fact, it is possible to see a different pattern for the estimates of the parameters and for the estimates of the standard errors. With respect to the parameter estimation, PC and EMB fits were more similar to one another than to the PWP fit. Lower levels of censoring provided the most similar results for PC and EMB.

Note also that these two methods were more similar in the case of non-constant variance of frailty than in the constant. With respect to the estimates of the standard errors EMB seemed to provide the most efficient estimates. In fact, the PC fits were always more inefficient than the rest, even in relation to the model that did not explicitly model the frailty, i.e. PWP. Again, EMB was more efficient in lower censoring and in non-constant variance of frailty cases. In addition, although the S-plus macro for the estimation of the PC gave an estimation of what it called the variance of the frailty (Therneau & Grambsch, 1998), we are not sure that it was in fact such variance. In this sense, note the discrepancies with the EMB results of the estimation of such variance.

Table 1. Results of the simulation. 500 data-sets of four possible designs (constant frailty and fixed covariates 1A; constant frailty and time-varying covariates 1B; non-constant frailties and fixed covariates 2A; non-constant frailties and time-varying covariates 2B). 100 individuals and two recurrences. Three levels of censoring (60%, 80% and 95%).

	n=100, censoring = 60%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.00984	0.06272	0.09305	0.16261	0.15457	0.17256
Beta2	0.01491	0.03160	0.01490	0.03278	0.01081	0.02789
s.e. Beta1	0.18456	0.19418	0.19418	0.19832	0.18281	0.18649
s.e. Beta2	0.00672	0.00691	0.00726	0.00745	0.00652	0.00668
Var frailty			0.00626	0.04133	0.92705	0.91868
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.00130	0.07330	0.00865	0.17326	0.06499	0.13064
Beta2	0.01378	0.03408	0.01367	0.03474	0.04596	0.03441
Beta3	0.01799	0.02471	0.01783	0.02530	0.06693	0.06126
s.e. Beta1	0.18670	0.19338	0.01946	0.19689	0.19305	0.18157
s.e. Beta2	0.00671	0.00691	0.00726	0.00739	0.00678	0.00663
s.e. Beta3	0.01063	0.01050	0.01073	0.01081	0.01060	0.01007
Var Frailty			0.00455	0.02825	0.95122	0.91837
	n=100, censoring = 80%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.08614	0.14580	0.11143	0.20428	0.10390	0.16684
Beta2	0.00553	0.02439	0.00597	0.03056	0.00400	0.05279
s.e. Beta1	0.34517	0.35449	0.36676	0.39950	0.30799	0.29621
s.e. Beta2	0.01443	0.01435	0.01327	0.01415	0.01262	0.01217
Var frailty			0.29096	0.69528	0.90931	0.91146
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	0.07811	0.15113	0.10246	0.21342	0.09290	0.17243
Beta2	0.00682	0.02737	0.00597	0.03428	0.00247	0.02815
Beta3	-0.12416	-0.10825	-0.11837	-0.10280	-0.11970	-0.10655
s.e. Beta1	0.34398	0.35390	0.36665	0.39631	0.30853	0.31750
s.e. Beta2	0.01476	0.01479	0.01338	0.01414	0.01293	0.01273
s.e. Beta3	0.01883	0.01854	0.02037	0.02143	0.01711	0.01691
Var Frailty			0.29377	0.67242	0.90686	0.87601
	n=100, censoring = 95%					
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	-0.48964	-0.37113	-0.49018	-0.37202	-0.50812	-0.62366
Beta2	0.02062	0.04942	0.02063	0.04946	0.01756	0.11219
s.e. Beta1	0.55103	0.54926	0.58883	0.57913	0.53947	0.53414
s.e. Beta2	0.01934	0.01936	0.02241	0.02228	0.01897	0.02875
Var frailty			0.00449	0.00495	0.93732	0.93564
	PWP1A	PWP1B	PC1A	PC1B	EMB1A	EMB1B
Beta1	-0.51248	-0.34955	-0.53836	-0.36148	-0.52390	-0.35898
Beta2	0.02647	0.05148	0.02510	0.05211	0.02469	0.06804
Beta3	0.50120	0.48832	0.52027	0.49929	0.50609	0.10052
s.e. Beta1	0.57865	0.56923	0.61556	0.59845	0.56670	0.53548
s.e. Beta2	0.01964	0.01918	0.02277	0.02226	0.01936	0.01810
s.e. Beta3	0.02754	0.02750	0.03108	0.03038	0.02667	0.02822
Var Frailty			0.16467	0.10965	0.92469	0.94274

PWP denotes Prentice, Williams and Peterson model; PC denotes penalised Cox model and EMB the AG gamma frailty model fitted using the EMB algorithm. The variances of the parameters in the PWP and the EMB were robustly estimated.

Application to nosocomial infection in an intensive care unit study

As mentioned above, in Barceló and Saez (2001) we tried to determine which factors, those associated to patients (intrinsic risk factors for the nosocomial infection, NI) or those related to the intensive care unit, ICU (extrinsic risk factors), were more relevant in the explanation of the occurrence of nosocomial infections in a tertiary-level hospital in Girona, Spain, during the second quarter of 1999.

The dependent variable (episode of infection hereinafter) consisted of either the time from the admission to the ICU to the onset of the infection (originated by a micro-organism, bacteria or fungi) or the time between the onset and the end of the infection. It was possible for patients to be infected more than once during their ICU stay. In the definition of our dependent variable, only ICU nosocomial infections were considered. Community-acquired infections, infections from other hospitals and infections from other hospital areas were not included under this definition. The beginning of the study did not always coincide with the patients' admission to the ICU, but with their admission in the hospital. Therefore, delayed entry was allowed. We also considered the possibility of (right) censoring because some patients would not get infected during their ICU stay.

Following the medical literature possible risk factors considered were classified as either intrinsic or extrinsic risk factors for NI. The former contained those directly related to the patient, such as gender and age, as well as those originating outside the ICU, such as previous infections (either community-acquired, or from another hospital or from another hospital area), severity of disease at admission and urgent surgery. Extrinsic risk factors for NI considered were location, mechanical ventilation, catheterism (central venous intravascular and arterial), tracheotomy, probes (urinary and nasogastric) and antibiotic treatment (antimicrobial used, duration and dosage). The effect of extrinsic risk factors was evaluated using the days of exposure to such a risk factor in a particular patient. The exposure was limited to the days prior to the onset of infections. Further details on data, variables and

additional results can be found in Barceló and Saez (2001).

Results of the fit of the model by PWP, PC and EMB are shown in Table 2. PWP was used here for comparative purposes. In this regard, note in Figure 1 that deviance residuals were not symmetrically distributed around zero and, above all, did not present a constant dispersion. Furthermore, in Figure 2, we show the estimates of the variance of the frailty (computed as shown in [A1]). The variance was not fixed between or within individuals. Some explanatory variables could explain such variability. See for instance in Figure 3 the relationship between the variance and one intrinsic factor (community-acquired previous infection) and between the variance and an extrinsic factor (mechanical ventilation).

In the implementation of the S-plus macro for fitting the PC, we used a design matrix consisting of independent clusters of observations, i.e., $Z_{ij} = 1$ if recurrence j belongs to individual i , and zero otherwise. With respect to EMB we needed to specify a model for the variance of the frailty (see equation (8)). We tried a forward stepwise strategy. We started with a single explanatory variable and included another one only if the AIC diminished. When we had a preliminary specification, and in order to check its robustness, a backward strategy with all the variables included was also tried. The final model for the variance is shown in Table 2.

From Table 2 we can see that the best fit was obtained from the EMB. In this sense, compare the EMB and the PC fits. The overdispersion (47.9% in EMB and 71.6% in PC); the AIC (162.41 and 180.42, respectively) and the pseudo R-square (Nagelkerke, 1991) (0.529 and 0.442, respectively) were lower in the case of EMB. Furthermore, the estimates obtained using our proposed modification were the most efficient. Note, in addition, that the estimate of the variance of the frailty provided by the PC S-plus macro was close to zero, meaning that there was no frailty in the model. This result contradicted the variability shown in Figures 1 to 3. Note also the varying behaviour of the variance of the frailties in Figure 4 (derived from the EMB fit).

Figure 1. Plot of the deviance residuals of the PWP model.

Figure 2. Plot of the estimate of the variance of the frailty. PWP model.

Figure 3. Plot of the estimate of the variance of the frailty vs. an intrinsic risk factor (community-acquired previous infection) and an extrinsic risk factor (mechanical ventilation). PWP model.

Table 2: Results of the estimation of the Prentice, Williams and Peterson, PWP; Penalised Cox and EMB models. PWP and EMB with robust estimation of the variance.

	PWP			Penalised Cox			EMB		
	Hazard Rate	lower .95	upper .95	Hazard Rate	lower .95	upper .95	Hazard Rate	lower .95	upper .95
Intrinsic risk factors									
Gender (male)	0.631	0.222	1.788	0.631	0.133	2.999	0.578	0.221	1.512
Age	1.009*	0.985	1.033	1.009*	0.981	1.037	1.009*	0.990	1.028
Previous infections (non)									
Community-acquired	4.56**	1.435	14.477	4.558*	0.954	21.772	6.923**	2.578	18.594
Other infections	0.56	0.201	1.557	0.560	0.118	2.648	0.418*	0.162	1.077
CDC (stable)									
Unstable intens. Care	54.468*	0.786	3776.406	8025.27**	1740	36968.77	49.012**	1.111	2162.135
Unstable shock	219.967**	3.187	15182.609	NA	NA	NA	257.040**	5.997	11017.03
Urg. surg. (non)	0.668	0.258	1.731	0.668	0.221	2.021	0.579	0.268	1.251
Extrinsic risk factors									
Location (rest of beds)									
Bed 4	2.255	0.545	9.340	2.255	0.346	14.707	2.247	0.621	8.135
Bed 5, 10, 11	0.946	0.381	2.348	0.946	0.257	3.478	0.784	0.369	1.667
Mechanical Vent. (non)									
≤ 3 days	0.011*	0.000	1.175	NA	NA	NA	0.006**	0.000	0.314
4-10 days	61.929**	1.325	2894.463	64.406**	1.35	3065.471	59.981**	2.105	1709.126
>10 days	129.534**	4.005	4189.427	134.714**	3.06	5934.619	98.713**	4.260	2287.525
Venous catheter (≤ 3d)	1.065	1.007	1.775	NA	NA	NA	1.082	1.011	1.822
Arterial catheter (non)									
> 0 days	1.833	0.579	5.805	1.833	0.495	6.782	1.629	0.638	4.162
Tracheotomy (non)									
≤ 6 days	0.231	0.028	1.924	0.231	0.030	1.768	0.243*	0.044	1.355
> 6 days	0.351	0.072	1.705	0.351	0.057	2.171	0.303*	0.075	1.224
Urinary probe (non)									
≤ 4 days	2.453	0.018	334.750	2.453	0.024	251.814	3.922	0.066	231.811
5-12 days	6.355	0.306	132.032	6.355	0.137	294.936	7.942*	0.630	100.190
> 12 days	1.566	0.079	31.014	1.566	0.035	70.518	1.998	0.159	25.104
Nasogastric probe (non)									
≤ 9 days	1.066	0.310	3.660	1.066	0.285	3.990	1.063	0.377	3.000
> 9 days	0.133**	0.023	0.748	0.133**	0.027	0.649	0.129**	0.030	0.550
Antibiotic treat. (non)									
≤ 7 days	0.415	0.112	1.539	0.415	0.093	1.858	0.329**	0.122	0.883
> 7 days	0.038**	0.004	0.385	0.038**	0.005	0.327	0.028**	0.004	0.216
Antibiotic dose-DDD	1.035	0.948	1.129	1.035	0.972	1.101	1.037	0.982	1.096
Deviance (degrees freedom)	130.427 (76)			130.422 (76)			112.4127 (76)		
Overdispersion	1.716144737			1.716078947			1.479114474		
AIC	180.427			180.422			162.4127		
Pseudo-R ²	0.441886			0.441889			0.529338327		
Var frailty				5e-007			1.124881		
Model for the variance							β	s.e.(β)	
Mechanical Vent. (non)									
≤ 3 days							0.051	0.03602	
4-10 days							0.056	0.03617	
>10 days							0.086	0.03210	
Tracheotomy (non)									
≤ 6 days							0.008	0.026	
> 6 days							-0.006	0.023	
Urinary probe (non)									
≤ 4 days							0.038	0.045	
5-12 days							-0.011	0.042	
> 12 days							-0.019	0.046	
Nasogastric probe (non)									
≤ 9 days							-0.0002	0.021	
> 9 days							-0.0059	0.021	

Figure 4. Plot of the estimate of the variance of the frailty. EMB model.

Summarizing the results, both, intrinsic and extrinsic factors were predictors of NI. In this sense an intrinsic factor such as CDC classification (unstable patients) and an extrinsic one like mechanical ventilation (more than 3 days) presented the highest hazard rates. For any type of infection all the (statistically significant) intrinsic variables were risk factors for NI. In decreasing order of importance we could mention CDC classification (unstable) and previous community-acquired infections. Most of the extrinsic factors were also risk factors. In this sense, and again in decreasing order of importance, we can list mechanical ventilation (more than 3 days), urinary probe (5-12 days), location (bed 4), the presence of an arterial catheter, and central venous catheter. Only three of the extrinsic factors were protective, presence of tracheotomy, nasogastric probe (more than 9

days), and, in particular, antibiotic treatment (days of treatment).

The interpretation of the model for the variance of the frailty is also worth while. Note that only extrinsic factors (mechanical ventilation, tracheotomy and probes, urinary and nasogastric) explained the variance of the frailty. In this sense, the sources of heterogeneity, both between and within individuals, could be attributed to the medical procedures that the ICUs use, whereas the effect of those factors related to the susceptibility of the patients could only be marginal.

Conclusion

Our purpose was to present a modification of the AG gamma frailty model. In particular we proposed to directly model the variance of the

frailty by means of a GLM and also to modify the EM algorithm, using the EMB algorithm, in order to simultaneously estimate a semiparametric model for the failure times and a model for the variance of the frailty.

In both the simulation and in the application to multivariate data from a nosocomial infection study, the EMB fit turned out to be better than the fit obtained from a marginal model (PWP) and from a conditional one (penalized Cox model). In this sense, the EMB provided the best fit (being the least overdispersed and having the highest AIC and the highest pseudo-R square) and estimated the parameters most efficiently. We think, therefore, that our proposed method is able to take into account and to estimate unobservable random effects in semiparametric models.

Two shortcomings, however, should be mentioned. First, as in the rest of frailty models, we introduce frailties into the marginal hazards in order to explicitly model the dependence between recurrences. The problem is that frailties are in fact capturing two different, although related, sources of variation, that is heterogeneity (or false contagion) and serial dependence (or true contagion) (Aalen, 1994). The former, the original use of frailty, is a consequence of unobserved individual covariates that are not included in the study either because of practical circumstances or because they are not known to be risk factors. The latter is in fact a consequence of unobserved common covariates that are integrated out (Petersen, 1998).

It seems, at any rate, that frailty models successfully capture heterogeneity but permit a considerable amount of non-controlled serial dependence. A possible but partial solution tried here is to stratify according to the recurrences, as in the PWP model, thus allowing the hazard to vary between them. The second shortcoming, also shared with the rest of frailty models, is the lack of methods with which to assess the goodness-of-fit of our method. At any rate, we are sure that these shortcomings deserve further research.

References

- Aalen, O. O. (1994). Effects of frailty in survival analysis. *Scandinavian Methods in Medical Research* 3, 227-243.
- Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10, 1100-1120.
- Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Statistical Methods in Medical Research*, 1, 297-315.
- Andersen, P. K., Borgan, O. Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York, Springer-Verlag.
- Andersen, P. K. Klein, J. P., Knudsen, K. M. & Tabanera-Palacios, R. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics*, 53, 1475-1484.
- Barceló, M. A. & Saez, M. (2001). Factors determining the incidence of nosocomial infections in an Intensive Care Unit of a public hospital in Girona, Spain. Research Group on Statistics, Applied Economics and Health (GRECS), University of Girona, Spain
- Behrman, J. R., Sickles, R., Taubman, P., & Yazbeck, A. (1991). Black-white mortality inequalities. *Journal of Econometrics*, 50, 183-203.
- Clayton, D. (1978). A model for association in bivariate life tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141-151.
- Clayton, D. & Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A*, 148, 82-117.
- Clayton, D. (1988). The analysis of event history data: a review of progress and outstanding problems. *Statistics in Medicine*, 7, 819-841.
- Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 34, 187-220.
- Fleming, T. R., & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York, John Wiley & Sons, Inc.

Gill, R. D. (1985). Discussion of multivariate generalizations of the proportional hazards model, by Clayton and Cuzick. *Journal of the Royal Statistical Society, Series A* 148, 82-117.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators of the von Mises method (Part 1). *Scandinavian Journal of Statistics*, 16, 97-128.

Klein, J. P. (1992). Semiparametric estimation of random effect using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58, 619-678.

Lin, D. & y Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, (408), 1074-1078.

Lindsey, J. (1999). A review of some extensions to generalized linear models. *Statistics in Medicine*, 18, 2223-2236.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Association, Series B* 44, 226-233.

Nagelkirke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

Nelder, J. A. (1998). A large class of models derived from generalized linear models. *Statistics in Medicine*, 17, 2747-2753.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, 72, 221-232.

Nelder, J. A., & Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, 7, 107-120.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19, 25-43.

Petersen, J. H. (1998). An additive frailty model for correlated life times. *Biometrics*, 54, 646-661.

Pierce, D. A., & Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, 81, 396, 977-986.

Prentice, R. L., Williams, B. J., & Peterson, A. L. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68, 2, 373-79.

Therneau, T. M., & Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine*, 16, 2029-2047.

Therneau, T. M., & Grambsch, P. (1998). *Penalized Cox Models*. Mayo Clinic.

Vaupel, J. W., Manton, K., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439-454.

Wassell, J. T., & Moeschberger, M. L. (1993). A bivariate survival model with modified gamma frailty for assessing the impact of interventions. *Statistics in Medicine*, 12, 241-48.

A Conversation With R. Clifford Blair On The Occasion Of His Retirement

Shlomo S. Sawilowsky
Evaluation and Research
Wayne State University

An interview was conducted on 23 November 2003 with R. Clifford Blair on the occasion on his retirement from the University of South Florida. This article is based on that interview. Biographical sketches and images of members of his academic genealogy are provided.

Keywords: R. Clifford Blair, nonparametric, Wilcoxon rank sum, rank transform, multivariate permutation tests, step-down multiple comparison test, comparative statistical power

“In the last 30 years there have been important changes in the canons of good statistical practice or data analysis. Until recently, and thanks to the work of J. V. Bradley and R. C. Blair among others, it is no longer heresy to say that distribution-free tests – such as the Wilcoxon-Mann-Whitney are preferable to their normal theory alternative – the t test.” (Bruno D. Zumbo & Donald W. Zimmerman, 1993, *Canadian Psychology*, 34(4), p. 441)

Background

JMASM: What are some of the memorable events from your childhood?

RCB: I was born in an area that is now Tampa, Florida. It was rural – I remember the cows, chickens, and pigs. We lived on a tiny, dirt road. We were poor; once my mother boiled an onion for three of us “youngins.” She told us we were having onion soup, but no, she really didn’t like onions so she would be having only the broth.

The people who lived in our area were farmers who came from southern Georgia. When the depression came along, they moved into the cities looking for work. My parents worked in a cigar factory.

This was in the time before machines were used, so my mother worked with hand-rollers. She went as far as the 7th grade. Her family lived in the Lake Okeechobee area, where they picked vegetables and hunted sea turtles. Her father, my grandfather, was a part-time Baptist minister and part-time moon-shiner.

My father, who died when I was nine years old, made it to the 2nd grade. He was a mechanic in the cigar factory. When he came into contact with the Spanish of the Cuban community he fell in love with the language. Eventually, he learned how to speak and read Spanish, and he especially enjoyed reading Mexican classics.

I went to a school where the girls wore shoes, but most boys didn’t. Actually, there were two kinds of students – “by the dayers” and “by the weekers.” The dayers were children who turned in their twenty cents lunch money day by day. The weekers were the upper class; those who had the entire week’s lunch money on Monday. The boys among the weekers had shoes, but those of us who were dayers, the lower members of the social hierarchy, didn’t have shoes. (My hobby is writing short stories,

Shlomo Sawilowsky is Wayne State University Distinguished Faculty Fellow, Professor of Evaluation and Research, and “data analyst.” Contact him at shlomo@wayne.edu.

and I've written on my experiences at that little school, Thomas A. Edison Elementary.)

I went on to Memorial Junior High School and then Hillsborough High School. I was in a class for the trainable and mildly retarded. There had always been good programs for the severely retarded, where the children were taught how to tie their shoes and other functional skills. However, the school system relied on informal programs for the mildly retarded. Our classes were taught by a basketball coach. He wasn't credentialed; he had the job because he had spare time.

My favorite class was personal hygiene. The coach taught us that we should wash under both of our arms. And then, we would have a test. The questions would be something like "1. You should wash under how many arms? (a) only your left arm, (b) only your right arm, (c) both arms", and my preferred answer, "(d) none of the above." Alas, this was too much for me, and I failed eleventh grade. So, I ran away from home. I went to Atlanta for a few weeks, hung out with some bums, almost starved to death, and had no choice but to come back.

My mother thought that my vision played a role in my lack of attention at school. It had been checked, but the doctor hadn't made the proper diagnosis. The degenerative eye disease I have is extremely rare in juveniles. When I returned from Atlanta, my mother decided to have my vision checked again. I was taken to a specialist, who determined I was nearly blind. I was bundled up and sent off to the Saint Augustine State School for the Blind, where I was viewed as being mildly retarded and having a severe visual impairment.

After I graduated, the Bureau of Blind Services sent me to a rehabilitation center in Daytona. It was popular at that time to give blind people jobs in a hospital or post office. They would run a small concession stand, selling candy, coffee, and cigarettes. It wasn't clear if I could be taught how to make change. I spent a lot of time sitting at a table with giant paper dollars, and large disks representing quarters, dimes, nickels, and pennies. The teacher would say, "I am buying two candy bars, they're seven cents a piece, and I am giving you a dollar. How much change do you give me back?" I would tear one of the giant paper

dollars in half and give it as change. After a while, it was determined that perhaps I might be more suitable for a career in manual labor.

They enrolled me in a class where we were taught how to work with plants in a nursery. I was in a horticulture class where everyone, including the teacher, was completely blind. Obviously, even though I have severely limited vision, it was sufficient to make me the king of the class. An important event occurred at that time, which was to change my life.

In an effort to help me make change for a dollar, the rehabilitation center had given me a magnifying glass. We were outside working with the plants, and I started complaining about the firebugs that were biting me. I said "These firebugs are eating me up, are they bothering you boys?", but of course they said "No, they aren't bothering us." Then, I started swatting all around me, making a lot of noise in doing so, for the entire day.

The next day, when we were working outside, I took out my magnifying glass and focused it on the back of their necks, so they would feel it burning. One classmate slapped his neck and said, "Damn Cliff, they're getting me now. I can feel them biting me all over the back of my neck." They really thought we were being attacked by firebugs!

I entertained myself doing that for quite a while, but then got sent to the school nurse, who was the disciplinarian. She called me in and said, "Cliff, this is not a discussion. I'm going to tell you this only once: All the talk about the outbreak of firebugs will cease immediately. You are dismissed."

The school officials decided, because of this incident, that I was a bit too precocious, so they gave me a quick screening IQ test, which was the first such test I had taken. My scores didn't match my academic profile. They called in a paid intern who was a doctoral student in psychology from University of Florida, who gave me another test. On that basis, he decided to take me to Gainesville, to visit the Department Chair, who gave me a complete battery of tests. So, I went from washing under (both) your arms to enrolling in college.

Now, I figured I was going to do higher mathematics in college, so I set about memorizing all of my nines tables! Then, I

enrolled in some classes. I was so scared that I would be sick to my stomach, brush my teeth, and go to class. There were no disabilities offices in those days; you either made it or you didn't. I do recall one professor who made it clear he wanted me out of his class. After a while, I decided he was right, and left school.

I got a job in a factory in Tampa emptying trash cans. After being there for about a week, one of the ladies on the assembly line asked me what I did. "I'm the trash man." I felt good about that reply, so I continued telling my co-workers that I was the trash man. After a while, an elderly gentleman called me over. He said, "I hear you've been telling everybody that you're the trash man. You better get this straight. You are *not* the trash man. *I'm* the trash man. You're the *assistant* trash man." So, I guess I promoted myself a little bit when I called myself the trash man, because I was, in fact, only the assistant trash man.

After a few years, I decided to try college again and came to the University of South Florida (USF). I continued to read and catch up. I would find big words in the dictionary, and try to use them in a sentence. My first success was, in fact, in an English class. The assignment pertained to a story about funerals, funeral homes, and death. I wrote a very somber, thoughtful, introspective yet reflective, spiritual essay. It was an intellectual breakthrough for me, and I was quite pleased with my effort.

The professor came into class with the graded papers. He said, "I have a paper here that was a delight. It's probably one of the finest examples of humorous satire I have read from a student. I was reading it to the passengers in my carpool, and the driver laughed so hard he drove off the road. We thought we would be killed. The student who wrote this is Cliff Blair; Cliff, congratulations!" My very first college success: I got an A+. I went on to graduate from college!

JMASM: What role did humor play in your youth?

RCB: Although in retrospect, I was somewhat depressed as a child, life became very funny to me. I would take a closer look at what was happening to me and laugh.

Maybe it started back at that rehabilitation center. I had been out with a couple of guys one night. We had been drinking beer, and at curfew they went back. I had decided to stay for a few more hours, and then I tried to sneak into the dormitory from the rear entrance. I had not been in the back, it was dark, and I have very limited vision. I came upon a fence, and in climbing over it I got stuck. I was caught upside down! I continued to struggle, and after a long while I finally broke free.

The next day the supervisor announced we had been vandalized that night. "It was awful," he said. "Someone trashed my wife's rose garden." It turned out I wasn't caught in a fence; it was a rose trellis. If I had only walked either a little to the right or to the left I would have avoided it. The incident was very funny, and I began to see the world as being a bit odd – as if I was viewing it upside down.

Research

JMASM: What interested you in statistics?

RCB: After the long journey through the undergraduate program, I decided I wanted to get a Master's degree. I didn't have the money, but there was a new program in ageing studies that had stipends for students. So, I decided to obtain a Master's degree in that subject. I took a course in social and behavioral science measurement theory from Professor John Neel, who's now at Georgia State University.

He had introduced Chebyshev's theorem, which certainly caught my attention, because it was way over my head. In the context of that lecture, he mentioned to the class that he was proud that the department had just obtained a programmable calculator made by Wang. It was programmed in pseudo-assembler, with two-digit numbers. I was very impressed with the device, but of course none of the other students demonstrated any interest – they were happy enough to get through the course. John offered to give me a closer look at it, and he showed me how it was programmed.

We had recently learned about Pearson's product-moment coefficient of correlation, so I asked and obtained permission to try and write a program to compute it on the Wang. I spent probably about a week working on it, but finally

coded the correlation coefficient. I showed John how I could input data for the X and Y variable and it would produce the result.

Professor Neel seemed impressed with my efforts, and in our ensuing conversations, he discovered the paucity of my math background. He gave me his ninth grade algebra book. I studied it, and I liked it. So, I started taking more of his classes in measurement and statistics. I completed the Master's degree and enrolled in the Ph. D. Measurement, Evaluation, and Research program in the College of Education.

To prepare, I took algebra, trigonometry, and introduction to calculus. Then, I took a Fortran class in the College of Engineering. (Later, as an Assistant Professor, you and I took a three course calculus sequence together.

JMASM: I enjoyed the refresher. I enrolled in the course so I could take notes for you, because by then you weren't able to see writing on the chalk board.

RCB: Correct.) Eventually, it became time for me to do my dissertation, and by then I was primarily interested in statistics. I was looking for a statistics topic, but the focus in the department was on measurement, evaluation, and research methods. Therefore, I went to the math department and met Professor James J. Higgins. We discussed various statistical topics.

Jimmy was really trained as a probabilist, but had become a statistician. I asked him to chair my dissertation committee, and Bruce Hall, the measurement expert from the College of Education, was the co-chair.

One of the things that fascinated me when I first started college was footnotes. Due to my vision, either I had never seen them before or I simply ignored them. They were tiny markings that I hadn't recognized as letters of the alphabet. As I went through college, therefore, I made it a point to read them.

I read some footnotes in statistics books regarding the comparative power of nonparametric hypothesis tests. Book after book that I read indicated that nonparametric tests have the advantage of not needing the specification of the population (i.e., normality), but the unfortunate shortcoming was that they

lacked statistical power as compared with parametric tests. Nonparametric tests were often described as rough, crude, quick, and dirty.

However, by this time I had read about asymptotic relative efficiencies (AREs) – I first came across it in a footnote. I saw a quote from William Mendenhall who said something like “Don't pay much attention to these things, because asymptotic relative efficiencies deal with infinite sample sizes and infinitesimal treatment effects” which has little application in the real world. Nevertheless, the ARE's indicate that a test such as the Wilcoxon Rank Sum (WRS), for example, should have a huge power advantage over the independent samples t test. But, textbook authors claimed it doesn't. I wondered about what point the WRS loses power or what its comparative power would be for small samples.

I made it my practice to check things out empirically, because I had taken that Fortran course. For example, when I heard the claims about the central limit theorem, I wrote a computer program to see what happens to the distribution of sample means as either the sample size increases or the number of resamplings increases for a fixed sample size. So, I began to do the same thing to check the power comparisons between the two tests.

We didn't have personal computers at that time, so I had to go across campus to the computer center. It was still in the days when we had to use a key punch machine to punch cards. I can remember the evening, just before they closed, that I obtained the first power results. The power of the Wilcoxon Rank Sum (WRS) test was *way above* that of the t test for certain nonnormal distributions. I checked and rechecked the code carefully, so I knew the results were correct.

Then, my hands began to shake, and I couldn't see the results even with my high powered loupe. Gradually, though, it dawned on me that hundreds of books that I had read were wrong. Authors explained the WRS must be less powerful because when original scores are converted to ranks, information in the data set is lost, and there is a resulting loss in statistical power. The explanation is logical, but wrong.

JMASM: There was a departmental library in the (former) College of Education building at USF. It was a sizable collection of statistics, measurement, research, and evaluation books. I checked out a book by Glass and Hopkins. At the place where they indicated nonparametric tests were less powerful, someone had written in the margin, “poo-poo.” Are you the culprit?

RCB: Yes, that was me. That was the standard thing that I wrote in the margins on this issue. By that time I had gotten enough preliminary results to know the statement was wrong. I got full of myself. I realized that I was right, and the big guys were wrong.

In the late 1970s, I tried to publish a couple of papers, but the editors would reject them. The reviewers indicated there must be a problem with the computer program, and that would be why the WRS has more power.

I read C. Alan Boneau’s work. He published some Monte Carlo simulations in 1962 in *Psychological Review*. (His more famous article appeared in *Psychological Bulletin* in 1960.) Unlike other authors who claimed the WRS was less powerful, at least he contended there wasn’t any difference. But, he had only investigated limited study conditions, such as very small sample sizes.

JMASM: The power advantages of the WRS over the *t* test, under departures from population normality – but not homoscedasticity – for a shift in location alternative, increases as the sample size increases. Yet, the recommendation in many textbooks is the opposite: As the sample size gets smaller, the security blanket of the central limit theorem is lost, so that is when one should turn to a nonparametric test. However, the recommendation of when to use a nonparametric test is being incorrectly dictated by the limitation of the *t* test, when in fact it should be based on the properties of the WRS. Do you agree?

RCB: Yes. To get the huge power advantages in nonparametric tests, certainly use them when there are large sample sizes!

Then, I saw the 1972 article by Glass, Peckham, and Sanders in *Review of Educational Research*. Their view was that the parametric

tests are robust enough so that there’s never a need to turn to the less efficient and less powerful nonparametric tests. They referred to James V. Bradley’s work. But, they discounted it because they claimed we now understand more about the robustness of the parametric test.

JMASM: They said, “applied statistics experienced an unnecessary hegira to nonparametrics”. A hegira means to escape danger!

RCB: They believed “the flight to nonparametrics was unnecessary”. They said Bradley’s work threatened the “safety of the herd.” I took that as demeaning. Apparently those of us who dabble in statistics have the mentality of being part of a herd.

I communicated frequently with James Bradley, and I was greatly influenced by his work. So, I took it upon myself to respond to Glass et al. My article appeared in *Review of Educational Research* in 1980.

This brings up the issue as to how I wrote my early manuscripts. When I first started writing I didn’t have a mentor. I would find a journal that I wanted to target, and read some articles that had been published in it. Unfortunately, in the controversies in the literature raging at that time, many combatants wrote harsh statements about one another. I was given to understand, therefore, that this was the scientific manner of publishing an article. So, I wrote my manuscripts in a fashion that raked various supporters of parametric procedures over the coals. I made of lot of people mad with me.

It turned out to be helpful, though, as reviewers were so angry with me, that instead of just rejecting my work, they spent considerable energy in response. I got very important tutoring from some of the best researchers in the field that way. They cited reference after reference, and, I would look each one up. It was a very valuable experience. Of course, it didn’t help that I was writing articles touting the benefits of nonparametric procedures, and reviewers figured I was wrong anyway.

After a while, though, my articles began to get published. I was gratified to see that in subsequent editions of many of the textbooks I referred to earlier, the authors made changes to

the text. Sometimes they quoted me, and sometimes they didn't. Later, they quoted me and you, and sometimes they didn't. But, the main point is that they changed their texts. They finally recognized that nonparametric tests are often far more powerful than parametric tests under the commonly found conditions discussed in many of our articles.

Because I hadn't had a proper background in mathematical statistics, I had to rely mainly on the computer programs. This had the advantage of requiring me to consider practical issues and conditions that, quite frankly, sometimes are overlooked or discounted by mathematical statisticians. I learned a lot about the real properties of statistics this way. Also, by this time, Jimmy was giving me articles to read, and that also helped a lot.

One day, in 1983, a doctoral student approached me about a dissertation topic. We discussed the rank transform. Ronald Iman – former President of the American Statistical Association – and William Conover, who presented the procedure, based their recommendation to use it on Monte Carlo evidence, but it was our contention that their support was insufficient. They had primarily examined the rank transform in the context of independent two sample and one-way layouts. Although they examined its properties under a factorial design, it was in nonrealistic contexts, such as the presence of only main effects, only an interaction, or very small main and interaction effects.

Therefore, I suggested the student examine the robustness and power properties of the rank transform in the context of the $2 \times 2 \times 2$ layout, with the presence of small, medium, and large higher-order interaction, lower-order interactions, and main effects. I thought the rank transform was a neat idea, but they hadn't sold it completely. Many sources indicated it was much more difficult to preserve robustness with respect to Type I errors when normality is violated – and similarly to detect – interaction effects, as compared with main effects.

That doctoral student wrote the Fortran program. By now the key punch machines were being replaced with terminals, so the process of coding, compiling, executing, debugging, and so forth was much faster. Soon, results began to

appear. I got a call late one night, and the student was concerned. He had gone over the program many times, but was still not getting good results for the rank transform. He was telling me that the Type I error rate for the test for interaction at a particular sample size and effect size had ballooned from 0.05 to 0.35. He was telling me that matters got *much* worse as the sample size got larger! We concluded that the statistic was flawed. This student and I then went on to write a number of articles on the rank transform. You were that student - remind me about what happened when we tried to publish those results.

JMASM: The main results were sent to a certain prestigious journal in 1985. After about six months, the Editor advised us that the paper was lost and to supply another copy. About nine months latter, we received a letter wherein a reviewer had requested a complete set of printouts – this was in the day of green and white 132 column-wide fan-fold computer paper, and the results were contained in a stack several feet thick. We mailed the printouts for the primary results immediately, but the manuscript was kept in review for almost two and a half years.

The article was rejected. The Editor based the decision on the weight of a single reviewer. That reviewer said that although he could find nothing wrong in the study conditions of the Monte Carlo, the procedures we used, or in the reporting of the results, he recommended the paper be rejected because it contradicted what well-known people had already written on the subject.

That well-known person the reviewer was quoting was, in fact, himself. Although the signature line had been blocked out to preserve anonymity, the editorial assistant had inadvertently failed to block out the affiliation.

Eventually, in 1989, Juliet Popper Shaffer published the primary dissertation results in the *Journal of Educational Statistics* – [now *Journal of Educational and Behavioral Statistics*]. But earlier, in 1987, the secondary results from the dissertation were published by Donald B. Owens in *Communications in Statistics* – remember our concluding sentence in that article? Subsequently, the literature

review was published by Penelope L. Peterson in *Review of Educational Research* in 1990.

RCB: It's the nature of the beast.

The rank transform is essentially worthless in the context of factorial ANOVA.

JMASM: Two of my doctoral students further examined the rank transform in 1997. Michael J. Nanna found it to work well in the context of the two independent samples Hotelling's T^2 . However, Todd C. Headrick demonstrated that as poor as the rank transform performs in the context of the two dependent samples t test and in factorial ANOVA, it performs even worse in the context of factorial ANCOVA.

RCB: It has always amazed me at how long people hold on to procedures that don't work. Years after all this was published, people still publish articles stating that it is a controversial topic. So-and-so say it does work, but Sawilowsky and Blair say it doesn't work.

JMASM: So would you say the jury is still out on the rank transform when it's Type I error rate goes to 1.00? This could be useful if you can't find a way to get a new drug to the market.

RCB: Yes, when nominal alpha is 0.05, 1.00 is a little high for a Type I error rate. It is my recollection that certain statisticians were on the pharmaceutical dog and pony circuit. If you had a drug you couldn't take to market, here's a statistic that guarantees rejection of the null hypothesis. It went so far that a major statistics software company, SAS, advised in their user manual to run the data through PROC RANK, and do the normal theory test on the ranks.

JMASM: Perhaps, many of the older textbook authors that you contradicted were not alive, and those who were alive were not in front of you to confront you. However, weren't you afraid to take on the scholarship of the discipline; afraid that you were taking on something bigger than you?

RCB: I considered myself to be a minor character, a tiny speck. I was once told that a prominent person in the field was asked, at a

conference, about some of my work that appeared to refute his work. He said, "It is too trivial for comment." At first, I was devastated by those remarks. Then, two colleagues explained that my work must be hitting the mark, otherwise it wouldn't be characterized as trivial, but as being wrong.

JMASM: You were once invited to speak at a national conference on a panel discussion regarding the rank transform. Another invitee (from my generation, not yours) spoke favorably on the procedure. A member of the audience complained that a lot of time and money was spent attending these conferences to obtain a "take-home" message, and yet the question remained why the two of you obtained different results. Your answer was perhaps the other person's work was based on different study parameters, different conditions, etc. The other person's reply: "Blair obviously is wrong".

RCB: Yes, I recall that. You and I had studied the rank transform in the context of the $2 \times 2 \times 2$ and 3×4 layouts. He had only examined its properties in the less complicated 2×2 , and even there, he only modeled very small main and interaction effects. The problems with the rank transform get much worse in a hurry.

JMASM: Nevertheless, I thought you were slighted, because his response wasn't about your work, but about you. Anyway, he was safe in saying that the bad results on the rank transform you were reporting were wrong.

RCB: Yes, for some reason people wanted to ignore the poor properties of the rank transform.

These experiences led to something that changed my perspective. I had read articles where people had gotten into confrontations, and that they were using coded words for "stupid." I went down that road myself, and used harsher language raised to the third power.

But, one day, I was sitting in my office when I was on the faculty at Johns Hopkins, and the phone rang. It was Boneau, who I had raked over the coals more than a decade prior. He had just come from his retirement dinner, or something of that nature. He said, "Did I really do such poor work?"

Before I got that call, I had already come to the conclusion that much of the problem with his work was he was limited to late 1950s – early 1960s computer equipment (IBM 650), and that is why he examined such small samples (e.g., $n = 5$), and other such limited study parameters. It took $\frac{1}{2}$ hour to generate a thousand random samples. It was a major undertaking for him to do the study he did.

So, when he asked, “Did I really do such poor work?” I felt like I had a stake thrust into my heart. I never felt so bad about anything in my life. There was a real voice out there that I had caused considerable pain. Until that point, all I had done was an academic exercise; from then on I realized that there are real people behind published research.

I told him that I would be forever grateful and in his debt if he could understand that the style with which I had been writing was attributable to the exuberance of my youth. He seemed to indicate that I was forgiven. I swore at that time, that I would never write another article with harsh language. And I didn’t, for at least six months!

JMASM: However, in deference to you, this is a role of an Assistant Professor. The role is primarily oversight and critique. While full Professors are professing and philosophizing, someone has to do the grunt work and check the details. That is one way Assistant Professors make their mark. This was you in the role of Assistant Professor back then. Perhaps, you might have written in a kindler, gentler fashion.

RCB: Agreed. There is no joy or anything to gain in putting others down. Yes, we were providing an important service in keeping an eye on the reporting of bad statistical work.

I was excited about results that I knew no one else knew. That is what gave me satisfaction. I recall in the defense of my dissertation, we had to have an outside person as the moderator. I got the best statistician to serve, because I was confident of my results, and I wanted to be put through the flames, knowing that if I could survive, my work would be correct.

JMASM: Do you recommend doctoral students follow your example and put themselves through the same thing?

RCB: No! Anyway, I thought the role of nonparametric statistics was an important one. I had some insights into the problem, and I had results that I knew no one else had.

My initial interest was on nonparametrics, the rank transform, and later multivariate permutation and step-down comparison tests. After the Boneau incident, however, I realized that a person could spend a career critiquing bad advice in statistics textbooks, or in statistics journals. I had gotten to the point where I could spot it easily. In fact, you and I published an article in *Biometrics* in 1993, where we had spotted such a problem.

JMASM: I share your concern (and some of your skill) in spotting flaws in published research – and why not? After all, I was your student. There is a related question, and perhaps you’ve given it some thought. The literature you have been referring to is important. People turn to the peer reviewed journals to find solutions, to solve the problems of our society. Do you value the literature in helping to solve the woes of humanity? You are retiring from a College of Public Health at the University of South Florida, where issues are studied because lives are at stake. And, even if the lives are not at stake, certainly the quality of life is at stake. Along with many of our colleagues, we could pick apart (not for the fun of it, even though we might enjoy it) the validity of study findings in a hurry. Should we, then, turn to the literature to help solve our problems?

RCB: Yes, but first there needs to be a lot more replication of research before the literature can be considered useful. Doctoral students come along and ask if a certain topic might be viable, and get it turned down because it has already been done. Yet, the study has never been replicated. They should not be discouraged.

JMASM: Isn’t that the primary role of a Master’s thesis?

RCB: I don't see a problem with a major thrust of a doctoral dissertation being a replication of an important study.

I used to tell my students when they took their statistics courses that it would enable them to read the applied literature with a more discerning eye. [Ending my career in a College of Public Health], I pointed out to them that frequently in the medical literature, it will be reported that a statistic was computed, and $p < .001$. But, nowhere in the article is sufficient information revealed to judge what was done, much less if it was done correctly.

For example, I'm aware of the background of a specific article on an aspect of diabetes published in the literature about twelve years ago. The author had learned how to use SAS. He would flip through the user's manual and try and find a statistic with a data set that looked like his. He had a repeated measures design, but didn't recognize it as such. I was first amazed, and then disappointed, that the article was accepted and published.

However, all is not lost. I've been consulting with Roy Beck, a Professor of Ophthalmology and Epidemiology at USF, for about ten years now. The quality of his work in the Optic Neuritis Treatment Trial is pristine. The randomizations are conducted properly, the researchers are masked (I used to say "blinded", but I don't anymore), and the quality of the research methodology is so high, when they publish work – it is valid. I believe there are certain other groups where the scholarship is superior. But, unfortunately, most of what gets published is junk.

JMASM: Can you speak about the Monte Carlo?

RCB: We have both worked primarily using Monte Carlo methods. One thing that has concerned me was that mathematicians and mathematical statisticians so look down on it that it is difficult to get work published using the methodology. I recall you tried to publish an article where the reviewer remarked that anyone with a computer on their desk could have come up with the statistic. Yes, anyone could have come up with it if they had the insight you had, but none did before you. Unfortunately,

important results were not considered publishable in that journal because of the methodology to get the results. The reviewer focused on the method you used. But, there was no closed form mathematical expression to solve the problem. It could have only been handled with Monte Carlo and related methods.

Similarly, I had a manuscript I sent to a certain prestigious journal that the Editor refused to send out for review. The reason was Monte Carlo methods had been used. The point of the manuscript was to show that a procedure previously published in that journal wasn't that bad, but here was a superior technique. The results were obtained via Monte Carlo methods, and the Editor couldn't get past that.

JMASM: A mathematician colleague of mine once said that he finds little value in Monte Carlo methods, other than it was a notable mathematician – von Neumann in 1949 – who coined the phrase in taking a procedure previously conducted by hand and successfully applying it to machines. His rationale: Suppose I wanted to determine the value of a certain function, and did so using Monte Carlo methods. I might run 1,000 repetitions and get a certain value. But, I could then run 1,001 repetitions, and presumably get a better estimate. Or better yet, I could run one million iterations.

I countered that Newton–Raphson, Cauchy, and Riemann are also estimation or approximation procedures. His argument seemed to be that an estimate obtained from the labor of the human mind is legitimate, but from a machine is not.

RCB: Monte Carlo results will never produce *the* answer to a problem. However, if an estimate is acceptable, I don't see the difference between the Newton–Raphson and the Monte Carlo result.

Recently, I was building a table of critical values for a new statistic I've developed. Each critical value was being obtained via permutation methods. I needed to produce over 146 trillion permutations to obtain each value. I realized I would never be able to complete the table this way. So, instead of getting all possible permutations, I took a million random permutations to produce an estimate. I checked

the approximate randomization results with several fully articulated critical values, and I had accuracy to more decimal places than I was reporting in the table.

JMASM: So why does Eugene Edgington say in a number of places in his 1980 book *Randomization Tests* that if one conducts an approximate randomization procedure, don't report this technical detail, so as not to confuse the reader? Why was his advice to hide this?

RCB: Throughout the ages, when new things were discovered, various disguises were used until the public learned to accept them. For some reason, especially in mathematics, there is the tendency to get hung up on the method, rather than the answer.

But, the practical value in using this method is obvious to anyone who, for example, needs to build a table of critical values. The results are correct, and they work.

I recall telling you many years ago, that if it could be shown to work reliably and produce valid results, I would gladly give up Monte Carlo methods in favor of waving gourds and feathers over a pile of goat guts – although I suppose I would have to draw the line at doing it while nude.

I read a book recently about the Indian mathematician Srinivasa Ramanujan, who was self-taught. Later, when he came to England with the assistance of G. H. Hardy to study number theory, he said, to paraphrase, "I hope and pray that no one finds a practical application for my work, because it would belittle it." He felt a practical application would detract from the beauty and elegance of what he had accomplished in mathematics. I suppose many mathematicians have the same fear, for mathematics should be viewed through the artistic and philosophical lenses.

That is fine for the mathematician. But, the line is crossed, for example, in civil engineering. The bridge stands up with mathematical models that are only simulated (and perhaps crudely at that), even if not elegantly derived. Indeed, perhaps the models can never be properly solved mathematically, but the bridge is useful and is still needed.

Another danger is when we take our values and insights from our discipline and try to carry it over to another discipline in order to criticize it. I had a student, many years ago (before the advent of personal computers), who I was trying to teach Fortran. I said, " $A = A + B$." In computer language, this statement simply means that the value in the register representing A is to be incremented by the value held in register B. The student, who was working on a Master's degree in mathematics, said, "A and B must both be zero".

She became irate when I wouldn't accede to her point. I was using symbols that she recognized, but not in the same fashion that she was accustomed to seeing them being used. I learned then that one doesn't casually or easily take the symbols and rules of one discipline and apply them in a critical fashion to another discipline.

JMASM: Perhaps with time, Monte Carlo work will become more acceptable. I noticed that Monte Carlo work in the past would appear in the final section of a journal article, only to buttress the primary results. But, of late, I've noticed the main findings are obtained via Monte Carlo methods, and the latter section contains squiggles in support.

RCB: Younger statisticians have more abilities and faith in Monte Carlo. Previously, a lot of reliance was placed on asymptotic theory, and the question of how that worked wasn't investigated too closely, except to say, it is "asymptotically chi-squared" or "asymptotically normal". Today, researchers are finding results based on small samples Monte Carlo studies, and when large samples are impractical, such as in permutation work, they rely on asymptotics to show the results should hold for larger samples.

JMASM: As time has passed, I've noticed that the algorithm is usually more important than the code. There are a lot of books available showing important Monte Carlo techniques, but the compiler for the language used hasn't been available, or updated, in decades.

In my opinion, the best platform for Monte Carlo work is still Fortran, even though many consider it a dead language. It executes

powers of ten faster (if the program written in another language or package will even execute) than code written in higher level programming languages such as S-plus, R, SAS IML; or statistical packages, such as SAS, SPSS, and Minitab. Was it the development of algorithms to simulate reality, the Monte Carlo method, or the Fortran that you found fascinating?

RCB: It was the whole ball of wax. If I just need a result, I may use an inefficient algorithm if it happens to be quicker to code. Other times, I might get caught up in making the code look pretty or elegant. I took to simulations with Monte Carlo because I never had an electric train as a child.

JMASM: You were never much involved in the social aspects of the American Statistical Association, the American Educational Research Association, and other professional organizations. What positive or adverse effect did that have on your career?

RCB: I was never ambitious. Learning what I discovered with these Monte Carlo studies was the reward for me. The big thrill was demonstrating that nonparametric tests can be more powerful than classical procedures. It was nice when other people recognized this, and found my work worthy of being published. I was most excited by the discovery of new knowledge.

JMASM: Do you believe not hob-knobbing in the social settings of the profession prevented you from receiving fellowships, grants, awards, or other types of recognition?

RCB: I really didn't care about those things. It wasn't important to me. In retrospect, though, I probably had much to learn from many people in the profession, and perhaps had I had more contact with them, my career might have gone in other directions.

At critical moments, though, I have been able to connect with established mathematical statisticians. I would have an idea, I would work it out, and I would enlist the assistance of someone who could help me with the details necessary to build a rigorous argument.

JMASM: In the days of Sir Ronald Fisher, E. J. G. Pitman, and Sir Maurice Kendall, apparently the world was not ready for rank-based nonparametric statistics. Frank Wilcoxon said, to paraphrase, "I've got an approximate, rapid procedure", or a "quick and dirty" procedure, perhaps inadvertently setting the tone for the ensuing battle.

Nonparametric rank tests gained steam with the publication of Sidney Siegal's *Nonparametric Statistics for the Behavioral Sciences* in 1956 (a top 15 cited work on Thompson's Web of Science), and both Donald Fraser's *Nonparametric Methods in Statistics* and Tate and Clelland's *Nonparametric and Shortcut Statistics* in 1957. However, there was an immediate backlash. Gaito, in an article in *Psychological Review* in 1960, said, "It is encouraging to note that some individuals have been reluctant to embrace wholeheartedly the nonparametric technique", and cited an article by Grant in the *Annual Review of Psychology* the year prior, who said, "Some much needed negative thinking has recently appeared on nonparametric techniques". The big debate throughout the 1970s – 1980s, that you participated in, was on the comparative power of rank based nonparametric statistics.

The 1980s brought the robust descriptive statistics' movement into the inferential statistics arena. The 1980s – 1990s, with the advent of inexpensive and powerful personal computers, puts us in the era of practical permutation and exact statistics.

I have colleagues who proclaim that even if there was a time for nonparametric rank tests, that time has passed. So, I ask you, "Was there ever a time, or better, will there ever be a time for nonparametric rank tests?"

RCB: I've seen the argument for permutation tests – we have a PC so why convert to ranks and do a rank based test when the permutation test can be done? This is the problem that we've discussed already, and unfortunately, it seems few people understand this. If you examine Monte Carlo results, it will be learned that permutation tests give virtually the same power as their parametric counterparts. For example, the permutation t test gives almost identical power as the two independent samples t

test. The reason for turning to a rank based nonparametric test, such as the WRS, is because of its power advantages!

Motivated by concerns about robustness of parametric tests, colleagues ask, “Why convert to ranks when it is now possible to do a permutation test on the original scores?” The answer is that I’m not turning to an alternative test because of the *t* test’s robustness; its Type I error rates are adequate under nonnormality. I’m turning to rank based nonparametric tests because of their power advantages.

For example, I’m currently working with visual acuity scores with a skew coefficient of 3. With that level of skew, the WRS will have four to ten times the power over the *t* test and the permutation *t* test. That is the reason for selecting a rank based nonparametric test.

Jim Higgins and I wrote a letter to the *American Statistician* in 2000 in response to someone promoting permutation tests. His point was you don’t have to lose power anymore by converting to ranks because you can now do a permutation test. Our response was to cite our research that showed the opposite – power is gained by converting to ranks. Their reply was there might be a theoretical reason to believe that, but in applied research those considerations don’t apply.

I took some of Roy Beck’s data and replicated a number of the studies we previously published, such as the article you and I published in *Psychological Bulletin* in 1992 using Ted’s [Theodore Micceri] real education and psychology data sets. The same four to ten times the power advantage accrued to the rank based nonparametric test. I started to write a retort to their reply, but I decided it was to no avail. This battle is endless.

JMASM: So there never has been a good time, according to the experts and masters, to do a rank based nonparametric test?

RCB: It never had its time, except perhaps briefly before calculators were invented. It was a quick way to analyze data. If you were working with sixty countries’ Gross National Products, the numbers would be too large to sum and square, but in converting to ranks it became manageable.

One reason why it never had its time was because rank based nonparametric statistics were always presented as a way to control Type I error in the absence of normality.

JMASM: That reminds me of the time you sent me to the library to retrieve Jeffrey Rasmussen’s 1985 article in *Evaluation Review*. He was critical of your work, and set out to refute it.

He constructed a study where he first applied a data transformation designed to maximize homoscedasticity and stabilize within-group normality before conducting the *t* test, but he failed to do any type of data cleansing before conducting the Wilcoxon Rank Sum test. He concluded your work showed phantom power advantages.

It was not a fair comparison. The Type I error properties of the WRS are invariant under departures from population normality, but its power properties are not! He should have also conducted some equitable form of data cleansing prior to conducting the WRS. As you say, people view the role of nonparametrics only as a method to control Type I error, forgetting about power considerations. I was so upset about this, by the time I returned to your office with the paper, my knuckles were white from grasping the article so tightly.

RCB: That’s why to this day I root against the Purdue Boilermakers football team.

For a time, you and I, and others working in this area, had an impact as far as what textbook authors wrote on rank based nonparametric tests and on the rank transform. But, as time passes, authors seem to be drifting back. I suppose we must leave it to the next generation to rediscover the power of rank based nonparametric tests.

Teaching

JMASM: In terms of classroom teaching, your *c.v.* indicates you’ve won many awards, and some of them multiple times. How did you make the transition from scholarship to teaching? Are there students at the end of your words, or are you directing your lectures to the discipline?

RCB: My focus is on the students. I’ve developed certain ideas regarding teaching.

They are based on my experiences as a college student. I would often listen to a professor and wonder, "Isn't there a clearer way to say this?"

There were certain areas of statistics that I found fuzzy and difficult to grasp. I wanted to find a way to transfer information in a better way. Richard Taylor – a student in one of my classes and now a faculty member – and I spend a lot of time on this issue in developing our biostatistics textbook that is to be published by Prentice-Hall. We take the material apart, piece by piece, to make sure each concept has a logical flow, and is understandable.

In my experience, a lot of what went on in the classroom, on the part of the professor, was done for reasons other than promoting learning. There is a strutting factor, to show you my importance, how much more I know than you, or how powerful I am that I can ruin your career with a low grade. If the focus is on the learning process, instead of all that, it changes the classroom dynamic and environment.

I was once criticized by a faculty member (who later was turned down for tenure). He bragged about how much more difficult the students found his course. He would point out that I was teaching statistics courses, which everyone knows should be more difficult than subject matter courses, and yet my students found that, after doing the required reading and homework, the course was rather easy. His courses, however, were received as being very difficult, and he took great pleasure in that. I explained to him that was what I spent most of my time doing in developing my lectures – to find ways to make the material understandable and obtainable, not difficult and incomprehensible.

Daniel Purdom [Professor of Educational Leadership and Higher Education at USF in the 1970s through 1990s] used to say no learning takes place without pain. I think about that statement all the time. For some students, and perhaps in some disciplines, that may be the case. But in my teaching experience it is not true, nor is it necessary. A lot may be learned in statistics without pain.

JMASM: The reactions to "statistics", when responding to people who ask what subject I

teach, are "that was my worst subject", "that was my hardest subject", or "I hated it."

RCB: Or, "that was my worst teacher." Unfortunately, we are overrun with bad teachers of statistics. When I first started teaching I used to say, "One of the things we desperately need is more dead statisticians."

I remember a certain statistics course I took. It was taught by the meanest, nastiest person I ever met. He was very full of himself, and the main point of his lectures was to demonstrate how smart he was and how dumb we were. He wrote a formula on the chalk board. He used "N-1", explaining that was the way to unbiased the estimate. He pointed to me and asked, "Cliff, what is N-1?"

I knew how to determine "4-1" or "2-1". But, I didn't know how subtraction was done when mixing letters and numbers. It didn't make sense to me. Most professors will go on to the next person. But, he wouldn't let it go, and he continued to grill me. "Come now, Cliff," he barked. "I just went over this. What is N-1?" I started sweating and was very nervous. Finally, a revelation came to me: "*M*" I yelled! He suggested I drop the class, which of course I did. If I was going to be a teacher, I knew then what kind of professor I didn't want to be.

It is vital to know when a little bit of pressure may be applied, and when a little bit of pressure must be released. I try to "take the temperature" of the class. I can tell when things start to get tense, and that is when I put aside the prepared lecture and launch into a story to make the same point. I let my students see my fingers wiggle when I'm adding or subtracting. It changes the atmosphere from drudgery to pleasure.

Administration

JMASM: Why did you accept an administrative post?

RCB: I became Chair of the Department for two reasons: (1) it was experiencing some difficulties and needed help, and (2) the Associate Dean asked me to do it as a personal favor. I hated every minute of it, as I knew I would, and I would never do it again. I didn't accept it for only altruistic reasons; I was offered

a Sabbatical leave the year following my tenure as Chair of the Department.

JMASM: Would you recommend someone who has recently acquired tenure to aspire to the Chair's position?

RCB: Only if that person likes working with budgets, or having faculty members in your office wringing their hands about all the things that afflict faculty members. I soon found the key to success as Chair, and I believe I had the reputation of being a successful administrator.

When people came to me for money for travel, equipment, etc., I said yes. Then, when the business manager would want to set up an appointment with me, I would find ways to put it off. I knew the date when my tenure as Chair would be up, and I was going immediately to the Sabbatical, leaving the finances in the capable hands of the next Chair. As a result, even today people talk about how remarkable it was that I was able to fulfill their every request.

Advice to Junior Faculty

JMASM: What is your advice to the new assistant professor?

RCB: Get out now while you can! I came into this business exactly at the right moment. What I enjoyed most about being a professor was my degrees of freedom.

I could chase the Wilcoxon test, the rank transform test, and the permutation step-down test. Back then, if I needed to do a Monte Carlo, I only needed the capabilities of a Tandy Radio Shack Model 80 personal computer to do the work.

However, today, in many universities, it is almost not possible to get tenure without bringing in federal dollars. And, I mean specifically federal grants, because state and local money doesn't provide sufficient indirect or overhead. The professor has to tailor the research agenda to meet the funding initiatives.

Active pursuit of a half million dollar federal grant, or more, is paramount in the life of

junior faculty. Very little consideration is given to what happens in the classroom, and hardly anyone cares about the quality of research if the number of publications is sufficiently high. Unfunded research, even if it wins the Nobel Prize, does not bring in dollars to the university.

To be fair, universities with this orientation make this clear to new assistant professors, and I imagine in places where they don't, the faculty figure it out for themselves. In the contact I have with some faculty struggling with this, I see that they are not pursuing what they really love; about what motivated their careers into academe, but rather, the pursuit of money for the university or for their laboratories.

Read the 1982 book *Betrayers of the Truth* by William Broad and Nicholas Wade, and the more recent *The Baltimore Case* by Daniel Kevles. They opened my eyes about funded research, although I've suspected that type of thing for many years.

There are universities officials who proclaim an ambition to create grant mills, a production line to capture federal dollars. Much of the fraud in research comes from this mindset. Perhaps the Principal Investigator didn't commit the fraud but was under pressure, and put so much pressure on junior faculty, fellows, postdocs, and graduate students that they committed the fraud. I've concluded that the quality of research decreases when the primary purpose for conducting it is to obtain research dollars instead of answering a research question.

If a faculty member is interested in pursuing a topic, and seeks funding for it – that's great. However, a study conducted primarily for the sake of providing the university it's indirect will be problematic. In order to get the grant renewed there are certain outcome expectations. It obviates the ability to do large scale, high quality research when the driving force is money instead of truth and new knowledge.

If an assistant professor asked what should be concentrated on to get promoted and tenure, I would respond to go after grant money. What I had for thirty years, the pursuit of new knowledge for the sake of new knowledge, in many universities, no longer exists.

Journal Articles v Textbooks

JMASM: The impression I got in my early years from you was that the success of a research agenda should be judged by peer-reviewed publications. I got the impression that people who write textbooks do so because they can no longer conduct research worthy of publication in peer-reviewed journals and periodicals. Twenty years later, you've given me the "William Mendenhall" maneuver: the supposition that there will never be a statistic or procedure named after me, so why not write a textbook, and indeed, turn out a dozen flavors of the same textbook for a dozen different markets. I see that your first project in retirement is the completion of your biostatistics book.

RCB: Your perception is correct. In my early years as a professor, I only wanted to generate new knowledge. I wasn't interested in setting down the same material that everyone else knows. I was in hot pursuit of questions I wanted to know the answers, such as how does the power function of this procedure compare with a competitor under realistic, applied conditions?

A few years ago, Richard Taylor and I started having conversations about what takes place in the classroom. It led to the desire to write a book that followed along the lines of my quest in research: write a textbook that uses new, and hopefully better, methods to communicate statistical knowledge. I would have never pursued writing *yet another* statistics book, but I thought I had enough ideas on improved pedagogy, materials, and methods to write a worthy new textbook. This, then, became a challenge to me. Therefore, I viewed writing this type of textbook as an extension of my initial reasons for being a professor.

The biostatistics book is turning out to be a different type of book, and at this time I don't really know how it will be received. When I sent the manuscript to a prospective publisher, the reviewers said it was terrible and should not be published. I've had enough papers rejected over the years that my first thought was perhaps it was not the best outlet, as opposed to being crushed that the text was worthless. And, upon closer inspection, I noted the reviewers said this textbook failed to use the standard approach in

presenting this concept, failed to promote the standard analysis in that context, failed to use the standard examples, assignments, and so forth. I was gratified about those comments, because that was what I had set out to do; write a textbook that didn't follow the standard approach, but represented new knowledge and new methods. I look forward to it coming out soon as a Prentice-Hall title.

Retirement

JMASM: John W. Tukey purportedly published more after he retired than prior to his retirement. What's in store for R. Clifford Blair?

RCB: There are a number of projects I would like to pursue. I recently presented a poster at the Society for Clinical Trials. They are concerned with, for example, the impact of adding ten patients to a trial. The MRI and doctor's fees can amount to \$20,000 per patient. I showed, keeping the power level constant, what happens to the required sample size in terms of how much smaller samples need to be when using nonparametric rank tests. There was considerable excitement; people were running around hollering and waving their arms to come view the poster. This made me think about going back and re-fighting some of the old battles on nonparametrics.

Or, redo the old studies, which were conducted in the context of hypothesis tests, but conduct them again in the confidence interval paradigm. Of course, the results – in terms of the length of the interval being smaller for the nonparametric rank test as compared with the parametric counterpart – will be the same. For some reason, in turning from hypothesis testing to confidence intervals, all that you and I, and our like-minded colleagues, have accomplished is lost, and needs to be demonstrated once again.

I would like to return to a study I started with Dennis Boos at North Carolina State University some years ago. It pertained to permutation multiple comparisons. I believe there are a couple of other papers still left in me, and perhaps a textbook to replace the Pedhazur linear models book.

We've just moved to a small, rural community in Butler, Tennessee. It's an isolated place where everyone takes care of one another. It is a modest, quiet place. If you have a desire to see a traffic light you will have to go out of your way to find one.

They say mountain people don't warm to outsiders, but they've welcomed us with open arms. Life there is about family reunions, blue grass music, picnics, and school activities.

My wife, Cathy, who was a Teacher of the Year in Florida, teaches in a small school with 126 students. She has eight students in her classroom. It's the type of school where classes are let out early because the bus driver has a dental appointment, and the Principal raises money to assist in building indoor plumbing for the poorer families.

We have several pieces of property there, including a small cabin on a river. I will be happy there.

R. Clifford Blair
January, 2005

Table 1. Publications of R. Clifford Blair as of January, 2005.

-
- Book
1. Blair, R. C., & Taylor, R. A. (2005, in press). *Biostatistics for the health sciences*. NY: Prentice Hall.
- Book Chapters
2. Blair, R. C., & Karniski, W. (1994). Distribution-free statistical analyses of surface and volumetric maps. In R. W. Thatcher, E. R. John, & M. Huerta (Eds.), *Functional neuroimaging: Technical foundations*, (pp. 19-28). San Diego, CA: Academic Press, Inc.
 3. Holmes, A. P., Blair, R. C., Watson, J. D. G., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. In R. Myers, V. Cunningham, D. Bailey, & T. Jones (Eds.), *Quantification of brain function using PET*. San Diego, CA: Academic Press, Inc.
- Peer Reviewed Publications
4. Durso, M. D., & Blair, R. C. (1978). A comparative study of achievement of 393 high school students enrolled in first year Century 21 and Gregg shorthand. *NABTE Review*, 5, 96-103.
 5. Blair, R. C., & Higgins, J. J. (1978). Tests of hypotheses for unbalanced factorial designs under various regression/coding method combinations. *Educational and Psychological Measurement*, 38, 621-631.
 6. Blair, R. C., & Higgins, J. J. (1978). Comments on 'Contrast coding in least squares regression analysis.' *American Educational Research Journal*, 15, 149-151.
 7. Blair, R. C. (1978). I've been testing some statistical hypotheses...can you guess what they are? *Journal of Educational Research*, 72, 116-118.
 8. Blair, R. C. (1979). A reaction to 'The right-test but the wrong occasion.' *NABTE Review*, 6, 63-64.
 9. Blair, R. C. (1979). A comment on the relative usefulness of the two independent means t-test and Wilcoxon's rank-sum test for the analysis of educational data. *Florida Journal of Educational Research*, 21, 97-110.
 10. Hurley, B. F., Glasser, S. P., Phelps, C. P., Anderson, D., Blair, R. C., & Riggs, C. E. (1980). Cardiovascular and sympathetic reactions to in-flight emergency responses among base fire fighters. *Aviation Space and Environmental Medicine*, 58, 788-792.
 11. Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, 5, 309-335.
 12. Blair, R. C., & Phillippy, S. W. (1980). Another look at the robustness of the two independent means t-test to population non-normality. *Florida Journal of Educational Research*, 22, 1-13.
 13. Blair, R. C., & Higgins, J. J. (1980). The power of the t and Wilcoxon statistics: A comparison. *Evaluation Review*, 4, 645-650.
 14. Blair, R. C., Higgins, J. J., & Smitley, W. D. S. (1980). On the relative power of the U and t-tests. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.
 15. Blair, R. C. (1981). A reaction to 'Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance.' *Review of Educational Research*, 51, 499-507.
 16. Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t-test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology*, 34, 124-128.

17. Blair, R. C. (1982). An examination of some commonly held attitudes regarding the nature and usefulness of nonparametric tests. *Florida Journal of Educational Research*, 24, 29-35.
18. Blair, R. C., & Lawson, S. B. (1982). Another look at the robustness of the product-moment correlation coefficient to population non-normality. *Florida Journal of Educational Research*, 24, 11-15.
19. Blair, R. C., Higgins, J. J., Topping, M. E. H., & Mortimer, A. L. (1983). An investigation of the robustness of the t-test to unit of analysis violations. *Educational and Psychological Measurement*, 43, 69-80.
20. Blair, R. C., & Higgins, J. J. (1984). Some comments on 'The unit of analysis: Group means vs. individual observations.' *Florida Journal of Educational Research*, 26, 5-20.
21. Phillips, E. R., & Blair, R. C. (1984). Investigating variables related to sex differences in students' abilities to solve word problems in algebra. *Focus on Learning Problems in Mathematics*, 5, 47-61.
22. Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon's signed ranks statistic. *Journal of Education Statistics*, 10, 368-383.
23. Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t-tests to that of Wilcoxon's signed rank test under various population shapes. *Psychological Bulletin*, 97, 119-128.
24. Blair, R. C., & Higgins, J. J. (1986). Comment on 'Statistical power with group mean as the unit of analysis.' *Journal of Educational Statistics*, 11, 161-169.
25. Karniski, W., Blair, R. C., & Vitucci, J. S. (1987). The use of the growth index and age-correction to refute the concept of 'catch-up' growth in premature infants. *American Journal of Diseases of Children*, 141, 520-526.
26. Blair, R. C., Sawilowsky, S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics*, 16, 1133-1144.
27. Karniski, W., & Blair, R. C. (1988). Topographical and temporal stability of the P300. *Electroencephalography and Clinical Neurophysiology*, 72, 373-383.
28. Karniski, W., & Blair, R. C. (1988). Stability of the global field power. *Psychophysiology*, 25, 460-460.
29. Sawilowsky, S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, 14, Fall.
30. Kanarek, K. S., Kuznicki, M. B., & Blair, R. C. (1990). Infusion of total parenteral nutrition via the umbilical artery. *Clinical Research*, 38(4), A996-A996.
31. Kanarek, K., Kuznicki, M. B., & Blair, R. C. (1991). Infusion of total parenteral nutrition via the umbilical artery. *Journal of Parenteral and Enteral Nutrition*, 15, 71-74.
32. Blair, R. C. (1991). New critical values for the generalized t and generalized rank-sum procedures. *Communications in Statistics — Simulation and Computation*, 20, 981-994.
33. Roetzheim, R. G., Brownlee, H. J., Pamies, R. J., Vandurme, D. J., Herold, A. H., Woodard, L., & Blair, R. C. (1992). Reverse targeting in a media-promoted breast cancer screening project. *Cancer*, 70, 1152-1158.
34. Kanarek, K. S., Williams, P. R., & Blair, R. C. (1992). Concurrent administration of albumin with total parenteral nutrition in sick newborn infants. *Journal of Parenteral and Enteral Nutrition*, 16, 49-53.
35. Karniski, W., Wyble, L., Lease, L., & Blair, R. C. (1992). The late somatosensory evoked potential in premature and term infants. II. Topography and latency development. *Electroencephalography and Clinical Neurophysiology*, 84, 44-54.
36. 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, 352-360.
37. Blair, R. C., & Morel, J. G. (1992). On the use of generalized t- and generalized rank-sum statistics in medical research. *Statistics in Medicine*, 11, 491-501.

38. Blair, R. C., & Morel, J. G. (1992). On the use of generalized t- and generalized rank-sum statistics in medical research. Rejoinder. *Statistics in Medicine*, 11, 507-509.
39. Blair, R. C., & Thompson, G. L. (1992). A distribution-free rank-like test for scale with unequal population locations. *Communications in Statistics*, 21, 353-371.
40. Kromrey, J. D., Chason, W. M., & Blair, R. C. (1992). Permute—a SAS algorithm for permutation testing. *Applied Psychological Measurement*, 16, 64-64.
41. Roetzheim, R. G., Van Durme, D. J., Brownlee, H. J., Herold, A. H., Woodard, L. J., & Blair, R. C. (1993). Barriers to screening among participants of a media-promoted breast cancer screening project. *Cancer Detection and Prevention*, 17(3), 367-377.
42. Powers, P. S., Boyd, F., Blair, R. C., Stevens, B., & Rosemurgy, A. (1993). Psychiatric Issues in Bariatric Surgery. *Obesity Surgery*, 2, 315-325.
43. Higgins, J. J., Blair, R. C., & Tashtoush, S. (1993). The aligned rank transform procedure. *Proceedings of the 1990 Kansas State University Conference on Applied Statistics in Agriculture* (pp. 185-196). Manhattan, Kansas: Kansas State University.
44. Herold, A. H., Young, D. L., Wightman, J. K., Anderson, M. J., Roetzheim, R. G., & Blair, R. C. (1993). Comparison of the cytology brush with the Dacron swab for detecting Chlamydia Trachomatis by enzyme immunoassay in female university students. *Journal of American College Health*, 41(5), 213-216.
45. Herold, A. H., Roetzheim, R. G., Young, D. L., Anderson, M. J., Blair, R. C., & Mockler, B. (1993). Effect of undocumented and excluded vaccinations on measles immunity in a university population. *Journal of the Florida Medical Association*, 80(3), 173-177.
46. Blair, R. C., & Sawilowsky, S. (1993). A note on the operating characteristics of the modified F test. *Biometrics*, 49, 935-939.
47. Blair, R. C., & Sawilowsky, S. (1993). Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls. *Statistics in Medicine*, 12, 2223-2243.
48. Blair, R. C., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, 30, 518-524.
49. Karniski, W., Blair, R. C., & Snider A. D. (1994). An exact statistical method for comparing topographical maps with any number of subjects and electrodes. *Brain Topography*, 6(3), 203-10.
50. Dryjski, M., Driscoll, J., Blair, R. C., McGurrin, M., Dagher, F. J., Ceraolo, M. J., O'Donnell, & Blackshear, W. M. (1994). The small abdominal aortic aneurysm: The eternal dilemma. *The Journal of Cardiovascular Surgery*, 35, 95-100.
51. Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research*, 29, 141-164.
52. Blair, R. C., & Beck, R. W. (1994). Analysis of visual-field thresholds via permutation based step-wise multiple comparison procedures. *Investigative Ophthalmology & Visual Science*, 35, 1509-1509.
53. Sawilowsky, S., Kelley, L., Blair, R. C., & Markman, B. S. (1995). Meta-analysis and the Solomon four-group design. *Journal of Experimental Education*, 62, 361-76.
54. Blair, R. C., & Hochberg, Y. (1995). Improved Bonferroni procedures for testing overall and pairwise homogeneity hypotheses. *Journal of Statistical Computation and Simulation*, 51, 281-89.
55. Algina, J., Blair, R. C., & Coombs, W. T. (1995). A maximum test for scale: Type I error rates and power. *Journal of Educational and Behavioral Statistics*, 20(1), 27-39.
56. Bright, P. E., Arnett, D. K., Blair, R. C., & Bayona, M. (1996). Gender and ethnic differences in survival in a cohort of HIV positive clients. *Ethnicity and Health*, 1(1), 77-85.
57. Blair, R. C., Troendle, J. F., & Beck, R. W. (1996). Control of familywise errors in multiple endpoint assessments via stepwise permutation tests. *Statistics in Medicine*, 15, 1107-1121.

58. Beck, R. W., Moke, P., Blair, R. C., & Nissenbaum, R. (1996). Uveitis associated with topical β -blockers. *Archives of Ophthalmology*, 114, 1181-1182.
59. Holmes, A. P., Blair, R. C., Watson, J. D. G., & Ford, I. (1996). Nonparametric analysis of statistical images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, 16, 7-22.
60. Optic Neuritis Study Group (1997). Visual function five years after optic neuritis: Experience of the Optic Neuritis Treatment Trial. *Archives of Ophthalmology*, 115, 1545-52.
61. Optic Neuritis Study Group (1997). The five year risk for multiple sclerosis after optic neuritis: Experience of the Optic Neuritis Treatment Trial. *Neurology*, 49, 1404-1413.
62. Herpetic Eye Disease Study Group (1997). A controlled trial of oral acyclovir for the prevention of stromal keratitis or iritis in patients with herpes simplex virus epithelial keratitis. *Archives of Ophthalmology*, 115, 703-12.
63. Troendle, J. F., Blair, R. C., Rumsey, D., & Moke, P. (1997). Parametric and non-parametric tests for the overall comparison of several treatments to a control when treatment is expected to increase variability. *Statistics in Medicine*, 16(23), 2729-2740.
64. Herpetic Eye Disease Study Group (1998). Acyclovir for the prevention of recurrent herpes simplex virus eye disease. *New England Journal of Medicine*, 339, 300-6.
65. Cole, S. R., & Blair, R. C. (1999). Overlapping confidence limits (letter; comment). *Journal of the American Academy of Dermatology*, 41, 1051-1052.
66. Higgins, J. J., Blair, R. C. (2000). Power of rank-based nonparametric tests (letter; comment). *The American Statistician*, 54, 86.
67. Herpetic Eye Disease Study Group (2000). Psychological stress and other potential triggers for recurrences of herpes simplex virus eye infections. *Archives of Ophthalmology*.
68. Long, D. T., Beck, R. W., Moke, P. S., Blair, R. C. Kip, K. E., Gal, R. L. Katz, B. J. & The Optic Neuritis Study Group (2001). The SKILL card test in optic neuritis: Experience of the Optic Neuritis Treatment Trial. *Journal of Neuro-Ophthalmology*, 21, 124-131.
69. Pediatric Eye Disease Investigator Group (2001). The amblyopia treatment study visual acuity testing protocol. *Archives of Ophthalmology*, 119, 1345-1353.
70. Dubnika, S. R., Blair, R. C. & Hettmansperger, T. P. (2002). Rank-based inferences for mixed paired and two sample designs. *Journal of Modern Applied Statistical Methods*, 1, 32-41.
71. Blair, R. C. & Cole, S. R. (2002). Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods*, 1, 139-142.
72. Blair, R. C. (2002). A distribution-free maximum test of location for two independent samples. *Journal of Modern Applied Statistical Methods*, 1, 13-18.
73. Beck R. W., Moke, P. S., Turpin, A. H., Ferris, F. L., SanGiovanni, J. P., Johnson, C. A., Birch, E. E., Chandler, D. L., Cox, T. A., Blair, R. C., & Kraker, R. T. (2003). A computerized method of visual acuity testing: Adaptation of the early treatment of diabetic retinopathy study testing protocol. *American Journal of Ophthalmology*, 135, 194-205.

Non-Peer Reviewed Publication

74. Blair, R. C. (1984). A review of Tandy-Graph and the Multi-Pen Plotter. *Advanced Computing*, 2, 20-22.

Figure 1. Academic Genealogy of R. Clifford Blair via Higgins, Loève, and Hadamard.

Figure 1. Academic Genealogy of R. Clifford Blair via Higgins, Loève, and Hadamard.

Erhard Weigel 1625 – 1699	Gottfried Wilhelm von Leibniz 1646 – 1716	Jacob Bernoulli 1654 – 1705

Johann Bernoulli 1667 – 1748	Leonhard Euler 1707 – 1783	Joseph Lagrange 1736 – 1813	Pierre Laplace 1749 – 1827

Siméon Poisson 1781 – 1840	Gaspard Prony 1755 – 1839	Joseph Liouville 1809 – 1882	Louis Paul Émile Richard 1795 – 1849

Charles Hermite 1822 – 1901	Jules Tannery 1848 – 1910	C. Émile Picard 1856 – 1941	Vito Volterra 1860 – 1940

Paul Pierre Lévy 1886 – 1971	Michel Loève 1907 – 1979	Lucien Le Cam 1924 – 2000	Julius Rubin Blum 1922 – 1982

David Lee Hanson	James Higgins	Bruce W. Hall	R. Clifford Blair

Figure 2. Academic Genealogy of R. Clifford Blair via Higgins, Loève, and Volterra.

	<p>Pietro Paoli 1759 - 1839</p>	<p>Vincenzo Brunacci 1768 - 1818</p>
<p>Ottaviano F. Mossotti 1791 - 1863</p>	<p>Enrico Betti 1823 - 1892</p>	<p>Vito Volterra 1860 - 1940</p>	<p>Jacques S. Hadamard 1865 - 1963</p>
<p>Paul Pierre Lévy 1886 - 1971</p>	<p>Michel Loève 1907 - 1979</p>	<p>Lucien Le Cam 1924 - 2000</p>	<p>Julius Rubin Blum 1922 - 1982</p>
<p>David Lee Hanson</p>	<p>James J. Higgins</p>	<p>Bruce W. Hall</p>	<p>R. Clifford Blair</p>

Figure 3. Academic Genealogy of R. Clifford Blair via Higgins, Le Cam, Zaremba, Darboux, & Poisson.

<p>(For continuation through Poisson, see Figure 1. For continuation through Monge, see Figure 4.)</p>	<p>Siméon Poisson 1781 – 1840</p>	<p>Gaspard Monge 1746 – 1818</p>	<p>Michel Chasles 1793 – 1880</p>
<p>J. Gaston Darboux 1842 – 1917</p>	<p>C. Émile Picard 1856 – 1941</p>	<p>Stanislaw Zaremba 1863 – 1942</p>	<p>Georgy Voronoy 1868 – 1908</p>	<p>Wacław Sierpiński 1882 – 1969</p>
<p>Jerzy Neyman 1894 – 1981</p>	<p>Lucien Le Cam 1924 – 2000</p>	<p>Michel Loève 1907 – 1979</p>	<p>Julius Rubin Blum 1922 – 1982</p>
<p>David Lee Hanson</p>	<p>James J. Higgins</p>	<p>Bruce W. Hall</p>	<p>R. Clifford Blair</p>	

Figure 4. Academic Genealogy of R. Clifford Blair via Higgins, Le Cam, Zaremba, Darboux, and Monge.

Jean d'Alembert 1717 – 1783	Charles Bossut 1732 – 1806	Gaspard Monge 1746 – 1818	Siméon Poisson 1781 – 1840	Michel Chasles 1793 – 1880
Gaston Darboux 1842 – 1917	Émile Picard 1856 – 1941	Stanislaw Zaremba 1863 – 1942	Georgy F. Voronoy 1868 – 1908	Wacław F. Sierpiński 1882 – 1969
Jerzy Neyman 1894 – 1981	Lucien Le Cam 1924 – 2000	Michel Loève 1907 – 1979	Julius Rubin Blum 1922 – 1982
David Lee Hanson	James J. Higgins	Bruce W. Hall	R. Clifford Blair	

Figure 5. Academic Genealogy of R. Clifford Blair via Higgins, Le Cam, and Voronoy.

<p>Joseph von Littrow 1781 – 1840</p>	<p>Nikolai D. Brashman 1796 – 1866</p>	<p>Pafnuty Chebyshev 1821 – 1894</p>
<p>Andrei A. Markov 1856 – 1922</p>	<p>Georgy F. Voronoy 1868 – 1908</p> <p>Stanislaw Zaremba 1863 – 1942</p>	<p>Wacław F. Sierpiński 1882 – 1969</p>
<p>Jerzy Neyman 1894 – 1981</p>	<p>Lucien Le Cam 1924 – 2000</p> <p>Michel Loève 1907 – 1979</p>	<p>Julius Rubin Blum 1922 – 1982</p>
<p>David Lee Hanson</p>	<p>James J. Higgins</p>	<p>Bruce W. Hall</p>
<p>R. Clifford Blair</p>		

Figure 6. Academic Genealogy of R. Clifford Blair via Hall and Stoker's Doctoral Advisor.

Figure 7. Academic Genealogy of R. Clifford Blair via Hall and Stoker's Master's Committee.

Figure 8. Selected Title Pages from the Academic Genealogy of R. Clifford Blair.*

Vincenzo Brunacci
1768 – 1818

Siméon Poisson
1781 – 1840

Ottaviano Mossotti
1791 – 1863

Michel Chasles
1793 – 1880

Figure 8 (con't). Selected title pages.

Joseph von Littrow
1781 – 1840 Joseph Liouville
1809 – 1882 Eugène Catalan
1814 – 1894 Charles Hermite
1822 – 1901

Figure 8 (con't). Selected title pages.

Gaston Darboux
1842 – 1917 Jules Tannery
1848 – 1910 Charles Émile Picard
1856 – 1941 Vito Volterra
1860 – 1940

Figure 8 (con't). Selected title pages.

Jacques Salomon Hadamard
1865— 1963

Wacław F. Sierpiński
1882 – 1969

Hermann H. Remmers
1892 – 19??

E. F. Lindquist
1914 – 1998

Figure 9. Philatelic, Numismatic, and Bank Note Images
from the Direct and Broader Academic Genealogy of R. Clifford Blair*

<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>
<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>
<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>
<p>Gottfried Wilhelm von Leibniz 1646 – 1716</p>	<p>Jacob Bernoulli 1654 – 1705</p>	<p>Leonhard Euler 1707 – 1783</p>
<p>Leonhard Euler 1707 – 1783</p>	<p>Leonhard Euler 1707 – 1783</p>	<p>Leonhard Euler 1707 – 1783</p>

Figure 9 (con't): Philatelic, Numismatic, and Bank Note Images.

<p>Leonhard Euler 1707 – 1783</p>	<p>Leonhard Euler 1707 – 1783</p>	<p>Joseph-Louis Lagrange 1736 – 1813</p>
<p>Gaspard Monge 1746 – 1818</p>	<p>Gaspard Monge 1746 – 1818</p>	<p>Pierre-Simon Laplace 1749 - 1827</p>
<p>Pierre-Simon Laplace 1749 - 1827</p>	<p>André Marie Ampère 1775 - 1836</p>	<p>André Marie Ampère 1775 – 1836</p>
<p>Evariste Galois 1811 – 1832</p>	<p>Urbain J. J. Le Verrier 1811 – 1877</p>	<p>Urbain J. J. Le Verrier 1811 – 1877</p>
<p>Pafnuty Chebyshev 1821 – 1894</p>	<p>Pafnuty Chebyshev 1821 – 1894</p>	<p>Pafnuty Chebyshev 1821 – 1894</p>
<p>Jules Henri Poincaré 1854 - 1912</p>	<p>Stanislaw Zaremba 1863 – 1942</p>	<p>Wacław F. Sierpiński 1882 – 1969</p>

*Scanned from the personal collection of Shlomo S. Sawilowsky and from internet sources (see references below).

Figure 10. R. Clifford Blair, early mentor, former doctoral students, and former graduate assistants.

John H. Neel
Early Mentor

Shlomo S. Sawilowsky
Former Doctoral Student,
Graduate Teaching Assistant, and
Graduate Research Assistant

Theodore Micceri
Former Doctoral Student and
Graduate Teaching Assistant

R. Clifford Blair
Karen N. Perrin, former Graduate Assistant
Richard A. Taylor, former Graduate Assistant

Table 2. Descendents in the academic genealogy of R. Clifford Blair, including doctoral candidates at the dissertation stage, as of January, 2005.

1. R. Clifford Blair. (Ph. D.), A comparison of the power of the two independent means t test to that of the Wilcoxon's Rank-Sum Test for samples of various sizes that have been drawn from a variety of non-normal populations." 131 pp., 1980.

Doctoral Students of R. Clifford Blair

2. Shlomo S. Sawilowsky. (Ph. D.), "Robust and power analysis of the $2 \times 2 \times 2$ ANOVA, rank transform, random normal scores, and expected normal scores transformation tests." 159 pp., 1985.

3. Theodore Micceri. (Ph.D.), "Testing for normality and evaluating the relative robustness of location estimators for empirical distributions derived from achievement tests and psychometric measures." 239 pp, 1987.

Doctoral Students of Shlomo S. Sawilowsky

4. Joyce Washington. (Ed. D.), "Health education and measuring the effects of minority student self-concept as it relates to school performance." 104 pp., 1993.

5. Sharonlyn Morgan-Harrison. (Ph. D.), "Some construct validation evidence for two new measures of self-determination." 89 pp., 1994.

6. Deborah L. Kelley. (Ph. D.), "The comparative power of several nonparametric alternatives to the analysis of variance in a $2 \times 2 \times 2$ layout." 214 pp., 1994.

7. Dennis J. Mullan. (Ph. D.), "An investigation of a residential customer satisfaction model at an electric utility." 102 pp., 1995.

8. Uju P. Eke. (Ph. D.), "A construct validation of a Self-Determination instrument: Using adult substance abuse consumers in residential settings." 79 pp., 1996.

9. Patrick D. Bridge. (Ph. D.), "The comparative power of the independent-samples t-test and Wilcoxon rank sum test in nonnormal distributions of real data sets in education and psychology." 113 pp., 1996.

10. Margaret P. Posch. (Ph. D.), "Comparative properties of nonparametric statistics for analyzing the $2 \times c$ layout for ordinal categorical data." 78 pp., 1996.

11. Thilak Gunasekera. (Ph. D.), "Effects of pretest sensitization associated with cooperative learning strategies on the achievement level of adult mathematics students." 97 pp., 1997.

12. Todd C. Headrick. (Ph. D.), "Type I error and power of the rank transform analysis of covariance (ANCOVA) in a 3×4 factorial layout." 355 pp., 1997.

13. Michael J. Nanna. (Ph. D.), "Analysis of Likert scale data in disability and medical rehabilitation research." 220 pp., 1997.

14. Anil N. F. Aranha. (Ph. D.), "Modeling self-determination among the elderly: A psychometric study of health care decision-making." 102 pp., 1998.

15. William Cade. (Ph. D.), "Sampling procedures and Type I error rates (for nonnormal populations)." 81 pp., 1998.

16. Cynthia Creighton. (Ph. D.), "Critical thinking skills and learning styles of first-year students in weekend occupational therapy programs." 80 pp, 1999.

17. Michael Wolf-Branigin. (Ph. D.), "Point pattern analysis in measuring physical inclusion of people with developmental disabilities." 182 pp., 1999.

18. Gail Fahoome. (Ph. D.), "A Monte Carlo study of twenty-one nonparametric statistics with normal and nonnormal data." 519 pp., 1999.

19. Joe Musial. (Ph. D.), "Comparing exact tests and asymptotic tests with colorectal cancer variables within the National Health and Nutrition Examination Survey III." 189 pp., 1999.

20. Juanita M. Lyons. (Ph. D.), "Methodology for the determination of the reliability of database derived data." 115 pp., 2000.

22. Jim Gullen.(Ph. D), “Goodness of fit indices as a one factor structural equation model.” 61 pp., 2000.

23. Karen Crawforth. (Ph. D.), Measuring the interrater reliability of a data collection instrument developed to evaluate anesthetic outcomes.” 144 pp., 2001.

24. Scott Compton (Ph. D.), “Type I error and power properties of seven two-sample tests when treatment affects location and scale.” 276 pp., 2001.

25. Kathy R. Peterson. (Ph. D.), “A study of six modifications of the ART (aligned rank transform) used to test for interaction.” 361 pp., 2001.

26. Rimma Novojenova. (Ed. D.), “Measurement of teacher’s personalization in the classroom environment.” 152 pp., 2002.

27. Jennifer M. Bunner. (Ph. D.). “Forming a bracketed interval around the trimmed mean: Alternatives to S_w . 112 pp., 2003.

28. Bruce R. Fay (Ph. D.), “A Monte Carlo computer study of the power properties of six distribution-free and/or nonparametric statistical tests under various methods of resolving tied ranks when applied to normal and nonnormal data distributions.” 528 pp, 2003.

29. Karen Lee (Ph. D.), “Parametric and nonparametric IRT models for assessing differential item functioning.” 113 pp., 2003.

30. Stephaine Krol-Jersevic. (Ed. D.), “Measuring oral communication apprehension in children.” 89 pp., 2004.

31. Jack Hill (Ph. D.). “The effects of pseudorandom number generator and initial seed selection on Monte Carlo simulations.” 251 pp., 2005.

32. Amittai ben Ami. (doctoral candidate)

33. Holly Atkins. (doctoral candidate)

34. Tana Bridge. (doctoral candidate)

35. Dave Fluharty. (doctoral candidate)

36. Roberta Foust. (doctoral candidate)

37. Calvin Holt. (doctoral candidate)

38. Kevin Lawson. (doctoral candidate)

39. Saydee Mends-cole. (doctoral candidate)

40. Kundisai Ndhelela (doctoral candidate)

41. Bulent Ozkan. (doctoral candidate)

42. Patricia Pelavin. (doctoral candidate)

43. Candice Pickens. (doctoral candidate)

44. Carol Piesko. (doctoral candidate)

45. Andree’ Sampson (doctoral candidate)

46. Lori Shingledecker. (doctoral candidate)

47. Boris Shulkin. (doctoral candidate)

48. Piper Farrell-Singleton (doctoral candidate)

49. Andrew Tierman. (doctoral candidate)

50. Michele Weber. (doctoral candidate)

51. Keith Williams. (doctoral candidate)

Doctoral Cognate (2nd advisor) Student

52. Mary Pratt Cooney. (Ph. D.), “Process drama and actor training.” 130 pp., 1999.

Doctoral Student of Todd C. Headrick

53. Simon Y. Aman. (Ph.D.), “An empirical investigation of nonparametric alternatives of Hottelling’s T^2 under non-normality.” 2005.

Doctoral Students of Gail Fahoome

54. Franklin Harrell. (doctoral candidate)

55. Sia Robinson. (doctoral candidate)

Biographical Sketches

Brief descriptions of members of R. Clifford Blair’s academic genealogy are provided below. Information in these synopses was obtained from a variety of sources, including Abailard and Berg (1970). (Considerable material from that reference is available verbatim in the online MacTutor History of Mathematics.) Other references included Burton (1997), James (2002), Temple (1981), and the Mathematics Genealogy Project (<http://www.genealogy.ams.org/>).

André Marie Ampère’s (1775 – 1836) biographical sketch appears here even though he is not in the direct academic lineage, because he was an influential instructor of Joseph Liouville, who took his course in mechanics at École Polytechnique and later his course in electrodynamics at the Collège de France. Ampère is primarily known for his work in chemistry and physics (e.g., light, heat, magnetism, electricity). However, he conducted considerable research in probability, which led to *The Mathematical Theory of Games*, and also a

text on the calculus. In 1814, he was elected to the Institut National des Sciences. This was a remarkable honor for the home-schooled and non-degreed Ampère, as he was elected over Augustin Louis Cauchy (1789 – 1857), one of the greatest mathematicians of the 19th century.

Dominique François Jean Arago's (1786 – 1853) biographical sketch appears here even though he is not in the direct academic lineage, because he was an influential instructor of Joseph Liouville. Arago was a Professor of Analytical Geometry at the École Polytechnique, and subsequently became Director of the Paris Observatory. Along with Louis Paul Émile Richard, one of his students was Urbain Jean Joseph Le Verrier (1811 – 1877). His research was on light, electricity, and magnetism. He served many years as the Secretary of the Académie des Sciences.

Jacob (Jacques or James) Bernoulli (1654 – 1705), following the wishes of his parents, reluctantly studied philosophy at the University of Basel and obtained the Master's in 1671, and then earned the licentiate in theology in 1676. After graduating, his travels led him to studying mathematics with Robert Boyle (1627 – 1691), Robert Hooke (1635 – 1703), Johann van Waveren Hudde (1628 – 1704), and Nicolas Malebranche (1638 – 1715). He started a private school for mathematics in Basel in 1682, and the following year he obtained a teaching position in mechanics at the University in Basel. He became Professor and Chair of Mathematics there in 1687. His early publications were on logic, algebra, and geometry. When his younger brother Johann sought his assistance in the study of mathematics, Jacob became a disciple of Leibniz. He published extensively in the newly established *Acta Eruditorum*, expounding on the calculus of Leibniz. Bernoulli's name is associated with the famous law of large numbers that is pervasive in probability theory. Bernoulli numbers made their appearance posthumously in *Ars Conjectandi* published in 1713, which contained the fundamentals of permutation and combinatorial theory.

Johann (John) Bernoulli (1667 – 1748), as with his brother Jacob, reluctantly followed his parent's wishes, and was employed in the family business as a salesman. He approached his brother to tutor him in mathematics. In 1695, he was appointed Professor of Mathematics at Groningen. Upon the demise of Jacob in 1705, he assumed the Professorship and Chair in Mathematics at Basel. Along with his brother Jacob, Johann published extensively in *Acta Eruditorum* on the calculus of Leibniz. Some work attributable to Johann was published in the name of his employer, Guillaume François Antoine, the Marquis de L'Hôpital (1661 – 1704). An example is the limit theorem commonly called L'Hôpital's rule. In Johann's correspondence with Leibniz, the phrase "integral calculus" was coined, and Johann adapted his brother's prior use of the elongated "s" for the integral symbol "∫." Later in his life, Johann was to help convince the parents of one of his students that their son should pursue mathematics instead of theology. That student was Leonhard Euler.

Enrico Betti (1823 – 1892) was a student of Ottaviano Mossotti at the Università di Pisa, and succeeded him in 1864 as the Chair of Mathematical Physics. Betti obtained his doctorate in 1846. He was a secondary school teacher, and later served at Università di Pisa as a faculty member and Rector. He was also the Director of the teaching college at Scuola Normale Superiore, Pisa. In addition to Vito Volterra, another one of his students was Luigi Bianchi (1856 – 1928). Betti played an important role in the development of mathematics in schools in the new Kingdom of Italy, translating classical texts (e.g., Euclid's *Elements*) into Italian, and similarly, in the world-wide transition from classical to modern algebra. His research interests were in algebra and topology. His 1871 topology work, which benefited from correspondence with Bernhard Riemann (1826 – 1866), provided the basis for what are called Betti numbers. Betti's theorem, a law of reciprocity in elasticity theory, was developed in 1878. He was Undersecretary of State for Education in 1874, and served as Senator in the Italian parliament in 1884.

Raymond Clifford Blair obtained his Bachelor's degree in International Studies in 1970, Master's of Arts in Gerontology the following year, and the Ph.D. in Measurement and Research in 1979, at the University of South Florida (USF). He became an Instructor at USF in 1976, and then accepted a position as an Assistant Professor in Evaluation and Research in 1979. He rose through the ranks, and became a full Professor in 1984. In 1987, he accepted the position of Coordinator of Measurement, Research, and Statistics, and Associate Professor at The Johns Hopkins University. He returned to USF the following year, accepting the joint position of Associate Professor in the Department of Pediatrics, College of Medicine, and the Department of Epidemiology/Biostatistics, College of Public Health. He was promoted to full Professor in the Department of Epidemiology and Biostatistics in 1997. He served as Deputy Chair from 1997 - 2000, and Interim Chair from 2000 - 2002. He was appointed Professor Emeritus in 2004. In 1993, he was awarded a grant by the IBM Corporation to develop pseudo-random number generators for the IBM RT PC computer. He published 70 articles, which appear in Table 1. His theoretical research was primarily on nonparametric rank tests, permutation statistics, multivariate statistics, and multiple comparison procedures. He published applied articles in biostatistics, public health, and medicine. Along with Shlomo S. Sawilowsky, his former doctoral student, he won the 1986 Distinguished Researcher Award of the Florida Educational Research Association and a 1987 Distinguished Paper, State and Regional Associations, of the American Educational Research Association. He won the 1995 and 1998 Distinguished Teacher awards of the USF Public Health Student Association. In 1996, he was honored as the USF Outstanding Teacher.

Julius Rubin Blum (1922 – 1982), in his youth, was sent by his parents from Germany to the United States. They perished in the Nazi holocaust before they could follow. He attended the University of California, Berkeley, was a member of Phi Beta Kappa, and obtained the Ph. D. in 1953. Officially, he was a student of Michel Loève. According to Professor Jane-Ling

Wang, “Le Cam was Blum’s thesis adviser in reality, but the university did not allow him to be the official adviser as they had been concurrent students at Berkeley. Le Cam graduated before Blum and supervised his thesis. Le Cam told me, and many others, this interesting story” (personal communications. Dr. Wang is Professor of Statistics, University of California, Davis (UCD), and received the Ph. D. in 1982 as a student of Le Cam at the University of California, Berkeley). Blum took a faculty position at UCD, and became the Chair of the Department of Statistics. In 1963, he became Professor and Chair of Mathematics at the University of New Mexico. In 1974, he joined the mathematics faculty at University of Wisconsin, Milwaukee. He returned to UCD as Associate Dean in 1979. His research interests, over 80 publications, were in stochastic approximation, multivariate generalization, ergodic theory, and nonparametric statistical inference. He co-authored the popular textbook *Probability and Statistics* in 1972 (with Professor Judah I. Rosenblatt, formerly of Case Western Reserve University and now with the University of Texas Medical Branch), which is available online at:
<http://www.bioinfo.utmb.edu/rosenblatt/index.html>.

Charles Bossut (1732 – 1806), a student of d’Alembert, was a Professor of Mathematics at Mézières, and then a Professor of Hydrodynamics at the Louvre. His two textbooks on mathematics and mechanics were widely used. He was awarded several prizes by the Académie des Sciences, and was elected member in 1768. In addition to Gaspard Monge, his students included Jean Charles de Borda (1733 – 1799) and Charles Augustin de Coulomb (1736 – 1806).

Nikolai Dmetrievich Brashman (1796 – 1866) was a teacher of mathematics at the University of Kazan, before accepting the position of Professor of Applied Mathematics in Moscow in 1834. He won the Demidov Prize from the Russian Academy of Sciences in 1836 for work in mechanics and mathematics. He founded the Moscow Mathematical Society.

Del Cavaliere Vincenzo Brunacci's (1768 – 1818) early mathematics training was under Stanislao (Sebastiano) Canovai (1740 – 1811). In 1785, he studied medicine at the University of Pisa. His mathematics instructor was Pietro Paoli. In 1788, he received a degree in medicine. He was appointed Professor of Nautical Mathematics in 1790. He joined the faculty at the University of Pavia in 1801, and eventually became its Chancellor. He published many books and articles, primarily on analysis and integral calculus. In 1806, he was awarded Knight of the Iron Crown and Inspector of Waters and Roads, and was elected to the Italian Society of Sciences. He became Inspector General of Public Education of Italy.

Lucien Le Cam (1924 – 2000) was an applied statistician working at Electricité de France for five years, and he was a graduate student at the Sorbonne in 1948, when Jerzy Neyman brought him to the University of California, Berkeley. Le Cam promptly flunked his doctoral qualifying exam. This humble beginning masked achievements he was to obtain in a career spanning about a half century at the University. After completing the Ph. D. in 1952, he was hired as an Instructor, rose through the ranks to full Professor of Statistics in 1960, and served as the Chair from 1961 – 1965. He published about 90 articles on topics relating to maximum likelihood, statistical decision functions, stochastic processes, asymptotic normal distributions, and applied cancer research. He co-edited a number of publications with Neyman (e. g., *Bernoulli-Bayes-Laplace Anniversary Volume* in 1965, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* in 1967 and 1972), and was the Associate Editor of *Zeitschrift für Wahrscheinlichkeits-theorie u. v. Gebiete* and *Polish Journal of Probability and Mathematical Statistics*. Among his students were Grace Lo Yang, Stephen Mack Stigler, and Jane-Ling Wang. Le Cam was President of the Institute of Mathematical Statistics in 1973, and was elected to the American Academy of Arts and Sciences (1976) and the American Association for the Advancement of Science (1977).

Eugène Charles Catalan (1814 – 1894) was a student of Joseph Liouville at École Polytechnique, but was expelled in 1833. He returned in 1835, and after graduating, accepted a faculty position at the Châlons sur Marne. He returned to Polytechnique as a Lecturer in 1838. He assisted Liouville in producing the *Journal de Mathématiques*. His solution to dissecting a polygon into triangles led to the discovery of Catalan numbers.

Michel (Floréal) Chasles (1793 – 1880), following a failed attempt at becoming a stockbroker, published a book in 1837 on the history of geometry. He became a professor at École Polytechnique in 1841, teaching astronomy, geodesy, and mechanics. Chasles obtained a simultaneous appointment as Chair of Higher Geometry at the Sorbonne in 1846. He published on projective geometry, conic sections, and synthetic geometry, emphasizing the history of mathematics. Hubert Anson Newton (1830 – 1896) was his student, whose student was E. H. (Eliakim Hastings) Moore (1862 – 1932), whose students were George David Birkhoff (1884 – 1944) and Oswald Veblen (1880 – 1960). Chasles was elected to the Académie des Sciences in 1851, a Fellow of the Royal Society of London in 1854, and to the London Mathematical Society in 1867. He was awarded the Copley Medal in 1865.

Pafnuty Lvovich Chebyshev (Tchebychev or Tschebyshev) (1821 – 1894) was lame and had a speech impediment. This was no obstacle to a brilliant career. He obtained his undergraduate degree in mathematics from Moscow University in 1841, his Master's in 1846, and his doctorate in 1849. His first two degrees were influenced by his mentor, Nikolai Brashman. He published on multiple integrals, Taylor series, law of large numbers, integration by logarithms, number theory, prime numbers, and orthogonal polynomials. He generalized the beta function, and his name is associated with Chebyshev polynomials and the Bienaymé-Chebyshev inequality, today referred to as the Chebyshev inequality. In addition to Andrei Markov, another of his students was Aleksandr Mikhailovich Lyapunov (1857 – 1918), whose student was Vladimir Andreevich Steklov (1864

– 1926). Cheybshev was given the title of Extraordinary Academician by the St Petersburg Academy of Sciences in 1856, elected to the Société Royale des Sciences of Liège in 1856, the Société Philomathique in 1856, the Berlin Academy of Sciences in 1871, the Bologna Academy in 1873, the Royal Society of London in 1877, the Italian Royal Academy in 1880, and the Swedish Academy of Sciences in 1893. Among many other prizes and titles, Chebyshev was awarded the French Légion d'Honneur.

Jean Le Rond d'Alembert (1717 – 1783) came from the classic ignoble beginning, as he was an illegitimate child left on the doorsteps of an orphanage. Fortunately, his identity was not kept secret, and while his father was alive he supplied financial support, which was until d'Alembert was nine years old. He was educated at the Jansenist Collège des Quatre Nations. He was admitted to the Paris Academy of Science in 1741. One of d'Alembert's major achievements was co-editing the 28 volume *Encyclopédie Diderot et d'Alembert* with Denis Diderot (1713 – 1784).

Jean Gaston Darboux (1842 – 1917) received his Ph. D. in Mathematics from École Normale Supérieure in 1866. He held academic posts at Collège de France in 1866, Lycée Louis le Grand the following year, École Normale Supérieure in 1872, and at the Sorbonne beginning in 1873. He taught higher geometry, became the Chair in Geometry in 1880, and Dean of the Faculty of Science from 1889 – 1903. His primary area of research was in differential geometry, but he also published on topics in algebra, function theory, and kinematics and dynamics. The Darboux integral bears his name. In 1884, he was elected to the Académie des Sciences, and in 1902 to the Royal Society of London. He was awarded the Sylvester Medal (James Joseph Sylvester, 1814 – 1897, founder of the *American Journal of Mathematics*) in 1916.

Robert L. Ebel obtained his Master's and Ph. D. from the University of Iowa. He was a high school teacher for nine years, and a school principal for three years. His was on the faculty of the University of Iowa from 1947 –

1957. He was a Vice President at the Educational Testing Service in Princeton, New Jersey, from 1957 – 1963. He returned to academia in 1963 at the Michigan State University (MSU), with an appointment to the faculty of Educational Counseling and Psychology, and also served as Assistant Dean. He authored numerous articles and textbooks in educational measurement, testing, and psychometric theory. He was the Editor of the *Encyclopedia of Educational Research* published by the American Educational Research Association (AERA). He was elected President of the National Council on Measurement in Education in 1957, and President of Division 5 (Evaluation, Measurement and Statistics) of the American Psychological Association in 1971. He won the AERA – American College Testing Program (ACT) “E. F. Lindquist Award” in 1989. His name is associated with a \$6,000 MSU College of Education Endowed Scholarship.

Leonhard Euler (1707 – 1783) obtained his doctorate at the Universität Basel in 1726 under Johann Bernoulli. The Euler and Bernoulli families were long time friends. Leonhard's father was a collegiate classmate of Johann Bernoulli; when Euler attended university at the age of 14, Johann provided him with reading lists. Later, when Leonhard accepted his first post at the St. Petersburg Academy in Russia (offered after the demise of Nicolaus Bernoulli, II, 1695 – 1726), he resided with Daniel Bernoulli (1700 – 1782). After a seven year stint in the Russian navy, Leonhard developed severe health problems, losing one eye and having poor vision in the other. Nevertheless, he won the 1738 and 1740 Grand Prize of the Paris Academy. Due to the Russian political climate, Euler left for the Berlin Academy of Sciences in 1741, where he published over 375 articles and books. He returned to St. Petersburg in 1766, by which time he was totally blind. This had little effect on his productivity, as he continued to publish almost as many manuscripts as he had prior to losing his vision, making him perhaps the most published mathematician in history. He wrote seminal articles on calculus, differential geometry, and number theory. He developed the

discipline of mathematical analysis and laid the foundation of analytical mechanics. He discovered the beta and gamma functions. The notation of “ $f(x)$ ” for a function, “ e ” for the base of natural logarithms, “ i ” for the imaginary number representing $\sqrt{-1}$, “ π ” for pi, “ Σ ” for summation, and many more were due to Euler. He also published important works in astronomy, cartography, mechanics, and fluid mechanics. In 1739, he published a delightful, but complex treatise on the relationship between mathematics and music.

Jacques Salomon Hadamard (1865 – 1963) received his Docteur ès Sciences in 1892 at the École Normale Supérieure. Emile Picard and Jules Tannery are indicated as his doctoral advisors, but he also took courses with Jean Gaston Darboux, Paul Emile Appell (1855 – 1930), and Edouard Jean-Baptiste Goursat (1858 – 1936). Hadamard was initially a school teacher, and later served on the mathematics faculty at Lycée Saint-Louis, Lycée Buffon, University of Bordeaux, Sorbonne, Collège de France, École Polytechnique, and finally, École Centrale des Arts et Manufactures. He published books on dimensional geometry, functional analysis, linear partial and hyperbolic differential equations, and about 300 scientific and pedagogy articles and books for general audiences. His research achievements included proving the famous prime number theorem, the most important result in number theory. In addition to Paul Lévy, his students included Maurice René Fréchet (1878 – 1973) and Szolem Mandelbrojt (1899 – 1983, who succeeded Hadamard at the Collège de France). Hadamard received the Bordin Prize of the Academy of Sciences in 1896, and the Prix Poncelet Prize in 1898. He was a member of the Academy of Sciences of the United States, the Royal Society of London, the Accademia dei Lincei, and the Soviet Academy of Sciences. He was elected President of the French Mathematical Society in 1906, and the Academy of Sciences in 1912.

Bruce Wendell Hall is Professor Emeritus in Educational Measurement and Research in the College of Education at the University of South Florida. He obtained his Ed.

D. from Florida State University in 1969. He was appointed to the faculty at USF later that year, and rose through the ranks to full Professor in 1979. He served as Chair of Educational Measurement and Research from 1976 to 1982 and again from 1990 to 2002. In addition to R. Clifford Blair, Hall chaired 30 students’ doctoral dissertations. He published 34 articles, made 142 paper presentations, and wrote 73 technical reports on educational research methods, instrument development, test reliability and validation, teacher attitudes, teacher attributions, teacher efficacy beliefs, classroom assessment, and school violence. He co-edited a volume on school testing programs published by the National Council on Measurement in Education (NCME) in 1976. He was twice elected President of the Florida Educational Research Association (1987 and 2003). He won the USF Provost’s Award in 1996, and the USF Professorial Excellence Award in 1998.

David Lee Hanson obtained the B. S. from the Massachusetts Institute of Technology, and the M. A. and Ph. D (1960) from Indiana University. His first position was with the IBM Research Center. Subsequently, he was employed at the Sandia Corporation until 1963, when he was appointed to the faculty of the Department of Statistics, University of Missouri – Columbia and the Department of Mathematics. He rose through the ranks to full Professor in 1967, and became Department Chair of Statistics in 1971. He joined the Department of Mathematical Sciences, State University of New York at Binghamton in 1973. He was Department Chair for 16 years, and currently is Professor of Probability and Mathematical Statistics. He was Program Director for Probability and Statistics at the National Science Foundation in 1979. In addition to James J. Higgins, his former doctoral students include Ralph P. Russo. Hanson’s publications include work on ergodic theory, the behavior of sums of random variables, Wiener processes (Norbert Wiener, 1894 – 1964), stochastic approximation, the theory of risk aversion, concave and monotonic regression, and hazard rates. He was an Associate Editor of *Annals of Mathematical Statistics*, *Annals of Probability*, and *Annals of*

Statistics. Hanson was elected Fellow of the Institute of Mathematical Statistics in 1966.

Charles Hermite (1822 – 1901) was a student of Louis Paul Émile Richard from 1840 – 1841 at the Collège Louis-le-Grand, who called him “un petit Lagrange.” Hermite was privately tutored by Eugene Catalan from 1841 – 1842. He was initially dismissed from École Polytechnique due to a physical disability that required him to walk with a cane, and graduated in 1847 elsewhere with a Baccalauréat. He returned to Polytechnique as a member of the faculty in 1848 where he remained until 1876. In 1856, he barely survived after having contracted small pox. He had a simultaneous appointment at the Sorbonne beginning in 1869. His primary contributions were in number theory, orthogonal polynomials, elliptics, and quadratic forms. In 1873, he proved e is a transcendental number. In addition to his doctoral students Jules Tannery and Henri Jules Poincaré, he taught Paul Emile Appell (1855 – 1930), Félix Edouard Justin Emile Borel (1871 – 1956), Marie Ennemond Camille Jordan (1838 – 1922), Paul Painlevé (1863 – 1933), as well as Darboux, Hadamard, and Picard. Hermite was elected to the Paris Academy in 1850, and to the Académie of Sciences in 1856. His name is associated with Hermite polynomials, Hermite differential equations, and Hermitian matrices.

Albert N. Hieronymus obtained his Master's (1946) and Ph. D. (1948) from the University of Iowa. He was a member of the Phi Delta Kappa honor society for over a half century. He became Professor Emeritus in 1987 at the University of Iowa, culminating his academic career that began at the College of Education in 1948. He became the second director of the Iowa Basic Skills Testing Program in 1948. He focused on infusing technology into standardized testing. He authored over 35 major standardized tests. His research areas were in learning theory, test development, and test validation. He was awarded the National Council on Measurement in Education (NCME) Career Award in 1991.

James J. Higgins obtained the Ph. D. in Statistics at the University of Missouri-

Columbia, in 1970. His first academic post was at the University of Missouri-Rolla, followed by his appointment at the University of South Florida from 1974 – 1980. Subsequently, he joined the faculty at Kansas State University in 1980, and is a full Professor. He served as the Head of the Department of Statistics from 1990 – 1995. His areas of theoretical research include mathematical statistics, nonparametric statistics, and reliability and life-testing. He also has published applied work on statistical education, correlated single subject designs, visitation patterns of animal foraging, and stochastic models for the synthesis of chemical compounds in red blood cells. To date, he has published a textbook on stochastic modeling and probability, a textbook on nonparametric statistics, and about 85 articles. In addition to serving as doctoral advisor to R. Clifford Blair and doctoral cognate advisor to Shlomo S. Sawilowsky, one of his former doctoral students was Sallie Keller-McNulty, who is President-elect of the American Statistical Association Board of Directors. Higgins received the College of Arts and Sciences Teaching Award in 1989, and was elected Fellow of the American Statistical Association in 1999.

Joseph-Louis Lagrange (Giuseppe Lodovico Lagrangia or Luigi De la Grange Tournier) (1736 – 1813) never met Leonhard Euler. Lagrange was mostly self-taught. However, in 1754, he began a life-long correspondence regarding his mathematical development with Euler. The following year he was appointed Professor of Mathematics at the Royal Artillery School in Turin at the age of only 19. In 1756, on Euler's recommendation, Lagrange was elected to the Berlin Academy. He was appointed Director of Mathematics at the Berlin Academy in 1766, which was Euler's post, on the latter's return to the University of St. Petersburg. Lagrange published on astronomy, dynamics, fluid mechanics, mechanics, number theory, probability, and of course, on the foundations of the calculus. The Lagrange multiplier, Lagrange integral, and Euler-Lagrange differential equation bear his name. He became a member of the Académie des Sciences in 1790. He was the inaugural Professor of Analysis at the École Polytechnique

in 1794, and was required to accept a joint appointment the following year at the newly established École Normale. His teaching skills did not reach the heights of his research skills, and Jean Baptiste Joseph Fourier (1768 – 1830) was assigned as his teaching assistant. He was awarded the Legion of Honour and Count of the Empire in 1808, and the Grand Croix of the Ordre Impérial de la Réunion in 1813.

Pierre-Simon Laplace's (1749 – 1827) advanced mathematical education was directed by Jean Le Rond d'Alembert, and through his efforts, Laplace obtained a position at the École Militaire. He quickly published over a dozen articles on minima and maxima, integral calculus, and differential equations, which led to his election to the Académie des Sciences in 1773. Laplace became an examiner at the Royal Artillery Corps in 1784, and in the following year he tested the 16 year old Napoleon Bonaparte (who passed). He was later (1812) to dedicate *Théorie Analytique des Probabilités* to Napoleon. He was appointed to the Bureau des Longitudes in 1795. Perhaps he was more scientist than mathematician; along with the chemist Antoine Lavoisier (1743 – 1794), he discovered the nature of respiration, then developed his nebular hypothesis, and subsequently published extensively on the most important physics topics of the time. He did considerable work in probability theory (including the sub-discipline due to Thomas Bayes, 1702 – 1761) and the theory of errors. The Laplace transform, Laplace integral, and Laplace operator bear his name. In 1806, he was elevated to Count of the Empire, and to Marquis in 1817.

Gottfried Wilhelm von Leibniz (1646 – 1716) obtained philosophy degrees from the University of Leipzig (undergraduate) in 1663 and the University of Jena (Master's) the following year. He studied mathematics under Erhard Weigel while at Jena. He completed his studies for the Doctoral degree in Law, but was denied, apparently, because he was too young. Therefore, he left for the University of Altdorf, where he received the Doctorate in Law in 1667. He studied mathematics with Christiaan Huygens (1629 – 1695) in Paris in 1672. He was

elected Fellow of the Royal Society of London the following year on the promise of developing a calculating machine (called a Stepped Reckoner, which was completed in 1694). Within four years, Leibniz was to develop his version of the calculus, and he published most of its elementary concepts, rules, and symbols in *Acta Eruditorum* by 1684. Although Sir Issac Newton (1643 – 1727) previously discovered the principles of the calculus in 1671, for a variety of reasons he never published them. Charges of plagiarism were launched in both directions. The matter was heard before Newton's home court – the Royal Society – where he had been its President since 1703. The Society commissioned a committee consisting primarily of Newton's British colleagues, such as Edmond Halley (1656 – 1742), with the notable exception of the French Abraham de Moivre (1667 – 1754). Leibniz' seemingly sole support was from his disciple Johann Bernoulli, who was not on the committee. The Society's conclusion was political, not scientific, and does not bear repeating. Newton and Leibniz can be considered co-discoverers of the calculus.

Paul Pierre Lévy (1886 – 1971) was a third generation mathematician. He matriculated at École des Mines in Paris, while simultaneously attending lectures from Jean Gaston Darboux and Charles Émile Picard at the Sorbonne. His doctoral advisor was Jacques Salomon Hadamard, who also served as examiner with Picard and Henri Jules Poincaré in 1912. The Mathematics Genealogy Project also lists Vito Volterra as his doctoral advisor, and indeed, functional analysis was Lévy's first research interest. (This concurs with Hadamard's work on Volterra's "line function calculus", which Hadamard renamed as Volterra's "functional calculus".) He served on the faculty of Écoles des Mines for a year, and then for 39 years at École Polytechnique. His former doctoral student, Michel Loève, stated Lévy had few students because he did not teach probability theory at Polytechnique. However, Lévy certainly had a generation of students who benefited from his 10 books and 278 articles, primarily written on probability. He also published on functional analysis, partial differential equations, Brownian motion, and

geometry. Lévy was elected honorary member of the London Mathematical Society in 1963, and in 1964 to the Académie des Sciences.

E. F. Lindquist (1914 – 1998), a native of Gowrie, Iowa (population about 1,000), was a psychometrician and statistician. He was a research assistant at the University of Iowa's College of Education in 1925. He became concerned with the process of assigning student grades based on casual and informal observations, or on subjective and unreliable opinions. This led him to the position of Director of the Iowa Testing Programs from 1930 – 1969. He co-invented the first electronic test scoring machine in 1955. He was also the co-founder of the American College Testing program (ACT) in 1959. He was the original developer of the Iowa Test of Basic Skills, and its first Director. In 1973, the University of Iowa dedicated the E. F. Lindquist Center for Measurement. The American Educational Research Association (AERA) and ACT co-sponsor an annual award in his name for outstanding theoretical research in testing and measurement. He was awarded the 1967 Distinguished Contributions to Research in Education Award by AERA.

Joseph Liouville (1809 – 1882) obtained his doctorate in 1827 from École Polytechnique. His examiners were Siméon Denis Poisson and Gaspard Clair François Marie Riche de Prony. He took several courses from André Marie Ampère and Dominique François Jean Arago at Polytechnique. He taught at Collège de France and École Centrale. Liouville launched the *Journal de Mathématiques Pures et Appliquées* in 1836. It became known as the Journal de Liouville, and it was an alternative to the previously established *Crelle's Journal* (August Leopold Crelle, 1780 – 1855). Liouville was elected to the Académie des Sciences in 1839, and the Bureau des Longitudes in 1840. In 1846, he published Evariste Galois's (1811 – 1832) hastily written final expositions prior to his death by duel. In politics, Liouville was elected to the Constituting Assembly in 1848. His work on the boundary value problem in differential equations resulted in the Sturm-Liouville theory (Charles-François Sturm, 1803

– 1855), an approach used in solving integral equations. He published about 200 articles on fractional calculus, integration of algebraic functions, transcendental numbers, and quadratic reciprocity. His work in differential geometry provides some of the foundations of statistical mechanics and measure theory.

Joseph Johann von Littrow (1781 – 1840) was a Professor of Astronomy at the University of Crakow, and served as the director of the Crakow Observatory from 1808 – 1810. Due to the campaign of Napoleon, Littrow hastily repaired to a Professorship in Astronomy at the University of Kazan in Russia. In 1816, he became the co-Director of the Pest Observatory in Hungary. He became Professor of Astronomy at the University of Vienna in 1819, and directed the Viennese Observatory. His areas of research were in astronomy, chronometry, geometry, optics, and physics. About 1840, he proposed digging ditches 20 miles in diameter in the Sahara, fueling them with kerosene, and igniting them to communicate with extraterrestrial life. On December 11, 1972, Apollo 17 landed at the southeastern rim of Mare Serenitatis in the Taurus – Littrow valley at 20.19080° N latitude, 30.77168° E longitude, a lunar surface named after Joseph von Littrow. He was knighted by the Emperor of Austria in 1837.

Michel Loève (1907 – 1979) was born in Yaffa, Israel, and eventually immigrated to France. He was naturalized as a United States citizen in 1953. While in France, he was awarded the title Actuaire I. S. F. A. (l'Institut de Science Financière et d'Assurances) by the Université de Lyon in 1936, and obtained his Doctorate in Mathematical Sciences from the Sorbonne in 1941. He held appointments at the Centre National de la Recherche Scientifique, was the Chargé de Recherches at the Institut Henri Poincaré of the Université de Paris, and briefly served on the faculty at the University of London. After completing a visiting Professorship at Columbia University, he became Professor of Mathematics at the University of California, Berkeley. He obtained appointments as Professor of Statistics in 1955 and Professor of Arts and Sciences in 1967. His lectures on probability theory were published in

textbook form as a volume in The University Series in Higher Mathematics in 1954. It became one of the most popularly used textbooks on modern probability theory. In addition to Julius Rubin Blum, Emanuel Parzen was one of Loève's doctoral students, who also wrote a classic textbook on the same subject. Loève was named Professor Emeritus in 1974. His wife and the University of California established the \$30,000 Line and Michel Loève International Prize in Probability.

Andrei Andreyevich Markov (1856 – 1922) graduated from St. Petersburg University, Russia, in 1878, and became a Professor in 1886. He published on analysis, approximation theory, number theory, limits, and converging series. He is noted for his work on stochastic processes and probability theory. His name is associated with Markov chains, a sequence of random variates wherein a predicted value is independent, but based on the current value.

Theodore Micceri obtained the Ph. D. in Measurement and Research from the University of South Florida (USF) in 1987. He was R. Clifford Blair's second and final doctoral student. Bruce W. Hall was co-advisor of his dissertation. He is a researcher in the USF Office of Institutional Effectiveness. He has 20 refereed publications on real data distributions, robustness of statistics, and instrument validation. Micceri has published over 375 technical reports on the evaluation of teacher practices, courseware design, and data base design. He is a Church Deacon and a Wood Badge trained Boy Scout leader.

Gaspard Monge, Comte de Péluse (1746 – 1818) graduated from the Collège de la Trinité in 1764. The following year he became a draftsman at École Royale du Génie, Mézières, where he came into contact with Charles Bossut. When Bossut took another post in 1769, Monge replaced him as Professor of Mathematics, and the following year he held a simultaneous position as Instructor in Physics at the École Royale du Génie. While at École Polytechnique, one of his teaching assistants was Jean Baptiste Joseph Fourier (1768 – 1830). Monge published frequently at the Académie des Sciences on

calculus of variations, infinitesimal geometry, partial differential equations, and combinatorics. He played an important role in creating École Polytechnique, and eventually became its Director. His support of Napoleon Bonaparte, even after his defeat at Waterloo, made Monge persona non grata in his latter years.

Ottaviano Fabrizio Mossotti (1791 – 1863), a student and later research assistant of Vincenzo Brunacci, obtained his degree in Engineering and Architecture at the University of Pavia in 1811. There is some evidence he took courses, and was influenced by Louis Gaspard Brugnattelli (1761 – 1818) and Alessandro Volta (1745 – 1827). He interned under Francesco Carlini (1783 – 1862) at the Royal Astronomical Observatory of Brera in Milan. An offer as Chair in Algebra and Geometry at Pavia was withdrawn when the university decided not to hire foreigners. In 1822, he was elected to the Società Italiana delle Scienze residente in Modena. He went to England for political reasons, returning later to become a Professor of Celestial Physics at the University of Pisa. In 1848, he fought in the Battle of Tuscany at Curtatone and Montanara, successfully leading a battalion of university students. In 1863, he was elected Senator of the Kingdom of Italy.

Jerzy (Splawa-)Neyman (Yuri Czeslawovich) (1894 – 1981), suffering from poor eye sight and tuberculosis, obtained his undergraduate degree from Kharkov University in 1947 and remained there as a Lecturer of Mathematics. He was influenced by his coursework in statistics, taken under Sergei Natanovich Bernstein (1880 – 1968). He met Wacław Sierpiński in Poland, and was motivated to study under him for his doctorate, which he received in 1924. He was examined by Sierpiński and Stefan Mazurkiewicz (1888 – 1945). Neyman became a teacher at Warsaw University and the College of Agriculture. As is well known, Neyman won a Rockefeller Fellowship to work with (Carl) Karl Pearson (1857 – 1936) in London in 1925, but was disappointed with Pearson's training in mathematics. He took a second year's fellowship to study with Félix Edouard Justin Emile Borel

(1871 – 1956) and Henri Léon Lebesgue (1875 – 1941) in Paris. Neyman returned to Poland in 1928, and set up and became the Director of the Biometric Laboratory at the Nencki Institute for Experimental Biology in Warsaw. He then joined Egon Sharpe Pearson (1895 – 1980), Karl's son, as an Associate Professor at University College in London, who he had met in 1925. Neyman accepted a position as Professor of Mathematics at the University of California, Berkeley. In 1955, he founded and became the Director of the Department of Statistics. Neyman and Egon Pearson collaborated on a number of articles, and they modified Sir Ronald Aylmer Fisher's (1890 – 1962) fiducial theory of statistics into the frequentist approach known as the Neyman – Pearson or “Bernoullian” paradigm of statistics. Neyman published on experimental design, generalized chi-square, hypothesis testing, optimal asymptotic tests, probability, and survey sampling. One of Neyman's greatest achievements was the development of the confidence interval, making him the father of modern statistics. He published applied research in meteorology and carcinogenesis toward the end of his career. Among his students were Erich Leo Lehmann (whose students included Madan Lal Puri, Peter John Bickel, Kjell Andreas Doksum, Gouri Kanta Bhattacharyya, Frank Rudolf Hampel, Howard Joseph Michael D'Abrera), George Bernard Dantzig, Frank Jones Massey, Jr., and Joseph Lawson Hodges, Jr. (whose student was Jerome Hamilton Klotz). Neyman won the Royal Statistical Society Guy Medal in 1966, the United States Medal of Science in 1969, and the 1973 Medal of the Copernicus Society of America. He was elected Fellow of the Royal Society in 1979. Neyman's slogan was “Statistics is the servant to all sciences”.

Pietro Paoli (1759 – 1839) taught mathematics at the University of Pavia. His two volume *Elements of Algebra* was a classic text used in Italy. His research was on analytic geometry, calculus, partial derivatives, and differential equations. In addition to Vincenzo Brunacci, his students included Giovanni Taddeo Farini (1778 – 1822).

Charles Émile Picard (1856 – 1941) obtained his Ph. D. in 1877 from École Normale Supérieure. He served on their faculty, and later at the University of Paris, Toulouse, and the Sorbonne. His areas of expertise were in analysis, function theory, differential equations, and analytic geometry. He discovered the Picard group transformations on a linear differential equation. He published numerous books, and served as an Editor of Liouville's journal from 1885 – 1941. He was elected to the Académie des Sciences in 1889 and the Académie Française in 1924. He received the Poncelet Prize in 1886, Grand Prix des Sciences Mathématiques in 1888, Grande Croix de la Légion d'Honneur in 1932, and the Mittag-Leffler Gold Medal in 1938. He served as President of the International Congress of Mathematicians in 1920.

Henri Jules Poincaré (1854 – 1912) was a student of Charles Hermite. He was an influential instructor of Paul Pierre Lévy, and served on his examination committee. Although Poincaré suffered greatly from various childhood illnesses, leaving him with muscular dysfunctions and poor eye sight, he was able to graduate from the École Polytechnique in 1875. He received his Doctorate in Mathematics from the University of Paris in 1879. His dissertation defense was less than stellar: His “*thesis is a little confused and shows that the author was still unable to express his ideas in a clear and simple manner.*” He accepted a professorship at the University of Caen, where it was revealed that his teaching skills were underdeveloped. Despite his disabilities, lackluster thesis, and sub par teaching skills, he is considered to be one of the greatest geniuses in history. The road to success began with an appointment in 1881 to the Faculty of Science, then as Chair of Mathematical Physics at the Sorbonne in 1886, and eventually to the École Polytechnique. He became the father of algebraic topology, analytic functions of several complex variables, and along with Magnus Gösta Mittag-Leffler (1846 – 1927), his work led to chaos theory. As impressive as were these accomplishments, they pale in comparison to his co-discovery of special relativity, along with Hendrik Antoon Lorentz (1853 – 1928) and Albert Einstein (1879 – 1955). He was elected to the Académie des

Sciences in 1887, and became its President in 1906. He is the only person elected to every division of the Académie (geography, geometry, mechanics, navigation, & physics). He was elected to the Académie Française in 1908.

Siméon Denis Poisson's (1781 – 1842) lack of fine motor coordination played a role in his decision not to pursue a career in medicine, and when he turned to mathematics, to avoid descriptive geometry that required drawing finely detailed charts. Nevertheless, under the tutelage of both Pierre-Simon Laplace and Joseph-Louis Lagrange, his work was considered so brilliant that his dissertation was accepted without the traditional examination. He was immediately offered his first position at École Polytechnique in 1800. His appointments blossomed as an astronomer at the Bureau des Longitudes in 1808 and inaugural Chair of Mechanics at the Faculté des Sciences in 1809. He published major treatises on astronomy, heat, electricity, physics, and nearly 400 tracts on mathematics. His name is associated with the Poisson integral, Poisson distributions, Poisson differential equation brackets, Poisson elasticity ratio, and the Poisson constant in electricity. In addition to Michel Chasles, another of his doctoral students was Johann Peter Gustav Lejeune Dirichlet (1805 – 1859) (with Jean-Baptiste Joseph Fourier, 1768 – 1830, serving as 2nd advisor). [Dirichlet had many notable academic descendants: Rudolf Otto Sigismund Lipschitz (1832 – 1903), followed by Felix C. Klein (1849 – 1925), and Wilhelm v. Behrens and Ludwig Bieberbach (1886 – 1982). Continuing through Bieberbach were Heinz Hopf (1894 – 1971), Beno Eckmann, and Peter Jost Huber.] Dominique François Jean Arago quoted Poisson to have said, “Life is good for two things: researching mathematics and teaching mathematics.”

Gaspard Clair François Marie Riche de Prony (1755 – 1839) graduated in 1776 with a degree in engineering from the École des Ponts et Chaussées, where he was subsequently employed and eventually became its Director in 1798. His work on the Louis XVI Bridge (Pont de la Concorde) elevated him to the position of Engineer-in-Chief in 1790. The following year,

working with Adrien-Marie Legendre (1752 – 1833), Lazare Nicolas Marguérite Carnot (1753 – 1823), and over six dozen assistants, he commenced producing the *Cadaastre*, an exhaustive book of logarithms and trigonometric functions. He wrote several text books on mechanics. He was a member of the Bureau de Longitude. de Prony promoted reforming curriculum toward applied mathematics, but Augustin Louis Cauchy's (1789 – 1857) firm stance on pure mathematics prevailed.

Hermann Henry Remmers (1892 – 1977) obtained his Ph. D. from the University of Iowa. He was a Professor of Education and Psychology at Purdue University for about 30 years, and served as the Director of the Division of Educational References. In 1935, he co-founded what was to become the Indiana Student Financial Aid Association. He was the originator of the Purdue Opinion Panel, which led to his noted book, *The American Teenager*, in 1957. He authored textbooks on educational psychology, educational measurement and evaluation, and about 200 articles and monographs on teaching, survey methods, testing, and evaluation. He was elected President of the Division of Educational Psychology of the American Psychological Association in 1951, and the President of the American Educational Research Association. His name is associated with Purdue University's \$1,000 H. H. Remmers Award for African American Studies.

Louis Paul Émile Richard (1795 – 1849) served on the faculty of the Collège de Pontivy, Collège Saint-Louis, and Collège Louis-le-Grand. In addition to Charles Hermite, his students included Urbain Jean Joseph Le Verrier (1811 – 1877), Joseph Alfred Serret (1819 – 1885), and Evariste Galois (1811 – 1832).

Shlomo Noach (Stephen Ram) Sawilowsky obtained the M. A. (Counselor Education, 1981) and Ph. D. (Measurement, Evaluation, and Research, 1985) from the University of South Florida (USF). He was R. Clifford Blair's first doctoral student, graduate teaching assistant for two years, and graduate research assistant for two years. James J.

Higgins was his dissertation 2nd advisor. Bruce W. Hall was his measurement instructor. Sawilowsky was a Visiting Assistant Professor at USF from 1985 – 1987, and accepted a position in the College of Education at Wayne State University (WSU) in 1987. He rose through the ranks to full Professor of Evaluation and Research in 1997, and has served as Department Chair since 1998. He accepted simultaneous teaching appointments with the faculty of Curriculum and Instruction in 1998 and Counselor Education in 2000. Sawilowsky and Blair's work on the rank transform won the 1986 Distinguished Researcher Award of the Florida Educational Research Association, and a 1987 Distinguished Paper Award, State and Regional Associations, of the American Educational Research Association (AERA). Sawilowsky has won many WSU teaching honors, including the 1995 University President's Award, 1997 College of Education Award, 1998 Graduate Mentor Award, and the 1999 Faculty Mentor Award. A list of his doctoral students, descendants in R. Clifford Blair's academic genealogy, is compiled in Table 2. He was awarded WSU Distinguished Faculty Fellow in 2000. Along with Sharon Field and Alan Hoffman, he obtained over \$3.5 Million in extramural funding for research on self-determination for students with and without disabilities, and co-authored a battery of standardized tests on self-determination. He has published over 80 articles on nonparametric rank tests, permutation and robust methods, classical measurement theory, and construct validity. He co-authored a textbook on statistics via Monte Carlo methods with Gail Fahoome, a former doctoral student, and he is the Editor of a volume on real data analysis to be published by the AERA Educational Statisticians. He founded the *Journal of Modern Applied Statistical Methods* in 2000, and serves as Editor. In a simultaneous career, he obtained his undergraduate degree in 1979 at the Rabbinical College of America. He served as the emissary of the *Lubavitcher Rebbe AMv"R* Menachem Mendel Schneerson, *ZTzVKLLH" H*, in Tampa (1980 – 1985) and St. Petersburg, Florida (1985 – 1987). Since 1987, he has taught *Talmud* (Rabbinical jurisprudence), *Halacha* (Jewish law), and *Chassidut* (philosophy) at various

synagogues in Michigan. In 2004, Sawilowsky obtained a second Rabbinical ordination in Jerusalem, Israel, from *HaRav* Dovid Ostroff of *Pirchei Shoshanim*.

Waclaw F. Sierpiński (1882 – 1969) obtained his undergraduate degree in 1904 from the Department of Mathematics and Physics at the University of Warsaw while it was under Russian occupation. He was a student of Georgy Voronoy at that time. He won a prestigious university prize for his work on number theory. In his memoirs, Sierpiński revealed that he deliberately left the answers to his final examinations blank to protest the Russian occupation of Poland and the University. This put the University in the position of denying the degree to a prize-winning student. Ultimately, however, he received the degree. He became a student of Zaremba (who was Voronoy's student) at the Jagiellonian University, Crakow. He obtained the doctorate in 1908, and in the same year accepted an appointment at the University of Lvov in 1908. Later, he served as the Dean of the Faculty of the University of Warsaw. One of his students was Stefan Mazurkiewicz (1888 – 1945). He published many books and articles, primarily on set theory, theory of irrational numbers, and point set topology. The Sierpiński curve bears his name. He founded the journal *Fundamenta Mathematicae*. He was elected to the Polish Academy, Vice Chair of the Warsaw Scientific Society, and the Polish Mathematical Society.

Howard Stoker obtained his Master's degree in 1950 at the University of Iowa. Albert N. Hieronymus was his thesis advisor, and Robert L. Ebel served on the Master's committee. Stoker received his Ph. D., as a student of Hermann Henry Remmers, in 1957, from Purdue University. He obtained his first academic appointment at Florida State University, where he taught from 1957 – 1984. He was awarded Professor Emeritus in 1985. From 1984 – 1988 he was the Head of Instructional Development and Evaluation in the Department of Education at the University of Tennessee, Memphis. From 1988 – 1992 he held his third professorship, this time at the University of Tennessee, Knoxville. He was

awarded Professor Emeritus from the University of Tennessee in 1992. He co-authored a two volume edited text on educational measurement in 1996. His research focused on standardized testing, test validity, and measurement theory.

Jules Tannery (1848 – 1910) obtained his Ph. D. in 1874 at École Normale Supérieur. He served as a member of the mathematics faculty at Lycée Saint-Louis, Sorbonne, École Normale Supérieur, École Normale – Sèvres, and Faculty of Sciences – Paris. He authored books on the history and philosophy of mathematics, and was an Editor of the *Bulletin des Sciences Mathématiques* from 1876 – 1910. He played an important role in the revising of mathematics curriculum in France. He was elected *member libre* of the Académie des Sciences in 1907.

Vito Volterra obtained his Doctorate in Physics at the University of Pisa under Enrico Betti in 1882. His initial appointment, the following year, was Professor of Mechanics. He assumed the Chair of Mathematical Physics after Betti's demise. Subsequently, he served on the faculty at the University of Turin and the University of Rome. He published on partial differential equations, celestial mechanics, elasticity, and biometrics. His name is associated with Volterra functional calculus or Volterra type integrals. He became a Senator of the Kingdom of Italy in 1905. He was decorated with the War Cross for his services as a veteran of the air forces group in the corps of engineers in World War I, and was credited with developing mounted guns in airplanes. He was the first to propose replacing hydrogen with helium in airships. He received honorary knighthood from King George V of England in 1921. Volterra fought against the Fascist take-over of the Italian Parliament in 1930, resulting in his dismissal the following year from the University of Rome. He was President of the Academia dei Lincei, and after his dismissals from Italian scientific societies by the Fascist government, he was elected to the Pontifical Academy of Sciences by Pope Pius XI in 1936.

Georgy Fedoseevich Voronoy (1868 – 1908) was a member of the Faculty of Physics and Mathematics at the University of St.

Petersburg, and Warsaw University, even while working on his undergraduate and Master's degrees. He obtained his Doctorate in Mathematics at the University of St. Petersburg. His dissertation won the Bunyakovsky Prize (Viktor Yakovlevich Bunyakovsky, 1804 – 1889). His area of research was number theory: algebraic numbers and the geometry of numbers. There is a discipline of art referred to as “Voronoi Paintings”, where the design is based on cells interacting directly with its neighbor and indirectly elsewhere. Samples may be viewed at the Trayecto Gallery, Vitoria, and currently at: <http://www.lxxl.pt/veado.html>.

Erhard Weigel (1625 – 1699) was Professor of Mathematics at Jena University, where he taught from 1653 - 1699. He was an inventor, educator, and advocate of the Gregorian calendar. His aim was to meld mathematics with philosophy. He sought to teach the sciences to the public, and in that effort, created a celestial instruction globe made of copper, brass and wood that is held at the National Maritime Museum in London. Gottfried Wilhelm von Leibniz was his student. Gottfried Kirch (1639 – 1710) and Samuel Pufendorf (1632 – 1694) are indicated as his students, but their ages suggest they may have studied under Weigel prior to his tenure at Jena. The Mathematics Genealogy Project lists Christoph Vogel (Doctorate of Philosophy in 1652) and Theophilus Wildius (Ratisbonensis, Doctorate of Philosophy in 1654) as Weigel's students, but the dates of their doctorates are similarly problematic.

Stanislaw Zaremba (1863 – 1942) attended the Sorbonne, where he obtained his doctoral degree in 1889. After teaching in France for a decade, he returned to his native Poland to accept a Chair at Jagiellonian University, Crakow. He was elected as the inaugural President of the Mathematical Society of Crakow, and was the Editor of the *Annals of the Polish Mathematical Society* for many years. His primary areas of research were in partial differential equations and potential theory, but he also published articles on mathematical physics and crystallography. He was elected to the Soviet Academy in 1925.

References

Abailard, P., & Berg, S. (1970). *Dictionary of scientific biography*. (Charles Coulston Gillispie, Ed.). NY: Charles Scribner's Sons.

Burton, D. M. (1997). *The history of mathematics: An introduction*. (3rd ed.). NY: McGraw-Hill.

James, I. (2002). *Remarkable mathematicians: From Euler to von Neumann*. Cambridge: The Mathematical Association of America.

MacTutor History of Mathematics. (www-gap.dcs.st-and.ac.uk/~history/index.html)

Temple, G. (1981). *100 years of mathematics: A personal viewpoint*. NY: Springer-Verlag.

Internet Photo Credits and Resources

Photos were obtained from public internet websites and are assumed to be in the public domain. The url's change frequently and many links are not stable. Please email the author for a complete listing of links valid at the time the photos were downloaded.

Images of some stamps were obtained from jeff560.tripod.com. Other copyright information: E. F. Lindquist: <http://www.uiowa.edu/~counspsy/program.htm>. Joseph Johann Littrow: Lithographie nach einer Büste von A. Dietrich, © Copyright by Bildarchiv der Österreichischen Nationalbibliothek, Wien. Lucien Le Cam: <http://stat-www.berkeley.edu/users/rice/LeCam/index.html>. Erhard Weigel: <http://www.astro.uni-bonn.de/~pbrosche/aa/acta/vol07/>.

Statistics and Technology: Reflections on 35 Years of Change

James J. Higgins
Department of Statistics
Kansas State University

From the days when statistical calculations were done on mechanical calculators to today, technology has transformed the discipline of statistics. More than just giving statisticians the power to crunch numbers, it has fundamentally changed the way we teach, do research, and consult. In this article, I give some examples of this from my 35 years as an academic statistician.

Introduction

When I began my undergraduate studies at the University of Illinois in 1961, the state of the art hand-held calculating device was the slide rule. I paid twenty-five dollars for mine, which was a lot of money in those days. The first statistical calculations I did were on a mechanical calculator, and the first book that I taught out of had a table of square roots in the appendix. I've seen mainframe computer centers and punch cards come and go. Now, powerful personal computers are commonplace, and a large fraction of the population has access to the internet.

All of this has fundamentally changed the discipline of statistics. It has changed what we teach and how we teach it. It has given statistical research a genuine experimental side to go along with theory, and it has changed the role of the statistical consultant. I've chosen examples from my experiences to illustrate the

changes that have occurred and in some cases to suggest directions that the discipline might go in the next decade or so.

The Introductory Pre-Calculus Undergraduate Course Past and Present

The approaches of two very successful authors, William Mendenhall and David Moore, capture the essence of the changes that have occurred in the introductory undergraduate statistics course.

Mendenhall began publishing introductory statistics books in the 1960s at the time when there was rapid growth in the demand for introductory statistics as a general education course. He successfully took material that was previously only accessible to students with calculus and brought it to the pre-calculus audience. His organization, which has been replicated by many authors, includes elementary descriptive statistics, axiomatic probability and probability distributions, and a systematic treatment of inference (e.g., one-sample, two-sample, regression, analysis of variance).

Moore's books epitomize the changes in thinking that began to take place in the late 1970's and 1980's. His *Statistics: Concepts and Controversies* (Moore, 2001a) begins with sampling and experimental design. It then has a discussion of descriptive statistics that includes

James J. Higgins is a Professor of Statistics at Kansas State University. He has published a textbook on stochastic modeling and probability and a textbook on nonparametric statistics. He is an elected Fellow of the American Statistical Association. Contact him at jhiggins@ksu.edu.

contingency tables, correlation, and simple linear regression all done without reference to statistical inference. Only enough probability is covered to deal with inference, and then just the basics of inference are discussed.

The analysis of real data has become the standard for good exposition. Students can be expected to have, at a minimum, a two-variable, hand-held calculator so that the drudgery of computing means, standard deviations, and regression equations is eliminated. Most books display and interpret elements of computer output, and many instructors expect students to be proficient with some statistical software. The student can now concentrate on what it means rather than on how to compute the answer.

Beyond Data Analysis

Although I do a lot of data analysis in my introductory courses, data analysis is not the most important thing I do as a consulting statistician. Rather it is in the planning of studies that I think I have the greatest impact. In DeGroot (1987), C. R. Rao had this to say:

“I believe that the two great methodologies in statistics are sample surveys, which is essentially collecting existing information, and design of experiments, where you generate observations to provide information on some given questions. Different types of data analysis are, of course, then applied depending upon what the statistician thinks is the right thing to do. They are not as fundamental as the data which are collected through principles of design and sample surveys.”

Box (1990) was critical of the notion that statistics is a branch of mathematics. He wrote, “Statistics is, or should be, about scientific investigation and how to do it better...” His commentary is very thought provoking.

The difficulty in trying to teach design, sampling, or better scientific investigation in an introductory undergraduate course is that most undergraduates haven't had the opportunity to be involved in the process of scientific discovery. At most they may have done

laboratory demonstrations that illustrate some scientific principle. The majority will not have dealt with a problem where they don't know the answer and have to take data to find it out. Thus they do not appreciate the most important reason to learn statistics, namely, scientific discovery.

Many instructors assign projects that illustrate discovery through data. Some projects are short so that they can be done in class, but they lack the complexity of real studies. Others are more extensive, but must of necessity take several weeks or even a semester to complete, see Hunter (1977). Here is where we could take the next step in the use of technology. I would like to see us merge video-game technology with our ability to simulate data from scientific studies to come up with interesting software that would invite student to conduct their own experiments in a computer lab.

Imagine, for instance, software that would simulate agricultural experiments. Students might have several varieties of corn that they could choose to plant, several options of fertilizer to use, a choice of whether to irrigate or not, several ways to control pests, different environments in which to plant the corn, different plots of ground upon which to do the experiment, and several responses to consider such as yield, plant damage, plant dry weight and the like. With computer graphics and animation showing a researcher planting the corn, applying the treatments, and harvesting afterward (all controlled by the student), the software would invite students to plan and carry out a scientific study in a way not unlike they might do in practice. It would be rather like using a flight simulator to teach the basics of flying an airplane. Students could be presented with many different scenarios that could be dealt with in a safe environment before they are turned loose to deal with the real world of scientific investigation.

Planet X

A few years ago our department was given the opportunity to design a studio classroom for one of our large introductory courses. The classroom that we came up with has 20 computers which are arranged on octagonal tables where students work in pairs. There are lots of opportunities for students to

interact with each other. Unlike a lab, students are in the classroom for every class period. My colleague Deb Rumsey designed the classroom, and we set about developing a curriculum that would take advantage of it.

For years we had been using computers to simulate data for class examples, but we wanted to try something more elaborate. We decided to build a large database representing characteristics of individuals who belong to some population. We wanted to put together a story to give interest to the database, and we wanted students to have a lot of flexibility in terms of what data they could collect and what questions they could ask. Finally, we wanted to put some graphical and animation elements into the program to give it visual appeal.

We considered modeling a small city perhaps using census data to populate our database. I think this has merits, but it presents some pedagogical problems, too. Students may have preconceived notions that would taint their analysis of the data. For instance if a student asks questions of the data about race or gender, their biases might not only affect their interpretation of the data, but they could also lead to a class discussion that goes beyond statistics and into the realm of sociology where the instructor may not wish to go.

Such concerns led us to create Planet X, a place that is like Earth but with differences to be discovered. There are 4 ethnic groups on Planet X, 50 cities, 9 governmental regions, coastal and inland cities, etc. The database has 500,000 inhabitants with 31 variables on each one representing various physical and social characteristics. Students can sample from the entire population or from various sub-populations. Computer animation shows a spaceship flying off to Planet X and going into orbit around the planet. Our students make contact with the inhabitants, gather data, and fly back to Earth where they do the analysis and write a report.

The philosophy behind Planet X is contrary to the conventional wisdom that it takes real data to engage students in statistics. I believe that data just need to be engaging, and whether the data are real or simulated is immaterial. Some students enjoy Planet X a lot. Others think it is hokey. Many students with a

little guidance write reasonably good reports about what they've found out from their data analysis. The fact that they have something concrete to write about gives focus to their writing. Evaluations indicate that students develop a level of comfort with survey methodology that we do not necessarily find in our traditional classes.

The impediment in developing this is having someone with the time and technical expertise in graphics and animation programming to do the work. We were fortunate to have someone who knew enough about this to get something to work at K-State although it proved not to be transportable to other locations for various technical reasons. Ultimately it will take professional software developers to put together a sufficiently complex set of simulated scientific studies to make possible a true test of the usefulness of this type of technology in introductory statistics courses. I will simply offer the opinion that the potential there.

Statistics as an Undergraduate Discipline

Once, at a seminar by a statistician from the pharmaceutical industry, I asked the speaker whether his company hired undergraduate statistics majors to manage the large databases that his company maintains. He admitted that although they might do that, most of those they hired had little statistics background. His company simply hired those that had some computing. I thought what a lost opportunity not only for the company but also for statistics as an undergraduate discipline.

With a few notable exceptions, statistics lacks visibility as an undergraduate discipline in colleges and universities. See Minton (1983). Having taught in the Florida university system for 6 years, I noted with dismay that the new Florida Gulf Coast University, which was established in 1997, did not have a statistics program; let alone a statistics department. Even though we tout the importance of statistics in the information age, statistics wasn't even a blip on the radar screen of this modern university.

Part of this has to do with how statistics departments came into being. Almost all began at major universities with the primary mission to produce M.S. and Ph.D. statisticians. See Bancroft, et al. (1958) for an account of the state

of the statistics profession in the 1950s. Entry to graduate school in statistics even today does not require an undergraduate degree in statistics.

Because statistics does not have a tradition as an undergraduate discipline in the same sense that mathematics does, there is not a clearly defined notion of what an undergraduate program in statistics about. This has troubled me for some time. What ideas and coursework are at the core of undergraduate statistics? Can these ideas be successfully taught in mathematics departments or departments of mathematical sciences where the majority of the undergraduate statistics programs now reside? What coursework would make a career path for the undergraduate statistics major? In the article "Nonmathematical Statistics: A New Direction for the Undergraduate Discipline", I attempted to answer these questions (Higgins, 1999).

Nonmathematical activities are very much a part of what a practicing statistician does and what customers of statistics need. They include things like managing large databases, planning studies in a team-oriented environment, ensuring protocol compliance, providing internet access to databases, and providing descriptive and graphical summaries of data (apart from the usual inferential statistics). I suggested eight courses that deal with these things that are not courses that would fit well within a traditional mathematics or mathematical sciences program. The titles are listed below. The article elaborates on the topics.

- (1) The Scientific Process
- (2) Planning and Managing Surveys
- (3) Planning and Managing Scientific Studies
- (4) Statistical Software for Data Management
- (5) Statistical Graphics
- (6) Computer Science in Statistics
- (7) Communicating Statistical Ideas
- (8) Management Principles for Statistics

These courses along with courses in inference could form the basis for a professional degree program in statistics. Students with this type of coursework could serve as "data specialists". It is not difficult to find job descriptions in industry, business, and

government that require the skills of a data specialist. The very technology that enables these organizations to gather massive amounts of data also creates a potential bonanza of opportunities for the undergraduate statistics major with the right type of education.

There have been some good efforts to bring a common understanding of what the undergraduate statistics degree program is about. Articles by Bryce, et al. (2001) and Ritter, et. al. (2001) represents one of the latest attempts to do this. These articles made recommendations about the curriculum some of which incorporated ideas from my 1999 article. In the same issue, Moore (2001b) took on a different problem: how to grow undergraduate programs. His conclusion was that economic considerations compel statistics and mathematics to work together.

If Moore is correct, then undergraduate statistics cannot reach its full potential. The mathematical discipline by its very definition is not structured to support the kinds of nonmathematical courses that a professional undergraduate statistics program would need. The best that we could hope for in this case is that statistics would be a liberal arts degree option that could be fulfilled by students getting a degree in mathematics and taking a handful of courses in statistics.

Distance Education

Since the days of radio, colleges and universities have had some form of distance education. Kansas State University for years supported a radio station as part of its agricultural outreach. Modes of delivery have evolved from radio, to television, to video courses on demand, and finally to the internet.

In 1989, I was asked to have my introductory graduate methods course video taped to be used as a distance education course for a program for food inspectors. The course was taped in a special classroom that had a camera at the back of the room. I just did my thing teaching as I always would. The only concessions that I made for the camera were that I wrote with big chalk on the board so that my writing would show up on camera, and I wore long-sleeved shirts and ties. The production was very primitive, but it was also cost effective to

produce which is an important consideration in putting together distance education material.

As it turned out, an educational television company that had national cable outlets obtained rights to the course for their distance education degree programs. Soon after that I began hearing from people from around the country who saw the course. My sister, who lives in Illinois and who did not know that I had done this, was clicking through the channels one evening when I came on the screen. Needless to say she was surprised. To my amazement those that saw the course seemed to like it. It found an audience that also surprised me: graduate students in nursing programs. Even though I talked about pigs, cows, wheat, and corn the course met a need for these students.

Later I redid the course for the internet. Here I made three choices that turned out to be right even though I had no previous experience with this form of teaching.

(1) I decided to make the presentation “linear” as I would in an ordinary classroom setting. I avoided the temptation to put in a lot of links and connections that would allow students to roam around and get away from the central flow of the material. I reasoned that if I were placed in the middle of a forest, I would not want to be given a lot of options for getting out, some of which might be dead ends. Rather I would like for someone to point to a single path as the way to go knowing at the end that I would be out of the forest.

(2) I divided the material that I would ordinarily teach in one class period into two parts, each with its concepts, reading assignments, and homework problems. Students have told me that they like this feature a lot.

(3) I presented the material in detailed outline form using PowerPoint slides rather than writing an online text. This allowed me to put in graphics and gizmos to give the pages some visual appeal while making the essential points as succinctly as I could. I require a textbook that students can refer to if they need additional explanation.

I have had over 1,500 students take this course. It is self-paced although I encourage students to finish within the semester that they sign up. We have 40 or more students a semester sign up for the course, and we offer it fall,

spring, and summer. I use the Excel spreadsheet for computing because most students have access to it although I am well aware of its limitations. It is very satisfying knowing that this course is accepted by many universities and colleges around the country. I’m sure that a major part of the success of the course is the high level of motivation of the students who take it. I recently developed an undergraduate internet course for business majors. It is too early to tell how well my style will work with these students.

The use of the internet technology comes at a price. It took me over a year working part-time to develop each of my courses. Thus, internet courses are only cost effective if they can be rerun several times. I would not recommend anyone doing this without extra compensation or release time to do the work. Our department is reimbursed for my time by the Division of Continuing Education. Some of that comes to me indirectly as discretionary funds that I can use for travel, computer equipment, graders, and the like.

Mathematics, Computing, and Research

I took a pretty good dose of probability, analysis, and measure theory to go along with my statistics Ph.D. coursework. My dissertation was “Convergence Rates for Weighted Sums of Independent Random Variables” under Dave Hanson’s direction. I chose the University of Missouri at Columbia in large part because I thought I would get a good background in mathematics to go along with statistics, and it has served me well even though now I consider myself to be an applied statistician. It was at the University of South Florida that I got my first significant exposure to applied research. Chris Tsokos, to whom I owe a great deal, directed me toward reliability theory which is an area that I have worked in since.

I’ve seen less emphasis on mathematics in statistics Ph.D. programs over the years. When I was in graduate school some version of measure theory was rather standard for Ph.D. students. Now I would say that it is far less common. I’m not sure whether this is good or not. Jacob Wolfowitz, who spent his last years at the University of South Florida, made it clear to me at one particular meeting of the

curriculum committee that our students needed more mathematics not more applied statistics. I was never completely convinced by that, but who was I to argue.

What I think I can safely say is that computing has changed our expectations of what constitutes research in statistics. Tukey (1986) had this witty but profound insight about the role of computing:

“In a world in which the price of calculation continues to decrease rapidly, but the price of theorem proving continues to hold steady or increase, elementary economics indicates that we ought to spend a large and larger fraction of our time on calculation.”

In the same sense that physics has theoretical and experimental sides, statistics has these two sides too thanks to the capability to do computer simulations.

One of the courses that I took in my first year at Missouri was a course on computer simulation taught by Bill Bulgren. Although I have to confess that I was, and still am, a lousy programmer, I was really taken by the power of the Monte Carlo method to readily provide answers to difficult questions that could not be touched with standard analytical methods. The ideas that I learned in that course have influenced my research and teaching throughout my career.

A number of my papers have dealt with small sample properties of statistical methods, something that can be investigated with well-designed computer simulation studies. Work by Blair and Higgins (1980) shed light on some long-standing misconceptions about the power of nonparametric methods in the social and behavioral sciences. Specifically, an influential paper by Glass et al, (1972) concluded that nonparametric methods have low power and are not suitable for serious data analysis. Nothing could be further from the truth as asymptotic theory shows, but unfortunately even today these wrong ideas persist. Ironically the wrong ideas about rank tests arose from poorly designed simulation studies.

Advances in statistical methodology often involve the interplay of applications, experimental statistics, and theoretical statistics. The rank-transform methodology, which was first proposed by Iman, as a student of Conover, at Kansas State University, is such an example. At first it seemed to hold promise an easy way to do nonparametric statistics for the types of designed experiments that one typically encounters in practice. Simply replace observations by ranks and do the same linear models analysis on ranks that one would do on normally distributed data. See Conover and Iman (1981) for an overview. Unfortunately, the simulations that supported its use did not pick up problems in testing for interaction in factorial experiments. Simulations studies such as Sawilowsky, et al. (1989) and theoretical studies such as Thompson (1991) showed the deficiencies. Akritas and Arnold (1994) clarified the nonparametric hypotheses actually tested by the rank-transform methodology. The research has come full circle for our department as we just hired one of Akritas's students, Haiyan Wang, who is doing research along these lines.

Textbook Writing

Technology has had a significant effect on the content of my two textbooks, not to mention the fact that without a word processor I would never have had the patience to write the books.

My first book written jointly with Sallie Keller-McNulty was *Concepts in Probability and Stochastic Modeling* (Higgins & Keller-McNulty, 1995). Sallie, who was recently elected president of the American Statistical Association, was an M.S. student of mine at the University of South Florida and a colleague at Kansas State prior to becoming head of the statistics group at Los Alamos Laboratories. Our book came from a course that we developed for our computer science department. We decided to use modeling rather than inference as the theme around which to organize the material. In particular we included Markov chains and some elementary queuing theory in the course and did so early enough that it would not be treated as an after thought.

To make topics like this accessible to students who were not strong mathematically

but who had programming skills, we made computer simulation of random events a key feature of the book. With this, one can ask students to investigate empirically some rather mathematically complicated random phenomena. For instance, it is a trivial matter to examine the zero crossings of a random walk by simulating 5000 or 10,000 tosses of a coin. Students find it surprising that so few crossings occur. One can approximate M/M/k queuing processes by simulating what we call Bernoulli queuing processes. Again the programming is nothing more than simulating tosses of multiple biased coins. Moreover with very little modification one can simulate non-homogeneous queuing processes and other rather complex systems. I now do the programming for the course with a spreadsheet where I not only can generate the data but graph it as well.

My other book *Introduction to Modern Nonparametric Statistics* (2004) was written for our nonparametric methods course. The audience is undergraduates and beginning graduate students in statistics and students from other areas, primarily biology, who need nonparametric methods for their research. Here again computing had a great deal to do with the approach that I took.

Many of the methods under the heading of nonparametric statistics are variations of permutation tests. The Wilcoxon rank-sum test with or without ties, the signed-rank test, the Kruskal-Wallis test, the Spearman test for correlation, the log-rank test for censored data, and various exact tests for contingency tables use some form of a permutation distribution of a statistic as the reference distribution for determining significance levels. The StatXact software, which came out in the late 1980s, was the first to exploit this in a comprehensive way.

My choice of topics for the book goes quite a bit beyond traditional rank tests, but I believe this is in keeping with a broader understanding of what now constitutes nonparametric statistics. Where possible I presented methods as special cases of permutation tests applied to scores. To deal with more complicated data structures, I included some bootstrap methods and a brief treatment of

the rank-based, robust methods of Hettmansperger and McKean (1998).

Software is now catching up with the theory of nonparametric statistics although there is still a ways to go. In the early days, nonparametric methods were thought of as quick hand calculation methods suitable only for small data sets, but in fact many of the methods are computationally intensive. I believe that we are poised to see a rapid growth in the use of nonparametric methods now that exact methods and bootstrap methods are being included in several popular software packages. Scott Richter and I are working on a book that shows how to implement many of the popular nonparametric methods in SAS.

Consulting

I hold a joint appointment at Kansas State University with the College of Arts and Sciences and the College of Agriculture. For the agriculture part of my appointment I am one of six statistical consultants for Kansas State Research and Extension, formerly the Agricultural Experiment Station. Consulting is an integral part of what our department does, and even those who don't hold consulting appointments often are involved in consulting projects. It has been the source of research problems, classroom examples, and textbooks including the popular book *Analysis of Messy Data* by Milliken and Johnson (1984).

In the 1980's a large part of our consulting centered on statistical computing. We had a large computing lab, and most of those who needed statistical computing came to the lab to get their work done. Researchers now do their own computing on their desktop or laptop computers, and computer software supports more methods than ever before. This is both good and bad. It is good because statisticians can focus their efforts at the planning stages of a study as they should. It is bad because even good researchers may choose the wrong method for their analysis, and the statistician is not there to catch the error.

Because of the changing consulting role, the notion that a statistician is someone who provides statistical computing services at the behest of a client is not as prevalent as it once was. Most of the projects that I now deal with

involve substantial issues of experimental design or sampling. It is not uncommon for me to receive credit for my contribution by being included as a co-author on scientific papers. This is a significant change in the way things were when I first began consulting.

I must comment on the controversial issue of how to evaluate the contribution of the academic consulting statistician. Is it service or is it research? In most cases the significant contribution is not in the methods that end up being used. These are often standard. Rather the contribution comes when the consulting statistician is able to recast the applied problem in such a way that it becomes apparent what methods should be applied. Even very good researchers in content areas have difficulty doing this. We should not discount the contribution of the statistician as mere service just because he or she has the education and experience to get it right. Many areas have a tradition of multiple-author papers and give due credit for them. In my opinion, we should do the same in statistics.

The Future

I don't suppose that statisticians as a group are any better equipped to discern the future than anyone else. If anything we are perhaps more cautious than most knowing the uncertainties inherent in extrapolating too far beyond the data. Thus, let me just offer an observation that many others have made. The ability of technology to produce huge amounts of high-dimensional data presents challenges for statisticians that cannot all be met with the methods that we now have. The need is apparent in such areas as engineering, genetics, space exploration, medicine, retailing, and homeland security. Even something as basic as creating data archives that can be accessed in a variety of usable forms presents significant technological and organizational challenges. In agricultural research, for instance, lack of data archiving results in a tremendous loss of information as data are discarded or lost after experiments are done and results are published. Whatever may emerge, methods for managing and analyzing large, high-dimensional databases will become increasingly important to society and one would hope to the discipline of statistics.

References

- Akritis, J. G., Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *Journal of the American Statistical Association*, 89, 336-343.
- Bancroft, T. A., Futransky, D., L., Harshbarger, B. (1958). Demand and supply of statisticians, now and in the future. *The American Statistician*, 12, 20-24.
- Blair, R. C., Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum test to that of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics*, 5, 309-335.
- Box, G. (1990). Commentary. *Technometrics*, 33, 251-252.
- Bryce, G.R., Gould, R., Notz, W. I. (2001). Curriculum guidelines for bachelor of science degrees in statistical science. *The American Statistician*, 55, 7-13.
- Conover, W. J., Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, 124-129.
- DeGroot, Morris H. (1987). A conversation with C.R.Rao. *Statistical Science*, 2, 53-67.
- Glass, G. V., Peckham, P. D., Sanders, J.R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, 42, 237-288.
- Hettmansperger, T. P., McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. Arnold, London.
- Higgins, J. J., Keller-McNulty, S. (1995). *Concepts in Probability and Stochastic Modeling*. Duxbury Press, Belmont, CA.
- Higgins, J. J. (1999). Nonmathematical statistics: a new direction for the undergraduate discipline. *The American Statistician*, 53, 1-6.
- Higgins, J. J. (2004). *An Introduction to Modern Nonparametric Statistics*. Thompson-Books/Cole, Pacific Grove, CA.
- Hunter, W. G. (1977). Some ideas about teaching design of experiments with 2^5 examples of experiments conducted by students. *The American Statistician*, 34, 140-145.

Milliken, G. A., Johnson, D.E. (1984). *Analysis of Messy Data, Volume I: Designed Experiments*. Chapman & Hall, London.

Minton, P. D. (1983). The visibility of statistics as a discipline. *The American Statistician*, 72, 284-289.

Moore, D. S. (2001a). *Statistics: Concepts and Controversies, 5th Edition*. W.H Freeman, New York.

Moore, D. S. (2001b). Undergraduate programs and the future of academic statistics. *The American Statistician*, 55, 1-6.

Ritter, M. A., Starbuck, R. R., Hogg, R.V. (2001). Advice from prospective employers on training BS statisticians, 55, 14-18.

Sawilowsky, S. S., Blair, R. C., Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, 14, 255-267.

Thomson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, 86, 410-419.

Tukey, J. (1986). Sunset salvo. *The American Statistician*, 40, 72-76.

Mentoring Doctoral Students: A Personal Perspective

Bruce W. Hall
College of Education
University of South Florida

In this brief essay, I reflect on the mentoring process based on advising over thirty doctoral students in measurement, evaluation, and research. There is considerable cause for optimism, and it is among the professors' highest honor to mentor the doctoral student.

Introduction

During my 32 years as a professor of educational research involved in graduate education at the University of South Florida, I have been privileged to assist over 200 doctoral candidates in the pursuit of their advanced degree. For 34 of those students, I served as Major Advisor. My services to the remaining doctoral students were typically as a committee member providing advise and guidance with instrumentation, sampling, statistical analysis, and other method-related issues.

Over the years, my experiences as advisor and mentor to doctoral candidates have given me cause for great optimism, and also deep concern, about the future of educational research, its production and application. My enthusiasm for the mentor-mentee relationship has at times soared on the wings of a sublime interaction, and at other times crashed under the weight of an intractable position.

Bruce W. Hall is Professor Emeritus of Measurement, Evaluation, and Research. He was twice elected President of the Florida Educational Research Association (1987 and 2003). email: bhall@tempest.coedu.usf.edu.

I can honestly say that I have never had to deal with stupidity in my mentoring duties. But I have suffered the presence of some students who were naïve about the responsibilities of candidacy, others who were obstinate in the face of needed changes in their research, and still others who were manipulative of the mentor-mentee relationship in the sense of trying to turn it into a series of negotiations to win the "best deal". I have had students who wanted their hand held through every inch of the dissertation process, others who threatened to walk out on their supervisory committee if any substantive changes to their work were expected. I have even had students who, without my knowledge or consent, attempted to replace doctoral committee members in hopes of creating a "best fit", much like one who repeatedly tries on and discards shoes in search of the shoe that doesn't pinch.

And then there are the students who bring completely unexpected idiosyncrasies to the mentoring experience. I once worked with a candidate who quickly and repeatedly responded "OK" to every suggestion I offered; After discovering that none of my suggestions was ever acted on, I slowly came to realize that his "OK" responses were nothing more than an

affectation manifested whenever he felt stressed. Such behaviors can burden the development of nurturing, constructive interactions within the mentoring context, and can quickly affect the quality of the dissertation work.

Each type of candidate reaction described above can be terribly burdensome to any professor who aspires to the role of doctoral mentor. To me, however, the behavior most troubling within the mentor-mentee arena is one that I call “unconditional discouragement”. I am speaking of candidates who appear so lacking in confidence in their dissertation-related capabilities that every question raised by the doctoral advisor, every suggestion offered becomes the impetus, maybe the excuse, for expressions of despair and defeat. A low threshold for defeat may seem a strange coping mechanism for someone who has successfully navigated the complexities of doctoral work. Yet, I have seen it used, and more than once. Its effect is one of misdirection — instead of focusing on task relevant matters, the advisor becomes focused on bolstering the candidate’s spirits, and little else gets accomplished.

Before I leave the reader convinced that my mentoring career has been a series of unrelieved disasters, let me say that for every mentoring session that was forgettable or regrettable, there have been dozens that filled me with a sense of quiet accomplishment. An effective mentoring relationship requires a certain facility with role-playing. You have to be tutor, counselor, guide, critic, coach and confidante, and you often have to assume these roles in quick succession. It also requires a profound belief in the potential of every student placed in your care.

By its nature the relationship is dynamic, continuously changing. At times it may even be intense, especially if either your student or you hold to strong positions on procedures, topics or issues. At its best, mentoring requires an openness to dialogue, the willingness to permit a free flow of ideas between the candidate and you. That necessitates a field of play on which each of you perceives the other as equal. When everything works, nothing is more stimulating. And it has worked for me many times.

Of course, the candidate must do her or his part. The interactions between doctoral advisor and candidate constitute a genuine professional linkage, the connections between the two being cemented by the candidate’s growing expertise within the field of study. With this understood, the candidate bears a significant responsibility for the success, i.e., the productivity, of the mentoring relationship. The paramount rules of mutual trust and respect must hold sway. The esteem and regard directed toward the candidate must also be directed back toward the advisor. Above all, the working relationship must rest on a foundation of honesty; if the candidate is unable to be forthright about difficulties encountered or confusions arising in her dissertation work, the advisor’s usefulness and effectiveness will be seriously compromised.

Within Greek mythology, the goddess Athena used Odysseus’s friend, Mentor, as a guise through which she became the guardian and teacher of Odysseus’s son, Telemachus. In much the same sense today, we as doctoral mentors serve as a guise through which our institutions of higher learning become entrusted with the academic care and nurturing of much of our nation’s intellectual offspring. There is no greater honor to be accorded a professor than the honor of mentor.

“Teaching” in Honor of Cliff Blair

Howard Stoker
Florida State University
The University of Tennessee

In this article, I conceptualize teaching as the profession of facilitating and stimulating learning. As “teachers”, we help students acquire learning skills that they may expand on later in their life. I review fifteen principles that facilitate effective learning.

Introduction

I don’t recall when I met Cliff Blair, and I doubt he would remember, either. It had to be in the context of an Annual Meeting of the Florida Educational Research Association. I was based in Tallahassee; he was based in Tampa. So, our paths did not cross, except for professional meetings. Hence, I cannot say that I knew him; I can only say that I knew who he was.

We carried similar responsibilities, the instruction of graduate students in the ins and outs of statistical analysis, design of experiments and the measurement of achievement. – boring topics to those who were not involved in the field. Some of us changed text books, revised notes, etc., in an attempt to keep the learners involved.

In the mid-1970s, Dr. George Aker became my boss. I think he was called the Director of some unit at FSU. I may have been Program Leader at that time. George was a

leader in Adult Education, recognized on campus, in Florida and around the world. In one publication, he was called “a rare kind of leader, who is both loved *and* respected.” It was during our association that I came to accept his beliefs, that if one is going to make a contribution to the learning of others, one must be a facilitator, or stimulator of learning, not a teacher. I bought into his philosophy, because it agreed with my own. I don’t know if Cliff ever met George, or read his papers, but I like to believe that Cliff was a supporter of that kind of philosophy.

George believed, as do I, that what we are about is growth and development in our lives. We are all individuals, have different likes and dislikes. We differ in potential for growth and differ in learning abilities. We have different attitudes, different prejudices, and different cultural preferences. Most of these are the result of prior learning, prior experiences, and environments. We are what we have learned.

Much of what we have learned, particularly in our early years, has little to do with what we need to learn in later years. In Algebra, we learned to solve quadratic equations and, yes, I “taught” that class, too. Why do you learn that skill? Well, it is important for a small group of individuals, but not for the masses.

Howard Stoker is Professor Emeritus of Educational Measurement and Evaluation at both Florida State University and the University of Tennessee. He remains active in the field. Email him at hstoker@utk.edu

Perhaps, along the way, we learn to solve problems. Perhaps there is something called transfer. The question is will my helping you learn one skill help you learn other skills?

As a learning facilitator, one must learn to recognize these differences in students, and make plans to have the learner bring those differences into play in the regular class meetings. They must feel free to express their differences in front of the facilitator and the other learners.

Although the members of a given class may have different backgrounds, they typically come to the class with a similar need to learn, or need to know. I never fooled myself as to why the students were in my class. I knew they had been *sent* there! I told them I was there for the money – sometimes producing a chuckle from the students. I told them that I had heard tales that some students had been sent to my class – or it could have been Cliff’s class – on the assumption that if they could not pass the class, then their faculty advisors would be saved from the task of flunking them.

I made a promise to students like that, that if they would *let* me, I would help them prove their advisors wrong. In essence, I was telling them that my role was that of helping them learn, not trying to “teach” them something. Most students accepted my invitation. There were a few who did not.

I was exposed to behavioral objectives, such as *The Taxonomy of Educational Objectives* and similar writings of the times. I examined them with the intent of seeing where they might fit into my philosophy of facilitating learning. I was not, perhaps, as successful as I might have been.

I did construct course objectives, and distributed them on the first day of class. One course for which I had responsibility was entitled “Analysis of Variance and the Design of Experiments.” I had 2 objectives for that class. One was that at the end of the term, the students would be able to analyze the data from an experiment.

The second was that, given the description of a problem, question(s) to be answered, the student would be able to design an experiment to answer the questions(s). (Some of the readers may remember those objectives.) I

did try to remain current in new developments, invoking computer analysis in place of desk computer analyses, as soon as it became practicable.

Meeting the objectives called for the student to learn skills beyond knowledge and comprehension in the Taxonomy classification scheme of things. For most students, the learning of the analysis skills was easily facilitated – crunching numbers comes easily for many folks. However, thinking through the techniques for designing an experiment required higher level thinking. Facilitating those skills was harder to do.

At one point, I had the class assembled in a room, for the dreaded Final Exam. I would have dispensed with it, but it was a University requirement. At the end of the period, one young lady came to the front of the room, after most everyone else had left. She literally threw her text book down on the table and said, “Show me, where in the book is the answer to this question.” I think it was Charlie Brown who said, “In the book of life, the answers are not all in the back.”

And, near the end of my tenure at FSU, during one of the last classes of the term, when I was talking about how one might apply some of the lessons that I had hoped had been learned, a young man (a doctoral student) asked, “Is this going to be on the final exam?” Alas, he did not do well on the final exam.

Dr. Aker proposed a list of givens for effective facilitators of learning. I thank him for the list, and recommend it for your consideration.

1. Try to see things as seen by the student.
2. Use reward, seldom use punishment, and never ridicule.
3. Have a deep sense of your responsibility, enjoy your work, *and* like people.
4. Feel secure in your own abilities, yet believe that you can do better.
5. Respect the dignity and worth of each individual.
6. Have a keen sense of fairness and objectivity in relating to others.
7. Accept, or try out new things and ideas.
8. Have a high level of patience.

9. Recognize the uniqueness and strengths of each individual.

10. Be sensitive to the needs, fears, problems and goals of the learners.

11. Reflect on the experiences of the learners, and try to analyze them in terms of their successes and failures.

12. Be humble in regard to your role and avoid the use of any assumed power.

13. Do not pretend to have all the answers – enjoy learning along with others.

14. Continue to expand your range of interest.

15. Be committed to your own life-long learning.

In summary, I have often been asked, “Are you a teacher?” I respond to that in the negative. In my best days, I might have been able to help you learn, but I could not teach you much of anything. Oh, I could “teach” you how to perform some skill, but with respect to statistics, educational measurement and related topics, the best that I, or Cliff, could do, was help you learn.

Appendix

Howard Stoker was introduced to educational measurement in 1949-50, when he enrolled in a test construction course, led by Dr. Robert Ebel, at the University of Iowa. In 1955, he enrolled at Purdue University, in a doctoral program in Educational Measurement, under the direction of Dr. H. H. Remmers.

In 1957, he joined the faculty at Florida State University, in the newly-formed Department of Educational Research and Testing, joining Dr. Hazen Curtis and Dr. Russ Kropp. That Department grew, in a few years, to include Drs. Jacob Beard, Robert Gagné, Gary Foster, and F. J. King, among others.

In 1984, he took early retirement from FSU, and joined the faculty of the University of Tennessee. He worked in Memphis, TN, with Dr. Raoul Arreola for four years, and then moved to Knoxville, and joined the Center for Assessment in the College of Education. He retired from University of Tennessee in 1994.

In 1996, together with Annie W. Ward and Millie Murray-Ward, they published *Educational Measurement: Origins, Theories and Explications* (University Press of America). He continues to be active in the profession.

A New Goodness-of-Fit Test for Item Response Theory

John H. Neel
Educational Policy Studies
Georgia State University

Chi-square techniques for testing goodness-of-fit in item response theory are shown to give incorrect results. A new measure, CB, based on cumulants is proposed which avoids the arbitrary nature of interval creation found in chi-square techniques. The distribution of CB is estimated using Monte Carlo techniques and critical values for testing goodness-of-fit are given.

Key Words: Goodness-of-fit, item response theory, item fit

Introduction

Item response theory (IRT) posits a functional relationship between the probability of success on a test item and an unobserved latent variable. Although one may wish for robustness, how well the many applications of IRT function is determined at least in part, and certainly in some cases completely, by how well the model fits observed data. Model fit to data on a particular test item has been judged by various chi-square techniques. Yen (1981) reviewed these techniques, found similarity between several, and recommended Q1. Modifications of Q1 have been implemented in various computer programs such as Bilog (Mislevy, R.J. & Bock, R.D., 1990) and BilogMG-3 (Zimowski et al, 2004).

In this article, I review the use of chi-square in examination of item fit and show that the chi-square statistic is misleading in that it shows items to not fit when one might in fact consider the items to fit well and that it shows items to fit when one might in fact consider the item to not fit well. Next, I explain why the

various variants of chi-square have these difficulties. Then I propose a new measure of item fit based on cumulants, show why this new technique is not susceptible to the problems of the chi-square techniques, and find critical points for this technique via Monte Carlo investigation of their distribution. Finally, I list some remaining research needs on this technique.

Use of Chi-square in Item Fit

Stone (2000) summarized the typical procedures for testing fit of IRT models: “(a) Item and ability parameters are estimated; (b) A small number of ability subgroups are formed (e.g., 10) to approximate the continuous ability distribution; (c) An observed score response distribution is constructed by cross-classifying examinees using their ability estimates and score responses. Using the IRT model, the item parameter estimates and an ability level representing the discrete ability subgroups (e.g., midpoint of ability subgroup), an expected score response distribution across score categories for an item is obtained; (e) These predictions are then compared with the observed score response distribution. This comparison generally involves computing a goodness-of-fit or chi-square statistic for each individual item (e.g. Bock

John H. Neel is Department Chair in Educational Policy Studies. His research interests are in statistical power, item response theory, and C programming. Contact him at epsjhn@gsu.edu

1972;Yen 1981), and/or an examination of residuals (Hambleton & Swaminathan, 1985)."

Notation

Following common notation θ is defined as ability and $P_i(\theta_j)$ as the probability of passing item i for ability θ_j . The three-parameter logistic model and its variant two- and one-parameter models are assumed for $P_i(\theta_j)$ throughout this article:

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-1.702a_i(\theta - b_i)}} .$$

Further, U_{ij} is defined as 1 if examinee j has a correct answer to item i and 0 if not. Some additional notation is:

N - number of examinees

n_j - number of examinees with common ability θ_j

K - the number of unique ability levels

See Hambleton, Swaminathan and Rogers (1991) for further model and notation explanation.

Chi-square techniques are misleading

Like many statistical techniques the goodness-of-fit technique is susceptible to increasing sample size. As sample size increases, the tests become ever more powerful and more and more items are rejected. Figure 1 is a histogram showing the upper tail p-values associated with chi-square tests of goodness-of-fit for 1000 items. These tests come from simulated data on 20 tests of 50 items each. A three-parameter model with a lognormal distribution for b , the logistic model location parameter; an exponential distribution for a , the logistic model slope parameter, a beta distribution for c , the lower asymptote, and ability normally distributed with mean 0 and standard deviation 1 was used to create item responses for 2000 examinees on each test. Discussion and justification for the use of these distributions may be found in Baker (1992).

A three-parameter model was then fit using BilogMg. The p-values are the values from the chi-square goodness-of-fit for the items. It is clear from the figure that the p-values have positive skew. There should have been 50 ($1000 \times .05 = 50$) p-values less than .05,

however, there were 123, almost $2\frac{1}{2}$ times as many as expected. Applying a test for proportions to these data to test whether the observed proportion, .123, of p-values less than .05, differs from the expected value of .05, we find a z value of 10.59 ($p < .0000000000000001$). In the sense that the data were created from the given model, we can view all items as fitting the model. The technique clearly rejects many more items as not fitting than should have been rejected. Similarly, testing at the .01 level we would expect to reject only 10 items but 40 would have been rejected for data that has adequate fit. Other conditions, for example, number of parameters in the IRT model, distribution of ability, size of calibration sample, will affect how many the items chi-square technique incorrectly identifies. In some cases the proportion of errors can be quite large. An exploration of these conditions is not the purpose of this study. Here it is only shown that the technique can in fact err on the side of identifying too many items that do not fit. The chi-square test thus does show items not to fit when one might in fact consider the items to fit well; i.e., 123 rejections when only 50 were expected.

That the chi-square techniques can show items to fit when the items do not fit can occur when proportions passing the items are different within the same interval on the ability scale. When this happens in the same interval, proportions that are too high are combined with proportions that are too low and the items thus seem to fit. This is discussed somewhat further in the next section.

Why are the chi-square techniques misleading?

Moore (1986) lists reasons that the chi-square techniques have problems. Among these are the "arbitrariness introduced by the necessity to choose cells" and "the discarding of information within the cells". The arbitrariness of the cells is one of the main problems in the use of chi-square. As used in such statistics as Q1, equal intervals are created along the ability scale and a value of $P_i(\theta_j)$ is selected to represent the probability of success throughout the interval.

Figure 1. Upper tail p-values associated with chi-square tests of goodness-of-fit for 1000 items.

How these intervals are created is arbitrary as is the length of the interval. In a particular case, the intervals that give a particular value of chi-square might give a different value if the intervals were either of a different length, began at a different point, or were both of a different length and began at different points.

A second problem is that Q1 uses the P_i value of the midpoint of the interval on the θ scale (other values, such as the maximum, minimum, or mean, might and have been used). In using this single value to represent all points in the interval, the possibly different probabilities throughout the interval are ignored. Treating all points in the interval as having the same $P_i(\theta_j)$ discards the information from the unequal $P_i(\theta_j)$ that exist across the interval due to the different values of θ_j . This is only worsened when intervals are combined, due to low sample size as is often done in chi-square goodness-of-fit tests, because a single value of $P_i(\theta_j)$ must then represent an even larger interval across the $P_i(\theta_j)$ scale.

Moreover, differences in observed proportions passing can be masked by the selection of intervals. This can happen if the first of two adjacent regions on the ability scale show a low proportion passing while the second shows a high proportion passing. If these two successive regions are included in the same interval, the total proportion passing could be very close to the appropriate and correct value.

Proposed Measure

In an attempt to bypass the difficulty of Q1 and similar grouped statistics, the modeled cumulative proportion passing an item is contrasted to the observed cumulative proportion passing. Consider that a given test was taken by N examinees resulting in ability estimates that are arranged in order from the smallest to the largest. Some of these ability estimates may be equal for different examinees and thus we might consider that we have J unique ability estimates and that we label these as $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_J; J \leq N$ with the general element being labeled as $\hat{\theta}_j$. We then let n_j be the number of equal ability estimates at $\hat{\theta}_j$; n_j will often be 1. Using the

appropriate IRT model fit from the data, $P_i(\hat{\theta}_j)$ is the modeled probability of a correct response on item i at $\hat{\theta}_j$ and $n_j P_i(\hat{\theta}_j)$ is the modeled expected number of correct answers at $\hat{\theta}_j$. The cumulative modeled expected number of correct responses up to and including $\hat{\theta}_j$ is

$$\sum_{k=1}^j n_k P_i(\hat{\theta}_k).$$

In order to bring this cumulative modeled expected number of correct responses into a common range regardless of the difficulty of the item or the number of examinees taking the test, each of these values is divided by their maximum value,

$$MAX = \sum_{k=1}^j n_k P_i(\hat{\theta}_k),$$

thus setting the range of these values from 0 to 1 and these values represent the modeled cumulative proportion passing the item, MCP_j :

$$MCP_j = \frac{\sum_{k=1}^j n_k P_i(\hat{\theta}_k)}{MAX}.$$

MCP_j can be compared to the observed cumulative proportion passing, $OCPP_j$, by counting the number of examinees who got the item correct at each ability level, cumulate these counts at the ability levels, and divide by the MAX . Note that dividing by MAX only brings the maximum value of $OCPP_j$ to one if the total number of observed correct responses to the item is exactly equal to the cumulative modeled expected number of correct responses. This is unlikely in practice. Thus, the maximum value of $OCPP_j$ will be less than one when fewer than the total number of correct responses is obtained and it will be greater than one when more than the total number of correct responses is obtained.

The proposed measure is based upon comparisons of the differences between MCP_j and $OCPP_j$. The basic idea is to examine the area between two lines. One line is formed by plotting MCP_j at each level of ability and then

connecting these points with straight lines. The second line is formed by plotting $OCPP_j$ at each level of ability and this second set of points is also connected using straight lines. Thus two lines are created each formed from a series of straight lines. The area between the lines is then taken as a measure of how much the lines diverge. If the area between the lines is zero, the two lines must coincide everywhere. In that case $MCPP_j$ equals $OCPP_j$ at every value of $\hat{\theta}_j$. As the lines diverge from each other, the area will grow larger. This is illustrated in Figure 2. The points in Figure 2 were selected for illustration purposes. In practice, the values of θ would not be evenly spaced and would likely not have integer values. For a typical test, there would be hundreds or thousands of unequally spaced θ values. In Figure 2, there are six areas bound by the vertical lines at -3, -2, -1, 0, 1, 2, and 3. These areas are of 3 types:

Trapezoid – bounded by (-3,-2) & (1,2)

Triangle – bounded by (-2,-1), (-1,0), & (2,3)

Two triangles – as bounded by (0,1)

Formulas for the areas of these figures are well known. The only thing perhaps not well known is to find the point where the two triangles touch in the interval (0,1). This is a simple process of the simultaneous solution of the two intersecting lines, usually a topic in a beginning algebra course. A caution is to be sure that any area calculated is given a positive sign. Some areas could become negative if in finding a length of a side or an altitude, a larger value were subtracted from a smaller one. In any case, with this caution to pay attention to the signs of numbers, finding the area between the lines is a simple application of formulas for the areas of two common figures, trapezoids and triangles. The individual areas can be found and then added to obtain the total area between the two lines. I have labeled this area as CB, for Clifford Blair or the area Caught Between the lines.

I define this measure by two sources. First, CB is an area measure similar to the DIF measure defined by Raju (1988). Second, CB is an area measure that combines information from each ability level. There is no discarding of

information and there is no arbitrariness of interval location or length because there are no intervals. The discarding of the intervals has been managed by the use of the cumulants.

An Example

Table 1 lists some created data to be used as an example to illustrate the proposed techniques. Table 1 contains 7 unique values of $\hat{\theta}_j$ with 20 examinees distributed across the $\hat{\theta}_j$ values. The number of examinees at each value of $\hat{\theta}_j$ is listed under n_j . The 20 examinees were distributed across the 7 ability levels to be suggestive of a normal distribution. $P_i(\hat{\theta}_j)$ is

tabled for each value of $\hat{\theta}_j$ using a one-parameter model with $b=0$. The expected number of passes at each ability level is the number of examinees at that ability level times the probability of success at the ability level. These are listed under

$$n_j P_i(\hat{\theta}_j)$$

The cumulative expected number of passes at each ability level is the sum of the expected number of passes up to that ability level. These are listed under

$$\sum_{k=1}^j n_k P_i(\hat{\theta}_k)$$

As discussed earlier these values are divided by their maximum value, MAX, which is the last value of

$$\sum_{k=1}^j n_k P_i(\hat{\theta}_k)$$

The u_{ij} values listed in Table 1 were selected for the subjects so that the observed number of passes was always within one unit of the expected number of passes. In the sense that the observed number of passes could not be made any closer, we can say that these data fit the model. The observed cumulative proportion of passes, $OCPP_j$, was found by cumulating the number of passes up to and including an ability level and then dividing by MAX.

Table 1. Data Illustrating Good Fit.

j	$\hat{\theta}_j$	n_j	$P_i(\hat{\theta}_j)$	$n_j P_i(\hat{\theta}_j)$	$\sum_{k=1}^j n_k P_i(\hat{\theta}_k)$	$MCP_j = \frac{\sum_{k=1}^j n_k P_i(\hat{\theta}_k)}{MAX}$	u_{ij}	$\sum_{k=1}^j u_{ki}/MAX$	AREA
1	-3	1	.0000	.0060	.0060	0.0006	0	.0000	
2	-2	3	.0322	.0965	.1025	0.0106	000	.0000	0.006
3	-1	4	.1542	.6168	.7194	0.0745	0010	.1036	0.020
4	0	5	.5000	2.5000	3.2194	0.3335	10101	.4143	0.055
5	1	3	.8458	2.5374	5.7567	0.5963	101	.6215	0.053
6	2	3	.9678	2.9035	8.6602	0.8970	111	.9322	0.030
7	3	1	.9940	0.9940	9.6542	1.0000	1	1.0358	0.036
J=7			MAX =				CB= .200		
			$\sum_{k=1}^j n_k P_i(\hat{\theta}_k) =$						
			9.6542

Figure 3 represents a plot the MCPP_j and OCPP_j on the vertical axis with $\hat{\theta}_j$ on the horizontal axis. The lines are formed by connecting the MCPP_j and OCPP_j Points.

Figure 3 represents rather good fit of the data in that the observed number of passes was selected to be within one unit of the expected number of passes for each ability level. This is in contrast to Figure 4. Figure 4 was created from the data of Table 2 just as Figure 3 was created from Table 1.

Table 2 presents data created to show poor model fit by changing the u_{ij} values while keeping the same abilities and one-parameter model as Table 1. The u_{ij} values at the first three levels were selected to represent more passes than the model indicates. Accordingly the areas for the two situations differ. The CB area is found in both Table 1 and Table 2 by finding the area for the various trapezoids and triangles and then adding these areas for the CB area. The CB area for the good fit of Table 1 and Figure 3 is .200 while the CB area for the poor fit of Table 2 and Figure 4 is 2.40. This is in the direction expected. CB should be less when the fit is good and greater when the fit is poor. Comparing these two areas brings up the question of when is the fit good and when is it poor? One answer to this question is to test the hypothesis that the fit is good. In order to test that hypothesis, the probability distribution of CB needs to be known. To determine the probability distribution of CB, the distribution of the area was simulated under known conditions.

Simulations of Null Distributions

Because each of the measures proposed here is based on cumulative passing rates, there is a dependence between the OCPP_j values and the MCPP_j values. This means that finding probability distributions of these statistics through an analytic solution is difficult because of the dependencies introduced by the cumulants. Consequently a Monte Carlo simulation of the probability distributions is often used to estimate percentage points of such distributions. See Stephens (1986) for such a study. A Monte Carlo study was conducted to estimate percentage points of the distributions of the statistic proposed here for its null

distributions; i.e. using data that were generated from known models under the null hypothesis that the data fit. Since the data were created from known models these data thus always fit the model so that the null hypothesis that the data fit was always true. I simulated data for one-, two- and three-parameter logistic IRT models over all combinations of the following numbers of items and number of subjects:

Numbers of items: 10, 20, 30, 50, 75, 100, 150, 300

Number of examinees: 100, 200, 300, 500, 800, 1000, 2000, 3000, 4000

There are 240 combinations of model, number of items, and number of subjects, $3 \times 8 \times 10$. The programming was done such that each of these 240 combinations could be run without intervention. Each combination was termed a "run". For each run data was simulated until 50,000 items were available. For each test, I created the 1- 0, pass-fail, item data for the given model, estimated item parameters using BILOGMG (Mislevy, R.J. & Bock, R.D, 1990), calculated CB, and saved these statistics along with appropriate identifying information to a file. I wrote a program to find the percentage points 1, 2, . . . , 99, 99.5, 99.9, and 99.99 from these files and tabled the resulting points.

In creating the 1- 0, pass-fail, data I used a standard normal distribution for abilities; a lognormal distribution for b , the logistic model location parameter; an exponential distribution for a , the logistic model slope parameter; and a beta distribution for c , the logistic model lower slope asymptote. I checked the accuracy of the implementation of these distributions by comparing sample values from each with values from the SPSS functions for these distributions. Agreement to 4 decimal places or beyond was found in each case.

I adapted a program by Wu (1997) to use as a random number generator. I added a 1000 number shuffling routine (Press et al, 1988) to the random number generator. Without shuffling, Wu's random number generator has a period of approximately 2.3×10^{18} , more than sufficient to not repeat for the numbers used here. Addition of the shuffler increases the

Table 2. Data Illustrating Poor Fit.

j	$\hat{\theta}_j$	n_j	$P_i(\hat{\theta}_j)$	$n_j P_i(\hat{\theta}_j)$	$\sum_{k=1}^j n_k P_i(\hat{\theta}_k)$	$MCP P_j = \frac{\sum_{k=1}^j n_k P_i(\hat{\theta}_k)}{MAX}$	u_{ij}	$\sum_{k=1}^j u_{ki}/MAX$	AREA
1	-3	1	.0000	.0060	.0060	0.0006	1	0.1036	
2	-2	3	.0322	.0965	.1025	0.0106	011	0.3107	0.202
3	-1	4	.1542	.6168	.7194	0.0745	0110	0.5179	0.372
4	0	5	.5000	2.5000	3.2194	0.3335	1010	0.8287	0.469
5	1	3	.8458	2.5374	5.7567	0.5963	101	1.0358	0.467
6	2	3	.9678	2.9035	8.6602	0.8970	111	1.3466	0.444
7	3	1	.9940	0.9940	9.6542	1.0000	1	1.4501	0.449
$J=7$					$MAX = \sum_{k=1}^j n_k P_i(\hat{\theta}_k) = 9.6542$	$CB = 2.40$			

period of the random number generator and, more importantly, removes lag correlation from the generated data.

Use of the Tables

Tables 3, 4, and 5 give the .05, .01, and .001 upper area points of CB for one-, two-, and three-parameter models. These values can be used for a hypothesis test for the goodness-of-fit at significance levels of .05, .01, and .001. To conduct the test, calculate CB for a given item and then compare the item to tabled value. If CB exceeds the tabled value, then fit is rejected at the significance level for that value. If CB does not exceed the value, then fit is not rejected. As

an example, if a 50 item test is calibrated on a sample of 1000 examinees and CB for an item is found to be .015, then fit for that item would be rejected at the .05 level ($CB_{.05} = .0142$), but would not be rejected at the .01 or the .001 levels ($CB_{.01} = 0.0186$, $CB_{.001} = .0255$). Complete tables for numbers of items equal to 10, 20, 30, 50, 75, 100, 150, and 300; calibration sample sizes of 100, 200, 300, 500, 800, 1000, 2000, 3000, and 4000; and for one, two, and three parameter models may be obtained from the author. These tables list the percentage points 1-99 (in increments of .01), 99.5, 99.9, and 99.99. Four point interpolation within the table should work well so that the tables should

Table 3. Monte Carlo Estimated Upper Area Points of CB for One-Parameter Models.

N	α	K							
		10	20	30	50	75	100	150	300
100	.05	.0502	.0368	.0339	.0348	.0371	.0394	.0447	.4275
	.01	.0643	.0456	.0416	.0430	.0471	.0508	.0638	.6450
	.001	.0804	.0555	.0510	.0540	.0640	.0702	.0883	.8136
200	.05	.0495	.0333	.0274	.0252	.0260	.0269	.0289	.2141
	.01	.0634	.0421	.0339	.0307	.0322	.0341	.0384	.5251
	.001	.0774	.0519	.0414	.0374	.0413	.0467	.0584	.7110
300	.05	.0494	.0325	.0250	.0215	.0215	.0221	.0233	.1839
	.01	.0632	.0410	.0308	.0262	.0266	.0280	.0314	.4283
	.001	.0782	.0525	.0384	.0316	.0334	.0372	.0461	.5183
500	.05	.0494	.0314	.0231	.0182	.0175	.0177	.0184	.0522
	.01	.0636	.0401	.0285	.0219	.0216	.0223	.0245	.1644
	.001	.0779	.0501	.0351	.0260	.0273	.0307	.0369	.2668
800	.05	.0490	.0312	.0219	.0160	.0149	.0147	.0149	.0356
	.01	.0629	.0396	.0271	.0194	.0184	.0188	.0205	.3474
	.001	.0765	.0502	.0327	.0233	.0235	.0272	.0319	.3727
1000	.05	.0493	.0309	.0217	.0152	.0139	.0136	.0137	.0173
	.01	.0635	.0393	.0262	.0183	.0172	.0178	.0191	.2408
	.001	.0783	.0490	.0318	.0219	.0228	.0247	.0312	.3012
1500	.05	.0490	.0307	.0211	.0140	.0124	.0119	.0118	.0130
	.01	.0630	.0388	.0258	.0166	.0155	.0159	.0173	.1137
	.001	.0768	.0482	.0307	.0194	.0212	.0228	.0276	.1954
2000	.05	.0494	.0307	.0207	.0133	.0115	.0111	.0108	.0116
	.01	.0633	.0384	.0250	.0159	.0147	.0149	.0162	.0234
	.001	.0771	.0470	.0297	.0187	.0191	.0221	.0266	.2245
3000	.05	.0491	.0306	.0204	.0126	.0106	.0101	.0098	.0097
	.01	.0636	.0385	.0244	.0149	.0135	.0139	.0150	.0164
	.001	.0762	.0466	.0287	.0171	.0183	.0212	.0244	.0301
4000	.05	.0490	.0306	.0203	.0122	.0101	.0096	.0092	.0091
	.01	.0626	.0386	.0240	.0142	.0129	.0137	.0148	.0163
	.001	.0760	.0466	.0282	.0165	.0181	.0211	.0243	.0285

N - Number of examinees in the calibration sample

α - Upper tail area

K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a one-parameter logistic IRT model with a calibration sample of size N on a K item test.

Table 4. Monte Carlo Estimated Upper Area Points of CB for Two-Parameter Models.

		10	20	30	50	75	100	150	300
N	α								
100	.05	.0464	.0344	.0362	.0416	.0458	.0493	.0536	.0594
	.01	.0600	.0446	.0487	.0585	.0666	.0720	.0798	.0923
	.001	.0768	.0639	.0689	.0872	.1009	.1094	.1293	.1480
200	.05	.0478	.0286	.0253	.0272	.0303	.0318	.0351	.0390
	.01	.0635	.0363	.0324	.0365	.0423	.0445	.0507	.0585
	.001	.0802	.0480	.0436	.0520	.0589	.0651	.0714	.0909
300	.05	.0489	.0277	.0221	.0220	.0242	.0258	.0277	.0310
	.01	.0652	.0356	.0278	.0289	.0328	.0364	.0394	.0458
	.001	.0846	.0486	.0358	.0380	.0448	.0504	.0561	.0686
500	.05	.0497	.0281	.0201	.0176	.0188	.0200	.0217	.0244
	.01	.0656	.0371	.0251	.0230	.0254	.0276	.0305	.0355
	.001	.0831	.0490	.0332	.0305	.0335	.0381	.0435	.0516
800	.05	.0503	.0284	.0191	.0151	.0158	.0165	.0176	.0198
	.01	.0662	.0377	.0245	.0195	.0213	.0228	.0253	.0297
	.001	.0848	.0492	.0323	.0265	.0294	.0319	.0365	.0443
1000	.05	.0505	.0288	.0191	.0142	.0145	.0152	.0163	.0182
	.01	.0659	.0378	.0247	.0186	.0199	.0214	.0237	.0272
	.001	.0837	.0485	.0332	.0255	.0280	.0322	.0361	.0408
1500	.05	.0504	.0296	.0190	.0131	.0129	.0136	.0141	.0160
	.01	.0657	.0384	.0245	.0173	.0178	.0196	.0211	.0244
	.001	.0812	.0509	.0322	.0253	.0258	.0298	.0336	.0392
2000	.05	.0508	.0292	.0190	.0126	.0121	.0124	.0133	.0149
	.01	.0665	.0387	.0245	.0170	.0167	.0180	.0205	.0232
	.001	.0824	.0490	.0339	.0251	.0253	.0278	.0328	.0386
3000	.05	.0510	.0295	.0187	.0120	.0115	.0116	.0121	.0134
	.01	.0667	.0384	.0243	.0167	.0168	.0174	.0190	.0215
	.001	.0838	.0483	.0328	.0258	.0258	.0287	.0316	.0368
4000	.05	.0512	.0296	.0190	.0118	.0109	.0112	.0121	.0131
	.01	.0661	.0385	.0244	.0167	.0161	.0176	.0194	.0212
	.001	.0832	.0498	.0318	.0252	.0255	.0277	.0326	.0369

N - Number of examinees in the calibration sample

α - Upper tail area

K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a two-parameter logistic IRT model with a calibration sample of size N on a K item test.

Table 5. Monte Carlo Estimated Upper Area Points of CB for Three-Parameter Models

N	α	K							
		10	20	30	50	75	100	150	300
100	.05	.0954	.0797	.0761	.0766	.0801	.0832	.0893	.1009
	.01	.1112	.0911	.0886	.0919	.0990	.1044	.1168	.1374
	.001	.1274	.1039	.1034	.1161	.1325	.1414	.1610	.1894
200	.05	.0903	.0687	.0589	.0519	.0510	.0525	.0559	.0620
	.01	.1068	.0821	.0703	.0617	.0624	.0660	.0743	.0836
	.001	.1240	.0950	.0832	.0746	.0834	.0938	.1007	.1150
300	.05	.0884	.0668	.0553	.0442	.0413	.0412	.0434	.0457
	.01	.1054	.0815	.0682	.0537	.0507	.0534	.0570	.0606
	.001	.1253	.0953	.0820	.0655	.0654	.3142	.0764	.0806
500	.05	.0873	.0655	.0537	.0397	.0337	.0325	.0321	.0311
	.01	.1048	.0805	.0680	.0499	.0419	.0419	.0411	.0392
	.001	.1253	.0955	.0826	.0639	.0566	.0550	.0573	.0573
800	.05	.0863	.0655	.0533	.0378	.0307	.0288	.0266	.0239
	.01	.1037	.0815	.0687	.0487	.0391	.0371	.0337	.0299
	.001	.1217	.0969	.0837	.0648	.0512	.0482	.0444	.0422
1000	.05	.0858	.0655	.0531	.0370	.0299	.0283	.0243	.0214
	.01	.1030	.0821	.0691	.0478	.0387	.0368	.0307	.0268
	.001	.1245	.0972	.0843	.0628	.0530	.2714	.0413	.0395
1500	.05	.0858	.0654	.0532	.0369	.0289	.0260	.0224	.0173
	.01	.1034	.0820	.0694	.0476	.0374	.0342	.0281	.0216
	.001	.1242	.0963	.0851	.0621	.0497	.2714	.0426	.0366
2000	.05	.0859	.0652	.0530	.0365	.0293	.0259	.0207	.0158
	.01	.1031	.0818	.0686	.0477	.0385	.0331	.0263	.0200
	.001	.1241	.0959	.0857	.0613	.0535	.0456	.0386	.0309
3000	.05	.0848	.0653	.0532	.0360	.0290	.0258	.0189	.0145
	.01	.1020	.0823	.0701	.0472	.0376	.0324	.0241	.0184
	.001	.1242	.0964	.0910	.0630	.0481	.0448	.0388	.0352
4000	.05	.0851	.0653	.0531	.0362	.0294	.0253	.0189	.0138
	.01	.1038	.0819	.0694	.0479	.0382	.0324	.0239	.0179
	.001	.1269	.0961	.0868	.0629	.0498	.3112	.0371	.0360

N - Number of examinees in the calibration sample

α - Upper tail area

K - number of items on the test

The tabled value is the Monte Carlo estimated point that cuts off an area of α in the upper tail of the distribution of CB when the item and ability parameters were estimated for a three-parameter logistic IRT model with a calibration sample of size N on a K item test.

provide adequate support for testing items on common sized tests and with common calibration sample sizes.

Future Research

How well these procedures work will depend on many factors. One such factor is how well the assumed distributions for ability and for the parameters of the one-, two-, and three-parameter logistic item response theory model match sample data. Accordingly, some studies of fit that examine CB for real data together with the distributional assumptions made here will be important. Although each set of three points, for $\alpha = .05, .01$, and $.001$, is based on 50,000 items, a simulation with more items might be necessary to obtain better estimated upper area points.

This could be time consuming for it took about 180 days of 400 megahertz computer time to complete the Monte Carlo portion of this study. Another factor will be how well CB compares in terms of power to other procedures such as the Q1 procedure. Studies comparing the power of such procedures will help.

Yet another factor is how well the interpolation will work. That would require comparison of interpolated points from this study with values that are found by simulation just as these values were found. Finally, given the ever increasing speed of modern computing, it is probably possible to simulate any given observed situation and estimate the required percentage points required for each test of goodness-of-fit.

For example, one might assume that the estimated ability levels in a given calibration sample were correct and then find the analogous points to those in this study for use in testing goodness-of-fit. The advantage of using the estimated abilities is that they should represent the distribution of ability and thus instead of assuming a distribution of abilities, such as was done in this study, the distribution of abilities is estimated from the observed data.

This should give a procedure that is stronger in the sense that it is not necessary to make one of the assumptions that was made here. It is also possible to make a similar use of the estimated logistic model parameters and obtain a similar benefit.

Conclusion

Weaknesses of traditional chi-square tests (e.g. Q1) of goodness-of-fit in item response theory are well known and have been shown here. An attempt to avoid these weaknesses was made by basing a statistic, CB, based on cumulants. Using cumulants avoided the arbitrary creation of intervals that causes difficulties in Q1 and thus might avoid the weaknesses of such chi-square statistics. Examples of CB were given under conditions of good and poor fit. Percentage points in the probability distribution of CB were estimated from a Monte Carlo study and an example given to show the use of these points. Suggestions were made regarding additional work with CB.

References

- Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-Fit techniques*. New York: Marcel Dekker.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Mislevy, R. J. & Bock, R. D. (1990). BILOG [Computer software]. Chicago: Scientific Software, Inc.
- Moore, D. S. (1986). Tests of chi-squared type. In D'Agostino, R. B. & Stephens, M. A. (Eds.), *Goodness-of-fit techniques* (pp. 63-96). New York: Marcel Dekker.
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Betterling, W. T. (1988). *Numerical recipes in C*. New York: Cambridge University Press.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.
- Snedecor, G. W. (1956). *Statistical Methods*. Ames, Iowa: Iowa State University Press.
- Stephens, M. A. (1986). Tests based on regression and correlation. In D'Agostino, R. B. & Stephens, M. A. (Eds.), *Goodness-of-fit techniques* (pp. 63-96). New York: Marcel Dekker.

Stone, C.A. (2000). Monte-Carlo based null distribution for an alternative fit statistic. *Journal of educational measurement*, 37, 58-75.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *applied psychological measurement*, 5, 245-262.

Wu, P. (1997). Multiplicative, congruential random-number generators with multiplier $\pm 2^k1 \pm 2^k2$ and modulus 2^p-1 . *ACM transactions on mathematical software*, 23(2), 255-265.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2004). BILOG-MG 3 [Computer software]. Chicago: Scientific Software, Inc.

Appendix

I met R. Clifford Blair during my first years as a professor in the College of Education at the University of South Florida. There are two incidents I would like to relate about Cliff that may give the reader some insight into his character. The first time he was a student in my class at the University of South Florida, he explained that he was legally blind and asked if he could record the classes. I, of course, consented and he routinely recorded every class. I was concerned as to how a student with limited vision would handle some of the basic statistical formulas and other mathematics in the class. I thought of it with my own limitations and how difficult it would be for me not to be able to see things. I was not very sophisticated as to how other people used alternative methods to learn.

One of the courses Cliff took from me was a course in test construction for teachers. Students in such a course soon consider themselves great experts at test construction and are often very critical of the tests they have in that course. When I returned the first test and went over it with the class, one student became very upset at a particular multiple choice item he had missed. He said that I had said a particular thing in class and that made the item choice he had selected correct. I replied that I would never have said that because it was clearly wrong and he must have misunderstood me. Another student jumped in and said that no, I had stated it just as the first student said and he had it in his class notes. The conversation went on a bit and I was beginning to think that I really had made an

error. At the time, I was too new to want to admit such a thing. I did not want to admit to myself that I had told the class anything wrong and certainly did not want to admit it to the class. Things were going worse for me as two other students began to support the first two when Cliff raised his hand and said, "Just a minute, I have it on tape here." Now I was really in difficulty. He had the evidence and I would have to hear it in front of everyone. He pressed the play button and there it was in my own voice: exactly what I told the students I had said. They had both written it down incorrectly. I have respected and appreciated Cliff Blair ever since.

I left the University of South Florida and came to Georgia State University. After a few years I took a trip back and went to see some old friends. There was a faculty lounge that was about the size of a large classroom. The door was near one corner of the room and Cliff was seated at a table in the far corner when I walked in. He had not known that I was coming but after two or three steps into the room, he stood up, greeted me, and invited me to sit down with him. After a bit of discussion, I reminded him that he had not seen me for several years and that he did not know I was coming. "How could you recognize me", I asked. He explained first that I was far enough away that his small area of useful vision could take in most of my body and that to him I have a characteristic walk and profile. From that, he recognized me.

Cliff is a surprising man who doesn't seem to have limits. He was always an excellent student and just as good a friend. In the test question incident, he identified the class (it was three classes back as I remember) that contained the discussion, found the tape, rewound it to the right point, and had it ready to play in a very short time. He was extremely well organized in both his recall of the situation and in his collection of tapes. In the lounge incident, he showed me how well he could use the abilities he had. He has used them well and has had a productive and profitable career. He is a respected and sought instructor. I am proud to have been around as he started that career. So I am naming this technique for him as others have done (Snedecor, 1956, p. 244) to thank him for the privilege of knowing him all these years.

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

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
Bartlett 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
ROC Curves*
Scatter Plots
Scatter Plot Matrix
Surface Plots
Violin Plots

Experimental Designs

Balanced Inc. Block
Box-Behnken
Central Composite
D-Optimal Designs
Fractional Factorial
Latin Squares
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
Hotelling'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.

"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

Introducing GGUM2004

Item Response Theory Models for Unfolding

The new GGUM2004 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 GGUM2004 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. GGUM2004 is compatible with computers running updated versions of Windows 98 SE, Windows 2000, and Windows XP. The software is accompanied by a detailed technical reference manual and a new Windows user's guide. **GGUM2004 is free** and can be downloaded from:

<http://www.education.umd.edu/EDMS/tutorials>

GGUM2004 improves upon its predecessor (GGUM2000) in several important ways:

- It has a user-friendly graphical interface for running commands and displaying output.
- It offers real-time graphics that characterize the performance of a given model.
- It provides new item fit indices with desirable statistical characteristics.
- It allows for missing item responses assuming the data are missing at random.
- It allows the number of response categories to vary across items.
- It estimates model parameters more quickly.

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2004 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

DataMineltSM announces PermuteltTM 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. PermuteltTM 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). PermuteltTM 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 PermuteltTM 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, PermuteltTM is the optimal, and only, solution.

To learn more about how PermuteltTM can be used for your enterprise, and to obtain a demo version, please contact its author, J.D. Opdyke, President, DataMineltSM, at JDOpdyke@DataMinelt.com or www.DataMinelt.com.

DataMineltSM is a technical consultancy providing statistical data mining, econometric analysis, and data warehousing services and expertise to the industry, consulting, and research sectors. PermuteltTM 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.

The easy way to find open access journals

www.doaj.org

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

Aims

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

Scope

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

In DOAJ browse by subject

Agriculture and Food Sciences
Biology and Life Sciences
Chemistry
General Works
History and Archaeology
Law and Political Science
Philosophy and Religion
Social Sciences

Arts and Architecture
Business and Economics
Earth and Environmental Sciences
Health Sciences
Languages and Literatures
Mathematics and statistics
Physics and Astronomy
Technology and Engineering

Contact

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

Funded by

www.soros.org

Hosted by

LUND
UNIVERSITY
www.lu.se

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!**

Further information including submission guidelines, subscription information and details of how to obtain a free sample copy are available at

www.blackwellpublishing.com/SIGN