

5-1-2009

Vol. 8, No. 1 (Full Issue)

JMASM Editors

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

Recommended Citation

Editors, JMASM (2009) "Vol. 8, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 8: Iss. 1, Article 34.
Available at: <http://digitalcommons.wayne.edu/jmasm/vol8/iss1/34>

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

Journal Of Modern Applied Statistical Methods

Shlomo S. Sawilowsky

Editor

College of Education
Wayne State University

Harvey Keselman

Associate Editor

Department of Psychology
University of Manitoba

Bruno D. Zumbo

Associate Editor

Measurement, Evaluation, & Research Methodology
University of British Columbia

Vance W. Berger

Assistant Editor

Biometry Research Group
National Cancer Institute

John L. Cuzzocrea

Assistant Editor

Educational Research
University of Akron

Todd C. Headrick

Assistant Editor

Educational Psychology and Special Education
Southern Illinois University-Carbondale

Alan Klockars

Assistant Editor

Educational Psychology
University of Washington

Journal Of Modern Applied Statistical Methods

Invited Articles

- 2 – 9 **Rand R. Wilcox,**
 Kathleen Costa Quantile Regression: On Inferences About
the Slopes Corresponding to One, Two or
Three Quantiles

Regular Articles

- 10 – 15 **Michèle Weber,**
 Shlomo Sawilowsky Comparative Power of The Independent t,
Permutation t, and Wilcoxon Tests
- 16 – 50 **T. Mark Beasley,**
 Bruno D. Zumbo Aligned Rank Tests for Interactions in Split-
Plot Designs: Distributional Assumptions
and Stochastic Heterogeneity
- 51 – 67 **B. W. Frankland,**
 Bruno D. Zumbo Quantifying Bimodality Part 2: A Likelihood
Ratio Test for the Comparison of a Unimodal
Normal Distribution and a Bimodal Mixture
of Two Normal Distributions
- 68 – 80 **Ross M. Gosky,**
 Sujit K. Ghosh A Comparative Study of Bayesian Model
Selection Criteria for Capture-Recapture
Models for Closed Populations
- 81 – 94 **Robert H. Lyles,**
 Reneé H. Moore,
 Amita K. Manatunga,
 Kirk A. Easley Covariate-Adjusted Constrained Bayes
Predictions of Random Intercepts and
Slopes
- 95 – 99 **James B. Hittner** Effects of Population Distribution, Sample
Size and Correlation Structure on Huberty's
Effect Size R
- 100 – 109 **Terry E. Dielman** A Note on Hypothesis Tests after Correction
for Autocorrelation: Solace for the
Cochrane-Orcutt Method?
- 110 – 121 **Nol Bendermacher,**
 Pierre Souren Beyond Kappa: Estimating Inter-Rater
Agreement with Nominal Classifications
- 122 – 131 **Stan Lipovetsky** Multiple Regression in Pair Correlation
Solution
- 132 – 146 **Krishna K. Saha** Quel Test for Two Linear Restrictions in the
Nonlinear Models

147 – 160	Terry E. Dielman	Least Absolute Value vs. Least Squares Estimation and Inference Procedures in Regression Models with Asymmetric Error Distributions
161 – 172	Oya Can Mutan, Birdal Şenoğlu	A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric
173 – 193	Tiffany A. Whittaker, Carolyn F. Furlow	The Comparison of Model Selection Criteria When Selecting Among Competing Hierarchical Linear Models
194 – 199	Sinan Calik, Cemil Colak, Ayse Turan	On the Expected Values of Distribution of the Sample Range of Order Statistics from the Geometric Distribution
200 – 207	Michael Vorburger, Breda Munoz	Approximations to Power When Comparing Two Small Independent Proportions
208 – 214	James F. Reed III	Improved Confidence Intervals for the Difference between Two Proportions
215 – 225	John Cuzzocrea, Shlomo Sawilowsky	Robustness to Non-Independence and Power of the I Test for Trend in Construct Validity
226 – 232	Hiral A. Shah, Sema A. Kalaian	Which is the Best Parametric Statistical Method for Analyzing Delphi Data?
233 – 236	Makarand Ratnaparkhi, Vasant B. Waikar, Fredrick J. Schuurmann	The Bootstrap Method for the Selection of a Shrinkage Factor in Two-Stage Estimation of the Reliability Function of an Exponential Distribution
237 – 247	Vincent A. R. Camara	A New Approximate Bayesian Approach for Decision Making About the Variance of a Gaussian Distribution Versus the Classical Approach
248 – 252	Liming Guan, John P. Wendell	Bias in Stabilized Sieve Sampling

253 – 265	Walid A. Abu-Dayyeh, M. S. Ahmed, R. A. Ahmed, Hassen A. Muttlak	Some Estimators for the Population Mean Using Auxiliary Information Under Ranked Set Sampling
266 – 272	Walid A. Abu-Dayyeh, Lana Al-Rousan	On the BLUE of the Population Mean for Location and Scale Parameters of Distributions Based on Moving Extreme Ranked Set Sampling
273 – 281	Said Ali Al-Hadhrami, Amer Ibrahim Al-Omari	Bayesian Inference on the Variance of Normal Distribution Using Moving Extremes Ranked Set Sampling
282 – 288	Edward L. McCombs, Matthew E. Elam, David B. Pratt	Estimating Task Duration in PERT Using the Weibull Probability Distribution
289 – 298	Oya S. Erdogan, Levent Ozbek	Industrialization in Animal Agriculture: A Kalman Filter Analysis
299 – 305	Michael Wolf-Branigin, Hyon-Sook Suh, Star Muir, Emily S. Ihara	Applying Census Data for Small Area Estimation in Community and Social Service Planning
306 – 315	Chinmoy K. Bose	Efficiency of Canonical Discriminant Function versus Mahalanobis Distance in Differentiating Groups: Screening Ovarian Cancer in a Multivariate System Analysis Using Enzyme Markers
316 – 321	Vance Berger	A Socratic Dialogue
<i>Emerging Scholars</i>		
322 – 336	Laura Mills, Robert A. Cribbie, Wei-Ming Luh	A Heteroskedastic, Rank-Based Approach for Analyzing 2 x 2 Independent Groups Designs
337 – 354	Jinsong Chen, Jaehwa Choi	A Comparison of Maximum Likelihood and Expected A Posteriori Estimation for Polychoric Correlation Using Monte Carlo Simulation

JMASM is an independent print and electronic journal (<http://tbf.coe.wayne.edu/jmasm>), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Editorial Assistant: **Julie M. Smith**
Internet Sponsor: **Paula C. Wood**, Dean, College of Education, Wayne State University

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

INVITED ARTICLES

Quantile Regression: On Inferences about the Slopes Corresponding to One, Two or Three Quantiles



Rand R. Wilcox

University of Southern California



Kathleen Costa

University of Southern California

The problem of testing hypotheses about the slope of a quantile regression line when the sample size is small is considered. A modified bootstrap method is suggested that is found to have certain advantages over the inverse rank method recommended by Koenker (1994). A method is suggested that simultaneously controls the probability of at least one Type I error when performing two or three tests corresponding to two or three specific quantiles. Using data from actual studies, it is illustrated that the new method can yield substantially shorter confidence intervals than the rank inverse method and, even with a large sample size, the choice of method can matter.

Key words: Tests of independence, familywise error, bootstrap methods, Porteus Maze Test, Olympic athletes.

Introduction

Consider the random variables X_1, \dots, X_p, Y having some unknown $(p+1)$ -variate distribution

and let Y_γ be the conditional γ quantile of Y given X_1, \dots, X_p . When using the Koenker and Bassett (1978) quantile regression method the goal is to estimate Y_γ assuming that

$$Y_\gamma = \alpha_\gamma + \beta_{1\gamma}X_1 + \dots + \beta_{p\gamma}X_p, \quad (1)$$

where the unknown parameters $\beta_{1\gamma}, \dots, \beta_{p\gamma}$ and α_γ are estimated based on the random sample $(X_{i1}, \dots, X_{ip}, Y_i)$, $i = 1, \dots, n$. The special case $\gamma = .5$ corresponds to what is called the least absolute value regression estimator, meaning

Rand R. Wilcox is a Professor of Psychology. He is the author of seven textbooks on statistics, the most recent of which is *Basic Statistics: Understanding Conventional Methods and Modern Insights* (2009, New York, Oxford University Press). Email: rwilcox@usc.edu. Kathleen Costa is in the Department of Kinesiology. Email: kcosta@usc.edu.

that the parameters are chosen so as to minimize the sum of the absolute values of the residuals. This special case predates ordinary least squares. For a summary of results relevant to $\gamma = .5$, see Birkes and Dodge (1993). A generalization of this method to other quantiles was first considered by Koenker and Bassett (1978). Since then, many new theoretical results have been published plus methods for computing confidence intervals for the parameters (e.g., Koenker, 1994; Koenker & Xiao, 2002). S-PLUS and R provide functions for estimating the parameters, which includes confidence intervals. Although some small-sample size results on the accuracy of these confidence intervals (plus the accuracy of several other methods) were reported by Koenker (1994), the results were limited to $p = 1$, $n=50$ and a Type I error probability of $\alpha = 0.1$. Moreover, his results were limited to symmetric distributions. Thus, a goal in this article is to comment on some situations not considered by Koenker (1994). The focus in this article is on testing

$$H_0 : \beta_{1\gamma} = 0.$$

Among the situations considered here, preliminary simulations indicated that the rank inversion method, recommended by Koenker (1994), continues to give fairly accurate confidence intervals for $n = 20$ (and $p = 1$) when testing at the .05 level with $\gamma = .5, .8$ and $.9$. However, for $\alpha = .01$, problems begin to emerge. For $\gamma = .5$, and when both X and Y have standard normal distributions, simulations indicate that now the actual Type I error probability is .002. For $\gamma = .8$ the estimate is .001. This is a concern when making inferences about two or more quantiles because if the goal is to control the probability of at least one Type I error using for example the Bonferroni inequality, having Type I error probabilities well below the nominal level could result in relatively poor power. Accordingly, one goal is to suggest an alternative approach that gives more satisfactory results for this special case.

Practical Reasons for Considering Quantile Regression

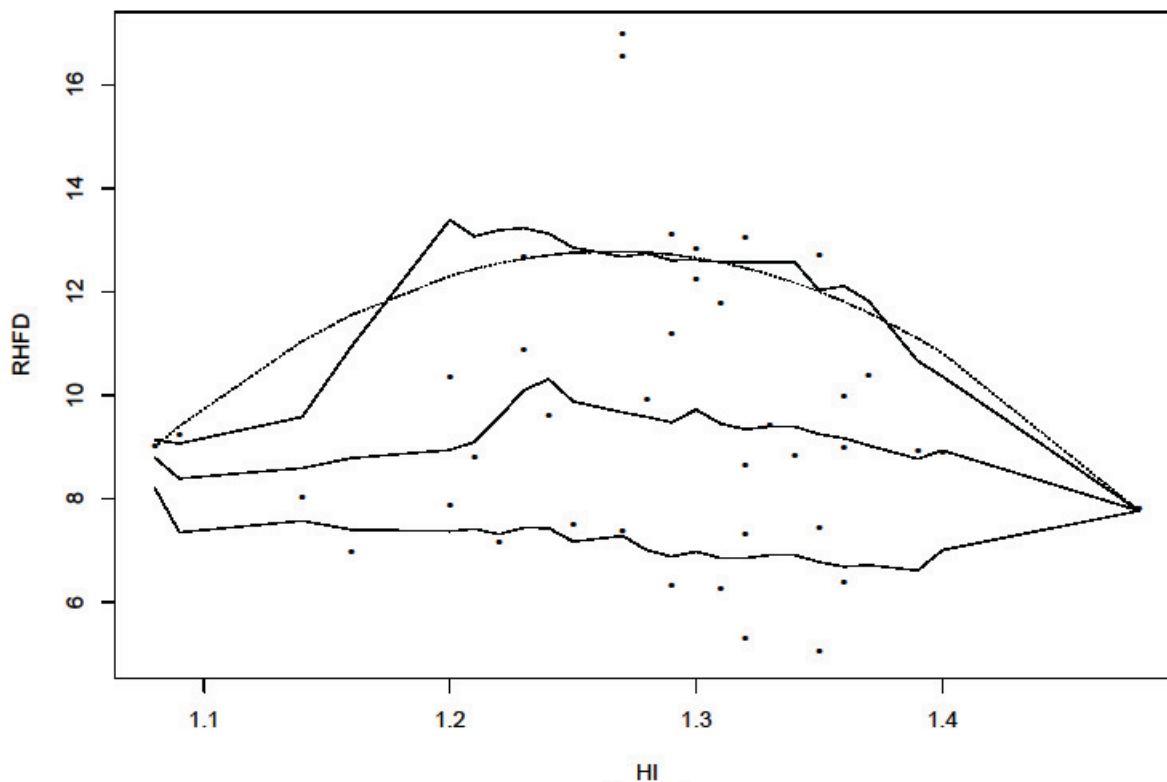
Well-known reasons exist for considering quantile regression, but two illustrations are provided that helped motivate this article. The first illustration stems from Costa (2004) where the goal was to study factors that influence increases in horizontal velocity of the body among Olympic athletes who compete in sprints. One issue of specific interest is the rate of horizontal force development (RHFD). Past studies (Henry, 1952; Payne & Blader, 1971; Mero, et al., 1983; Hafez, et al., 1985) indicate horizontal velocity at block departure is dependent on the horizontal impulse generated within the starting blocks. Faster horizontal velocities at the end of the first step out of the blocks are generated with larger net horizontal reaction forces during ground contact (Mero, 1988). These, and related results summarized in Costa (1994), led to the hypothesis that there is an association between horizontal impulse (HI) and RHFD during the first step out of the blocks.

The sample size is $n = 39$. Initial examination of the data, based on various smooths, hinted at a slightly non-linear association between HI and RHFD. Using, for example, the robust version of the smooth in Cleveland (1979), it appears that as HI increases, RHFD increases somewhat, up to a point, and then decreases. However, a test of the hypothesis that the association is linear, using the method in Stute, et al. (1998) in combination with a least squares fit, failed to reject at the .05 level. Replacing least squares with the more robust Theil (1950) and Sen (1968) regression estimator, again the hypothesis of a linear association is not rejected at the .05 level, but with only 39 pairs of points, this might be due to low power. Testing the hypothesis that the regression line is both straight and horizontal, using the wild bootstrap method in Wilcox (2005, section 9.5), the hypothesis of independence between HI and RHFD was detected at the .1 level. Simple analyses, such as Pearson's correlation and least squares regression, provided no indication of an association (Pearson's correlation is $r = -0.04$).

However, consider Figure 1, which shows a scatterplot of the data and three smooths indicated by the three solid lines. The top

QUANTILE REGRESSION SLOPE INFERENCES: ONE, TWO OR THREE QUANTILES

Figure 1: A plot of HI versus HFRD Plus the .8 and .2 Quantile Regression Lines



smooth is aimed at estimating the .8 quantile of RHFDR given HI. The middle smooth estimates the median of RHFDR and the bottom smooth is for the .2 quantile. The so-called running-interval smooth was used, as described for example in Wilcox (2005, section 11.4.4), in conjunction with the Harrell and Davis (1982) quantile estimator. (The S-PLUS function `runmq`, which comes with the library of functions in Wilcox, 2005, was used.) This suggests that as we move from the lower to the upper quantiles, a non-linear association begins to emerge. As is evident, for the .8 quantile, the association appears to be quadratic. If

$$Y_{.8} = \beta_{0,.8} + \beta_{1,.8}X + \beta_{2,.8}X^2,$$

then the estimates of $\beta_{0,.8}$, $\beta_{1,.8}$ and $\beta_{2,.8}$ are -162.63, 277.13 and -109.47, respectively. The dashed line in Figure 1 shows this fitted model, which appears to be in reasonable agreement with the corresponding smooth.

Thus, an issue is testing both $H_0 : \beta_{1,\gamma} = 0$ and $H_0 : \beta_{2,\gamma} = 0$ in a manner that controls the probability of at least one Type I error in a reasonably accurate fashion.

The second illustration demonstrates that even with n large, quantile regression can help provide a deeper understanding about any association that might exist. Williams, et al. (2005) conducted a study dealing generally with the Porteus Maze Test (PMT), which is used to evaluate intelligence and executive functioning and screen for intellectual deficiency. A portion of the study dealt with the association between the so-called Q score resulting from the PMT test and a measure of maladjustment for the participants in this study. The sample size is $n = 1063$. Pearson's correlation is $r = 0.109$, and using the usual Student's T test, the corresponding p-value is less than .001. The .5 quantile regression estimate of the slope is 0 indicating no association.

Figure 2 shows a plot of the data. The three straight lines starting from the bottom, are the .5, .8 and .9 quantile regression lines. So it appears that as we move from the median value of Y toward the higher quantiles, an association appears. Using either the inverse rank method or the method considered here, $H_0 : \beta_{1,.9} = 0$ is rejected at the .05 level. The least squares slope is estimated to be .0099, which is close to the estimated .8 quantile regression slope, which is .0098. The estimate of the slope for the .9 quantile is .029. So, although Pearson's correlation rejects at the .001 level, the quantile regression lines provide an interesting perspective on the nature of the association.

Methodology

The Koenker and Bassett (1978) quantile regression method arises as follows. For some γ , $0 < \gamma < 1$, let

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

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

$$\sum \rho_\gamma(r_i), \quad (2)$$

where $r_i = Y_i - b_{1\gamma}X_{i1} - \dots - b_{p\gamma}X_{ip} - a_\gamma$ are the residuals. Here, the values that minimize (2) were determined with the function rq that is included in the robust library included with the software S-PLUS.

The proposed method for dealing with very small sample sizes is based in part on a bootstrap estimate of the standard error. The idea of using a bootstrap estimate of the standard error is not new, but the $i = 1, \dots, n$ more obvious approximation of the null distribution of the test statistic, labeled U below, is already known to be unsatisfactory. (For general results

on bootstrap estimates of the standard error, see Buchinsky, 1994; Hahn, 1994.) More precisely, Koenker (1994) found that the actual probability coverage tended to be higher than the nominal level. Referring to his Table 2, when both X and Y have a Student's T distribution with degrees of freedom 1, 3 or 8, the actual probability coverage, when computing a .9 confidence interval, was estimated to be .920, .948 and .945, respectively. So, in terms of Type I error probabilities, the actual probability of a Type I error can be too low versus the nominal level. Very similar results were obtained here, as indicated in Section 3. One minor goal here is to expand upon Koenker's simulation study by considering sample sizes ranging between 20 and 200, a wider range of α values, and some alternative situations that include skewed distributions. A more major goal is to suggest an adjustment that helps correct the problem just described. And as previously indicated, another goal is to control the probability of at least one Type I error when two or three specific quantiles are of interest.

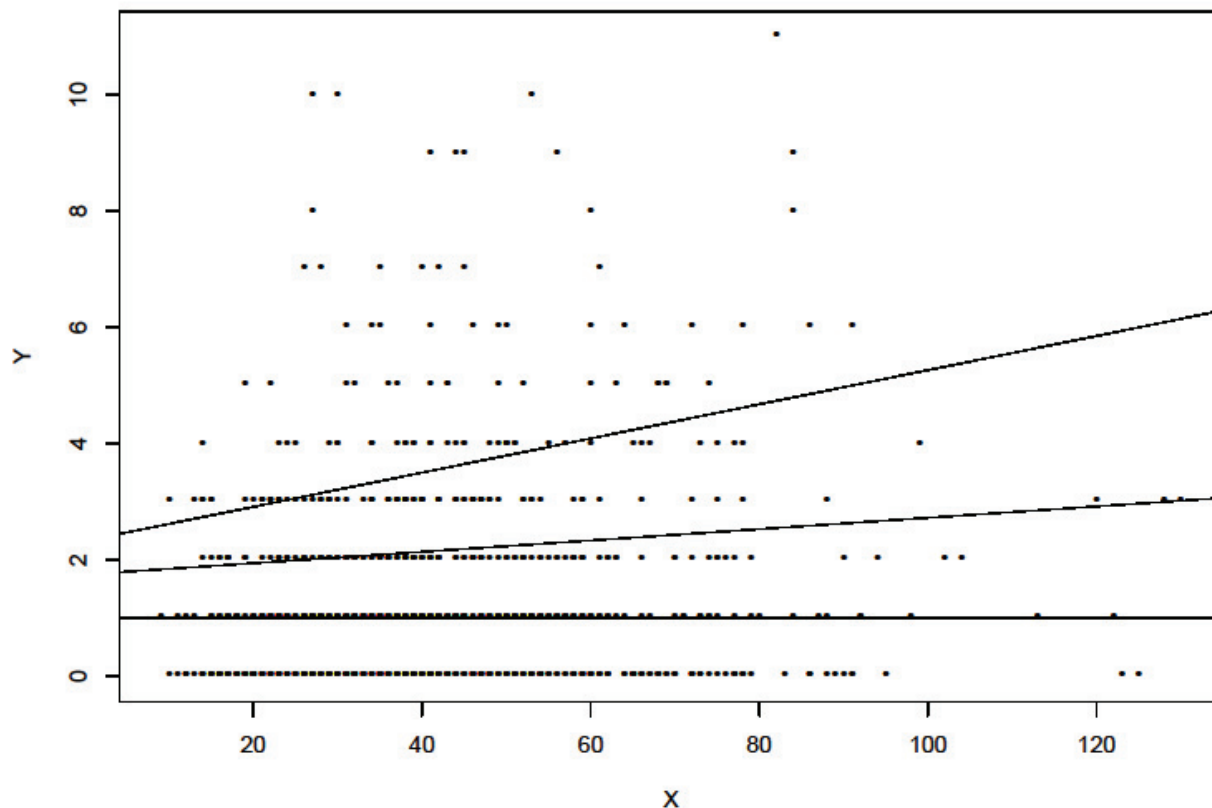
Let $(X_{i1}^*, \dots, X_{ip}^*, Y_i^*)$, $i = 1, \dots, n$, be a bootstrap sample obtained by randomly sampling, with replacement, n vectors of observations $(X_{i1}, \dots, X_{ip}, Y_i)$, $i = 1, \dots, n$. Given, γ label the resulting estimate of the slopes b_k^* , $k = 1, \dots, p$. Repeat this process B times yielding $b_{1k}^*, \dots, b_{Bk}^*$. Then from basic principles, an estimate of the squared standard error of $b_{k\gamma}$ is

$$S_k^2 = \frac{1}{B-1} \sum_{b=1}^B (b_{bk}^* - \bar{b}_k)^2,$$

where $\bar{b}_k = \sum b_{bk}^* / B$. So an approximate $1 - \alpha$ confidence interval for β_γ is $b_{k\gamma} \pm z_{1-\alpha/2} S_k$, where $z_{1-\alpha/2}$ is the $1 - \alpha / 2$ quantile of a standard normal distribution.

As previously indicated, preliminary simulations indicated that the actual probability coverage is larger than the nominal level. Here, a slight variation of Gosset's original strategy for

Figure 2: Q Scores versus a Maladjustment Score



deriving Student's T test is used in an attempt to reduce this problem. That is, assume X and Y are independent, standard normal random variables and use simulations to approximate the distribution of $U_{k\gamma} = \frac{|b_{k\gamma}|}{S_k}$. Letting $\hat{u}_{1-\alpha}$ be the

resulting estimate of the $1 - \alpha$ quantile of distribution of U, the $1 - \alpha$ confidence interval for $\beta_{k\gamma}$ is taken to be $b_{k\gamma} \pm \hat{u}_{1-\alpha} S_k$.

Consider $p = 1$; the $1 - \alpha$ quantile of the distribution of $U_{1,5}$ was estimated for $n=10, 20, 30, 40, 60, 100$ and 200 , and $\alpha = .1, .05, .025$ and $.01$ using simulations with 1,000 replications. Then a least squares estimate was fitted having the form

$$\hat{u}_{1-\alpha} = d_0 + \frac{d_1}{\sqrt{n}}$$

The resulting values for d_0 and d_1 are shown in

the top portion of Table 1. Results on how this approximation performs under non-normality are given in the next section of this paper. Still assuming $p = 1$, next consider the goal of making inferences about the slope corresponding to two different choices for $\gamma: .2$ and $.8$.

Furthermore, the goal is to control the probability of at least one Type I error (cf. Koenker & Machado, 1999). The strategy now is to approximate the null distribution of $\max(U_{1,2}, U_{1,8})$. This was also done for $n=10, 20, 30, 40, 60, 100$ and 200 , and $\alpha=.1, .05, .025$ and $.01$. The resulting values for d_0 and d_1 are shown in the middle portion of Table 1.

Finally, consideration was given to the goal where three choices for γ are to be used, namely, $.2, .5$ and $.8$. The idea is that any one choice for γ might miss an association that would be detected if a different choice were used, and again there is the goal of controlling the probability of at least one Type I error. The

resulting values for d_0 and d_1 are shown in the bottom portion of Table 1.

Table 1: Values for d_0 and d_1

α	d_0	d_1
$\gamma=0.5$		
.100	1.645	-1.19
.050	1.96	-1.37
.025	2.24	-1.18
.010	2.58	-1.69
$\gamma = (0.2, 0.8)$		
.100	1.98	-1.13
.050	2.37	-1.56
.025	2.60	-1.04
.010	3.02	-1.35
$\gamma = (0.2, 0.5, 0.8)$		
.100	2.14	-1.31
.050	2.49	-1.49
.025	2.86	-1.52
.010	3.42	-1.85

Simulation Study

Simulations were used to study the small-sample properties of the methods just described, where the critical value is taken to be $\hat{u}_{1-\alpha}$. The distribution for X was taken to be standard normal and the distribution for Y was taken to be one of four g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$Y = ((\exp(gZ) - 1) / g)\exp(hZ^2 / 2)$$

if $g > 0$, and $Y = Z\exp(hZ^2 / 2)$ if $g = 0$, has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0.0$), a symmetric heavy-tailed distribution ($h = 0.2, g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0, g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each

distribution considered. Additional properties of g-and-h distributions are summarized by Hoaglin (1985).

Table 2: Some Properties of the g-and-h Distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Table 3 shows the estimated probability of a Type I error when testing at the .05 or .01 level with $n = 20$. The estimates are based on 1,000 replications. From Robey and Barcikowski (1992), 1,000 replications is sufficient from a power point of view. More specifically, if we test the hypothesis that the actual Type I error rate is .05, and if we want power to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required. As is evident, all indications are that reasonable control over the probability of a Type I error is obtained. Similar results were obtained when for a fixed γ , the goal is to test $H_0 : \beta_{1,\gamma} = 0$ and $H_0 : \beta_{2,\gamma} = 0$, or the three hypotheses $H_0 : \beta_{1,\gamma} = 0$, $H_0 : \beta_{2,\gamma} = 0$ and $H_0 : \beta_{3,\gamma} = 0$, provided that when n is small, $.2 \leq \gamma \leq .8$; for brevity, the results are not reported.

Table 3: Probability of at Least One Type I Error, $n = 20$

g	h	$\alpha = 0.05$	$\alpha = 0.01$
0.0	0.0	0.061	0.011
0.0	0.2	0.056	0.008
0.2	0.0	0.064	0.011
0.2	0.2	0.056	0.007

QUANTILE REGRESSION SLOPE INFERENCES: ONE, TWO OR THREE QUANTILES

Note that when dealing with the case $p=2$ or 3 , the method used here can be used to control the probability of at least one Type I error when testing simultaneously the two hypotheses $H_0 : \beta_{1,\gamma} = 0$ and $H_0 : \beta_{2,\gamma} = 0$, or the three hypotheses $H_0 : \beta_{1,\gamma} = 0$, $H_0 : \beta_{2,\gamma} = 0$ and $H_0 : \beta_{3,\gamma} = 0$. It is noted that the simulations were repeated when testing these two hypotheses and it was found that the values in Table 3 can be used provided that, when n is small, $.2 \leq \gamma \leq .8$.

Comments and Illustrations Regarding Confidence Intervals

Based purely on simulations, there seems to be little separating the rank inverse method recommended by Koenker (1994) and the bootstrap method used here when $\alpha = .05$. It is illustrated, however, that when working with real data, the two methods can yield substantially different results.

Consider again the Olympic athlete data and the model $Y_s = \beta_{0,s} + \beta_{1,s}X + \beta_{2,s}X^2$. Using the rank inverse method, the .95 confidence intervals for $\beta_{1,s}$ and $\beta_{2,s}$ are $(-1.798(10)^{308}, 320.245)$ and $(-143.70, 4903.07)$, respectively. By contrast, using the bootstrap method, the .95 confidence intervals are $(25.95, 528.32)$ and $(-208.28, -10.66)$. Not only do the methods give different results when testing $H_0 : \beta_{1,s} = 0$ testing, the length of the confidence intervals differ substantially. A similar result is obtained when testing $H_0 : \beta_{2,s} = 0$, only now the difference between the lengths of the confidence intervals is less dramatic.

Returning to the Porteus maze data, consider the model $Y_s = \beta_{0,s} + \beta_{1,s}X$. The estimate of $\beta_{1,s}$ is zero and using the standard method, the .95 confidence interval is $(-0.235, 0.000)$. Using the bootstrap method studied here, the .95 confidence interval is $(-0.126, 0.126)$. So the standard method is unusual in the sense that the upper end of the confidence interval is equal to the estimated slope. For the .9 quantile, the

estimate of the slope is -0.133 and the standard method gives a .95 confidence interval of $(-0.247, -0.007)$. Now the point estimate is near the center of the confidence interval. The bootstrap confidence interval is $(-0.253, -0.014)$.

Conclusion

It is noted that some additional methods and situations were considered beyond those already described. Simulations were run with $\gamma = .1$, but now the null distribution of U was found to be rather unstable as a function of the distributions used to generate the data when the sample size is small. A percentile bootstrap method was considered, but it was found to be considerably less satisfactory in terms of probability coverage. The main point is that the adjusted bootstrap method considered here appears to perform reasonably well even under what would seem like extreme departures from normality. Moreover, both methods considered here seem to perform well when sampling from skewed distributions. Generally, when X and Y are independent, the choice between the two methods considered seems to make little difference, but when there is an association, this might no longer be the case, as was illustrated. Finally, R and S-Plus software is available from the author for applying the bootstrap method studied here. Ask for the function `qregci`.

References

- Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression*. NY: Wiley.
- Buchinsky, M. (1991). *The theory and practice of quantile regression*. Ph.D. Thesis, Dept. of Economics, Harvard University.
- Costa, K. E. (2004). *Control and dynamics during horizontal impulse generation*. Ph.D. Thesis, Dept. of Kinesiology, University of Southern California.
- Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, 11, 105-121.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes*, 461-515. NY: Wiley.

WILCOX & COSTA

Koenker, R. (1994). Confidence intervals for regression quantiles. In P. Mandl & M. Huskova (Eds.), *Asymptotic statistics*, 349-359. Proceedings of the Fifth Prague Symposium.

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

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

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

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45, 283-288.

Williams, N., et al. (2005). *Porteus' mazes and executive function in children: Standardized administration and scoring, and relationships to childhood aggression and delinquency*. Unpublished manuscript, Dept. of Psychology, University of Southern California.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing (2nd Ed.)*. San Diego, CA: Academic Press.

Comparative Power Of The Independent t, Permutation t, and Wilcoxon Tests

Michèle Weber
Private Scholar

Shlomo Sawilowsky
Wayne State University

The nonparametric Wilcoxon Rank Sum (also known as the Mann-Whitney U) and the permutation t-tests are robust with respect to Type I error for departures from population normality, and both are powerful alternatives to the independent samples Student's t-test for detecting shift in location. The question remains regarding their comparative statistical power for small samples, particularly for non-normal distributions. Monte Carlo simulations indicated the rank-based Wilcoxon test was found to be more powerful than both the t and the permutation t-tests.

Key words: t test, Wilcoxon, permutation, power.

Introduction

When testing for shift in location, Blair and Higgins (1985b) and Sawilowsky (1992; see also 1990) demonstrated that the nonparametric Wilcoxon Rank Sum test (also known as the Mann-Whitney U) is more powerful than the two independent samples Student's t test for data obtained from non-normal populations. For example, the Wilcoxon test can be up to four times more powerful than the t-test when the data are sampled from an exponential distribution (Sawilowsky & Blair, 1992).

Permutation techniques are also distribution-free (Bradley, 1968; Edgington, 1995; Maritz, 1981; Mielke & Berry, 2001). In this context, they require independence (Good, 1994; Maritz, 1981), exchangeability (Boik, 1987; Commenges, 2003; Good, 2002), continuity of the distributions (Edgington, 1995), and homogeneity of variance (Boik, 1987). Regarding their power properties, Good (1994), among many other authors, postulated that permutation methods are superior in terms of comparative power as compared with nonparametric procedures.

Adams and Anthony (1996) and Ludbrook and Dudley (1998) agreed with this view, and asserted that the reason permutation tests have higher power than nonparametric counterparts is because of the use of actual data instead of ranks. However, in a Letter to the Editor published in *The American Statistician*, Higgins and Blair (2000) demurred, and countered that statistical power is not lost via ranking data.

The same point was made previously by Blair (1985), "I have never seen an assertion of parametric power superiority accompanied by a citation to support the position. This is not too surprising since the statistical literature does not support such a position" (p. 4-5). This sentiment was echoed by Sawilowsky (1993) via an analogy:

Both an accomplished opera singer sings and an off-key beginning tuba player plays dots and dashes of the International Morse code. While some may consider the opera singer's notes to be sounds of music, there is, in fact, no more information in those dots and dashes than in the off-key notes of the beginning tuba player, with respect to the code. If the complexity and subtlety of what is often imagined to be included in interval scales is noise and not

Michèle Weber is a private scholar in San Jose, California. Email: mi.fatal-weber@att.net. Shlomo Sawilowsky is a professor of educational statistics, and editor of JMASM. Email: shlomo@wayne.edu.

signal, parametric tests will have no more information available than a rank test, and will be less efficient by trying to discriminate a signal from noise when in fact there isn't any. (p. 398)

Purpose of the study

Higgins and Blair (2000) opined that the Wilcoxon test is more powerful than the permutation t-test (and Student's t-test) when testing for shift in location. They postulated that the power properties of the permutation statistic follow the spectrum of the native test, not the nonparametric alternative. The purpose of this study, therefore, is to determine if indeed the permutation t-test follows the power properties of the two independent samples Student's t, or if it is fact superior to the nonparametric Wilcoxon Rank Sum test.

The resolution of this debate will have considerable impact on real data analysis with small samples in applied research. The rationale for selecting an optimum method for statistical analysis resides in the importance of detecting a treatment effect or naturally occurring condition, even it is subtle, assuming that it exists. The ability to detect the effect is quantified by the statistical power of the test. This makes the study of the comparative power properties of the permutation technique very important in applied research, where the effect size of treatments or interventions is oftentimes very small.

Methodology

A Fortran program was written to study the properties of the two independent samples Student's t test, the permutation t test, and the Wilcoxon Rank Sum test. Nominal alpha was set to $\alpha = 0.05$. The sample sizes studied were $n_1 = n_2 = 10$; $n_1 = 5, n_2 = 15$; $n_1 = n_2 = 20$; and $n_1 = 10, n_2 = 30$. Data were drawn from a normal distribution ($\mu = 0, \sigma = 1$), exponential distribution ($\mu = \sigma = 1$) and Chi-square distribution ($df = 1$).

The Type I error portion of the study was conducted by drawing samples with replacement for the various combinations of sample sizes and distribution, conducting the hypothesis tests, recording the results, and

repeating the experiment for one million repetitions per study parameter. The power portion of the study was based on 1,500 repetitions per experiment. The reduction in repetitions was required due to the CPU time necessary for permutation intensive computations. The means were shifted by $\mu = .2\sigma, .5\sigma, .8\sigma,$ and 1.2σ of the respective distribution.

Results

Type I Error Rates

The Type I error rates, which have been extensively studied elsewhere, are briefly repeated here to demonstrate the veracity of the Fortran program. All Type I error results replicated well-known characteristics of the tests. The Student's t-test yielded conservative Type I error rates under population non-normality. For example, the Type I error rates for the exponential distribution for $n_1 = 5, n_2 = 15$ was 0.0276. Similarly, the result for the Chi-square distribution ($df = 1$) was 0.0180. However, the Type I error rates for all conditions studied for the Wilcoxon Rank Sum test and the permutation t-tests were within sampling error of nominal alpha.

Power Results

The comparative power results for the normal distribution also replicated well-known results in the literature. The t and the permutation t-tests' statistical power were nearly indistinguishable. The Wilcoxon Rank Sum test's power was either the same, or slightly less, as noted, for example, in Figure 1. As suggested by asymptotic theory, the maximum power advantage of the two t-tests over the Wilcoxon test was only about 0.04.

The results for the exponential distribution ($\mu = \sigma = 1$) with the different shifts in location, as reflected in Figure 2, demonstrates the Wilcoxon test is more powerful than the t and permutation t-tests, of which the latter two have essentially the same power. As shown in Figure 3, the power properties for the Chi-square distribution ($df = 1$) indicates the same power advantages for the Wilcoxon Rank-Sum test, with the t-test and

POWER OF THE INDEPENDENT t, PERMUTATION tT, AND WILCOXON TESTS

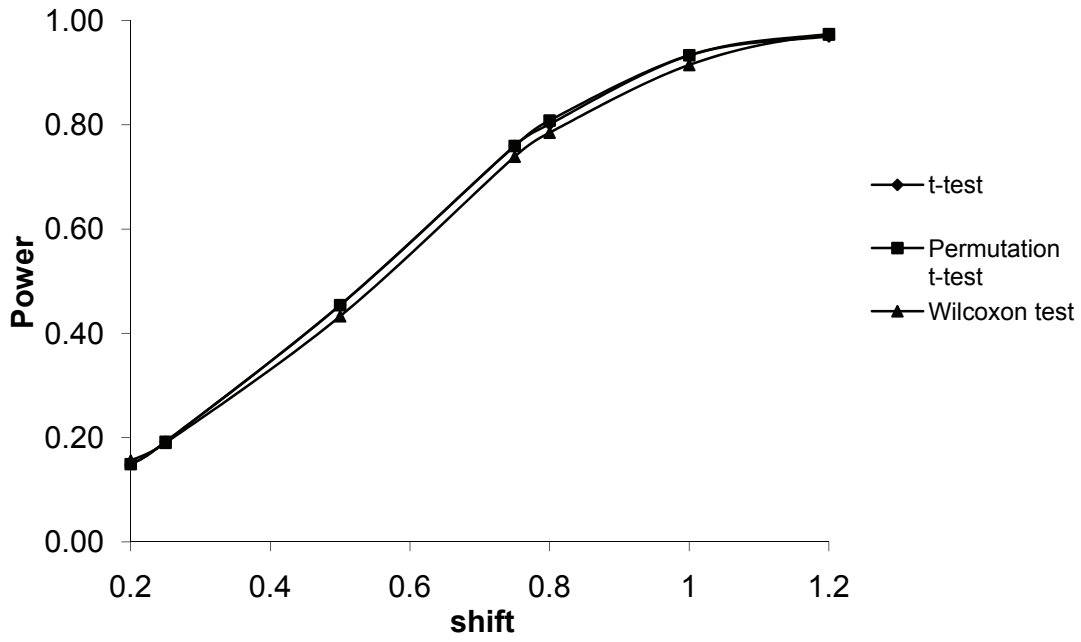


Figure 1: Shift vs. Power in the Normal Distribution for Sample Sizes $n_1 = n_2 = 20$

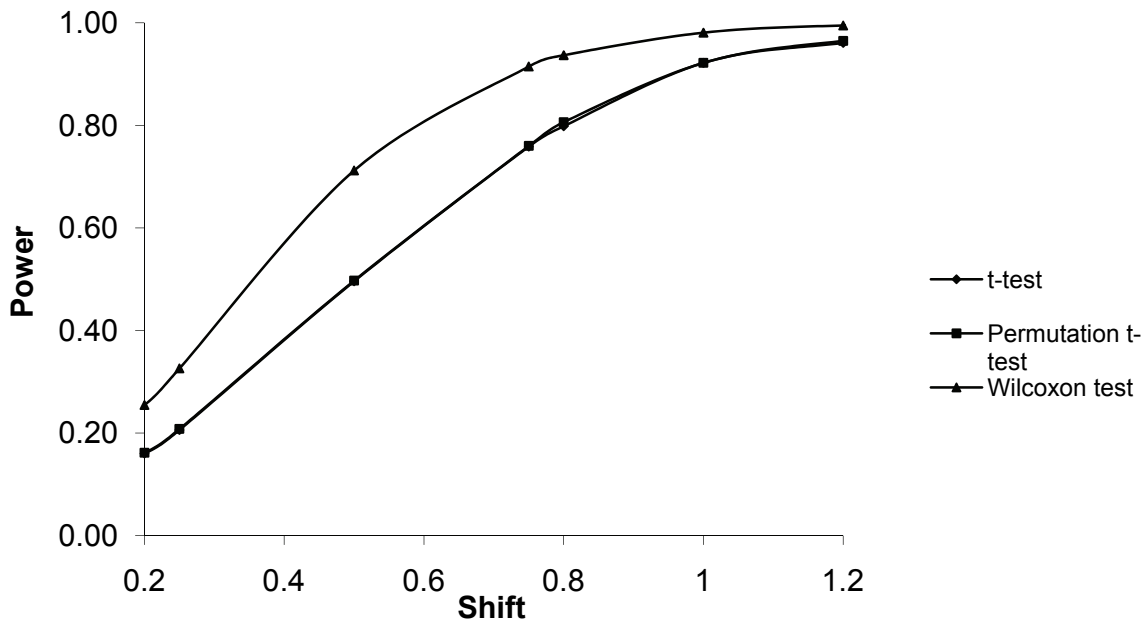


Figure 2: Shift vs. Power in the Exponential Distribution for Sample Sizes $n_1 = n_2 = 20$

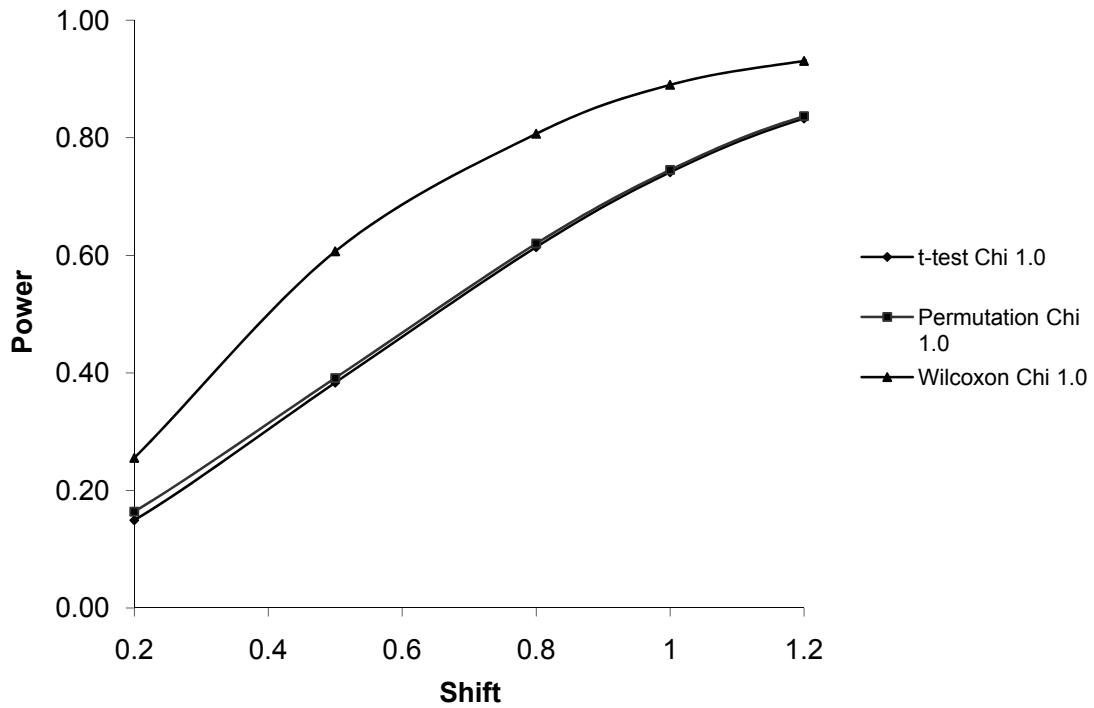


Figure 3: Shift vs. Power in the Chi-square Distribution ($df = 1$) for Sample Sizes $n_1 = n_2 = 10$

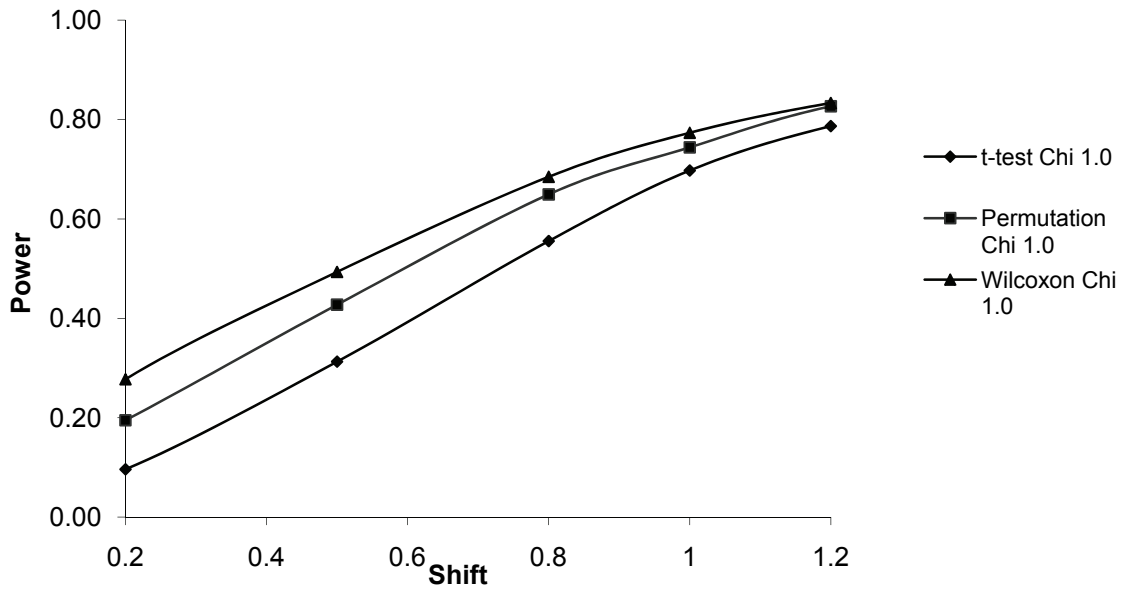


Figure 4: Shift vs. Power in the Chi-square Distribution ($df = 1$) for Sample Sizes $n_1 = 5$ & $n_2 = 15$

POWER OF THE INDEPENDENT t, PERMUTATION tT, AND WILCOXON TESTS

permutation t-test presenting nearly identical and substantially less statistical power. As indicated in Figure 4, the power results for the Chi-square distribution ($df = 1$) and unequal sample sizes indicated the permutation test became more competitive than the Student's t-test, but both tests remained considerably less powerful than the Wilcoxon Rank-Sum test.

Conclusion

Although Edgington (1995), Good (1994), and many others have presumed that the permutation t-test would be considerably more powerful than nonparametric tests, such as the Wilcoxon Rank-Sum test, the results of this Monte Carlo simulation did not support their opinion. These results pertain to the detection of a treatment modeled as a shift in location parameter, and of course, are based on the distributions, sample sizes, and the α level studied.

The primary answer provided by this simulation study is that the permutation test, in the context of the two independent samples layout, follows the depressed power spectrum of the Student's t-test, and not the superior spectrum afforded by the Wilcoxon test. Therefore, workers in applied research would be better served, when testing hypotheses of shift in location parameter, to use the nonparametric test instead of the permutation test.

Secondary results, interestingly, confirmed that the permutation t-test provides considerable power advantages over the Student's t-test for unbalanced sample sizes (e.g., Lu, Chase, & Li, 2001).

References

Adams, D. C. & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, 54(4), 733-738.

Blair, R. C. (1985, March 31-April 4). Some comments on the statistical treatment of ranks. Paper presented at the 1985 AERA/NCME annual meeting, Chicago, IL.

Blair, R. C. & Higgins, J.J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, 5, 309-335.

Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's sign-ranks test under various population shapes. *Psychological Bulletin*, 97, 119-128.

Blair, R. C., Higgins, J.J. & Smitley, W.D. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, 40, 26-42.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, 15(2), 171-185.

Edgington, E. S. (1995). *Randomization Tests*. (3rd ed). New York, NY: Marcel Dekker.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York, NY: Springer-Verlag.

Good, P. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, 1(2), 243-247.

Higgins, J. J. & Blair, R. C. (2000, February). Letter to the Editor. *The American Statistician*, 54, 86.

Hodges, J. & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*, 27, 324-335.

Lehmann, E.L. & D'Abrera, H.J. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York, NY: McGraw-Hill.

Lu, M., Chase, G. & Li, S. (2001). Permutation tests and other tests statistics for ill-behaved data: Experience of the NINDS t-PA stroke trial. *Communications in Statistics-Theory and Methods*, 30(7), 1481- 1496.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52(2), 127-133.

Maritz, J. S. (1981). *Distribution Free Methods*. London, England: Chapman and Hall.

Mielke, P. W. & Berry, K. J. (2001). *Permutation Methods: A Distance Function Approach*. New York, NY: Springer.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, 60(1), 91-126.

Sawilowsky, S. S. (1993). Comments on using alternatives to normal theory statistics in social and behavioral science. *Canadian Psychology*, 34, 398-406.

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111, 353-360.

Aligned Rank Tests for Interactions in Split-Plot Designs: Distributional Assumptions and Stochastic Heterogeneity

T. Mark Beasley
University of Alabama at Birmingham

Bruno D. Zumbo
University of British Columbia

Three aligned rank methods for transforming data from multiple group repeated measures (split-plot) designs are reviewed. Univariate and multivariate statistics for testing the interaction in split-plot designs are elaborated. Computational examples are presented to provide a context for performing these ranking procedures and statistical tests. SAS/IML and SPSS syntax code to perform the procedures is included in the Appendix.

Key words: nonparametrics, aligned ranks, split-plot design, repeated measures, stochastic heterogeneity.

Introduction

Measuring pre-treatment or baseline levels of behavior, aptitude, achievement, or pre-existing status is often necessary as a means of assessing the internal validity of applied research (Cook & Campbell, 1979). Therefore, repeated measures designs involving two or more independent groups (split-plot designs) are among the most common experimental designs in educational, psychological, developmental, and many other fields of scientific research (e.g., Keselman et al., 1998; Koch, Amara, Stokes, & Gillings, 1980). Various statistical procedures have been suggested for analyzing data from split-plot designs when parametric model assumptions are violated. The focus here is aligned rank procedures for testing the interaction.

The effects of ranking on data and the resultant test statistics for one- and two-factor designs involving only between-subjects factors

(e.g., Blair, Sawilowsky, & Higgins, 1987; Sawilowsky, Blair, & Higgins, 1989; Vargha & Delaney, 1998; Toothaker & Newman, 1994; Wilcox, 1993; Zimmerman, 1996) and single sample within-subjects designs (e.g., Agresti & Pendergast, 1986; Harwell & Serlin, 1994, 1997; Zimmerman & Zumbo, 1993) are well known. However, there have been fewer investigations concerning the effects of ranking in split-plot designs (e.g., Akritas & Arnold, 1994; Beasley, 2000, in press; Brunner & Langer, 2001; Higgins & Tashtoush, 1994; Koch, 1969).

Methodology

Parametric Models for Split-Plot Designs:
Univariate Approach

The univariate analysis of variance (ANOVA) approach to the split-plot design employs the following linear model:

$$Y_{ijk} = \mu_{***} + \beta_j + \pi_{i(j)} + \tau_k + \beta\tau_{jk} + \tau\pi_{k(j)} + \varepsilon_{ijk} \quad (1)$$

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 (repeated measures) factor, ε_{ijk} is a random

T. Mark Beasley is Associate Professor of Public Health in the Department of Biostatistics in the School of Business. Email: mbeasley@uab.edu. Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as a member of the Department of Statistics and the Institute of Applied Mathematics. Email: bruno.zumbo@ubc.ca.

error vector, and $N = \sum n_j$ is the total number of subjects. The interaction of the between-subjects (i.e., independent grouping or treatment variable) and the within-subjects (i.e., repeated measures) factors is of interest in many applications (Boik, 1993; Koch et al., 1980). In educational experiments, the interaction typically represents differential gains in achievement for a treatment group. In psychological and developmental research, the interaction indicates that independent groups do not have parallel profiles or do not exhibit identical growth curves (Winer, Brown, & Michels, 1991). In genetics experiments, the interaction typically indicates differential growth rates for organisms of different genotypes (Lynch & Walsh, 1998).

The interaction 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:

$$H_{0(J \times K)} : \sum_{j=1}^J \sum_{k=1}^K (\beta\tau_{jk})^2 = 0 \quad (2)$$

In using the parametric F -ratio for testing the interaction, the random error components (ϵ_{ijk}) are assumed to be independent and identically distributed with a mean of zero, a common variance (σ_ϵ^2), and normal shape for each of the JK cells (i.e., $NID[0, \sigma_\epsilon^2]$ for all j and k). By requiring identical error distributions, it can be assured that a rejection of the null hypothesis in (2) is due to shifts (differences) among location parameters. Furthermore, by assuming normal error distributions means as estimates of location will yield the maximum statistical power for rejecting (2).

For $K > 2$, there is an additional assumption concerning the sphericity of the pooled covariance matrix. If the pooled covariance matrix is non-spherical, the F -ratio is valid if the degrees-of-freedom (dfs) are corrected by a factor epsilon (see Huynh & Feldt, 1970). Methods for estimating epsilon have been investigated for over four decades (e.g., Box, 1954; Greenhouse & Geisser, 1959; Huynh & Feldt, 1970, 1976; Lecoutre, 1991). Also, general approximate methods to correct the dfs have been developed (Huynh, 1978).

However, these df -correction procedures tend to be less powerful than multivariate approaches to analyzing repeated measures designs (e.g., Algina & Keselman, 1998; Algina & Oshima, 1994; Keselman & Algina, 1996) and thus will not be elaborated.

Multivariate Approach

The multivariate approach to analyzing repeated measures designs (i.e., multivariate profile analysis) is often suggested because the multivariate tests do not require the additional sphericity assumption. This of great concern for repeated measures (e.g., longitudinal) designs because it seems unreasonable to make assumptions about the consistency of covariances (i.e., correlational structure) among measures taken over an extended period of time (Koch et al., 1980). One approach to conducting the multivariate profile analysis is to take pairwise differences among the K repeated measures in order to compute $(K-1)$ transformed scores, $\mathbf{Y}^* = \mathbf{YD}$, where \mathbf{Y} is the $N \times K$ data matrix of scores (Y_{ijk}) and \mathbf{D} is a $K \times (K-1)$ difference matrix of the general form:

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \quad (3)$$

These transformed scores are then submitted to a MANOVA with the following multivariate linear model:

$$\mathbf{Y}^*_j = \mathbf{M}^{**} + \mathbf{B}_j + \mathbf{E}_j, \quad (4)$$

where \mathbf{M}^{**} is a $(K-1)$ vector of grand means (centroids), \mathbf{B}_j is a $(K-1)$ vector of between-subjects effects, and \mathbf{E}_j is a random error matrix. Testing the null hypothesis ($H_{0(K)}: \mathbf{M}^{**} = \mathbf{0}_{(K-1)}$, where $\mathbf{0}_{(K-1)}$ is a $(K-1)$ vector of zeros) is equivalent to testing the repeated measures main effect. With the original scores expressed as difference scores, the multivariate model (4)

contains only between-subjects effects. Thus, the null hypothesis in (2) can be expressed as:

$$H_{0(j \times K)}: \mathbf{B}_1 = \mathbf{B}_2 = \dots = \mathbf{B}_j = \dots = \mathbf{B}_J, \quad (5)$$

where \mathbf{B}_j is a $(K-1)$ vector of between-subjects effects (i.e., mean differences) for the j^{th} group. Thus, the variables from the \mathbf{Y}^* matrix are defined as difference scores and the null hypothesis in (2) can also be expressed as:

$$H_{0(j \times K)}: (\mu_{1k} - \mu_{1k'}) = (\mu_{2k} - \mu_{2k'}) = \dots = (\mu_{jk} - \mu_{jk'}) = \dots = (\mu_{Jk} - \mu_{Jk'}), \text{ for } k \neq k'; k = 1, \dots, K. \quad (6)$$

To illustrate the assumptions underlying the multivariate approach to repeated measures data, define Σ_j as the $K \times K$ covariance matrix of \mathbf{Y}_j . The homogeneity of covariance assumption requires that the J covariance matrices (Σ_j) are equivalent so that they can be combined to form the pooled covariance matrix, Σ . Parametric tests for the multivariate model (4) assume that the random error components are independent and multivariate normal with means of zero and a common covariance matrix (i.e., $\text{NID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$).

In contrast to the univariate approach (1), the multivariate model (4) does not require homogeneity of the variances for each of the K repeated measures. That is, the multivariate approach does not require the diagonal elements of Σ to be equal. By taking difference scores this also translates into not requiring the $(K-1)$ transformed variables (\mathbf{Y}^*) to have the same variances. For example, with $K=3$ repeated measures, the variance of the first pairwise difference, $\sigma^2(Y_{j1} - Y_{j2})$, is not assumed to be equivalent to the variance of the second pairwise difference, $\sigma^2(Y_{j1} - Y_{j2})$, under the multivariate model (4); however, this variance homogeneity, which is equivalent to the sphericity requirement (see Winer, et al., 1991, pp. 240-243), is assumed implicitly in the univariate model (1).

Rank-Based Tests

Regardless of whether (a) the univariate ANOVA test with possible *df*-corrections (e.g., Huynh, 1978; Huynh & Feldt, 1976; Lecoutre, 1991), or (b) the multivariate approach to analyzing repeated measures design is employed, there are normality assumptions for parametric models. Unfortunately, the normality assumption is violated frequently in a variety of research fields including genetics (e.g., Allison et al., 1999) and behavioral research (e.g., Bradley, 1968; Cliff, 1996; Micceri, 1989; Zumbo & Coulombe, 1997).

Rank-based approaches can be used in order to relax the normality assumptions by assuming that the error components are random variables from some continuous distribution, not necessarily the normal. However, rank-based approaches cannot be simply applied due to violations of model assumptions. For example, Zimmerman and Zumbo (1993) demonstrated that rank transformed scores inherit the heterogeneity of variance in the original data. Likewise, ranks can also inherit the non-sphericity present in repeated measures data (Beasley & Zumbo, 1998; Harwell & Serlin, 1994). Thus, to test hypotheses concerning shifts in location parameters the assumptions of independence, homogeneity of variance, and identical shape must still preside (Serlin & Harwell, 2001).

Specifically, credible inferences about means require the assumption that the population distributions are symmetric (Koch, 1969; Serlin & Harwell, 2001); whereas, credible inferences concerning location parameters generally require the assumption that the population distributions are of identical shape, not necessarily symmetric (i.e., $\text{IID}[0, \sigma_\epsilon^2]$ or $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$). This frequently overlooked detail is one reason why so much attention has been given to rank-based procedures such as tests of stochastic homogeneity (Vargha & Delaney, 1998), distributional equivalence (Agresti & Pendergast, 1986; Beasley, 2000), or fully nonparametric hypotheses (Akritas & Arnold, 1994).

As a departure from parametric models that test differences among means, general

nonparametric models specifying only that observations in different cells which are governed by different distribution functions (Akritas & Arnold, 1994; Akritas, Arnold, & Brunner, 1997) have been developed for a variety of factorial designs including split-plot designs (Akritas & Arnold, 1994; Brunner & Langer, 2000). For a split-plot design, the fully nonparametric approach would involve ranking the data from 1 to NK and computing the appropriate test statistics (e.g., Serlin & Harwell, 2001).

Brunner, Domhof, and Langer (2002) warn that this practice should not be regarded as a technique for the derivation of statistics but rather as a property that can be useful for computational purposes. Therefore, fully nonparametric tests are not viewed as robust alternatives to normal theory methods, allowing direct inference concerning location parameters (Akritas, et al., 1997). Rather, statistically significant fully nonparametric tests are attributed to differences among any distributional characteristic (e.g., location, dispersion, shape). Hypotheses of this form reduce the risk of drawing incorrect conclusions about the likely sources of the significant interaction, but do so at the cost of not being able to characterize precisely how population distributions differ (Serlin & Harwell, 2001).

Rank-based tests, however, are especially sensitive to shifts in location parameters because they are computed using mean ranks. Therefore, even if assumptions concerning identical distributions and homogeneous variances are not tenable, the researcher may still conclude that one or more groups are stochastically dominant over another group(s). For an interaction in a multiple group repeated measures design, this concept of stochastic heterogeneity (Vargha & Delaney, 1998) implies that one or more groups tends to have higher scores on some measurement and that this stochastic dominance is not constant over the K measurements (Agresti & Pendergast, 1986; Brunner & Langer, 2000).

Aligned Rank Transform Procedures

Because the Rank Transform is monotonic, it is commonly believed that the null hypothesis for the parametric test of interaction

(2) from model (1) is similar to the null hypothesis for similar tests performed on ranks, except statistical inferences concern mean ranks (i.e., location parameters). However, interaction tests performed on ranked data from factorial designs have performed poorly compared with their normal theory counterparts. This is because the expected value of ranks for an observation in one cell has a non-linear dependence on the original means of the other cells (Headrick & Sawilowsky, 2000). For example, consider a two-factor model where ranks are assigned regardless of cell membership. The result is that if one of the effects is large then other effects must (because of the ranking) be small, thus producing distorted Type I and Type II error rates. Thus, a parametric test for interaction applied to ranks lacks an invariance property. Hence, interaction and main effect relationships are not expected to be maintained after rank transformations are performed (Blair, et al., 1987).

Headrick and Sawilowsky (2000) demonstrated computationally that in the presence of main effects the expected mean ranks for the cells in a factorial design can indicate an interaction when the original data do not. Moreover, Salter and Fawcett (1993) demonstrated conditions in which an interaction effect in the original data is lost in the ranking process. These situations illustrate that additivity in the original data does not imply additivity of the ranks, nor does additivity in the ranks imply additivity in the original data. Thus, Hora and Conover (1984) warned that simply ranking the data does not provide an adequate test for non-additivity (i.e., interaction) in the conventional sense of testing shifts among location parameters.

Several studies have shown that aligning the data before ranking yields better tests of the interactions among location parameters in factorial designs. Based on the work of Hodges and Lehmann (1962), 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. Higgins and Tashtoush (1994) and Koch (1969) have

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

proposed aligned rank procedures for testing interactions in split-plot designs. Based on Hollander and Sethuraman (1978), statistics for the Friedman (1937) model of ranks have been suggested as tests for interactions (Beasley, 2000; Rasmussen, 1989). Each of these procedures aligns the data in different ways.

Higgins and Tashtoush Alignment Procedure

Both the McSweeney (1967) and Hettmansperger (1984) 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. To elaborate, the data from a split-plot design has three nuisance parameters that must be removed in order to align the scores for ranking and subsequent analysis of interaction effects. Specifically, the three nuisance parameters from model (1) are the repeated measures main effect (τ_k), the between-subjects main effect (β_j), and subjects' individual differences effect that is nested in the between-subjects factor, $\pi_{i(j)}$. In terms of population effects, model (1) can be expressed as:

$$(Y_{ijk} - \mu_{***}) = \beta_j + \pi_{i(j)} + \tau_k + \beta\tau_{jk}$$

(see Winer, et al., 1991). Solving for the interaction yields:

$$\beta\tau_{jk} = (Y_{ijk} - \mu_{***}) - \beta_j - \pi_{i(j)} - \tau_k.$$

Using sample estimates of the effects yields:

$$\begin{aligned} \beta\tau_{jk} = & (Y_{ijk} - \bar{Y}_{***}) - (\bar{Y}_{*j*} - \bar{Y}_{***}) - \\ & (\bar{Y}_{ij*} - \bar{Y}_{*j*}) - (\bar{Y}_{**k} - \bar{Y}_{***}), \end{aligned} \quad (7)$$

where \bar{Y}_{**k} is the marginal mean of the k^{th} measure averaged over all N subjects, \bar{Y}_{*j*} is the marginal mean of the j^{th} measure averaged over all K measures and 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. Thus, to create scores aligned for effects other than the interaction ($\beta\tau_{jk}$) in model (1), equation (7) reduces to:

$$Y^*_{ijk} = [Y_{ijk} - \bar{Y}_{**k} - \bar{Y}_{ij*} + \bar{Y}_{***}], \quad (8)$$

These aligned scores have the nuisance effects removed so that a subsequent test performed on the ranks of Y^*_{ijk} will be sensitive only to detecting interaction effects. Higgins and Tashtoush (1994) proposed using this method of alignment and then ranking the aligned data from 1 to NK as follows:

$$A_{ijk} = \text{Rank}[Y_{ijk} - \bar{Y}_{**k} - \bar{Y}_{ij*} + \bar{Y}_{***}] \quad (9)$$

(see Table 1). Following Hettmansperger (1984), this alignment could also be accomplished by obtaining the residuals from a linear model regressing Y_{ijk} on a set of $(N-1)$ dummy codes that represent the subject effect ($\pi_{i(j)}$) and a set of $(K-1)$ contrast codes that represent the repeated-measures main effect (τ_k) from model (1). As can be inferred from (8) a set of $(J-1)$ contrast codes that represent the between-subjects main effect (β_j) is not necessary for the residualization.

Univariate Approach

Higgins and Tashtoush (1994) recommended applying the split-plot ANOVA from model (1) to the aligned ranks ($F_{(A)}$), thus replacing Y_{ijk} with A_{ijk} . As previously mentioned, many of the properties of the original data transmit to ranks, including heterogeneity of variance (Zimmerman & Zumbo, 1993) and non-sphericity (Harwell & Serlin, 1994). Therefore, it is possible that the aligned ranks could also inherit some of the distributional properties of the original data as well. Thus, when performing the split-plot ANOVA F on aligned ranks, df -correction methods (e.g., Huynh & Feldt, 1976) may be employed if the pooled covariance matrix is non-spherical or if the between-subjects covariance matrices are heterogeneous (e.g., Huynh, 1978). These methods performed on ranks hold the Type I error rate near the nominal

alpha but have low statistical power in a variety of conditions (Beasley & Zumbo, 1998).

Multivariate Approach

Agresti and Pendergast (1986) proposed a multivariate rank-based test for testing repeated measures effects in a single-sample design. Beasley (2002) extended this approach for testing the interaction in a split-plot design using aligned ranks (9). Define \mathbf{E} as a $K \times K$ pooled-sample cross-product error matrix with elements:

$$e_{kk} = \sum_{j=1}^J \sum_{i=1}^{n_j} (A_{ijk} - \bar{A}_{jk})(A_{ijk} - \bar{A}_{jk}). \quad (10)$$

Let \mathbf{E}^* be a $JK \times JK$ block diagonal matrix where the j^{th} block of the main “diagonal” for \mathbf{E}^* is defined as \mathbf{E}/n_j , and all other off-diagonal blocks are zero. That is, \mathbf{E}^* 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} , $\mathbf{E}^* = \mathbf{n} \otimes \mathbf{E}$. Also, define $\mathbf{A}_{JK} = [\bar{A}_{11}, \bar{A}_{12}, \dots, \bar{A}_{1K}, \bar{A}_{21}, \dots, \bar{A}_{2K}, \dots, \bar{A}_{J1}, \dots, \bar{A}_{JK}]'$ 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. For example, in a $J = 3 \times K = 4$ split-plot design, define:

$$\mathbf{C}_J = \begin{bmatrix} 2 & -1 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{C}_K = \begin{bmatrix} -3 & -1 & 1 & 3 \\ -1 & 1 & 1 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix}$$

$$\mathbf{C}_{JK} = \begin{bmatrix} -6 & -2 & 2 & 6 & 3 & 1 & -1 & -3 & 3 & 1 & -1 & -3 \\ -2 & 2 & 2 & -2 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -2 & 6 & -6 & 2 & 1 & -3 & 3 & -1 & 1 & -3 & 3 & -1 \\ 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 & 3 & 1 & -1 & -3 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 & 1 & -3 & 3 & -1 \end{bmatrix}.$$

It should be noted, however, that \mathbf{C}_J and \mathbf{C}_K need not be orthogonal, only linearly independent. For example, this matrix could be constructed by defining \mathbf{C}_J and \mathbf{C}_K as difference matrices in the general form of \mathbf{D} in (3), and thus,

$$\mathbf{C}_{JK} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Based on Agresti and Pendergast (1986), Beasley (2002) proposed the statistic,

$$H_{(A)} = (\mathbf{C}_{JK} \mathbf{A}_{JK})' (\mathbf{C}_{JK} \mathbf{E}^* \mathbf{C}'_{JK})^{-1} (\mathbf{C}_{JK} \mathbf{A}_{JK}). \quad (11)$$

It should be noted that $H_{(A)}$ is the Hotelling's (1931) trace for the interaction effect from a multivariate profile analysis of model (4) performed on A_{ijk} . Thus, this procedure could also be accomplished by computing $\mathbf{A}^* = \mathbf{A}\mathbf{D}$, where \mathbf{A} is the $(N \times K)$ data matrix of aligned ranks (9), and then replacing \mathbf{Y}^* with \mathbf{A}^* in the multivariate model (4).

Because it is a rank-based version of the Hotelling's trace, $H_{(A)}$ multiplied by $(N-1)$ should approximate a χ^2 distribution with $df = (J-1)(K-1)$, asymptotically. Consistent with Agresti and Pendergast (1986), transforming $H_{(A)}$ to an F -test may provide better control of Type I error rates as opposed to comparing $H_{(A)}(N-1)$ to a chi-square distribution with $df = (J-1)(K-1)$, especially with smaller sample sizes (Beasley, 2002; Harwell & Serlin, 1997). Based on Hotelling (1951), $H_{(A)}$ is transformed to an F approximation statistic by:

$$F_{H(A)} = [2(sn+1)/(s^2(2m+s+1))]H_{(A)}, \quad (12)$$

where $s = \min[(J-1), (K-1)]$, $m = [(K-J-1)/2]$, and $n = [(N-J-K)/2]$. This F approximation has numerator df_s of $df_h = [s(2m+s+1)] = [(J-1)(K-1)]$ and denominator df_s of $df_e =$

$[2(sn+1)]$. Alternatively, a critical value for $H(A)$ could be obtained from the sampling distribution of the Hotelling's trace using the s , m , and n parameters. This approach has been shown to maintain the expected Type I error rate better than the F approximate test (12) with a relatively small sample size of $N = 30$ (Beasley, 2002). Unfortunately, few multivariate texts have extensive tables of these critical values.

Koch Model of Ranking

In the Koch (1969) model, each of the K^2 paired differences among the repeated measures is ranked separately regardless of group membership. These ranks are then summed over the K levels of the repeated measures factor. To elaborate, for each of the K repeated measures, let $T_{ij(k,k')} = \text{Rank}[Y_{ijk} - Y_{ijk'}]$ using mid-ranks in case of ties. Thus, $T_{ij(k,k')}$ ranges from 1 to N , except when $k = k'$ in which case $[Y_{ijk} - Y_{ijk'}] = 0$, and thus, all values of $T_{ij(k,k)} = (N+1)/2$. Also, many of the K^2 ranked differences are reverse rankings so that the correlation between say $T_{ij(1,2)}$ and $T_{ij(2,1)}$ is -1. The final data set is defined as

$$Q_{ijk} = \sum_{k'=1}^K T_{ij(k,k')} \quad (13)$$

(see Table 2). This procedure aligns the data in a less explicit manner than the Higgins-Tashtoush method (9). Specifically, the subjects' individual differences effect that is nested in the between-subjects factor, $\pi_i(j)$ from model (1), is removed by computing pairwise differences. This is analogous to the manner in which $\pi_i(j)$ is removed from Y_{ijk} in model (1) by computing $\mathbf{Y}^* = \mathbf{YD}$ and submitting \mathbf{Y}^* to the multivariate model in (4), which only has between-subjects effects. Furthermore, by ranking each pairwise difference separately (i.e., $T_{ij(k,k')}$) before summing, the mean for each of the K measures and for all the Q_{ijk} values must equal $K(N+1)/2$. This eliminates the variance due to the repeated measures main effect (τ_k) from model (1).

To test the interaction, a univariate F -test on this ranked data $F(Q)$ could be performed (Iman, Hora, & Conover, 1984). However, Koch (1969, p. 495) proposed performing a nonparametric analog to the multivariate profile analysis, $V(Q)$. Let $\mathbf{Q}_{ij} = [Q_{ij1}, \dots, Q_{ijk}, \dots, Q_{ijK}]'$ be a $(K \times n_j)$ data matrix for the j^{th} group and let $\bar{\mathbf{Q}}_j$ be a K dimensional vector of means for the j^{th} group:

$$\bar{\mathbf{Q}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{Q}_{ij} = [\bar{Q}_{j1}, \dots, \bar{Q}_{jk}, \dots, \bar{Q}_{jK}]'. \quad (14)$$

Also, let $\tilde{\mathbf{Q}}_j = \{\bar{\mathbf{Q}}_j - K(N+1)/2\}$ be a vector of mean deviations and define the pooled covariance matrix as $\mathbf{S}_Q = 1/N [\mathbf{Q}_{ij} - K(N+1)/2][\mathbf{Q}_{ij} - K(N+1)/2]'$. The test statistic $V(Q)$ is computed as:

$$V(Q) = (N-1)/N [\tilde{\mathbf{Q}}_*' \mathbf{S}_Q^{-1} \tilde{\mathbf{Q}}_*], \quad (15)$$

where

$$\mathbf{S}_Q^* = \mathbf{n} \otimes \mathbf{S}_Q, \quad \tilde{\mathbf{Q}}_* = [\tilde{\mathbf{Q}}_1', \dots, \tilde{\mathbf{Q}}_2', \dots, \tilde{\mathbf{Q}}_J']',$$

and $\mathbf{n} = \text{diag}\{1/n_1, 1/n_2, \dots, 1/n_J\}$.

This test is a synthesis of a nonparametric multivariate statistic for the repeated measures main effect (Koch & Sen, 1968) and the Kruskal-Wallis test. In fact, it is computationally equivalent to the Pillai's (1960) trace (V) scaled by $(N-1)$. That is, a multivariate profile analysis performed on Q_{ijk} yields a Pillai's trace such that $V(Q) = V(N-1)$. Thus, this procedure could also be accomplished by computing $\mathbf{Q}^* = \mathbf{QD}$, where \mathbf{Q} is the $(N \times K)$ data matrix for the Koch model ranks (14), and then substituting \mathbf{Y}^* with \mathbf{Q}^* in the multivariate model (4).

$V(Q)$ is a permutationally distribution-free test. As sample sizes become large the number of permutations prohibits the computation of an exact test; however, the permutation distribution is χ^2 with $df = (J-1)(K-1)$ asymptotically. As an alternative approach to this statistic proposed by Koch

(1969), the Hotelling's trace could be used, thus calculating $H(Q)$, the statistic in (11), by replacing A_{ijk} (9) with Q_{ijk} (13). As before, $H(Q)$ could be transformed to an F approximation test by (12) or critical values from the multivariate referent distribution (e.g., Hotelling's trace; Pillai's trace) could be obtained in order to assess statistical significance.

Assumptions and Hypotheses for Interaction Tests Performed on Aligned Ranks

It is important to reiterate that statistically significant values of these tests performed on aligned ranks (e.g., $H(A)$, $V(Q)$) do not necessarily imply that the interaction is due to differences in location parameters unless additional assumptions are made. Strictly, statistical tests performed on aligned ranks involve inferences concerning the distribution of the original data. This is because the aligned ranks can be considered placeholders for the percentiles of the original raw score distribution (Y_{ijk}) with the nuisance location parameters removed (M. R. Harwell, personal communication, April 24, 2001). To elaborate, the univariate F -ratio performed on A_{ijk} or Q_{ijk} in a repeated measures design actually evaluates a null hypothesis of exchangeability or permutational equivalence:

$$H_0(J \times K):$$

$$\mathbf{G}_1(\mathbf{Y}_1) = \mathbf{G}_2(\mathbf{Y}_2) = \dots = \mathbf{G}_j(\mathbf{Y}_j) = \dots = \mathbf{G}_J(\mathbf{Y}_J), \quad (16)$$

where $\mathbf{G}_j(\mathbf{Y}_j)$ is the K -dimensional distribution function of the original scores for the j^{th} group (Agresti & Pendergast, 1986, p. 1418). This implies that not only are all J groups expected have identical distribution functions, the K repeated measures are also expected to have identical distribution functions (i.e., $\text{IID}[0, \sigma_\epsilon^2]$ for all j and k).

The multivariate procedures (11 or 15) test a broader null hypothesis of between-group marginal homogeneity:

$$H_0(J \times K):$$

$$\mathbf{G}_1(Y_{1k}) = \mathbf{G}_2(Y_{2k}) = \dots = \mathbf{G}_j(Y_{jk}) = \dots = \mathbf{G}_J(Y_{Jk}), \text{ for } k = 1, \dots, K, \quad (17)$$

where $\mathbf{G}_j(Y_{jk})$ is the one-dimensional distribution function of the k^{th} repeated measure for the j^{th} group (Y_{ijk}). Strictly, this is a null hypothesis of distributional equivalence across the J groups for each of the K repeated measures. That is, each of the K repeated measures may have different distribution functions, but as long as there are no distributional differences across the J groups, (17) is true. Thus, to obtain the asymptotic null distributions of the test statistics (11 or 15), it is only necessary to assume the null hypothesis (17) of between-group distributional equivalence (i.e., $\text{IID}[0, \sigma_\epsilon^2]$ for all j for each k separately or $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}]$) rather than to make stronger assumptions concerning joint (or permutational) distributions (i.e., common correlations between pairs of measures).

To illustrate, suppose that on the first and second measures in a $J = 2$ by $K = 3$ split-plot design, both groups are sampled from symmetric distributions with common variances (σ_1^2 and σ_2^2); however, both groups are sampled from identically skewed distributions with a common variance (σ_3^2) for the third repeated measurement. This situation would not violate the multivariate $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}]$ assumption; however, it would violate the univariate $\text{IID}[0, \sigma_\epsilon^2]$ assumption.

Shift Model for Aligned Ranks in Split-Plot Designs

The major purpose of the alignment process 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 processes (9) and (13) remove the mean values for the nuisance main effects, thus involving linear transformations of the data. However, both A_{ijk} and Q_{ijk} are monotone transformations of the aligned data. As a result, these aligned rank procedures do not guarantee that test statistics

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

performed on A_{ijk} or Q_{ijk} will reflect shifts in location parameters. Therefore in order to make a credible inference about a single parameter, assumptions about other parameters are necessary (Serlin & Harwell, 2001).

Assuming that all JK cells have identically shaped distributions with a common variance (i.e., $\text{IID}[0, \sigma_\varepsilon^2]$ for all j and k), then rejection of the null hypothesis (16) must be 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 in (16) as:

$$\begin{aligned} H_{0(J \times K)}: \\ G_1(\mathbf{Y}_1 - \mathbf{1}\Delta_1) = G_2(\mathbf{Y}_2 - \mathbf{1}\Delta_2) = \dots = G_J(\mathbf{Y}_J - \mathbf{1}\Delta_J) = \dots \\ = G_J(\mathbf{Y}_J - \mathbf{1}\Delta_J) \end{aligned} \quad (18)$$

where $\mathbf{1}$ is an $n_j \times 1$ vector of ones and $\Delta_j = [\delta_{j1} \ \delta_{j2} \ \dots \ \delta_{jk} \ \dots \ \delta_{jK}]$ is a $1 \times K$ vector of location parameters for the j^{th} group. To illustrate the shift model for the multivariate approach to the split-plot design, define the null hypothesis in (17) as:

$$\begin{aligned} H_{0(J \times K)}: \\ G_1(Y_{1k} - \delta_{1k}) = G_2(Y_{2k} - \delta_{2k}) = \dots = G_J(Y_{jk} - \delta_{jk}) = \dots \\ = G_J(Y_{Jk} - \delta_{Jk}), \text{ for } k = 1, \dots, K, \end{aligned} \quad (19)$$

where δ_{jk} is a scalar location parameter for the jk^{th} cell. It is important to note that 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 hypotheses in (2) or (6), 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 distribution for each of the K repeated measures is identical for each of the J groups; however, there is no assumption that the 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 (4) and the univariate approach (1) that requires the sphericity assumption (Agresti & Pendergast, 1986). Therefore, just as the null hypotheses for the univariate (2) and multivariate (6) parametric models are equivalent, differing only in the sphericity condition required by the univariate test, the same holds for the univariate (18) and multivariate (19) shift models for aligned ranks. Furthermore, note that the null hypotheses (18) and (19) are equivalent in terms of location parameters. Thus under either the univariate $\text{IID}[0, \sigma_\varepsilon^2]$ assumption or the multivariate $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$ assumption, the null hypotheses in (18) or (19), respectively, reduce to an interaction null hypothesis expressed in terms of location parameters:

$$\begin{aligned} H_{0(J \times K)}: \\ (\delta_{1k} - \delta_{1k'}) = (\delta_{2k} - \delta_{2k'}) = \dots = (\delta_{jk} - \delta_{jk'}) = \\ \dots = (\delta_{Jk} - \delta_{Jk'}) \text{ for } k \neq k'; k = 1, \dots, K, \end{aligned} \quad (20)$$

which is conceptually similar to a rejection of the parametric null hypothesis in (6). The difference between these null hypotheses is that the parametric models (1) and (4) require normally distributed error components, and thus, a rejection of (2) or (6) implies the effect must be attributed to differences among means. The shift models require identical, not necessarily normal, error distributions, and thus, a rejection of (20) implies that the effect can be attributed to differences among location parameters but not necessarily means (e.g., medians). It is important to note, however, that if (20) is false, then (18) and (19) are also false. However, a false (18) or (19) does not imply that (20) is necessarily false. That is, a significant test statistic may reflect differences in other distributional characteristics (i.e., variance or shape) rather than differences in location (Serlin & Harwell, 2001), unless these additional distributional assumptions are met.

Friedman Model of Ranks

For data from a repeated measures design, a researcher could employ the Friedman (1937) model and rank the data from 1 to K across the K levels of the repeated measures factor for each subject. The Friedman model of ranks has been applied to related samples data as well as to data originating from repeated measures designs (Zimmerman & Zumbo, 1993). The Friedman model has also been suggested when the assumptions of the split-plot ANOVA are violated (e.g., Beasley, 2000; Rasmussen, 1989). After applying the Friedman model of ranking to a split-plot design, all subjects have the same marginal mean of $(K+1)/2$. Thus, it is an attempt to eliminate the between-subjects variance (β_j) and the nested subjects variance ($\pi_{i(j)}$) in model (1) (Hollander & Wolfe, 1973, p. 143).

The Friedman model rank method does not remove the repeated measures main effect (τ_k) from model (1). Beasley (2000) demonstrated that test statistics for the Friedman model maintained the expected Type I error rate when a slight repeated measures main effect was present; however, without removing the repeated measures main effect through alignment, the statistics for testing the interaction suggested by Beasley (2000) can demonstrate low statistical power when a strong repeated measures main effect is present in each group. Aligning the data before applying Friedman ranks results in Type I error rates that are more consistent with the nominal alpha and a gain in statistical power, especially for a univariate approach (Beasley & Zumbo, in press).

To apply the Friedman ranks to data from a split-plot design, let R_{ijk} be the rank assigned to measure k for the i^{th} subject in group j after alignment (8). Also, let \bar{R}_{jk} be the mean of the ranks assigned to measure k by the subjects in group j , \bar{R}_{*k} be the mean of the ranks assigned to measure k averaged over all N subjects, and $\bar{R}_{**}=(K+1)/2$, which is the average of all NK ranks (see Table 3).

Univariate Approach

Based on Beckett and Schucany's (1979) multiple comparison tests, Beasley

(2000) demonstrated an omnibus test for the Friedman model with two or more independent groups of subjects. Based on the χ^2 analog of Scheffé's (1959) theorem (see Marascuilo, 1966), the Friedman model for $J \geq 2$ independent samples can be generalized as:

$$F(R) = \frac{\sum_{j=1}^J \sum_{k=1}^K n_j (\bar{R}_{jk} - \bar{R}_{*k})^2}{K(K+1)/12} \quad (21)$$

This test approximates a χ^2 distribution with $df=(J-1)(K-1)$, asymptotically (Beasley, 2000). However, with smaller samples sizes computing an F -ratio on R_{ijk} may be more appropriate if the covariance structure is spherical. Otherwise epsilon-adjusted tests or multivariate procedures are more appropriate (Beasley & Zumbo, in press).

Multivariate Approach

Hollander and Sethuraman (1978) developed a multivariate statistic to test for discordance in ranking patterns for $J = 2$ groups of raters. Beasley (2000) proposed an extension of this statistic for $J \geq 2$ groups. For the j^{th} group, let $\mathbf{m}_j = [(\bar{R}_{j1} - \bar{R}_{*1}), \dots, (\bar{R}_{jk} - \bar{R}_{*k}), \dots, (\bar{R}_{jK} - \bar{R}_{*K})]'$, for $j = 1, \dots, J$, be a K -dimensional column vector of deviations for the k^{th} measure for each group j . Let \mathbf{S}_R be the total sample covariance matrix of the ranks computed with ordinary least squares. Also, define \mathbf{S}_R^* as the Kronecker product of a diagonal matrix $\mathbf{n} = \text{diag}\{1/n_1, \dots, 1/n_J\}$ and \mathbf{S}_R , $\mathbf{S}_R^* = \mathbf{n} \otimes \mathbf{S}_R$. Then, the following statistic takes the general quadratic form:

$$V(R) = \mathbf{M}' \mathbf{S}_R^{*-} \mathbf{M} \quad (22)$$

where $\mathbf{M} = [\mathbf{m}_1', \dots, \mathbf{m}_j', \dots, \mathbf{m}_J']'$ is a JK column vector. Because the data matrix has a fixed mean of $(K+1)/2$, both \mathbf{S}_R and \mathbf{S}_R^* will be singular. Therefore, a generalized inverse must be employed to compute \mathbf{S}_R^{*-} . For computational purposes, it should be noted that $V(R)$ is the Pillai's trace (V) scaled by $(N-1)$. That is, a

multivariate profile analysis performed on the Friedman ranks (R_{ijk}) yields a Pillai's trace such that $V(R) = V(N-1)$, which approximates a χ^2 distribution with $df = (J-1)(K-1)$, asymptotically (Beasley, 2000). Thus, this procedure could also be accomplished by computing $\mathbf{R}^* = \mathbf{R}\mathbf{D}$, where \mathbf{R} is the $(N \times K)$ data matrix for the Friedman model ranks, and then substitute \mathbf{Y}^* with \mathbf{R}^* in the multivariate model (4). As an alternative approach to this statistic proposed by Beasley (2000), the Hotelling's trace could be used, thus calculating $H(R)$, the statistic in (11), by replacing A_{ijk} with R_{ijk} . As shown previously, $H(Q)$ could be transformed to an F approximation test by (12) or critical values from the multivariate referent distribution (e.g., Hotelling's trace; Pillai's trace) could be obtained in order to assess statistical significance.

Assumptions and Hypotheses for Interaction Tests Performed on Friedman Ranks

By using the shift model (18) and requiring the univariate model assumptions of $\text{IID}[0, \sigma_\varepsilon^2]$ for all j and k , a rejection of (18) using the univariate $F(R)$ test (21) implies that (20) is false (i.e., the interaction is due to differences in location parameters). Likewise, requiring the multivariate model assumption that the random error vectors ($\boldsymbol{\varepsilon}_{jk}$) are independent and identically distributed across the J groups for each of the K repeated measures separately (i.e., $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\boldsymbol{\Sigma}\mathbf{D}]$), a rejection of (19) using $V(R)$ implies that (20) is false. However, if these distributional assumptions are not tenable, inferences concerning shifts in location parameters are not credible. Therefore in the strictest sense, the null hypothesis in (20) applied to the Friedman model ranks implies the equality of ranking patterns across groups, which would involve a Chi-square test of homogeneity of ranking distributions in a $J \times K!$ contingency table. Analogous to the null hypotheses for aligned ranks, (20) does not imply that the probabilities of occurrence for each permutation of the ranks are equal in value across groups.

To elaborate, the univariate model null hypothesis of permutational equivalence (16)

and the multivariate model null hypothesis of distributional equivalence (17) can be formulated in terms of the probability of ranking patterns for R_{ijk} . Let ϕ_r be the r^{th} permutation of the K Friedman ranks ($r = 1, \dots, K!$). Let π_{rj} be the probability of the r^{th} permutation for subjects in the j^{th} group. Because the average rank for each individual equals $(K+1)/2$, the null hypothesis in (20) can be expressed in a form similar to (5):

$$H_0(J \times K): \Delta_1 = \dots = \Delta_j = \dots = \Delta_J, \quad (23)$$

where,

$$\Delta_j = \sum_{r=1}^{K!} \pi_{rj} \phi_r.$$

Thus, consistent with the null hypothesis in (16), the univariate $F(R)$ statistic approximates a chi-square distribution with $df = (J-1)(K-1)$ under the null hypothesis:

$H_0(J \times K)$:

$$\pi_{rj} = 1/K!, \text{ for } r = 1, \dots, K! \text{ and } j = 1, \dots, J. \quad (24)$$

Therefore, $F(R)$ (21) does not necessarily provide a test of (20) because a false (24) does not imply a false (20). It is also important to recognize that if (24) is true so are (16), (17), and (20), but (20) does not imply (24). That is, it is possible to have identical mean ranks without each permutation of ranks occurring with the same frequency. Therefore, using $F(R)$ as an approximate test may occasionally reject (20) incorrectly because (24) is false.

Likewise, $V(R)$ does not necessarily test the null hypothesis (20). The null hypothesis actually tested by $V(R)$ is:

$H_0(J \times K)$:

$$\pi_{r1} = \dots = \pi_{rj} = \dots = \pi_{rJ} \text{ for } r = 1, \dots, K! \quad (25)$$

The asymptotic distribution of $V(R)$ is χ^2 with $df = (J-1)(K-1)$ under (25) but not necessarily under (20). As with the univariate $F(R)$ test, it is

important to recognize that if (25) is true so is (20), but (20) does not imply (25). That is, it is possible for two groups to have identical mean ranks but different permutational distributions. Therefore, using $V(R)$ as an approximate test may occasionally reject (20) incorrectly because (25) is false.

It should be noted that if the univariate null hypothesis (24) is true so is the multivariate null hypothesis (25). However, if (25) is true, it does not imply that (24) is true. Likewise, a false (24) does not imply a false (25). Thus, the univariate $F(R)$ and the multivariate $V(R)$ statistics test two distinctly different, although conceptually related, hypotheses concerning the similarity of ranking patterns among multiple groups. Table 4 shows various scenarios in which these null hypotheses are true or false in a $(J=2) \times (K=3)$ split-plot design.

The multivariate model null hypothesis (25) is less restrictive than the univariate model null hypothesis (24) because $F(R)$ uses a fixed covariance structure (i.e., $K(K+1)/12$) in the denominator (Marascuilo & McSweeney, 1967), thus implying compound symmetry of the covariance matrix. Thus, the null hypothesis in (24) implies sphericity because it translates to the assumption that the errors are $\text{IID}[0, \sigma_{\epsilon}^2]$ for all j and k from the univariate model null hypothesis in (16).

Similarly, the null hypothesis in (25) translates into relaxing the assumption that all K repeated measures have identical distributions. This is analogous to the multivariate model null hypothesis in (17), which only assumes the random error components are independent and identically distributed across the J groups for each of the k measures separately (i.e., $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}]$; Hollander & Wolfe, 1973, p. 145). Thus, $V(R)$ as a multivariate test of the null hypothesis in (25) does not assume sphericity of the covariance matrix. This is because under the null hypothesis in (25) each group is not required to have $\pi_{rj} = 1/K!$, which implies a fixed covariance structure and thus sphericity.

If it is tenable to assume that the errors are $\text{IID}[0, \sigma_{\epsilon}^2]$ for all j and k , then rejections of (24) using the univariate $F(R)$ imply an interaction due to location parameters (i.e., a

false 20). Likewise, rejections of (25) using the multivariate $V(R)$ imply a false (20) if the errors are assumed to be $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}]$.

Although the univariate (24) and multivariate (25) null hypotheses for Friedman ranks can be expressed by different formulations than the univariate (18) and multivariate (19) null hypotheses for the shift model for aligned ranks, the concept of stochastic homogeneity applies to the Friedman ranks (Randles & Wolfe, 1979; Vargha & Delaney, 1998). However, if the additional distributional assumptions are not met, these statistics based on Friedman model ranks should strictly be considered test of stochastic homogeneity (Beasley, 2000; Serlin & Harwell, 2001; Vargha & Delaney, 1998).

Computational Example One

Table 1 shows hypothetical data and sample moments for a $J=2$ groups by $K=3$ repeated measures design. An educational psychology research application of this design could be a comparison of the forgetting rates over a three week period (e.g., recall measured at 7, 14, and 21 days) for children classified as slow ($j=1$) or fast ($j=2$) learners (e.g., Gentile, Voelkl, Mt. Pleasant, & Monaco, 1995). A medical psychology application would be a comparison of the addiction severity scores of opioid-dependent patients in a Day Treatment program ($j=1$) versus patients in an Enhanced Standard Methadone program ($j=2$) at three time points: Pre-treatment, Post-treatment, and Follow-up (e.g., Avants, Margolin, Sindelar, & Rounsaville, 1999).

Analyses of these data using the univariate model (1) show that the between-subjects effect was statistically significant, $F(Y)_{(1,16)} = 6.27, p = .023$. The covariance structure was non-spherical with a Greenhouse-Geisser epsilon estimate of .681. The Huynh-Feldt correction results in an epsilon estimate of .769. After a Huynh-Feldt correction to the df s, both the repeated measures main effect [$F(Y)_{(1.54, 24.61)} = 194.22, p < .001$] and the interaction effect [$F(Y)_{(1.54, 24.61)} = 12.20, p = .001$] were statistically significant. A multivariate profile analysis yielded similar findings. Both the Pillai's trace ($V(Y) = 0.936$)

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

and Hotelling's trace ($H(Y) = 14.706$) for the repeated measures main effect were statistically significant ($p < .001$). For the interaction effect, both the Pillai's trace ($V(Y) = 0.494$) and Hotelling's trace ($H(Y) = 0.977$) were statistically significant ($p = .006$) also.

Examining the moments for each of the $JK=6$ cells in Table 1, it is apparent that the data are skewed for many cells, thus potentially violating the normality assumptions of both the univariate (1) and multivariate (4) models. This provides a reason for employing rank-based tests. However, given that both the repeated measures and between-subjects main effects were statistically significant, it is necessary to align the data before ranking and subsequent analysis.

Table 1 also shows the aligned data (8) and the aligned ranks (9). Analysis of the aligned ranks showed a statistically significant interaction using the univariate model [$F_{(A)(2,32)} = 16.33, p < .001$]. The Greenhouse-Geisser epsilon estimate was .839 and the Huynh-Feldt correction was .984. Thus, any correction to the dfs would not affect statistical significance. The multivariate approach yielded a statistically significant Hotelling's trace [$H_{(A)} = 1.426$ from (11)], which multiplied by $(N-1)=17$ yields a chi-square approximate statistic of $\chi^2_{(A)(df=2)} = 24.242, p < .001$. Converting $H_{(A)}$ to an F approximate using (12) yields $F_{H(A)(2,15)} = 10.697, p = .001$.

Table 2 shows the Koch (1969) model of alignment and ranking. As was the case with the aligned ranks, the results show a statistically significant interaction with a Pillai's trace of $V(Q) = 0.574$ from (15), which multiplied by $(N-1) = 17$ yields a Chi-square approximate statistic of $\chi^2_{(Q)(df=2)} = 9.758, p < .01$. The Hotelling's trace for the Koch model ranks was $H(Q) = 1.345$ with an F approximate (12) of $F_{H(Q)(2,15)} = 10.091, p = .002$.

Table 3 shows the aligned data and the Friedman (1937) model of ranking applied to the aligned data. As was the case with the aligned ranks and the Koch ranks, the results show a statistically significant interaction. Analyzing a univariate model and calculating the multiple

group extension of the Friedman (1937) statistic (21) yields [$F_{(R)(df=2)} = 15.239, p < .001$]. The Huynh-Feldt correction of the Greenhouse-Geisser estimate of epsilon was 1.0. Thus, there are no corrections to the dfs . The multivariate approach yielded a statistically significant Pillai's trace of $V(R) = 0.624$ from (22), which multiplied by $(N-1) = 17$ yields a Chi-square approximate statistic of 10.608, $p < .005$. The Hotelling's trace for the Friedman model aligned ranks was $H(R) = 1.657$ with an F approximate (12) of $F_{H(R)(2,15)} = 12.426, p = .001$.

By further examination of the six cells in Table 1, the data at time $k = 1$ are positively skewed with similar means, variances, and kurtosis values for both groups. At time $k = 2$, the data for both groups are symmetric with similar variances, but group $j = 2$ has a higher mean. At time $k = 3$, there are still location differences, but the data for both groups are negatively skewed with similar variances and kurtosis.

In analyzing real data, it is difficult to trust sample statistics for skew and kurtosis, especially for small sample sizes. Therefore, judging whether the IID assumptions are tenable presents a conundrum. Although such practice is not advised, for the sake of illustration, suppose that these sample moments are valid estimates of population parameters. This data pattern then illustrates a situation in which there is a violation of the univariate shift model (18) distributional assumptions (i.e., IID $[0, \sigma_\epsilon^2]$ for all j and k); however, the multivariate shift model (19) assumption (i.e., IID $[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$) seems tenable. That is, the univariate model requires that all six cells have identical distribution functions; whereas, the multivariate model only requires the two groups to have identical distribution functions for each of the $K = 3$ measures separately. Given that all three multivariate aligned rank tests led to rejections of the interaction null hypothesis in (17), the interaction can be attributed to shifts in location parameters (i.e., a false 20). Furthermore, one may conclude that the stochastic dominance of one group over the other was not constant across the $K = 3$ repeated measures.

Computational Example Two

Table 5 shows the sample moments and the univariate and multivariate test statistics for the Original Data, Aligned Ranks, Koch Model Ranks, and Friedman Model Ranks for hypothetical data from $J = 3$ groups by $K = 4$ repeated measures design (see Appendix for data). A medical psychology research application of this design could be a comparison of the number of errors in recall over $K = 4$ trials for men with treated blood pressure elevation ($j = 3$), men with untreated elevated blood pressure ($j = 2$), and a group of normotensive males ($j = 1$) (e.g., Waldstein, et al., 1991). A genetic association research application would be an alcohol sensitivity study in which motor coordination of humans with $J = 3$ different genotypes (e.g., aa, AA, Aa) was measured once before ($k = 1$) and three times after ingesting a standard dose of alcohol (e.g., Boomsma, Martin, & Molenaar, 1989).

Suppose that these sample moments are valid estimates of population parameters, then upon examination of the Original Data, it can be seen that Group One has positively skewed data with minor changes in spread (variance) and location (mean and median) across the four measures. Similarly, Group Three also has positively skewed data with minor changes in variance over time. However, Group Three also exhibits significant increases in location over the four time periods. Thus, if this example only included Groups One and Three, even the more restrictive distributional assumptions of the univariate shift model (18) would be tenable. That is, the eight cells for Groups One and Three have similar variance and shape (i.e., IID[0, σ_ϵ^2] for all j and k) and differ only in location.

By contrast, Group Two has data that is positively skewed initially ($k = 1$). Subsequently, Group Two increases in location, fluctuates in spread, and changes from a positively skewed shape at $k = 1$ to a symmetric shape at $k = 2$ and then to a negatively skewed shape at the third and fourth measures. In comparing Group Two to the other groups, neither the univariate (18) nor the multivariate shift model (19) distributional assumptions are met. Therefore, the significant test statistics that result in rejections of the null hypotheses (16) or (17)

cannot be attributed to a single parameter. Thus, the rejection must be interpreted as the groups demonstrating stochastic heterogeneity in trends (growth curves). Namely, Group Two appears to be stochastically dominant over the other two groups at time points $k = 2$ and 3 and stochastically dominant over Group One at $k = 4$; however, contrast procedures are necessary to test this interpretation.

Multiple Comparison Procedures for Aligned Rank Procedures

Given that the three rank-based procedures are viable approaches to analyzing repeated measures data, then contrast procedures based on these methods should hold quite generally (Agresti & Pendergast, 1986; Beasley, 2000, 2002; Koch, 1969). The most typical form is a product interaction contrast (Hochberg & Tamhane, 1987, pp. 294-303; Marascuilo & Levin, 1970) defined as:

$$\hat{\psi} = a_1(b_1\bar{U}_{11} + b_2\bar{U}_{12} + \dots + b_k\bar{U}_{1k} + \dots + b_K\bar{U}_{1K}) + a_2(b_1\bar{U}_{21} + b_2\bar{U}_{22} + \dots + b_k\bar{U}_{2k} + \dots + b_K\bar{U}_{2K}) + a_j(b_1\bar{U}_{j1} + b_2\bar{U}_{j2} + \dots + b_k\bar{U}_{jk} + \dots + b_K\bar{U}_{jK}) + a_J(b_1\bar{U}_{J1} + b_2\bar{U}_{J2} + \dots + b_k\bar{U}_{Jk} + \dots + b_K\bar{U}_{JK}); \quad (26)$$

where \bar{U}_{jk} is a general term for the mean rank of the j^{th} group on the k^{th} repeated measure.

Define $\mathbf{a} = (a_1 + a_2 + \dots + a_j \dots + a_J)'$ as a vector of contrast coefficients that compares the J independent samples and $\mathbf{b} = (b_1 + b_2 + \dots + b_k + \dots + b_K)'$ as a vector of contrast coefficients that involves the K repeated measures with the restriction that $\sum a_j = 0$ and $\sum b_k = 0$. For comparing the J independent groups, a set of pairwise or group combination contrasts would most likely be of interest for defining \mathbf{a} . For comparing the K repeated measures either pairwise, polynomial, or trend contrasts would most typically define \mathbf{b} (Lix & Keselman, 1996; Marascuilo & McSweeney, 1967). In some cases, it may be desirable to normalize the trend coefficients, \mathbf{b} , so that the metric of the repeated measures variable will not

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

change, thus making confidence intervals more interpretable.

From a univariate perspective, a pooled squared standard error of a contrast in a split-plot design (see Kirk, 1982, pp. 516-518) can be calculated by defining:

$$SE_{\hat{\psi}}^2 = \sum_{j=1}^J \left(\frac{a_j^2}{n_j} \right) \frac{(\mathbf{b}' \mathbf{E} \mathbf{b})}{(N - J)}, \quad (27)$$

where \mathbf{E} is the error matrix (4) computed for U_{ijk} (i.e., any of the three ranking procedures). This approach assumes homogeneity of variance of the transformed scores:

$$U^*_{ij} = \sum_{k=1}^K b_k U_{ijk}. \quad (28)$$

This requirement of homogeneity of variance for transformed scores implies the sphericity of the pooled covariance matrix (4). Thus from the perspective of rank-based tests, this approach requires that the error components are IID[0, σ_ϵ^2] for all j and k .

From a multivariate perspective, a standard error that does not require homogeneity of variance of the transformed scores (i.e., sphericity) can be calculated by defining J separate Sums of Squares (SS):

$$SS_{U^*_j} = \sum_{i=1}^{n_j} (U^*_{ij} - \bar{U}^*_j)^2, \quad (29)$$

where \bar{U}^*_j is the mean for the j^{th} group for the transformed scores U^*_{ij} in (29). The standard error is calculated as:

$$SE_{\hat{\psi}}^2 = \sum_{j=1}^J \left(\frac{a_j^2}{n_j} \right) \frac{SS_{U^*_j}}{(n_j - 1)}, \quad (30)$$

A $(1-\alpha)\%$ confidence interval for the contrast of aligned ranks can be formed by:

$$\hat{\psi} \pm S (SE_{\hat{\psi}}). \quad (31)$$

The null hypothesis $H_0: \psi = 0$ is rejected if the confidence interval in (31) does not cover zero. If the univariate IID[0, σ_ϵ^2] assumption is tenable, $SE_{\hat{\psi}}$ can be defined as the square root of (26). However, $SE_{\hat{\psi}}$ should be defined as the square root of (30) if the transformed scores have heterogeneous variances (i.e., the sphericity condition does not hold).

The definition of S depends on the type of contrast conducted. For example, in the $J = 3$ by $K = 4$ design from Example Two, suppose that after rejecting the null hypothesis (17) the interest was in assessing whether the linear trend, $\mathbf{b}'_{\mathbf{L}} = \{-3 -1 +1 +3\}/\sqrt{20}$, of Group One is stochastically different from the linear trend of the other two groups combined, $\mathbf{a}'_1 = \{+1 -0.5 -0.5\}$, and whether the linear trends for Groups Two and Three are stochastically different, $\mathbf{a}'_2 = \{0 +1 -1\}$. In this case, the trend coefficients, $\mathbf{b}_{\mathbf{L}}$ were normalized so that the metric of the repeated measures variable was not changed, thus making subsequent confidence intervals more interpretable.

Also, consider the same group comparisons for the Initial Change from Time $k = 1$ to Time $k = 2$, $\mathbf{b}'_{\mathbf{C}} = \{-1 +1 0 0\}$. Thus, $c = 4$ post hoc tests would be conducted. To construct a post hoc confidence interval, S could be defined as a critical value from Student's t distribution using the Dunn-Sidak correction, $\alpha_{\text{DS}} = [1-(1-\alpha)^{1/c}]/2$:

$$S = t_{(1-\alpha_{\text{DS}}), df_e}. \quad (32)$$

For $c = 4$ contrasts, $\alpha_{\text{DS}} = .00637$; however, df_e for (32) differs for the univariate (27) and multivariate approaches (30). For the univariate pooled standard error (27), $df_e = (N-J)$; however, if the standard error in (30) is used then a Welch (1947) correction must be applied to df_e . For defining S in terms of the sampling distribution of the Hotelling's trace or other multivariate referent distribution, refer to Gabriel (1968) and Sheehan-Holt (1998).

For computational convenience, the interaction contrasts can be calculated by

transforming the data into a single variable: $\mathbf{U}\mathbf{b}$, where \mathbf{U} is the $N \times K$ data matrix and \mathbf{b} is the $K \times 1$ vector of trend coefficients. Then, the group contrasts, \mathbf{a} , can be performed on the transformed data. The univariate pooled standard error (27) can be computed from methods that assume equal variances, such as Fisher's LSD. The multivariate standard error (30) can be computed from methods that do not assume equal variances, such as Tamhane's (1979) $T2$.

It is debatable whether the multivariate (30) or univariate (27) approach is better in terms of robustness and power (Maxwell & Delaney, 2000), and thus, this issue should be investigated. However, the multivariate approach would be expected to yield more precise confidence intervals than the univariate approach, especially in situations where the pooled covariance matrix is non-spherical (Boik, 1981).

Conducting post hoc analyses is not generally suggested as an optimal procedure to adopt (Marascuilo & Levin, 1970). Rather, a defined set of planned contrasts with an appropriate adjustment for controlling Type I errors is often recommended, in which case the omnibus tests previously elaborated should be bypassed. For conducting multiple planned comparisons or simultaneous test procedures, there are several excellent references for both the univariate and multivariate approaches references (e.g., Hochberg & Tamhane, 1987; Gabriel, 1968; Lix & Keselman, 1996; Maxwell & Delaney, 2000; Sheehan-Holt, 1998).

Defining Confidence Intervals for Interpretable Parameters

Reasons for rejecting an interaction null hypothesis are of more interest than the simple conclusion that it is false; therefore, the contrast testing procedures detailed in the previous section are of great utility. Furthermore, there is a trend toward interpreting confidence intervals instead simply reporting p -values in a variety of research disciplines (Campbell & Gardner, 1988; Gardener & Altman, 1986; Serlin, 1993). Moreover, it is important to construct confidence intervals around interpretable parameters when possible. Thompson (2002) discusses a bootstrap methodology to compute confidence intervals

for effect sizes from parametric analyses. For location parameters less sensitive to skewness, confidence intervals for medians have been proposed (Bonett & Price, 2002; Campbell & Gardner, 1988; Hodges & Lehmann, 1963).

Unfortunately, aligned ranks have no inherent meaning except that they serve as placeholders for the percentiles of the original raw score distribution with the nuisance location parameters removed. Thus, the rank statistics previously discussed are useful for assessing the statistical significance of the interaction, but they do not provide direct information about the nature or magnitude of the effect. For this reason, Koch, et al. (1980) suggested that results from nonparametric omnibus tests should be accompanied by appropriate descriptive statistics (e.g., frequency distributions or percentiles) and nonparametric estimates for confidence intervals. Newson (2002) reviewed methods for computing confidence intervals for rank-based statistics, which convey estimates and boundaries for informative parameters such as Cliff's (1996) d and Somers' (1962) D .

Confidence Intervals for Aligned Ranks

The cell means for the aligned ranks provide descriptions of the degree to which the JK cells have different locations due to discrepancies from the marginal distributions (i.e., due to interaction). Thus, these cell means give information about interaction trends relative to main effects and which cells contribute more to the omnibus interaction effect. For repeated measures designs, Agresti and Pendergast (1986) suggested dividing ranks by $(NK+1)$. These values, $U_{ijk} = A_{ijk}/(NK+1)$, have a grand mean, $\bar{U}_{**} = 0.5$, that is equivalent to the median of the aligned scores. The cell means, \bar{U}_{jk} , provide the probability that a randomly selected observation from cell jk is larger than an independent observation selected at random from another cell after removing the main effects. This approach suggested by Agresti and Pendergast (1986) is consistent, though not identical, to Cliff's (1996) notion of dominance¹ and the computation of relative effects² (Brunner, et al., 2002). It is also similar to the Hodges and Lehmann (1963) median difference,

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

which estimates the typical difference between individual observations from different cells.

As noted, the interaction contrasts can be accomplished by transforming the data, \mathbf{U}_b , and performing the group contrasts, \mathbf{a} , on the transformed data. The upper panel of Table 6 shows the means and standard deviations for $\mathbf{U} = \mathbf{A}/(NK+1)$, the data transformed by the linear trend contrast, \mathbf{U}_{bL} , and the data differenced by the Initial Change contrast, \mathbf{U}_{bC} . The upper panel of Table 7 shows the univariate-based (27) and multivariate-based (30) 95% confidence intervals for the four contrasts previously discussed performed on the adjusted aligned ranks.

The cell mean for Group 1 at time $k = 1$ had the highest mean of 0.9476. This indicates that, after removal of the main effects, this cell had higher scores relative to the other cells and that a randomly selected observation from this cell has a very high probability (0.9476) of being larger than an independent observation selected at random from any other cell. Likewise, the cell mean for Group 1 at time $k = 4$ had the lowest mean of 0.1012, and thus, a randomly selected observation from this cell has a very low probability of being larger than an independent observation selected at random from any other cell.

Similar to Cliff's (1996) d -statistic, the difference in these probabilities can be used to judge the stochastic dominance of one cell over another. Thus, the aligned ranks for Group 1 have a descending trend in that relative to the main effects the observations in Group 1 tend to get stochastically smaller over time. By examining the original data in Table 5, Group 1 had a slight increase in means across the $K = 4$ time points. Therefore, the aligned ranks provide information about which cells have stochastically larger scores relative to the main effects. In other words, given that there was a repeated measures main effect with increasing means for all three groups combined, the trend for Group 1 was descending in a relative manner. This can be seen in the data transformed by the linear contrast coefficient, \mathbf{U}_{bL} , in which the probability of larger scores (i.e., stochastic dominance) for observations in Groups 1 tends to decrease at a rate of -.620 on average.

For Group 2, the probability of larger scores tends to increase at average rate of .336 relative to the main effects. For Group 3, the stochastic dominance of scores relative to the main effects increases at a slight lower rate (.202) as compared to Group 2. For comparing Group 1 to Groups 2 and 3 combined, the results show a value of $\hat{\psi}_{\mathbf{a}_1\mathbf{b}_L} = -0.8891$. This indicates that Groups 2 and 3 combined, as compared to Group 1, have a very high probability of having stochastically larger scores at time $k = 4$ and smaller scores at $k = 1$. To elaborate, suppose Case A is a randomly selected case from Group 2 or 3 and Case B is a randomly selected case from Group 1. The probability that Case A will have a steeper ascending (positive monotonic) trend across the $K = 4$ time points than Case B from Group 1 is 0.8891.

The univariate 95% simultaneous confidence interval indicates that plausible values range between -1.1276 and -0.6506. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -1.0474 and -0.7308. Note that the sign of the contrast value only indicates the direction of the stochastic dominance; it does not indicate a negative probability. Also, this approach can yield a bound on the confidence interval that exceeds 1 (-1 in this case), thus, an asymmetrical confidence interval with 1 (or -1) as the upper (or lower) bound may be constructed. Other methods create this bound and asymmetrical confidence interval by computing the standard errors in a different manner (see Endnotes 1 and 2; Brunner, et al., 2002; Cliff, 1996; Newson, 2001). The difference between Groups 2 and 3 is not statistically significant: both the univariate and multivariate 95% confidence intervals contained zero as a plausible value (see Table 7).

By examining the data transformed by the initial change contrast coefficient, \mathbf{U}_{bC} , it is observed that observations from time $k = 1$ tend to be stochastically larger than observations taken at $k = 2$, for Groups 1 and 3. For Group 2, the measures taken at $k = 2$ are stochastically larger than the scores from $k = 1$ and the probability of randomly selecting a larger score at $k = 2$ increases by 0.3657 relative to the main

effects. Thus, Group 2 has a tendency for scores to become stochastically larger from $k = 1$ to $k = 2$; whereas, Groups 1 and 3 have a tendency for scores to decrease relative to the main effects.

As compared to Group 1, Groups 2 and 3 combined have a higher probability of scores becoming stochastically larger from time point $k = 1$ to $k = 2$, $\hat{\psi}_{\mathbf{a}|bC} = -0.4824$. The univariate 95% simultaneous confidence interval indicates that plausible values range between -0.7236 and -0.2385. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -0.6531 and -0.3117. The contrast of Group 2 with Group 3 is statistically significant; thus, the probability that Group 2 has stochastically larger scores at $k = 2$ relative to $k = 1$ as compared to Group 3 is 0.8552.

To elaborate, suppose a randomly selected case from Group 2 and a randomly selected case from Group 3. The probability that the case selected from Group 2 will have a stochastically larger gain from time $k = 1$ to $k = 2$ as compared to the latter case from Group 3 is 0.8552. The univariate 95% simultaneous confidence interval indicates that plausible values range between .5833 and 1.1271. The multivariate 95% simultaneous confidence interval gives a wider band of plausible values that range between 0.4779 and 1.2326. As with previous analyses, a researcher may choose to construct an asymmetrical confidence interval with 1 as the upper bound or use other methods that compute standard errors in a different manner (Brunner, et al., 2002; Cliff, 1996; Newson, 2001).

Confidence Intervals for Koch Model Ranks

Using the logic of Agresti and Pendergast (1986), the Koch ranks can be transformed by:

$$U_{ijk} = [Q_{ijk} - (((N+1)/2))]/[(K-1)(N+1)].$$

These values have a grand mean of 0.5. The cell means provide descriptions of the degree to which the JK cells have different locations due to discrepancies from the marginal distributions. As shown in Table 6, the cell mean values for the Koch ranks (middle panel) are similar to the

aligned rank cell means (upper panel). Thus, it would seem that the Koch ranks could be interpreted in a similar manner, but whether they represent probabilities in the same sense that the aligned ranks is debatable.

In Table 7, note that the Koch model tends to give lower estimates of the contrast effects with smaller standard errors, thus, one may question the statistical power of the Koch model relative to the aligned rank procedure. For identically skewed (i.e., multivariate exponential) error distributions, Tandon and Moeschberger (1989) found the Koch model to have similar power as parametric procedures, whereas, Beasley (2002) found the aligned rank procedure to have more statistical power than parametric tests for interactions. It is debatable whether these differences are due to estimation bias, violations of assumptions, or differences in statistical power.

Confidence Intervals for Friedman Ranks

A different logic is used to standardize the Friedman Ranks:

$$U_{ijk} = [R_{ijk} - (((K+1)/2))]/[(K^2-1)/12].$$

For each subject, U_{ijk} has a mean of 0 and unit variance, which is similar in concept to Hettmansperger's (1984) standardization of ranks. As previously noted, the interaction contrasts can be accomplished by transforming the data, \mathbf{U}_b , and then performing the group contrasts, \mathbf{a} , on the transformed data. The lower panel of Table 6 shows the means and standard deviations for $\mathbf{U} = [R_{ijk} - (((K+1)/2))]/[(K^2-1)/12]$. To transform the data by the linear trend contrast, \mathbf{b}_L is standardized, rather than normalized, so that it also has a variance of one, rather than a sum of squares of one, $\mathbf{b}'_L = \{-1.3416 - 0.4472 + 0.4472 + 1.3416\}$. The values of $\mathbf{U}_b \mathbf{b}_L / K$ are a linear transformation of Page's (1963) L statistic and represent each individual's rank correlation with the linear trend coefficients (Lyerly, 1952). Thus, the mean values of $\mathbf{U}_b \mathbf{b}_L / K$ for each group represent the group's average concordance with the ordered alternative, in this case linear trend. The contrasts, \mathbf{a} , applied to these values will estimate how the groups differ

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

Table 1: Hypothetical Data and the Aligned Ranking Procedure for the $J = 2$ by $K = 3$ Split-Plot Design in Example One.

	Original Data				Aligned Data			Aligned Ranks		
	$k = 1$	$k = 2$	$k = 3$	\bar{Y}_{ij^*}	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Group One $j = 1$ Slow Learners or Day Treatment	1.1	6.2	7.2	4.83	.56	-.03	-.53	39	28	18
	2.2	4.8	6.1	4.37	2.13	-.96	-1.16	53	10	8
	2.3	7.1	8.0	5.80	.79	-.10	-.70	44	24	15
	2.4	8.1	9.4	6.63	.06	.07	-.13	30	31	23
	3.2	7.3	10.4	6.97	.53	-1.06	.54	37	9	38
	3.4	9.3	10.5	7.73	-.04	.17	-.13	26	34	22
	4.1	8.1	9.3	7.17	1.23	-.46	-.76	48	20	14
	10.1	10.4	10.2	10.23	4.16	-1.23	-.2.93	54	7	1
Mean	3.60	7.66	8.89	6.72	1.18	-.45	-.73	41.38	20.38	17.38
Median	2.80	7.70	9.35	6.80	.68	-.28	-.61	41.5	22.00	16.50
SD	2.78	1.75	1.62	1.84	1.39	.56	1.03	10.25	10.60	11.03
Variance	7.72	3.05	2.64	3.37	1.93	.32	1.06	105.13	112.27	121.70
Skew	2.24	-.06	-.76	.75	1.69	-.36	-1.47	-.23	-.12	.54
Kurtosis	5.65	-.09	-.75	1.08	2.88	-1.97	3.22	-1.22	-1.88	1.17
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	1.0	7.9	8.8	5.90	-.61	.60	0	16	41	29
	2.4	9.2	10.1	7.23	-.54	.57	-.03	17	40	27
	2.2	10.1	11.8	8.03	-1.54	.67	.87	3	42	45
	2.3	10.9	11.1	8.10	-1.51	1.40	.10	4	50	33
	3.1	10.1	13.2	8.80	-1.41	-.10	1.50	5	25	51
	3.3	9.9	12.1	8.43	-.84	.07	.77	12	32	43
	3.2	11.2	14.4	9.60	-2.11	.20	1.90	2	35	52
	4.4	12.3	13.1	9.93	-1.24	.97	.27	6	46	36
4.9	11.2	14.2	10.10	-.91	-.30	1.20	11	21	47	
9.2	13.1	14.3	12.20	1.29	-.50	-.80	49	19	13	
Mean	3.60	10.59	12.31	8.83	-.94	.36	.58	12.50	35.10	37.60
Median	3.15	10.50	12.60	8.62	-1.07	.39	.52	8.50	37.50	39.50
SD	2.26	1.50	1.89	1.74	.92	.59	.82	13.90	10.63	12.36
Variance	5.11	2.24	3.59	3.03	.85	.35	.67	193.17	112.99	152.71
Skew	1.85	-.06	-0.63	.31	1.63	.26	.05	2.35	-.32	-.74
Kurtosis	4.35	.21	-0.50	.76	3.90	-.56	-.53	6.22	-1.18	.075
Epsilon*	.769				.769			.984		

Note: * Based on the Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

BEASLEY & ZUMBO

Table 2: Hypothetical Example of Koch’s Model of Ranking for Interactions for Hypothetical Data in Table 1.

Koch’s Model for Analyzing Interaction Effects									
	$T_{ij(1,1)}$	$T_{ij(1,2)}$	$T_{ij(1,3)}$	$T_{ij(2,1)}$	$T_{ij(2,2)}$	$T_{ij(2,3)}$	$T_{ij(3,1)}$	$T_{ij(3,2)}$	$T_{ij(3,3)}$
Group One $j = 1$ Slow Learners or Day Treatment	9.5	12	13	7	9.5	12	6	7	9.5
	9.5	17	17	2	9.5	8	2	11	9.5
	9.5	13	14	6	9.5	14	5	5	9.5
	9.5	11	12	8	9.5	7	7	12	9.5
	9.5	14.	10	5	9.5	2	9	17	9.5
	9.5	10	11	9	9.5	11	8	8	9.5
	9.5	15	15	4	9.5	9.5	4	9.5	9.5
	9.5	18	18	1	9.5	18	1	1	9.5
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	9.5	6	8	13	9.5	14	11	5	9.5
	9.5	7	9	12	9.5	14	10	5	9.5
	9.5	4	3	15	9.5	6	16	13	9.5
	9.5	1	5.5	18	9.5	17	13.5	2	9.5
	9.5	5	2	14	9.5	3	17	16	9.5
	9.5	8	5.5	11	9.5	5	13.5	14	9.5
	9.5	2	1	17	9.5	1	18	18	9.5
	9.5	3	7	16	9.5	16	12	3	9.5
9.5	9	4	10	9.5	4	15	15	9.5	
9.5	16	16	3	9.5	9.5	3	9.5	9.5	
	Q_{ij1}			Q_{ij2}			Q_{ij3}		
Group One $j = 1$ Slow Learners or Day Treatment	34.5			28.5			22.5		
	43.5			19.5			22.5		
	36.5			29.5			19.5		
	32.5			24.5			28.5		
	33.5			16.5			35.5		
	30.5	\bar{Q}_{11}	37.00	29.5	\bar{Q}_{12}	24.94	25.5	\bar{Q}_{13}	23/56
	39.5	$SD(Q_{11})$	5.37	23.0	$SD(Q_{12})$	4.95	23.0	$SD(Q_{13})$	6.92
45.5	$Var(Q_{11})$	28.86	28.5	$Var(Q_{12})$	24.53	11.5	$Var(Q_{13})$	47.89	
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	23.5			36.5			25.5		
	25.5			35.5			24.5		
	16.5			30.5			38.5		
	16.0			44.5			25.0		
	16.5			26.5			42.5		
	23.0			25.5			37.0		
	12.5			27.5			45.5		
19.5	\bar{Q}_{21}	21.70	41.5	\bar{Q}_{22}	31.35	24.5	\bar{Q}_{23}	32.45	
22.5	$SD(Q_{21})$	8.08	23.5	$SD(Q_{22})$	7.76	39.5	$SD(Q_{23})$	8.93	
41.5	$Var(Q_{21})$	65.34	22.0	$Var(Q_{22})$	60.23	22.0	$Var(Q_{23})$	79.75	

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

Table 3: Friedman Model of Aligned Ranks for Hypothetical Data in Table 1.

	Aligned Data			Friedman Aligned Ranks		
	<i>k</i> = 1	<i>k</i> = 2	<i>k</i> = 3	<i>k</i> = 1	<i>k</i> = 2	<i>k</i> = 3
Group One <i>j</i> = 1 Slow Learners or Day Treatment	.56	-.03	-.53	3	2	1
	2.13	-.96	-1.16	3	2	1
	.79	-.10	-.70	3	2	1
	.06	.07	-.13	2	3	1
	.53	-1.06	.54	2	1	3
	-.04	.17	-.13	2	3	1
	1.23	-.46	-.76	3	2	1
	4.16	-1.23	-2.93	3	2	1
Mean	1.18	-.45	-.73	2.625	2.125	1.250
Median	.68	-.28	-.61	3.000	2.000	1.000
SD	1.39	.56	1.03	.518	.641	.707
Variance	1.93	.32	1.06	.268	.411	.500
Skew	1.69	-.36	-1.47	-.644	-.068	2.828
Kurtosis	2.88	-1.97	3.22	-2.240	.741	8.000
Group Two <i>j</i> = 2 Fast Learners or Enhanced Standard Methadone	-.61	.60	0	1	3	2
	-.54	.57	-.03	1	3	2
	-1.54	.67	.87	1	2	3
	-1.51	1.40	.10	1	3	2
	-1.41	-.10	1.50	1	2	3
	-.84	.07	.77	1	2	3
	-2.11	.20	1.90	1	2	3
	-1.24	.97	.27	1	3	2
	-.91	-.30	1.20	1	2	3
	1.29	-.50	-.80	3	2	1
Mean	-.94	.36	.58	1.200	2.400	2.400
Median	-1.07	.39	.52	1.000	2.000	2.500
SD	.92	.59	.82	.633	.516	.699
Variance	.85	.35	.67	.400	.267	.489
Skew	1.63	.26	.05	3.162	.484	-.780
Kurtosis	3.90	-.56	-.53	10.000	-2.277	-.146
Epsilon*	.769			1.000		

Note: * Based on the Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

Table 4: Hypothetical Population Distribution of Probabilities (π_{r*}) for Friedman Model Ranks with Descriptive Statistics for Each Element (R_{ijk}). in a $J = 2$ by $K = 3$ Split-Plot Design.

Permutation			Probability of r^{th} Permutation					Group Configuration		Status of Null Hypotheses		
R ₁	R ₂	R ₃	π_{r1}	π_{r2}	π_{r3}	π_{r4}	π_{r5}	$j = 1$	$j = 2$	H ₀ (23)	H ₀ (24)	H ₀ (25)
1	2	3	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{10}{24}$	$\frac{1}{12}$	$\frac{1}{6}$	π_{r1}	π_{r1}	True	True	True
1	3	2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{8}$	π_{r2}	π_{r2}	True	False	True
2	1	3	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{24}$	$\frac{1}{6}$	$\frac{1}{24}$	π_{r2}	π_{r3}	True	False	False
2	3	1	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{24}$	$\frac{1}{6}$	$\frac{1}{24}$	π_{r4}	π_{r4}	True	False	True
3	1	2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{24}$	$\frac{1}{4}$	$\frac{7}{24}$	π_{r4}	π_{r5}	True	False	False
3	2	1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{10}{24}$	$\frac{1}{4}$	$\frac{1}{3}$	π_{r3}	π_{r4}	False	False	False
		\bar{R}_1	2.000	2.000	2.000	2.333	2.333					
		σ_{R1}^2	0.667	0.500	0.917	0.556	0.806					
		\bar{R}_2	2.000	2.000	2.000	1.833	1.833					
		σ_{R2}^2	0.667	0.833	0.167	0.639	0.472					
		\bar{R}_3	2.000	2.000	2.000	1.833	1.833					
		σ_{R3}^2	0.667	0.667	0.917	0.639	0.556					
		ϵ	1.000	0.923	0.640	0.992	0.903					

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

Table 5: Sample Moment and Tests Statistics for Hypothetical Data from the $J=3$ by $K=4$ Split-Plot Design in Example Two.

Original Data	Group $j = 1$ ($n_1 = 8$) (e.g., Normotensive; aa)				Group $j = 2$ ($n_2 = 10$) (e.g., Untreated EBP; AA)				Group $j = 3$ ($n_3 = 8$) (e.g., Treated EBP, Aa)			
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Mean	5.80	6.11	6.88	7.21	5.80	11.79	13.29	15.84	5.88	7.05	9.59	13.70
Median	5.10	5.20	5.95	6.55	5.10	11.75	13.70	17.10	5.20	6.20	8.70	13.05
SD	2.63	2.93	2.73	2.86	2.82	1.55	1.77	2.72	2.85	2.91	2.80	2.78
Variance	6.90	8.56	7.43	8.19	7.98	2.40	3.12	7.39	8.15	8.46	7.82	7.70
Skew	2.34	2.45	2.33	2.37	2.09	0.10	-0.75	-0.78	2.13	2.59	2.36	2.40
Kurtosis	5.94	6.44	5.85	6.15	5.19	0.08	-0.65	-1.21	5.45	7.06	5.99	6.28
H-F = .795, $F_{(Y)(4.77, 54.86)} = 53.42, p < .001$; $H_{(Y)} = 10.06, p < .001$; $V_{(Y)} = 1.66, p < .001$												
Aligned Ranks												
Mean	99.50	58.25	33.69	10.63	15.20	69.50	69.90	67.65	57.75	22.25	38.75	83.88
Median	99.50	56.50	34.00	10.50	6.50	66.00	72.50	71.00	57.00	21.50	39.00	83.50
SD	2.45	6.88	5.81	3.32	24.53	16.08	18.44	26.67	12.14	4.30	8.89	6.22
Variance	6.00	47.36	33.78	11.05	602.0	258.5	340.0	711.2	147.4	18.50	79.00	38.70
Skew	0	1.09	0.21	-0.36	2.84	0.03	-1.75	0.00	0.01	1.19	-0.09	-0.06
Kurtosis	-1.20	0.37	-0.93	-0.53	8.46	-1.61	4.75	-1.95	-0.86	1.93	-2.20	-1.66
H-F = .893, $F_{(A)(5.36, 61.61)} = 43.10, p < .001$; $H_{(A)} = 8.50, p < .001$; $V_{(A)} = 1.61, p < .001$												
Koch Ranks												
Mean	80.25	60.75	47.38	27.63	32.35	62.25	59.55	61.85	54.81	36.94	53.69	70.56
Median	81.00	57.25	49.25	27.50	28.50	63.25	57.50	59.50	55.75	34.50	52.75	70.50
SD	7.16	8.22	8.27	7.66	12.56	10.15	12.37	12.89	7.20	8.17	7.28	7.81
Variance	51.29	67.57	68.41	58.70	157.7	103.0	153.1	166.1	51.78	66.82	53.00	61.03
Skew	-1.02	0.84	-0.35	0.70	1.72	0.08	0.37	0.86	-0.06	0.79	0.14	-0.05
Kurtosis	1.49	0.42	-0.28	0.68	3.39	-0.79	1.34	-0.04	0.01	-0.22	-1.87	-0.51
H-F = 1.00, $F_{(Q)(6, 69)} = 30.35, p < .001$; $H_{(Q)} = 7.52, p < .001$; $V_{(Q)} = 1.55, p < .001$												
Friedman Ranks												
Mean	4.00	3.00	2.00	1.00	1.30	2.80	2.90	3.00	2.75	1.00	2.25	4.00
Median	4.00	3.00	2.00	1.00	1.00	2.50	3.00	3.00	3.00	1.00	2.00	4.00
SD	0	0	0	0	0.95	0.92	0.88	0.94	0.46	0	0.46	0
Variance	0	0	0	0	0.90	0.84	0.77	0.89	0.21	0	0.21	0
Skew					3.16	0.47	-1.02	0.00	-1.44		1.44	
Kurtosis					10.00	-1.81	1.83	-2.13	0.00		0.00	
H-F = .931, $F_{(R)(df=5.59)} = 56.50, p < .001$; $H_{(R)} = 8.80, p < .001$; $V_{(R)} = 1.60, p < .001$												

Note: H-F = Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

BEASLEY & ZUMBO

Table 6: Sample Moment and Tests Statistics for Hypothetical Data from the $J=3$ by $K=4$ Split-Plot Design in Example Two.

	Group $j = 1$ ($n_1 = 8$) (e.g., Normotensive; aa)				Group $j = 2$ ($n_2 = 10$) (e.g., Untreated EBP; AA)				Group $j = 3$ ($n_3 = 8$) (e.g., Treated EBP; Aa)			
<i>Aligned Ranks</i> $U = A/(NK+1)$												
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Mean	.9476	.5548	.3208	.1012	.1448	.6619	.6657	.6443	.5500	.2119	.3690	.7988
SD	.0233	.0655	.0554	.0317	.2337	.1531	.1756	.2540	.1156	.0410	.0847	.0592
Linear U_{bL}	-.6201 $SD = .0194$.3359 $SD = .3157$.2020 $SD = .1132$			
Change U_{bC}	-.2778 $SD = .0592$.3657 $SD = .2199$				-.2391 $SD = .0910$			
<i>Koch Ranks</i> $U = Q_{ijk} - [((N+1)/2)]/[(K-1)(N+1)]$												
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Mean	.8241	.5833	.4182	.1744	.2327	.6019	.5685	.5969	.5100	.2894	.4961	.7045
SD	.0884	.1015	.1021	.0946	.1550	.1253	.1528	.1591	.0888	.1009	.0899	.0965
Linear U_{bL}	-.4727 $SD = .0666$.2369 $SD = .2184$.1767 $SD = .1094$			
Change U_{bC}	-.2407 $SD = .1848$.3691 $SD = .1759$				-.2207 $SD = .1504$			
<i>Friedman Ranks</i> $U = Q_{ijk} - [((N+1)/2)]/[(K-1)(N+1)]$												
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
Mean	1.3416	.4472	-.4472	-1.3416	-1.0733	.2683	.3578	.4472	.2236	-1.3416	-.2236	1.3416
SD	0	0	0	0	.8485	.8219	.7832	.8433	.41404	0	.41404	0
Linear $U_{bL/K}$	-1.0000 $SD = 0$.5200 $SD = .5750$.5000 $SD = .1852$			
Change $U_{bC/K}$ SD	-.3162 $SD = 0$.4742 $SD = .4104$				-.5535 $SD = .1464$			

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

Table 7: Results of Contrast Procedures for the $J = 2$ by $K = 3$ Split-Plot Design in Example Two.

Rank (Contrast)		Univariate Approach $df = 23; S = 2.701$			Multivariate Approach				
Aligned Ranks	$\hat{\psi}$	$SE(27)$	Lower Bound	Upper Bound	df^*	S	$SE(30)$	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-0.8891	0.0883	-1.1276	-0.6506	12.11	2.920	0.0542	-1.0474	-0.7308
$\mathbf{a}_2\mathbf{b}_L$	0.1339	0.0984	-0.1319	0.3997	11.73	2.936	0.1076	-0.2461	0.5139
$\mathbf{a}_1\mathbf{b}_C$	-0.4824	0.0903	-0.7263	-0.2385	18.22	2.762	0.0618	-0.6531	-0.3117
$\mathbf{a}_2\mathbf{b}_C$	0.8552	0.1007	0.5833	1.1271	12.53	2.903	0.1084	0.4779	1.2326
Koch Ranks	$\hat{\psi}$	$SE(27)$	Lower Bound	Upper Bound	df	S	$SE(30)$	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-0.6795	0.0655	-0.8564	-0.5026	20.27	2.732	0.0461	-0.8054	-0.5536
$\mathbf{a}_2\mathbf{b}_L$	0.0602	0.0730	-0.1369	0.2572	13.79	2.861	0.0792	-0.2102	0.3305
$\mathbf{a}_1\mathbf{b}_C$	-0.3150	0.0730	-0.5122	-0.1178	12.06	2.922	0.0758	-0.5365	-0.0935
$\mathbf{a}_2\mathbf{b}_C$	0.5898	0.0813	0.3702	0.8094	15.90	2.806	0.0770	0.3335	0.8461
Friedman Ranks	$\hat{\psi}$	$SE(27)$	Lower Bound	Upper Bound	df	S	$SE(30)$	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-1.5100	0.1592	-1.9400	-1.0800	11.24	2.958	0.0966	-1.7957	-1.2243
$\mathbf{a}_2\mathbf{b}_L$	0.0200	0.1774	-0.4591	0.4991	11.24	2.958	0.1933	-0.6696	0.7096
$\mathbf{a}_1\mathbf{b}_C$	-0.2767	0.1123	-0.5800	0.0266	11.82	2.932	0.0685	-0.4775	-0.0759
$\mathbf{a}_2\mathbf{b}_C$	1.0277	0.1251	0.6898	1.3657	11.82	2.932	0.1371	0.5443	1.5111

Notes: From (32) $\alpha_{DS} = .00637$. $\mathbf{a}_1 = \{+2 -1 -1\}$ is a comparison of Group One to a combination of Groups Two and Three. $\mathbf{a}_2 = \{0 +1 -1\}$ is a comparison of Groups Two and Three. $\mathbf{b}_L = \{-3 -1 +1 +3\}$ is a linear polynomial contrast. $\mathbf{b}_Q = \{+1 -1 -1 +1\}$ is a quadratic polynomial contrast. *The dfs for the Multivariate Approach were computed from the Welch (1947) correction.

in their concordance with the ordered alternative (i.e., linear trend) on average. As shown in the lower panel of Table 6, Group 1 has a perfect negative rank correlation with the linear trend with no variance, which means that relative to the main effects each person in Group 1 had a descending trend or was discordant with the ordered alternative. Groups 2 and 3 had rank correlations with the linear trend (concordance) of approximately 0.50. Comparing Group 1 to Groups 2 and 3 combined, it is apparent that there are strong differences in their average rank correlation, $\hat{\psi}_{\mathbf{a}_1\mathbf{b}_L} = -1.510$. The univariate 95% simultaneous confidence interval indicates that plausible values range between -1.9400 and -1.0800. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -1.7957 and -1.2243.

This type of interpretation can be used for any trend contrast that involves a linear combination of all K repeated measures by thinking of the trend in terms of ordered alternatives. These results can also be couched in terms of stochastic heterogeneity (Beasley, 2000; Vargha & Delaney, 1998) in that Groups 2 and 3 combined, as compared to Group 1, have a very high probability of yielding stochastic larger scores at time $k = 4$ and smaller scores at $k = 1$ (i.e., very high probability of having stochastically larger or steeper slopes). Group 2 did not significantly differ from Group 3 in terms of linear trend (i.e., the confidence interval contains zero).

To transform the data by the initial trend contrast is standardized, $\mathbf{b}'_C = \{-\sqrt{2} \quad +\sqrt{2} \quad 0 \quad 0\}$. The values of $\mathbf{U}\mathbf{b}_C/K$ represent each individuals rank correlation with this ordered alternative. The results in the bottom panel of Table 7 show that Group 1 does not significantly differ from Groups 2 and 3 combined (i.e., the confidence interval contains zero). However, the change from time $k = 1$ to $k = 2$ was positive for Group 2 and negative for Group 3 (see Table 6, lower panel). The difference in these rank correlations was -1.0277. The univariate 95% simultaneous confidence interval indicates that plausible values for the difference in rank correlation range between 0.6898 and 1.3657. The multivariate 95% simultaneous confidence

interval gives a wider band of plausible values that range between 0.5433 and 1.5111.

For analyses such as Initial Change contrast, $\mathbf{U}\mathbf{b}_C$, only two of the K repeated measures are used and thus interpretations reduce to the interpretations similar to the sign test. However, this approach includes information from the other time points; thus, these effects are relative to the other time points. If a more direct interpretation is desired, then the signs or signed ranks for the differences for the two measures could be computed and statistical analyses conducted to compare the groups. This is a methodology proposed by Cliff (1996) and is not detailed here.

Conclusion

Rank-based methods could be applied to the data in a multiple group repeated measures experiment because the normality assumptions of the split-plot ANOVA model in (1) are violated. In such a case, testing against the shift model null hypothesis (20) would be of interest because it seems conceptually similar to the differences among means in the parametric model hypotheses in (2) or (6). However, if aligned rank procedures are employed and tests of interactions are conducted, then (20) may be rejected incorrectly because some other hypothesis (i.e., 16, 17, 24 or 25) is false. That is, a statistically significant test statistic may be attributable to differences in other distributional characteristics (i.e., variance or shape) rather than reflecting solely differences in location, unless additional distributional assumptions are made (Serlin & Harwell, 2001).

In order to test against (20) and make inferences in terms of location parameters, distributional assumptions must be made. Credible inferences concerning location parameters (20) require the assumption that the population distributions are of identical shape (Serlin & Harwell, 2001; Vargha & Delaney, 1998). This may seem restrictive, however, because parametric statistical tests, which also require $\text{IID}[0, \sigma_\epsilon^2]$ or $\text{IID}[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$ with the additional restriction that the error distributions have a normal shape (Bradley, 1968) have been conducted for decades.

Unfortunately, these distributional assumptions present a conundrum for data analysis. Specifically, the sample estimates of skew and kurtosis are unstable, especially with small sample sizes. Therefore, it is difficult to judge the tenability of the IID assumptions. The choices are: (a) accept the assumptions without testing their tenability or (b) test the assumptions based on unstable estimates. Furthermore, estimates of skew and kurtosis are more reliable with larger samples sizes. However, parametric procedures are more likely to be robust with large samples sizes and the advantage of rank-based procedures over parametric methods in terms of statistical power is likely to decrease.

To circumvent this conundrum, Akritas and Arnold (1994) have argued that hypotheses should be expressed in a manner that does not place additional distributional assumptions on the data. These fully nonparametric hypotheses differ because statistically significant results are not attributed to location parameters alone but rather to any distributional difference. Vargha and Delaney (1998) and Beasley (2002) have suggested analyses of hypotheses related to stochastic heterogeneity. Similarly, Cliff (1996) has argued that rank-based and other nonparametric methods provide ordinal answers to ordinal questions, which are equivalent to results of stochastic heterogeneity and that these results correspond more closely to the goals of many researchers. These forms of hypotheses reduce the risk of drawing incorrect conclusions about the likely sources of the significant interaction, but do so at the cost of not being able to characterize precisely how population distributions differ (Serlin & Harwell, 2001).

The process of aligning the scores before ranking permits test statistics to focus on interactions among location parameters; by removing main effects, the aligned ranks should not inherit any effects due to marginal location differences (i.e., main effects). However, the alignment does not remove other marginal distributional effects; therefore, aligned ranks may still inherit the distributional properties of the original data (e.g., heterogeneity of variance). When the distributions have heterogeneous variances or have different shapes, the null hypothesis of equal location parameters (20) and the null hypothesis of

identical distributions are no longer equivalent. Therefore, as analogs to parametric procedures, aligned rank tests are likely to be sensitive to variance heterogeneity, especially with unequal sample sizes (Algina & Keselman, 1998; Kowalchuk, Keselman, & Algina, 2003; Lei, Holt, & Beasley, 2004).

Similarly, Wilcox (1993) noted that parametric tests are not robust to differences in skew when sample sizes are not equal; however, they are more sensitive to mean differences when there are differences in shape and equal sample sizes. Thus, it may be conjectured that the aligned rank procedures as tests of location parameters would be somewhat robust to heterogeneous variance and differences in shape when sample sizes are equal; however, Lei, et al. (2004) have shown that tests that correct for unequal variances (e.g., Huynh, 1978) performed on aligned ranks still detect distributional (i.e., variance) differences when location parameters do not reflect an interaction. Furthermore, with increasing disparity among sample sizes, aligned rank procedures become more sensitive to detecting any distributional difference and thus should strictly be considered tests of stochastic homogeneity.

Vargha and Delaney (1998) explicated this issue by showing that the null hypotheses of stochastic homogeneity and a null hypothesis of equal mean ranks are equivalent for non-identical, but symmetric distributions. They also demonstrated that stochastic homogeneity and a null hypothesis of equal location parameters (20) are equivalent for identical, asymmetric distributions. Therefore, statistically significant values for interaction tests performed on aligned ranks, and the subsequent rejections of the associated null hypotheses, typically imply a pattern in which one of the J groups is stochastically larger than the other(s) on at least one of the K repeated measures and that this stochastic dominance is not constant across all K repeated measures (Brunner & Langer, 2000; Vargha & Delaney, 1998).

To illustrate, imagine a $J = 2$ groups (e.g., Control and Treatment) by $K = 3$ repeated measures (e.g., Pretest, Posttest, Follow-up) design. Suppose that for the first measure ($k = 1$) the two groups are stochastically identical,

$G_1(Y_{11}) = G_2(Y_{21})$, which would be expected on a pretest if the groups were randomly assigned. Thus for all real values, u , the probability of scores larger than u is the same in both groups, $P(Y_{11} > u) = P(Y_{21} > u)$.

Now imagine that the posttest ($k = 2$) was measured after some treatment had been administered to second group ($j = 2$) while the first group remained a control. If the treatment worked, then the second group should have higher scores, and thus, $G_1(Y_{12}) \neq G_2(Y_{22})$. Because the Treatment group has scores (Y_{12}) that are stochastically larger than the scores for the Control group (Y_{22}), the between-group probabilities of scores larger than all real values (u) are no longer equal, $P(Y_{12} > u) \leq P(Y_{22} > u)$. This conclusion that the stochastic dominance of one group over another is not constant over time is consistent with the answers that aligned rank tests provide to the ordinal question: did the groups respond differently after treatment? Specifically, the treatment group tends to have stochastically larger gains than the control group.

Although statistically significant results may be attributed to other distributional differences, these aligned rank tests are especially sensitive to shifts in location parameters because they use mean ranks in their computation. Therefore, statistically significant test statistics performed on aligned ranks can generally be attributed to differences in location parameters (Marascuilo & McSweeney, 1977, pp. 304-305), which is fortunate because it is difficult to test the tenability of the IID assumptions associated with the shift models. Newson (2002) reviewed methods for constructing confidence intervals that are robust to between-group differences in parameters other than location (e.g., variance; skew). Technically, however, statistically significant tests performed on aligned ranks cannot be attributed solely to differences in location parameters. Given the difficulty of testing model assumptions especially with small samples, results from these procedures should be interpreted in terms of stochastic heterogeneity (Beasley, 2002; Varga & Delaney, 1998). Newson (2002) and Cliff (1996) suggest that

rank-based statistics are based on population parameters, related to Somer's (1962) D , which are extremely informative in terms of stochastic dominance and can be estimated using corresponding sample statistics. Thus, although aligned rank-procedures produce what may be considered a more ambiguous formulation of the underlying null hypothesis that is of interest conceptually, the conclusions are consistent with the ordinal answers that Cliff (1996) has extolled as the effect of actual interest to many researchers.

Notes

1. In a two-group Between-Subjects design, Cliff (1996) has shown that transforming the ranks by $[(2R_{ijk} - 1)/N]$ yields a rank mean difference equal to the d statistic. This transformation will only yield standard errors similar to Cliff's method asymptotically. This is because they are based on different counting procedures. Furthermore, this transformation does not necessarily extend to multiple groups and dependent measures. Thus, the transformation suggested by Agresti and Pendergast (1986) was used.

2. Brunner, et al. (2002) showed a linear transformation of unaligned ranks $[(R_{ijk} - \frac{1}{2})/NK]$, similar to the Agresti and Pendergast (1986) suggestion, will yield cell means that provide estimates of relative treatment effects. Test statistics performed on these values will provide valid tests of fully nonparametric hypotheses. According to Brunner, et al. (2002), however, these values cannot simply be used to compute standard errors, unless the sample size is large. Constructing accurate confidence intervals using the Brunner, et al. method involves a more complicated procedure of computing partial ranks and logit transformations. Whether the Brunner, et al. method can be applied to aligned ranks has yet to be investigated. Thus, for the sake of simplicity the transformation suggested by Agresti and Pendergast (1986) was used.

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

References

- Agresti, A., & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. *Communications in Statistics: 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.
- Akritis, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92, 258-265.
- 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.
- Allison, D. B., et al. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics*, 65, 531-544.
- Avants, S. K., Margolin, A., Sindelar, J., & Rounsaville, B. J. (1999). Day treatment versus enhanced standard methadone services for opioid-dependent patients: A comparison of clinical efficacy and cost. *American Journal of Psychiatry*, 156, 95.
- Beasley, T. M. (2000). Nonparametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs. *Journal of Educational & Behavioral Statistics*, 25, 20-59.
- Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*, 37, 197-226.
- Beasley, T. M., & Zumbo, B. D. (April, 1998). *Rank transformation and df-Correction Procedures for Split-Plot Designs*. Paper presented at the meeting of the American Educational Research Association. San Diego, CA.
- Beckett, J., & Schucany, W. R. (1979). Concordance among categorized groups of judges. *Journal of Educational Statistics*, 4, 125-137.
- Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in test for interactions. *Communications in Statistics: Simulation & Computation*, 16, 1133-1145.
- Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, 46, 241-255.
- Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics*, 18, 1-40.
- Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, 7, 370-383.
- Boomsma, D. I., Martin, N. G., & Molenaar, P. C. M. (1989). Factor and simplex models for repeated measures: Applications to two psychomotor measures of alcohol sensitivity in twins. *Behavior Genetics*, 19, 79-96.
- Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.
- Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. NY: Wiley.
- Brunner, E., & Langer, F. (2000). Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical Journal*, 42, 663-675.
- Campbell, M. J., & Gardner, M. J. (1988). Calculating confidence intervals for some non-parametric analyses. *British Medical Journal*, 296, 1454-1456.

- Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675-701.
- Gabriel, K. R. (1968). Simultaneous test procedures in multivariate analysis of variance. *Biometrika*, *55*, 489-504.
- Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.
- Gentile, J. R., Voelkl, K. E., Mt. Pleasant, J., & Monaco, N. M. (1995). Recall after relearning by fast and slow learners. *Journal of Experimental Education*, *63*, 185-197.
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95-112.
- 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: Simulation & Computation*, *26*, 605-618.
- Headrick, T. C., & Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics: Simulation & Computation*, *29*, 1059-1088.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. NY: Wiley.
- Higgins, J. J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World*, *1*, 201-211.
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. NY: Wiley.
- Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, *33*, 482-497.
- Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on ranks. *Annals of Mathematical Statistics*, *34*, 598-611.
- Hollander, M., & Sethuraman, J. (1978). Testing for agreement between two groups of judges. *Biometrika*, *65*, 403-411.
- Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. NY: Wiley.
- Hora, S. C., & Conover, W. J., (1984). The *F*-statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, *79*, 668-673.
- 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.
- Keselman, H. J., & Algina, J. (1996). The analysis of higher-order repeated measures designs. In *Advances in social science methodology*, Vol. 4, B. Thompson (Ed.), pp. 45-70. Greenwich, CT: JAI Press.
- Keselman, H. J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.
- Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*, 2nd ed. Belmont CA: Brooks-Cole.

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

- Koch, G. G. (1969). Some aspects of the statistical analysis of "split-plot" experiments in completely randomized layouts. *Journal of the American Statistical Association*, 64, 485-506.
- Koch, G. G., Amara, I. A., Stokes, M. E., & Gillings, D. B. (1980). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *International Statistical Review*, 48, 249-265.
- Koch, G. G., & Sen, P. K. (1968). Some aspects of the statistical analysis of the mixed model. *Biometrics*, 24, 27-48.
- Kowalchuk, R. K., Keselman, H. J., & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research*, 38, 433-461.
- Lecoutre, B. (1991). A correction for the 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.
- Lei, X., Holt, J., & Beasley, T. M. (2004). Aligned rank tests as robust alternatives for testing interactions in multiple group repeated measures designs with heterogeneous covariances. *Journal of Modern Applied Statistical Methods*, 3(2), 462-475.
- Lix, L. M., & Keselman, H. J. (1996). Interaction contrasts in repeated measures designs. *British Journal of Mathematical & Statistical Psychology*, 49, 147-162.
- Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. *Psychometrika*, 17, 412-428.
- Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.
- Marascuilo, L. A., (1966). Large-sample multiple comparisons. *Psychological Bulletin*, 65, 280-290
- 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.
- Marascuilo, L. A., & McSweeney, M. (1967). Nonparametric and post hoc comparisons for trend. *Psychological Bulletin*, 67, 401-412.
- Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks-Cole.
- Maxwell, S. E., & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Erlbaum.
- 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.
- Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*, 2(1), 45-64.
- Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216-230.
- Pillai, K. C. S. (1960). *Statistical tables for tests of multivariate hypotheses*. Manila: University of the Philippines, Statistical Service Center.
- Randles, R.H., & Wolfe, D.A. (1979). *Introduction to the theory of nonparametric statistics*. NY: Wiley.
- Rasmussen, J. L. (1989). Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity. *British Journal of Mathematical & Statistical Psychology*, 42, 91-102.
- Sawilowsky, S., S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform in factorial ANOVA. *Communications in Statistics*, 14, 25-267.
- Salter, K. C., & Fawcett, R. F. (1993). The ART test of interaction: A robust and powerful test of interaction in factorial models. *Communications in Statistics: Simulation & Computation*, 22, 137-153.

Scheffé, H. (1959). *The Analysis of Variance*. NY: Wiley.

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

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.

Sheehan-Holt, J. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational & Psychological Measurement*, 58, 861-881.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27, 799-811.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 24-31.

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.

Waldstein, S. R., et al. (1991). Learning and memory function in men with untreated blood pressure elevation. *Journal of Consulting & Clinical Psychology*, 59, 513-517.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34, 28-35.

Wilcox, R. (1993). Robustness in ANOVA. In *Applied analysis of variance in the behavioral sciences*, E. Edwards (Ed.), pp. 345-374. NY: Marcel Dekker.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*, (3rd ed.). NY: McGraw-Hill.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education*, 64, 351-362.

Zimmerman, D., & Zumbo, B. (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.

Appendix I: SAS Code

```
data egtwo;
input k1 k2 k3 k4 group;
cards;
3.90 4.20 5.10 5.10 1
4.10 4.00 5.00 5.20 1
4.30 5.00 5.40 6.10 1
5.00 5.10 6.00 6.00 1
5.20 5.30 5.90 7.20 1
5.80 6.00 7.00 7.00 1
6.10 6.20 7.30 7.10 1
12.00 13.10 13.30 14.00 1
3.00 9.20 10.10 11.30 2
4.10 10.10 11.00 12.20 2
4.00 11.20 11.90 13.00 2
4.20 12.30 13.10 17.20 2
5.20 11.20 14.30 15.20 2
5.00 11.30 13.40 18.30 2
6.00 12.20 14.90 17.00 2
6.20 12.40 14.00 17.90 2
7.30 13.50 15.20 18.20 2
13.00 14.50 15.00 18.10 2
3.00 4.90 7.70 11.60 3
4.10 6.10 7.70 11.70 3
5.10 5.90 8.10 13.20 3
5.00 5.80 8.70 12.80 3
5.30 6.30 9.70 13.90 3
5.90 6.30 8.70 13.00 3
6.10 7.00 9.90 13.10 3
12.50 14.10 16.20 20.30 3
;proc sort out=two;by group;
data three;options ls=120;
proc iml; use two;
```

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

```

read all var{k1 k2 k3 k4} into
Y;read all var{group} into
Group;
JJ=max(Group);K=ncol(Y);N=nrow(Y
);NV=j(JJ,1,0);
Q=j(N,K,0);FR=j(N,K,0); CK=j((k-
1),K,0);CJ=j((jj-1),JJ,0);
dfjk=(JJ-1)*(K-1);dfeu=N-JJ)*K-
1);
smv=min(JJ,K);smv=smv-1;
mmv=(ABS(K-JJ)-1)/2;
nmv=(N-JJ-K)/2;
dfem=2#((smv#nmv)+1);
do hh=1 to JJ;
do ii=1 to N;
if group[ii,1]=hh then
NV[hh,1]=NV[hh,1]+1;
end;end;
RMMEAN=Y[:,];RMMEAN=(j(N,1,1))*R
MMEAN;
PMEAN=Y[:,];GMEAN=PMEAN[:,];PMEA
N=PMEAN*(j(1,K,1));
AD=(Y-PMEAN-
RMMEAN)+GMEAN;AR=RANKTIE(AD);AR=
AR/((N*K)+1);
do hh=1 to K;
do ii=1 to K;
DX=Y[:,hh]-
Y[:,ii];RDX=RANKTIE(DX);Q[:,hh]=Q[
,ii]+RDX;
end;end;
Q=(Q-((N+1)/2))/((K-1)*(N+1));
do ii=1 to N;
FR[ii,]=RANKTIE(AD[ii,]);
end;
FR=(FR-((K+1)/2))/((K##2)-
1)/12);
do ii=1 to (K-1);
CK[ii,ii]=1; CK[ii,(ii+1)]=-1;
end;
do ii= 1 to (JJ-1);
CJ[ii,ii]=1;CJ[ii,(ii+1)]=-1;
end;
CJK=CJ@CK;
AMEANK=AR[:,];QMEANK=Q[:,];RMEAN
K=FR[:,];
do ii=1 to JJ;
if ii=1 then zz=1;else
zz=zz+NV[(ii-1),1];
if ii=1 then zzz=NV[ii,1];else
zzz=zzz+NV[ii,1];
do hh=zz to zzz;
if hh=zz then AJ=AR[hh,]; else
AJ=AJ//AR[hh,];
if hh=zz then QJ=Q[hh,]; else
QJ=QJ//Q[hh,];
if hh=zz then RJ=FR[hh,]; else
RJ=RJ//FR[hh,];
end;
MAJ=AJ[:,];DMAJ=MAJ-
AMEANK;DMAJ=DMAJ#(NV[ii,1]);
EAJ=AJ-
((j((NV[ii,1]),1,1))*MAJ);
if ii = 1 then AMEAN=MAJ; else
AMEAN=AMEAN//MAJ;
if ii = 1 then DEVA=MAJ; else
DEVA=DEVA|MAJ;
if ii = 1 then EA=EAJ; else
EA=EA//EAJ;
MQJ=QJ[:,];DMQJ=MQJ-QMEANK;
DMQJ=DMQJ#(NV[ii,1]);
EQJ=QJ-
((j((NV[ii,1]),1,1))*MQJ);
if ii = 1 then QMEAN=MQJ; else
QMEAN=QMEAN//MQJ;
if ii = 1 then DEVQ=MQJ; else
DEVQ=DEVQ|MQJ;
if ii = 1 then EQ=EQJ; else
EQ=EQ//EQJ;
MRJ=RJ[:,];DMRJ=MRJ-
RMEANK;DMRJ=DMRJ#(NV[ii,1]);
ERJ=RJ-
((j((NV[ii,1]),1,1))*MRJ);
if ii = 1 then RMEAN=MRJ; else
RMEAN=RMEAN//MRJ;
if ii = 1 then DEVR=MRJ; else
DEVR=DEVR|MRJ;
if ii = 1 then ER=ERJ; else
ER=ER//ERJ;
end;
EA=EA`*EA;TA=AR-
((j(N,1,1))*AMEANK);TA=TA`*TA;
EQ=EQ`*EQ;TQ=Q-
((j(N,1,1))*QMEANK);TQ=TQ`*TQ;
ER=ER`*ER;TR=FR-
((j(N,1,1))*RMEANK);TR=TR`*TR;
HTA=((CJK*(DEVA`))`)*(ginv((CJK*
(diag((1/nv)))@EA)*((CJK`))))*
(CJK*(DEVA`));
VA=
((CJK*(DEVA`))`)*(ginv((CJK*(di
ag((1/nv)))@TA)*((CJK`))))*(CJK
*(DEVA`));

```

BEASLEY & ZUMBO

```

HTQ= ((CJK*(DEVQ`))`)* (ginv((CJK*
((diag((1/nv)))@EQ)*((CJK`)))))*
(CJK*(DEVQ`));
VQ=
((CJK*(DEVQ`))`)* (ginv((CJK*((di
ag((1/nv)))@TQ)*((CJK`)))))* (CJK
*(DEVQ`));
HTR= ((CJK*(DEVR`))`)* (ginv((CJK*
((diag((1/nv)))@ER)*((CJK`)))))*
(CJK*(DEVR`));
VR=
((CJK*(DEVR`))`)* (ginv((CJK*((di
ag((1/nv)))@TR)*((CJK`)))))* (CJK
*(DEVR`));
FHA=HTA#(dfem/(smv#dfjk)); pvalam
=1-(probf(FHA,dfjk,dfem));
FHQ=HTQ#(dfem/(smv#dfjk)); pvalqm
=1-(probf(FHQ,dfjk,dfem));
FHR=HTR#(dfem/(smv#dfjk)); pvalrm
=1-(probf(FHR,dfjk,dfem));
FA=(((CJK*(DEVA`))`)* (ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)* (CJK*(DEVA`)))/(TRACE(EA))* (df
eu/dfjk);
pvalau=1-(probf(FA,dfjk,dfeu));
FQ=(((CJK*(DEVQ`))`)* (ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)* (CJK*(DEVQ`)))/(TRACE(EQ))* (df
eu/dfjk);
pvalqu=1-(probf(FQ,dfjk,dfeu));
FRC= ((CJK*(DEVR`))`)* (ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)* (CJK*(DEVR`)))/((K#(K+1))/12);
pvalru=1-(probchi(FRC,dfjk));

Print 'Univariate Tests';
Rowun={"Aligned Ranks F(A)",
"Koch Ranks F(Q)",
"Chi-Square - Friedman Ranks
F(R)"};
ColUN={"TEST" "DFh" "DFe" "p-
value"};
UPrt=(FA//FQ//FRC)|| (dfjk//dfjk/
/dfjk)|| (dfeu//dfeu//0)||
(pvalau//pvalqu//pvalru);
print UPrt [rowname=rowun
colname=colun];

Print 'Multivariate Tests';Print
'DFh =' dfjk;Print 'DFe =' dfem;

```

```

Rowmn={"Aligned Ranks (A)",
"Koch Ranks (Q)", "Friedman
Ranks (R)"};
ColmN={"Pillia Trace V(*)"
"Hotelling Trace H(*)" "F-
approx" "p-value"};
MpRt=(VA//VQ//VR)|| (HTA//HTQ//HT
R)|| (FHA//FHQ//FHR)|| (pvalam//pv
alqm//pvalrm);
print MPrt [rowname=rowmn
colname=colmn];

DLINE={-3 -1 1
3};DLINE=DLINE/(20##.5);
DCHNG={-1 1 0 0};
YL=Y*(DLINE`);YC=Y*(DCHNG`);
AL=AR*(DLINE`);AC=AR*(DCHNG`);
QL=Q*(DLINE`);QC=Q*(DCHNG`);
FL=(FR*(DLINE`)#2)/4;FC=(FR*((
DCHNG`)#(2##.5)))/4;

outx=Y||AR||Q||FR||YL||YC||AL||A
C||QL||QC||FL||FC||Group;
create xxx from outx[colname={k1
k2 k3 k4 ak1 ak2 ak3 ak4 qk1 qk2
qk3 qk4 fk1 fk2 fk3 fk4 yl yc al
ac ql qc fl fc group}];
append from outx;
data last;set xxx;
proc glm;class group;
model k1 k2 k3 k4=group/nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/summary;run;
proc glm;class group;
model ak1 ak2 ak3
ak4=group/nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
proc glm;class group;
model qk1 qk2 qk3 qk4 = group /
nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;

```

ALIGNED RANK TESTS FOR INTERACTIONS IN SPLIT-PLOT DESIGNS

```
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
proc glm;class group;
model fk1 fk2 fk3 fk4 = group /
nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
```

```
polynomial/ summary;run;
proc glm;class group;
model yl yc al ac ql qc fl fc =
group / nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
```

Quantifying Bimodality Part 2: A Likelihood Ratio Test for the Comparison of a Unimodal Normal Distribution and a Bimodal Mixture of Two Normal Distributions

B. W. Frankland
Dalhousie University

Bruno D. Zumbo
University of British Columbia

Scientists in a variety of fields are often faced with the question of whether a sample is best described as unimodal or bimodal. In an earlier paper (Frankland & Zumbo, 2002), a simple and convenient method for assessing bimodality was described. That method is extended by developing and demonstrating a likelihood ratio test (LRT) for bimodality for the comparison of a unimodal normal distribution and a bimodal mixture of two normal distributions. As in Frankland and Zumbo (2002), the LRT approach is demonstrated using algorithms in SPSS.

Key words: Bimodality, likelihood ratio test, mixture distribution, SPSS.

Introduction

Previously, a method for assessing bimodality using the non-linear algorithms in SPSS was presented (Frankland & Zumbo, 2002). It is a method for modeling complex mixture distributions with a unimodal normal distribution (with 2 free parameters) and with a bimodal mixture of two normal distributions (with 5 free parameters). The current work extends that previous work to the development of a likelihood ratio test (LRT) for bimodality. In this extension, the research question is: Does a bimodal mixture of two normal distributions represent a significantly better fit to the data than a unimodal normal distribution? Here, the fit of the data to the unimodal normal distribution is considered the null hypothesis. The fit of the data to the bimodal mixture of

two normal distributions is considered the alternative hypothesis. The null hypothesis is rejected if it provides a significantly poorer fit to the data.

As noted in Frankland and Zumbo (2002), the techniques developed herein are focused on putative mixtures of normal distributions; they can be applied, in principle, to the comparison of any set of theoretical distributions. Normal distributions were chosen as the focus because it is likely that the normal distribution is a reasonable approximation to the data, either as a single unimodal distribution, or as each component of the mixture of two distributions. It is admitted, a priori, that the solution offered is not an analytical solution to the question of bimodality. The point was to develop an accessible, flexible and, most importantly, accurate method that could be used to test any number of hypotheses. The procedure uses the commercially available statistical package SPSS (most statistical packages should be capable of comparable analyses) to accomplish a Monte Carlo simulation to generate the likelihood ratio distribution for the bimodal/unimodal comparison. Because it can be assumed that most researchers will use this technique to analyze a single (or limited number

B. W. Frankland is an adjunct professor in Psychology at Dalhousie University. Email him at: Brad.Frankland@dal.ca. Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as a member of the Department of Statistics and the Institute of Applied Mathematics. Email him at bruno.zumbo@ubc.ca.

TESTING BIMODALITY

of) data set, the application is demonstrated within that context.

The Likelihood Ratio Method

A set of empirically determined data, of size n (lower case n), is compared to two hypothetical population distributions. It is assumed that the data is represented as a histogram (hereafter, data histogram, or histogram; the term empirical data will refer to the original pre-binned data).

The histogram will define the number of bins, and their statistics (lower limit, center, upper limit) for the subsequent analyses (see Frankland & Zumbo, 2002). This determination should be made in the context of subsequent simulation. The sample size (n) is the most important factor for creating bins. The most efficient method is to determine the mean and standard deviation of the sample using a traditional method. These are estimates of the mean and standard deviation (μ , σ) of the corresponding normal population. Thereafter, the number of bins per standard deviation is set to accommodate the expected range and density of scores for any sample of size n , from this particular population, $N(\mu, \sigma)$. For example, given $n = 500$, one could use 10 bins per sd, with a full range of z-scores from -5.0 to 5.0. (This point will be discussed more fully later.) These bins can then be adjusted to fit the actual data.

Alternatively, the raw data can be converted to z-scores, and the likelihood ratio test can be conducted using z-scores. The likelihood ratio test is agnostic with respect to the original scale of the data. The use of z-scores is more convenient for testing multiple data sets. However, the fitted statistics for the unimodal and bimodal distributions are not obtained. Here, z-scores were used (see Frankland & Zumbo, 2002 for raw scores).

In the first step, the best-fit parameters for the unimodal and bimodal functions are determined (see Frankland & Zumbo, 2002). The data histogram is first compared to a function that describes a hypothetical unimodal normal distribution (hereafter, unimodal function). With the unimodal function, the free parameters to be determined are the mean (μ) and standard deviation (σ , or variance, σ^2):

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \quad (1)$$

The mean and standard deviation could be easily obtained using traditional methods, but for this application, the mean and standard deviation must be determined using a method that compliments that which is used for the bimodal values.

The histogram is compared to a function that describes a hypothetical bimodal mixture of two normal distributions (i.e., bimodal function). In this case, there is a mean (μ_1 , μ_2) and a standard deviation (σ_1 , σ_2) for each normal distribution, as well as, the mixture proportion (λ ; note that some authors use π , and others use α , for this parameter):

$$\begin{aligned} B(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \lambda) &= \\ \lambda * \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(X_i - \mu_1)^2}{2\sigma_1^2}} &+ \\ (1 - \lambda) * \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(X_i - \mu_2)^2}{2\sigma_2^2}} & \\ = \lambda * N_1(\mu_1, \sigma_1^2) + (1 - \lambda) * N_2(\mu_2, \sigma_2^2) & \quad (2) \end{aligned}$$

Means and variances with subscripts refer to those from the bimodal distribution. In addition to the best-fit parameters, the likelihood that the sample came from a unimodal population, and the likelihood that the sample came from a bimodal population are determined. These two likelihoods are converted to a ratio (the likelihood ratio, LR_{data}).

In the second step, a Monte Carlo simulation is used to create the likelihood ratio test (*LRT*) for bimodality. In this step, a normal distribution is defined, $N(0, 1)$. From this unimodal population, a sample is taken. This sample is converted to a histogram using the same bin parameters as defined previously. The bins used for the raw data must match the bins used for the simulation. The binned sample from the normal distribution is then fitted to a unimodal function and to a bimodal function. Finally, the likelihood ratio for the sample is computed. This process is repeated for a large

number of samples drawn at random from the defined unimodal normal population. From the set of samples, a distribution of the likelihood ratios is created. The likelihood ratio distribution provides a direct assessment of the probability of getting the original likelihood ratio (LR_{data}), if the data were drawn from a unimodal normal population. If that probability is low, then the original data is assumed to be bimodal. This is simply type 1 error rate, which is normally set to $\alpha = 0.05$.

In the simulation there are K samples, each being denoted by k . Each of the K samples is based on n data points, drawn from a normal distribution. Each data point is X_i . These n data points per sample are converted to a histogram: The number and the boundaries of bins are determined by the original data (i.e., bin centers and limits reflect the raw data, X). There are I_i bins ($I = 1$ to I). The initial definition of the bins should encompass the full plausible range of the data (i.e., ideally, the tails should stretch to infinity). For each sample, one likelihood-ratio statistic (LR_k) is produced. The distribution of K likelihood-ratio statistics (LR_k , $k = 1$ to K) statistics provides the test of likelihood ratio of the data (LR_{data}). Note that the original empirical data determines the sample size n . As will be discussed, the sample size is the primary determinate of the number of bins, I . Time, computational resources and desired accuracy determine K . The procedure is demonstrated with a specific example.

The Original Data

For this demonstration, a bimodal data set of $N = 500$ data points was created. The data set consisted of a mixture of two normal distributions. Each distribution was obtained using the SPSS command NORMAL, which generates standard Normal pseudo-random variates. The first distribution was $N(\mu_1, \sigma_1) = N(-1.0, 0.7)$ and the second was $N(\mu_2, \sigma_2) = N(1.0, 1.0)$. Note that the variances are different. The data set consisted of 60% from the first distribution and 40% from the second distribution (sd), and is notated as $B(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = B(-1.00, 0.71, 1.00, 1.00, 0.60)$. The raw data had a mean of .169, a standard deviation (sd) of 1.296, a skew of .412 + .109 and a kurtosis of .371 + .218. The median was .329.

These data were converted to z-scores and then binned. By design, there were 10 bins per sd and a full range of $-5.0 \leq z \leq 5.0$. Each bin had a width of .01 sd. There was a single bin centered at $z = 0$ (hence, the bin was defined as $-.05 \leq z \leq .05$). By design, there were 101 bins in total, with the last being $4.95 \leq z \leq 5.05$ and $-5.05 \leq z \leq -4.95$. However, the bins in the tails were widened to encompass the ranges $4.95 \leq z \leq 6.95$ and $-6.95 \leq z \leq -4.95$. This captures the skewness that can manifest in an empirical bimodal distribution (alternatively, one can use a larger range of bins). The resulting distribution is shown in Figure 1.

The z-scores in the raw data ranged from -2.09 to 3.07, and after binning, there were only 50 bins with non-zero counts (see Figure 1). However, the full range of bins must be provided, with zero counts for those that are empty. This is important for the subsequent simulations. There are ways to create empty bins in SPSS, but for a single data set, the manual method is about as fast as any other. The data do not appear to be bimodal, although they are not obviously normal either (it simply seems skewed). Based on counts per bin, the binned distribution, with $I = 50$, produced $(\mu, \sigma) = (-.002, 1.001)$. This is slightly altered from the original raw data. This alteration is important because all subsequent analyses are based on the binned data.

The subsequent analysis uses the bin lower limit (xl), bin center (xc) and upper limit (xu), so the SPSS data file is expected to contain the following variables:

- Binnum: bin number (not actually used, but useful for humans)
- Observed: observed count per bin (X)
- xl: bin lower limit in the original scores
- xc: bin center in the original scores
- xu: bin upper limit in the original scores
- Total: total counts (total number of data points, a constant)

Fitting the Original Data

As described previously (Frankland & Zumbo, 2002), when fitting the unimodal or bimodal functions, the algorithm determines the parameters for the unimodal, $N(\mu, \sigma)$, and

TESTING BIMODALITY

Figure 1: The Empirical Bimodal Distribution $B(-1.00, 0.71, 1.00, .1.00, 0.60)$

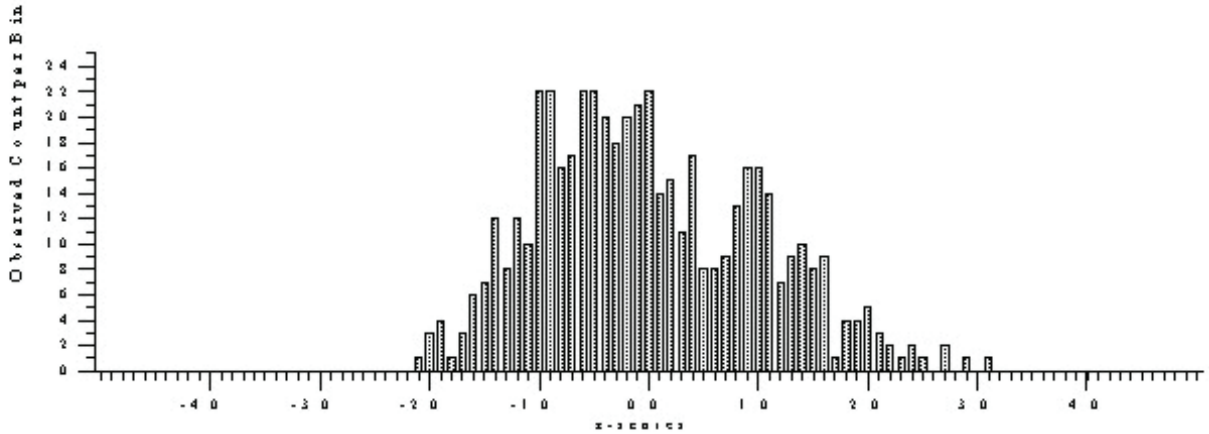


Figure 1: The empirical bimodal distribution, $B(-1.00, 0.71, 1.00, .1.00, 0.60)$.

bimodal, $B(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)$, distributions using an iterative, sequential-quadratic, search algorithm. The algorithm determines the values of μ and σ (or $\mu_1, \sigma_1, \mu_2, \sigma_2$, and λ) so that the predicted count per bin (\hat{Y}_i or Y^*_i) forms the best possible match to the by minimizing the sum of the squared deviations between theoretical count per bin and the actual count per bin (Y_i). Conceptually, the fitting procedure is the same as ordinary, unweighted, least-squares regression (OLS) with X_i being the center of the bin, and Y_i being the actual count per bin. The X_i are transformed non-linearly to create predicted bin count \hat{Y}_i (or Y^*_i).

The parameters of the functions are adjusted iteratively until \hat{Y}_i produces the best match to Y_i , ascertained by minimizing the sum of the squared deviations between the predicted and actual, $\Sigma e^2_i = \Sigma(\hat{Y}_i - Y_i)^2$. Relative to OLS, only the method of fitting is different. Note that, in this analysis, every bin has the same contribution to the final solution regardless of the number of scores per bin. The predictions, \hat{Y}_i , are not weighted by sample size per bin. This is the simplest approach, but a weighted approach could be developed (i.e., weighted by bin count or, equivalently, bin error).

Because each bin has an equivalent contribution to the final solution, one must

choose the bins carefully. After much trial and error, bolstered by post hoc rationalizations, it seems that empty bins for the full range of z-scores should be included in the analysis. This is not a χ^2 solution, though there are links to that methodology. Such data should not be dropped (or trimmed). That is, empty bins between bins with counts, and empty bins representing the tails of the distribution should be retained or added to the histogram. These empty bins in the tails can be combined if necessary.

First, as noted, after much trial and error, the inclusion of empty bins does not seem to make a lot of difference to the final solution. The fitted parameters do change, but the change is within the error of all approaches. However, the inclusion of empty bins has many benefits for the later simulations.

Second, the true functions that are being fitted technically stretch to $\pm\infty$. It is only by virtue of sample size that the data does not stretch to infinity. Having empty bins in the tails forces the functions to go to zero when they should go to zero. Alternatively, the inclusion of empty bins in the tails is equivalent to forcing the regression solution to go through the origin, which is reasonable (the distributions approach zero asymptotically). At this point, the goal is to

find the best fitting parameters to define the populations.

Third, bins in the tails (particularly those distant from the center of the data) represent real, though rare, data. Assuming that the original sampling that led to the data is truly random, deleting such outliers would be equivalent to lobbing off a part of the population. On the other hand, if retained, these outliers have high leverage in the solution. Empty bins in the tails also represent real information (low probability events). The inclusion of empty bins in the tails has the effect of reducing the leverage associated with the retention of outliers. That is, in the solution, a number of bins with $(\hat{Y}, Y) = (\sim 0, 0)$ will balance a few bins with $(\hat{Y}, Y) = (\sim 0, 1)$ or $(\sim 0, 2)$.

Last, bins in the tails can be combined. The fitting functions work by determining the probability of observed data per range of z-scores. The functions use numerical integration with a trapezoid rule. In the tails, the functions are relatively flat (or, at least, approaching linear). Hence, in the tails, the use of a trapezoid rule with wider bins would not introduce large distortions.

Bins should define a reasonable range of data that can incorporate the full range of the data, including the possible range that might occur in the subsequent simulation. A simple definition would use a range of $\pm 5\sigma$ (i.e., z-scores). More sophisticated estimates can be made, particularly with very large samples, but this seems to be a useful default value.

The fitting algorithms for the unimodal and bimodal functions are shown in Listing 1 (also see Frankland & Zumbo, 2002). Note that probabilities are actually computed using a two-trapezoid rule per bin, with three values (X_l , X_c , X_u) per bin (the routine uses proportions per bin, but it could be written to use actual counts).

Listing 1

```
compute      prop = observed/total.
model program mean= 0.0 sd = 1.0.
compute xa = abs(xl - xc).
compute xb = abs(xu - xc).
compute h1 = (.398942/sd)
              * exp(-(((xl-mean)**2) / (2*sd**2))) .
```

```
compute h2 = (.398942/sd)
              * exp(-(((xc-mean)**2) / (2*sd**2))) .
compute h3 = (.398942/sd)
              * exp(-(((xu-mean)**2) / (2*sd**2))) .
compute predun = .5 * (h1+h2) * xa
                + .5*(h2+h3)*xb.
cnlr      prop.
          /pred = predun
          /bounds sd gt 0.0001
          /save = predun residun.
```

```
model program mean1 = -1.0 mean2 = 1.0
          sd1 = 1.0 sd2 = 1.0 ratio = 0.5.
compute xa = abs(xl - xc).
compute xb = abs(xu - xc).
compute h1 = (.398942/sd1)
              *exp(-(((xl-mean1)**2)/(2*sd1**2))) .
compute h2 = (.398942/sd1)
              *exp(-(((xc-mean1)**2)/(2*sd1**2))).
compute h3 = (.398942/sd1)
              *exp(-(((xu-mean1)**2)/(2*sd1**2))).
compute h4 = (.398942/sd2)
              *exp(-(((xl-mean2)**2)/(2*sd2**2))) .
compute h5 = (.398942/sd2)
              *exp(-(((xc-mean2)**2)/(2*sd2**2))).
compute h6 = (.398942/sd2)
              *exp(-(((xu-mean2)**2)/(2*sd2**2))).
compute predbi = ratio *.5*(h1+h2)*xa
                + .5*(h2+h3) * xb
                + (1-ratio)*(.5*(h4+h5)*xa
                + .5*(h5+h6) * xb).
cnlr      prop.
          /pred = predbi
          /bounds sd1 gt 0.0001; sd2 gt 0.0001;
          1.0 ge ratio ge 0.0
          /save = predbi residbi.
```

The constraints (bounds) are placed on the values of variances and the ratio. A constant could be included in the equations. In practice, it seems to make little difference for the fit of either function. More precisely, other factors, particularly the width of the bins, have a greater effect. The routine produces the predicted proportion per bin, \hat{Y}_i (or Y'_i , notated as predun and predbi) and the residual, $e_i = Y'_i - Y_i$ (notated as residun and residbi). These variables are added to the data file. The sum of the residuals should be zero. The sum of the residuals-squared ($\sum e_i^2$) is equivalent to $SS_{Y,X}$ in

TESTING BIMODALITY

OLS (i.e., $SS_{Y,X} = \Sigma e_i^2$), which can be converted to the standard error of estimate, $s_{Y,X}$, and eventually R^2

For the current data, after creating bins that stretched to $\pm 7\sigma$, the analysis using the bimodal function produced $B(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) = B(-0.489, 0.681, 1.145, 0.497, 0.734)$, with $s^2_{Y,X} = 1.389 \cdot 10^{-5}$ and $R^2 = .933$. The standard errors on the parameters estimates are 0.046, 0.043, 0.075, 0.064, and 0.041 respectively. When converted back to raw scores using the inverse of the z-transform, these correspond to $B(-0.802, 0.881, 1.312, 0.643, 0.734)$. This compares acceptably with the parameters used to define the population. The analysis using the unimodal function produced $N(\mu, \sigma) = N(-0.163, 1.038)$, with $s^2_{Y,X} = 2.419 \cdot 10^{-5}$ and $R^2 = .880$. The standard errors on the parameters estimates are 0.044 and 0.036 respectively. Converted to raw scores, one has $N(-0.380, 1.343)$. The fitted functions are layered on top of the original data in Figure 2.

The s^2_Y used for the computation of R^2 is the variance of the counts, not the variance of the original data. The point here is not to compare the parameters returned by algorithm to those of the optimal solution. Rather, the point is to compare the fits using the unimodal and bimodal functions when computed using the same routine. Note that the change in fit is $\Delta R^2 = .933 - .880 = .053$.

For comparison purposes, if the histogram is cut off at the edge of the data (i.e., $-2.2 < z < 3.2$), but retaining the empty bins between those extremes, one obtains $B(-.486, .686, 1.150, .491, .739)$ with $s^2_{Y,X} = 2.764 \cdot 10^{-5}$ and $R^2 = .880$. Note that these are within the errors cited above. For the unimodal function, one gets $N(-.176, 1.054)$ with $s^2_{Y,X} = 5.720 \cdot 10^{-5}$ and $R^2 = .764$.

If all the empty bins are removed (even those between other non-empty bins), one obtains $B(-.486, .686, 1.150, .491, .739)$ with $s^2_{Y,X} = 2.948 \cdot 10^{-5}$ and $R^2 = .866$, and $N(-.175, 1.055)$ with $s^2_{Y,X} = 4.777 \cdot 10^{-5}$ and $R^2 = .769$. Clearly, all three methods produce equivalent fits and parameters.

As expected, in all cases the bimodal function produced the better fit between \hat{Y}_i and Y_i . when using the same method (smaller error, higher R^2). It is interesting that $s^2_{Y,X}$ is smaller when there are more bins (i.e., more X and Y points). This is counterintuitive, but it implies that the additional points – the empty bins – have very little error (so that the average error decreases). In addition, note that the choice of bin values does not affect the *relative* fits dramatically. The ratios $s^2_{Y,X,b}/s^2_{Y,X,u}$ are .662, .691 and .710 respectively, while the ΔR^2 are .055, .082 and .098 respectively.

Figure 2: The Unimodal Bimodal Curve Fits to the Empirical Bimodal Distribution

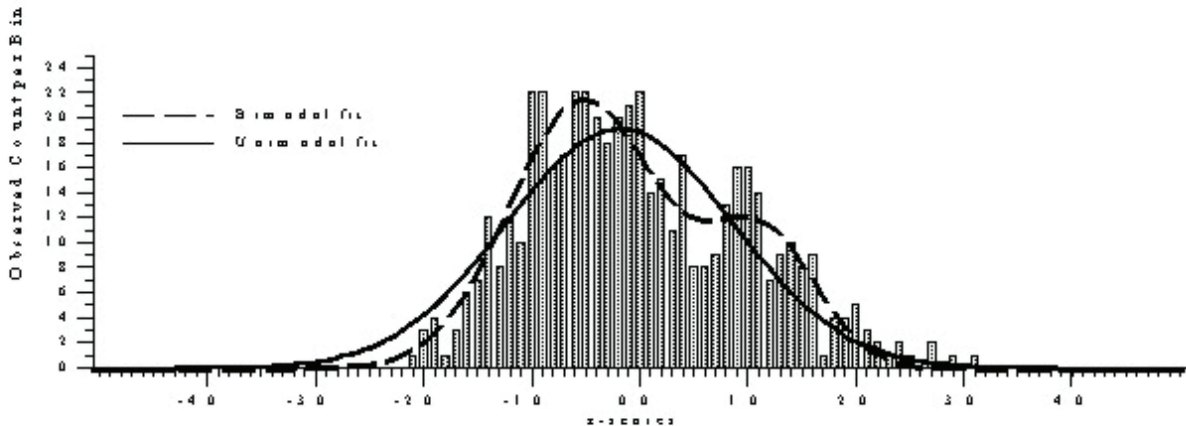


Figure 2: The unimodal and bimodal curve fits to the empirical bimodal distribution.

The Likelihood Ratio for the Original Data

These two fits cannot be directly compared (either R^2 or $s^2_{Y,X}$) because the population with greater number of free parameters will generally produce the better fit. The reason for this is somewhat oblique to the statistical analysis. The premise is that when comparing two theories (i.e., two populations) both theories will have been selected by past research to be reasonable fits to the data (even if only by eye). Hence, both functions will approximately match the data, so the function with the more flexibility (more df's) will generally fit better.

Instead, the likelihood that the data came from a unimodal population can be compared to the likelihood that the data came from a bimodal population (this is almost Bayesian). The probability, or likelihood, of getting the particular set of data if, in fact, that data came from the specified unimodal, $N(\mu, \sigma)$ is $L(N)$. It is also known as L_0 , since the simpler unimodal distribution will become the null hypothesis. Similarly, for the bimodal population, $B(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)$, the probability is $L(B)$ or L_A since the more complex bimodal population will become the alternative hypothesis. In each case, the probability is nothing more than the product of the probabilities for the individual bins, i ($i = 1$ to I).

$$L(N) = L_0 = \prod_i P_i(\mu, \sigma) \quad (3)$$

$$L(B) = L_A = \prod_i P_i(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda) \quad (4)$$

L_0 and L_A represent the maximum likelihood solutions for each population: They are, in some sense, the best possible fits between the data and the corresponding function, and therefore represent the maximum probabilities (likelihoods) for each hypothesis. The usual mean and standard deviation is, in fact, the maximum-likelihood solution for the normal distribution.

Usually, to compare the two hypotheses, the ratio of likelihoods is computed. This is the likelihood ratio (LR_{data} or Λ_{data}) for the data. In the likelihood ratio, for reasons that will be

obvious momentarily, the simpler or null hypothesis is placed in the numerator.

$$LR_{data} = \Lambda_{data} = L_0/L_A \quad (5)$$

This ratio will be bounded by (0.0, 1.0). A ratio near 0 indicates that the alternative hypothesis (the bimodal distribution) is a much better fit, and a ratio near 1 indicates that both hypotheses provide equivalent fits. A ratio much greater than 1 should be impossible since the unimodal should not be able to provide a better fit than the bimodal distribution.

Because there are many computational advantages, one usually works with the natural logarithm of the likelihood ratio. Hence, one usually has:

$$\begin{aligned} -2\ln(\Lambda_{data}) &= -2 \ln\left(L_0/L_A\right) \quad (6) \\ &= 2 \ln(L_A) - 2 \ln(L_0) \\ &= 2 \ln[\prod_i P_i(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)] - 2 \ln[\prod_i P_i(\mu, \sigma)] \\ &= 2 \sum_i \ln[P_i(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)] - 2 \sum_i \ln[P_i(\mu, \sigma)] \\ &= 2 \sum_i \{\ln[P_i(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)] - \ln[P_i(\mu, \sigma)]\} \\ &= LLR_{data} \end{aligned}$$

The value of logarithm of the likelihood ratio, $LLR = -2\ln(\Lambda)$, is bounded by $(-\infty, \infty)$, although very large positive or negative values (< -1000 , > 1000) would not be expected. A zero indicates equivalent fits, negative values imply that the unimodal is a better fit while positive values imply that the binomial is a better fit. Large positive values lead to rejection of the null hypothesis. The important point (for algorithms) is that by using $\ln(L_A)$ and $\ln(L_0)$, one converts the previous products and their ratio into a series of sums. Most importantly, the difference between the two hypotheses can be computed on a bin-by-bin basis, and then summed.

To find the ratio, the likelihood that the data comes from the best-fit unimodal distribution must be determined, along with the likelihood that the data comes from best-fit bimodal distribution. This has not been detailed

TESTING BIMODALITY

in previous work (i.e., Frankland & Zumbo, 2001). To obtain these probabilities, for each bin, the probability of an observed count, given the theoretical bin count must be determined. If it is assumed that, within any single bin, i ($i = 1$ to I), the counts per bin follow a Poisson distribution (a normal distribution per bin may also be assumed), the probability of any observed count (Y_i) can be obtained given the theoretically predicted count (\hat{Y}_i or Y'_i). Within each bin, the probability for any particular count is (subscripts have been dropped for clarity):

$$p(Y/\hat{Y}) = \frac{\hat{Y}^Y}{Y!}$$

$$\ln[p(Y/\hat{Y})] = Y \ln(\hat{Y}) - \ln(Y!) - \hat{Y} \quad (7)$$

Essentially, the predicted count, \hat{Y} , is nothing more than a non-linear transformation of the bin value (e.g., collectively, xl, xc, xu). Listing 2 provides a method for computing the probabilities in each bin, given predicted counts from the unimodal and bimodal functions. Note that since the Poisson distribution uses counts, not proportions, proportions (of Listing 1) are converted into counts. The loop simply computes the factorial. The loop should not be executed if the observed count is zero.

Listing 2: Poisson Probabilities per Bin: Unimodal and Bimodal Functions

```
compute expectun = predun * total.
compute expectbi = predbi * total.
compute poisun = (expectun**observed)
               * exp(-1*expectun).
compute poisbi = (expectbi**observed)
               * exp(-1*expectbi).
loop          #i = 1 to observed.
  compute poisun = poisun / #i.
  compute poisbi = poisbi / #i.
end loop.
compute lnpoisun = ln(poisun).
compute lnpoisbi = ln(poisbi).
compute llrdata = 2 * (lnpoisbi - lnpoisun).
```

Bin-by-bin, the probabilities (poisun and poisbi) are converted to logs and then subtracted. In the

final step, the differences would be summed to create the ratio LLR_{data} .

However, the use of logarithms has many benefits. As noted in Equations 6 and 7 (and Listing 2), the factorial depends on the observed count and as such, is the same for both the unimodal and bimodal functions on a bin-by-bin basis. When converted to logs, the factorials become sums that cancel in each bin. Hence, the pesky loop to compute the factorial is not needed, removing complications arising from bins with zero counts. The early transition to logarithms also prevents possible overflow errors in the event that there is a large difference between the observed and predicted, and underflow errors in the event that $e^{-\hat{Y}}$ is very small. Hence, Listing 2 is revised as follows:

Listing 2 Revised: LRT using Poisson Probabilities per Bin

```
compute expectun = predun * total.
compute expectbi = predbi * total.
compute lnpoisun = observed
               * ln(expectun) - expectun.
compute lnpoisbi = observed
               * ln(expectbi) - expectbi.
compute llrdata = 2 * (lnpoisbi - lnpoisun).
```

Comment Sum the ln(LRdata) using the simple frequencies command.

```
frequencies prop expectun residun lnpoisun
              expectbi residbi lnpoisbi llrdata
              /format = notable
              /statistics = mean stddev variance
              minimum maximum sum.
```

The sum of llrdata is $LLR_{data} = -2\ln(\Lambda_{data})$, easily obtained from the descriptives or frequencies command of SPSS

For the current data, when using bins in the full range of $-5.0 < z < 5$, the $LLR = 28.645$. Note that this is far from the value of zero that would be expected if the data were truly unimodal. However, this is not surprising given that the data was designed as bimodal.

With the more restrictive range $-2.2 < z < 3.2$ (i.e., cut the histogram at the edge of the data, but retaining intervening empty bins), the value is $LLR = 4.259$ (but the simulation distribution changes accordingly). When no

empty bins are included the value is $LLR = 2.363$. One then uses these values to determine whether the sample is more likely to have come from a unimodal or bimodal population.

The other variables in the frequencies command provide quick, but useful, checks on the analysis. The sums of the proportion per bin (prop) or the predicted per bin (predun, predbi) should be one. The sums of the observed, expectun and expectbi should equal the number of data points ($N = 500$). The residuals should sum to zero. In addition, the variance of the residuals is essentially the squared-standard error of estimate ($s^2_{Y.X}$) for each function. The output includes the variance of the bin counts. This is s^2_Y . From this, one can compute the correlation $R^2 \approx 1 - s^2_{Y.X}/s^2_Y$. Although the CNLR routine will provide R^2 directly, this computation is useful when computing R^2 in a simulation (herein it serves as a further check).

The Sampling Distribution of Log-Likelihood Ratio (LLR)

The last step is to decide whether or not the observed ratio, $LLR_{data} = -2\ln(\Lambda_{data})$, is reasonable if, in fact, the null hypothesis is true. This is the likelihood ratio test (LRT) or more properly the log-likelihood ratio test ($LLRT$). To make this decision, one needs the theoretical sampling distribution of $LLR = -2\ln(\Lambda)$. This theoretical distribution is focused on the possible values of LLR when the data is taken from the defined unimodal normal distribution (i.e., from $N(\mu, \sigma)$, or L_0). The empirical or theoretical bimodal population is irrelevant to the creation of this distribution.

To create this distribution, one defines a normal distribution, and then takes a sample (notated by k) from that distribution. That sample is fit with a unimodal function, with a bimodal function and then the LLR_k is determined. If the two functions provide equivalent fits, the LLR_k is expected to be near 0.0, but in fact, a value slightly greater than 0.0 is expected (if the data is unimodal, the bimodal function will provide a better fit given its greater flexibility). This sample has the same sample size as the original data (n ; called total in Listing 2).

The process is repeated for K samples to create the sampling distribution of LLR . The

mean of this distribution is expected to be slightly greater than 0.0. The value of K reflects the desired precision in the final likelihood ratio distribution, weighted by the amount of patience.

If the observed value of $LLR_{data} = -2\ln(\Lambda_{data})$ for the single sample under consideration is unlikely given that distribution, then the null hypothesis is rejected – it is then concluded that the data is bimodal. That is, the null hypothesis is that the data is unimodal. In that case, the LLR_{data} should be near 1.0 (the mean of the likelihood ratio distribution). If the data is actually bimodal, then the value of LLR_{data} will be unexpectedly large. The usual criteria regarding Type 1 Error Rate (α) can be employed as the basis for the decision. If the LLR_{data} is one of those values that is so large that it would only be expected to occur 5% of the time (if the null hypothesis, L_0 , were to be true), then it is unlikely and the associated L_0 is unlikely, and the null hypothesis is rejected.

A Monte-Carlo simulation is used to create the sampling distribution of LLR . There are many nuances that can be varied for the simulation, but the point here is to create a basic template that can be used broadly or adapted to specific situations.

For that simulation, there are a couple of important observations. Firstly, for the simulation, the scaling of the bin centers (or limits) is irrelevant to the issue of computing the sampling distribution of the LLR . The routine assesses the relative match of the Y_i to the predictions based on the best fitting unimodal function ($\hat{Y}_{i,0}$) and bimodal function ($\hat{Y}_{i,A}$). Both the unimodal and bimodal functions use the same bin centers. Both $\hat{Y}_{i,0}$ and $\hat{Y}_{i,A}$ are simply non-linear transforms of the same underlying bin centers. Hence, the data can be conveniently rescaled so that the bins are coded in terms of z-scores, with a certain number of bins per standard deviation. The data only needs to be coded in the original units for discussion of the actual unimodal or bimodal means, variances, as well as the λ .

Secondly, the routine must run unattended. This requires careful consideration of the bin definitions. When taking random samples from a population (unimodal or bimodal), every sample in the simulation will

TESTING BIMODALITY

produce different data. Each data point will be assigned to one bin in the histogram. Therefore, one must provide the full range of possible bins for the data to fall into. Most critically, there cannot be any missing bins within the range of the data. That is, each sample may not produce a non-zero count for every bin, but for the simulation as a whole, every individual data point must fall within some bin. Hence, all possible bins must be defined a priori. However, this is not difficult to do because one knows that the K samples are derived from a normal distribution with known mean and variance.

It is important to remember that sampling from the theoretical normal distribution may produce data that extends beyond the range of the original raw data. Hence, the creation of the likelihood ratio distribution should allow for bins that encompass far more range than that of the original data. The bins should extend as far as is reasonable given the theoretical normal population and the empirical data to be tested. It is also appropriate (or safe) to retain a wide bin for each tail to capture the occasional data point that goes beyond the expected range.

The bins in the simulation must match those used with the histogram for the original data: If not, the wrong sampling distribution is created. Note that the LLR_{data} previously computed depended on the types of bins used. Hence, the bins used to construct the histogram for the original data, and the bins used for the simulation must be the same. This is most easily accomplished using a fixed number of bins per sd (e.g., 5 or 10 depending on n), with a range of bins that is adequate for both the raw data and the simulation. Bins in the tails can be made wider without affecting the solution. This is the logic behind the aforementioned range of $-5.0 < z < 5.0$ with bin for the lower tail expanded to -5.0 to -7.0 and the bin for the upper tail expanded to 5.0 to 7.0 . For example, if one is working with original data that contains $n = 100$ data points, the range of z-scores should be about $+3.5$ standard deviations with 4 or 5 bins per sd (hence 28 to 35 bins in total). This should result in reasonable counts near the center of the distribution, while allowing for the increased spread that is characteristic of a bimodal distribution.

If an original sample of $n = 100$ should have a large proportion of data extending beyond 3.5 standard deviations, one should question the need for a test based on the null hypothesis of a unimodal normal distribution (i.e., the data is clearly not normal). That is, only 0.047% of a normal distribution is beyond 3.5σ , which, for a sample of 100, is no scores. For 500 data points, the range should be expanded to at least $+4$ bins (0.0063% of a normal distribution) or $+5$ (0.00057% of a normal distribution). Of course, wider limits are needed because nothing is truly normal, and one must have sufficient range to encompass the original histogram which is not likely normal.

Creating the Sampling Distribution

For the Monte-Carlo simulation, the only significant addition to the previously cited routines is the automated data generation. In the following, it is acknowledged that many of the routines can be simplified or streamlined. This presentation was chosen to maintain the clarity of the logic.

To find the distribution of the LLR , the Monte-Carlo simulation (MCS) uses the full range of bins defined by the data and simulation. These bins are notated by z-scores since this is convenient. The MCS then takes K samples, each of size n , from this population. This results in one large data file. That file contains individual data points. That large data file is split and each sample is analyzed separately (and automatically). Each sample is converted to a histogram using the aforementioned bin sizes. Again, the bin sizes for the theoretical distribution are perfectly matched to those used with the real data, and the real data must have defined and used a sufficient range of bins for the entire simulation. Then, for each sample, the LLR is computed. Finally, all samples are reduced to a single data file containing the distribution of LLR_k . This distribution can be plotted, or more simply the necessary critical values can be obtained.

The first part of the process is shown in Listing 3. This generates K samples of size n . There are a couple of tricks to be discussed momentarily. Note the random seed.

Listing 3: Generating K samples of Size n

```

set seed      random.
input program.
compute      #mean = 0.
compute      #std = 1.
loop         #K = 1 to 1000.
+ loop       #N = 1 to 601.
+ compute    K = #K.
+ compute    meanbin = 51.
+ compute    binpersd = 10.
+ compute    total = 500.
+ do if ( #N le 101 ).
+ compute    N = -1.
+ compute    zscore = 0.
+ compute    binnum = #N.
+ compute    xl = (binnum-meanbin - .5)
              / binpersd.
+ compute    xc = (binnum-meanbin)
              / binpersd.
+ compute    xu = (binnum-meanbin + .5)
              / binpersd.
+ end if.
+ do if ( #N gt 101 ).
+ compute    N = #N - 101.
+ compute    zscore = normal(#std) + #mean.
+ compute    binnum = rnd(zscore * binpersd)
              + meanbin.
+ end if.
+ end case.
+ end loop.
end loop.
end file.
end input program.
execute.

```

```

frequencies  binnum.
if ( binnum le 1 )    binnum = 1.
if ( binnum ge 101 ) binnum = 101.
if ( binnum eq 1 )    xl = -6.95.
if ( binnum eq 101 ) xu = 6.95.
if ( binnum eq 1 )    xc = (xu - xl) / 2 + xl.
if ( binnum eq 101 ) xc = (xu - xl) / 2 + xl.
execute.

```

First, to generate n data points, $n+101$ data points are generated. The extra 101 data points are a trick. They are place holders to ensure that every data set has the same range of bins. They define the bin sizes (in z-scores). The 101 comes from the desire to have a range of $-5.0 < z < 5.0$, with 10 bins per standard

deviation. Note the variable meanbin and binpersd. There is one odd bin at the center. This can be altered to suit the circumstances (i.e., a different number of bins per standard deviation; a range of z-score range of bins).

Second, bins are actually numbered from 1 to 101, rather than from -50 to 50. The variable meanbin defines the center bin. The values of xl, xc, and xu define the limits (lower, center, upper) of the bin in terms of z-scores. These are most useful for verifying the execution of the program. The frequencies command simply serves to check if any data exceeded the expected range of z-scores. Note that the tails are artificially widened after the data is created.

This routine creates a data file that contains the following variables per case:

K	sample number
meanbin	the center bin
binpersd	the number of bins per standard deviation
xl	the lower (left) limit of the bin, in z-scores
xc	the center of the bin, in z-scores
xu	the upper (right) limit of the bin, in z-scores
total	the total number of data points per sample
N	datum number (not actually used, but useful for humans) N = -1 indicates a bin place holder
zscore	the z-score of the created datum
binnum	the conversion of the zscore to a bin number

Note that some of the defined values are constants for all cases (for each data point). This is essentially the same as in original data.

The processing continues in Listing 4. This large data file is split into K smaller files for individual analyses. The SPSS SPLIT FILE function accomplishes this. The data is then sorted (within each sample is faster) by bin number, and collapsed by bin number using the AGGREGATE function. This creates a histogram, for each sample, by counting the number of times each binnum was presented in the data (the line observed = $n(\text{binnum})$). Other variables are collapsed as well. Note that meanbin, binpersd, total, xl, xc, and xu are all constants. Hence, taking the first occurrence

TESTING BIMODALITY

(within each sample) is the most efficient manner to get these values: It does not require any computations by SPSS. Also note that the breaking variables (K and binnum) are automatically included in each sample, while the variable zscore is dropped (one could take the mean of zscore to obtain the true bin center).

Listing 4: Converting Data to Histograms, then Cleaning

```
split file by K.
sort cases by k binnum.
aggregate  outfile = *
           /break = K binnum
           /meanbin binpersd total =
             first(meanbin, binpersd, total)
           /xl xc xu = first(xl, xc, xu)
           /observed = n(binnum).

execute.
compute   observed = observed - 1.
frequencies observed.
if (observed lt 0) observed = 0.
```

In addition, recall that the first 101 values of data only served to ensure that every bin existed (i.e., they were place holders used to define bins). This case would have been included in the count of values per bin number (binnum). Hence, every bin has one count (i.e., observed) too many, so one must subtract one from every value of observed . Note that if the range of bins was not defined sufficiently (the initial 101 bins), there will be a negative count in some bins. This would create havoc with the routines, so a check is used to force the count per bin (observed) to be greater than or equal to zero. The `frequencies` command is a better check. In fact, if there are negative bin counts (after subtracting one), the analysis should be re-run, or widen the tails still further. Technically, this would also require recomputing LLR_{data} because the bins used for the simulation must match the bins used for the data.

Listing 5 provides the fitting of the two functions and the computation of LLR_k . It is essentially a repeat of previous discussions (particularly Listing 1). Note the `set results none` command. This turns off the outputting of results which is very useful in a simulation. In addition, `split file processing` is still engaged.

This is the slowest part of the routine (get a large coffee).

Listing 5: Fitting Each Sample with the Bimodal and Unimodal Functions to Obtain LLR

`set results none.`

```
compute prop = observed / total.

model program mean=0.0 sd = 1.0.
compute xa = abs(xl - xc).
compute xb = abs(xu - xc).
compute h1 = (.39894228/ sd)
             *exp(-(((xl-mean)**2)/(2*sd**2))).
compute h2 = (.39894228/ sd)
             *exp(-(((xc-mean)**2)/(2*sd**2))).
compute h3 = (.39894228/ sd)
             *exp(-(((xu-mean)**2)/(2*sd**2))).
compute predun = .5*(h1+h2)*xa
                + .5*(h2+h3)*xb.

cnlr   prop
       /pred = predun
       /bounds sd gt 0.0001
       /save = predun residun
       /criteria iter 100.

model program mean1=-1.0 mean2=1.0
             sd1=1.0 sd2=1.0 ratio=0.5.
compute xa = abs(xl - xc).
compute xb = abs(xu - xc).
compute h1 = (.39894228/ sd1)
             *exp(-(((xl-mean1)**2)/(2*sd1**2))).
compute h2 = (.39894228/ sd1)
             *exp(-(((xc-mean1)**2)/(2*sd1**2))).
compute h3 = (.39894228/ sd1)
             *exp(-(((xu-mean1)**2)/(2*sd1**2))).
compute h4 = (.39894228/ sd2)
             *exp(-(((xl-mean2)**2)/(2*sd2**2))).
compute h5 = (.39894228/ sd2)
             *exp(-(((xc-mean2)**2)/(2*sd2**2))).
compute h6 = (.39894228/ sd2)
             *exp(-(((xu-mean2)**2)/(2*sd2**2))).
compute predbi = ratio *(.5*(h1+h2)*xa
                       + .5*(h2+h3)*xb)
                + (1-ratio)*(.5*(h4+h5)*xa
                              + .5*(h5+h6)*xb) .

cnlr   prop
       /pred = predbi
       /bounds sd1 gt 0.00001;
       sd2 gt 0.00001;
```

```

1.0 ge ratio ge 0.0
/save = predbi residbi
/criteria iter 100.

compute expectun = predun * total.
compute expectbi = predbi * total.
compute lnpoisun = observed * ln(expectun)
- expectun.
compute lnpoisbi = observed * ln(expectbi)
- expectbi.
compute llrdata = 2*(lnpoisbi - lnpoisun).
execute.

```

Finally, as shown in Listing 6, the data are collapsed once again (using the AGGREGATE function) to create one case (i.e., one line in the data file) per sample. This one case contains all the essential information for the entire sample. The most important is the LLR_k from which the sampling distribution of LLR can be created. The use of percentiles in the FREQUENCIES command provides the standard critical points directly, but the distribution can also be created.

Listing 6: The Sampling Distribution of LLR

```

aggregate outfile = *
/break k
/nbins = n(total)
/count = sum(observed)
/sumy predun residun predbi residbi=
sum(prop, predun, residun, predbi,
residbi)

/sdy sdresun sdresbi =
sd(prop, residun, residbi)
/llr = sum(llrdata).

compute R2bi = 1 - (sdresbi**2 / sdy**2).
compute R2un = 1 - (sdresun**2 / sdy**2).
compute chgR2 = (sdresun**2 - sdresbi**2)
/ sdy**2.

frequencies variables = llr chgr2
/percentiles = 90 95 99
/statistics = mean stddev variance
minimum maximum median
skewness seskew kurtosis sekurt
/order= analysis.

```

For the current data, using the bin sizes of the original data, with $n = 500$ and $K = 1,000$, one obtains the following distribution of LLR (see Figure 3).

From this, it can be determined that 5% of the distribution for LLR exceeded the critical value of .186, so the observed value of $LLR_{data} = 28.645$ is significant. The hypothesis that the data came from a unimodal distribution is rejected, using a type 1 error rate of $\alpha = .05$. This is not surprising given the population definition $B(-1.0, 0.7, 1.0, 1.0, .6)$, the large sample size ($N=500$), range of bins $-5.0 < z < 5.0$ and the 10 bins per standard deviation. Note that it is the sample size that allows for a large range of z , with a small z per bin. The 10% point was .153, and the 1% point was .226. The mean for the distribution was .002 and the standard deviation was .122 (skew: $-.573 + .077$; kurtosis: $.413 + .155$). Note that the mean is quite close to the expected value of zero.

The CNLR function does not allow the correlation (R^2) to be saved per sample. However, R^2 can be computed per sample from $R^2 \approx 1 - s^2_{Y.X}/s^2_Y$. This can also be converted to a distribution. Given the unimodal and bimodal R^2 per sample, one can create ΔR^2 , and create the distribution of ΔR^2 . The sample ΔR^2 can also be compared to this distribution, or this empirically determined sampling distribution of ΔR^2 can be compared to the theoretical distribution of ΔR^2 with $df_1=3$ and $df_2=n-5$.

For the current data, the change in fit was $\Delta R^2 = .053$. For the distribution of ΔR^2 , the critical points were .0000916 at 10%, .000115 at 5% and .000152 at 1%. The mean was .00000153 and the standard deviation .00000064. Given that the observed ΔR^2 was .053, the hypothesis that the data came from a unimodal distribution is rejected.

The standard deviation function in the AGGREGATE command (e.g., `sdresun = sd(residun)`) returns the inferential form of the standard deviation which in these simulations is $\Sigma e^2_i / (I-1)$ (where I = number of bins). However, the CNLR algorithm provides the standard error of regression ($s^2_{Y.X}$), and this is used to compute R^2 for each sample. Thus, the $s^2_{Y.X}$ cited in the output of the unimodal modal is $\Sigma e^2_i / (I-2)$, and the $s^2_{Y.X}$ cited for the bimodal modal is $\Sigma e^2_i / (I-5)$. Therefore, technically, the R^2 cited in the

TESTING BIMODALITY

Figure 3: The Likelihood Ratio Distribution ($K = 1,000, N = 500$)

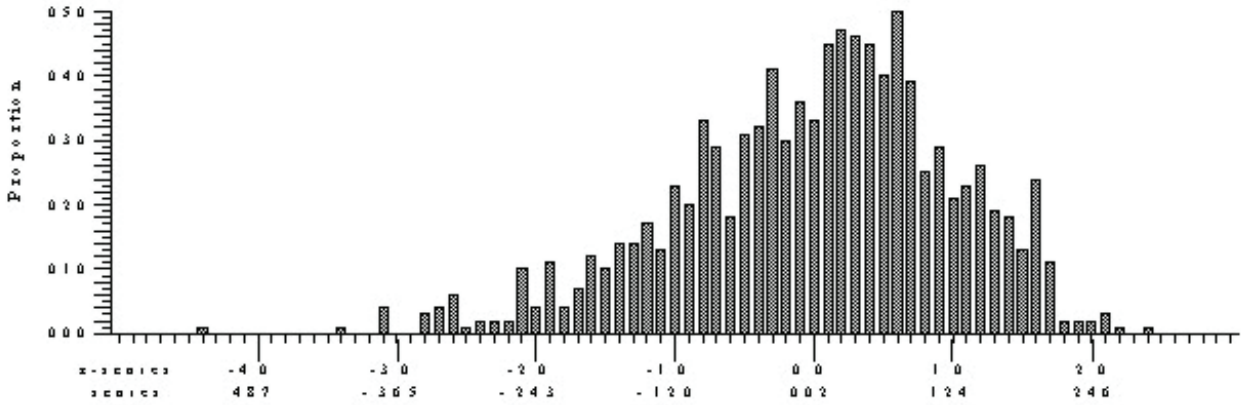


Figure 3: The Likelihood Ratio Distribution ($K = 1,000, N = 500$).

output of the CNLR for the single sample cannot be directly compared to the distribution of R^2 determined from $R^2 \approx 1 - s^2_{Y.X}/s^2_Y$. However, the difference is slight, and this entire process only estimates the distributions (i.e, it is not an analytic solution). A correction could be applied if desired ($I-1 / I-df_{\text{model}}$), which would be useful if generating very large simulations.

Extensions

The CNLR (or NLR) command, with the use of the subcommand /outfile= aaaa.bbb, allows various parameters from each sample to be saved for future analyses. For example, the fitted parameters can be obtained per sample (μ , σ) and (μ_1 , σ_1 , μ_2 , σ_2 , λ) so to map the parameter space. When examining these values, it should be kept in mind that the algorithm will occasionally flip the order of μ_1 and μ_2 , so, before computing any interesting statistics, one should insure that μ_1 is less than μ_2 (flipping μ_1 and μ_2 also requires flipping σ_1 and σ_2 , as well as inverting λ). The output file also contains the $SS_{Y.X}$, the number of cases and the split file number.

A χ^2 test of the fit can also be obtained, by computing $(\hat{Y}_i - Y_i)^2/\hat{Y}_i$, per bin before collapsing the data. This is not advocated because the sum can create overflow errors. The reduced χ^2 can also be used. It is interesting to

note that in the special case when the null is a normal distribution, and the alternative is a mixture distribution of two normal distributions with equal variances ($\sigma_1 = \sigma_2$), the sampling distribution of LLR is known to be related to the χ^2 distribution:

$$\begin{aligned} \chi^2_v &= -2 \ln(\Lambda) \\ &= -2 \ln\left(\frac{L_0}{L_A}\right) \\ &= 2 \ln(L_A) - 2 \ln(L_0) \end{aligned} \quad (8)$$

The df (v) for the χ^2 distribution is equal to the difference in the number of parameters fitted. In this special case, $v = 2$: There are two parameters for the unimodal normal distribution, $N(\mu, \sigma)$, and four for the bimodal mixture of two normal distributions $B(\mu_1, \sigma, \mu_2, \sigma, \lambda)$. Hence, in that special case, $LLR_{\text{data}} = -2 \ln(\Lambda_{\text{data}})$ can be compared to the χ^2_2 distribution (see McLachlan, 1987, for a more extensive discussion), though this equivalency assumes that the computation of the expected frequency per bin follows a Normal, rather than Poisson distribution. If the LLR_{data} exceeds the critical value for $\chi^2(2)$, then the null can be rejected. The χ^2_2 distribution can also be compared to the LLR distribution obtained herein. However, these constraints are

not acceptable in the general case. The current procedure allows the use of unequal variances, and the current procedure can be extended to any non-normal distributions.

Because the scaling of the bin centers (or limits) is irrelevant (i.e., X_i), z-scores can be used for any data set. When the data set is converted to a histogram, the important issues are the number of bins per standard deviation and the full range of bins. The actual scaling of the bins is irrelevant (to the computation of LLR_{data}). As such, any particular data set can be converted to a standard histogram, with a set number of bins and range of bins. The LLR_{data} can be determined for that standard histogram. This LLR_{data} could then be compared to tabulated values of critical $LLRs$ for particular values type 1 error rate (α). That is, using SPSS, tables of critical values can be created for various combinations of total sample size, bin size (bins per sd), and range of bins. This would avoid all the tedium of running this simulation for every data set. The simulation could be saved for non-tabulated situations. Arguably, it is still better to complete the entire simulation so that the bins can be carefully tailored to data. Note that the number of samples in the simulation (i.e., K) should not be an issue. That is, K reflects desired precision and reliability. Every simulation should produce approximately the same critical values (always remember that the fitting process is iterative, not algorithmic).

Power

The previous simulations can be used to compute power for any given exact alternative to the null. The exact alternative would specify the parameters in $B(\mu_1, \sigma_1, \mu_2, \sigma_2, \lambda)$, and then use this as the population for the simulation, in place of $N(\mu, \sigma)$. The proportion of sample LLR 's that exceeded the previously defined critical values for the null would then be determined. For this to work, the bimodal population must use the same range of bins and the same bin size as the corresponding null hypothesis.

Often, when computing power, a wider range of bins is needed because data in the tails are more common from a bimodal population. This implies that the LLR test for the unimodal population would need to be computed with a larger range of bins.

In addition, when the bimodal population is not symmetric (i.e., $\sigma_1 = \sigma_2$, and $\lambda = .5$), it is more difficult to get the population centered in the histogram. The middle of the histogram should correspond to the mean of the bimodal distribution, so that there is sufficient range in the tails. This is a pragmatic issue since the bimodal population, and hence every random sample from it, is fit with the unimodal and bimodal function on the basis of the same histogram.

Conclusion

This work has been a demonstration of the application of commonly available statistical software, in this case SPSS, to solving the problem of assessing putative mixture distributions, particularly decisions comparing a unimodal normal distribution to a bimodal mixture of two normal distributions. Routines were developed to enable anyone to determine the best-fit statistics for fitting data to a unimodal normal distribution or a bimodal mixture of two normal distribution, to then use those parameters to generate the LLR , and finally, to generate the sampling distribution of the LLR .

These routines have been developed and refined over a number of versions of SPSS from 6.0. to 11.5. In fact, the routines were initially developed within SPSS 4.0, running under VMS 8.0, on a VAX 4500. Different version might require minor modifications. In addition, routines have be developed and run on a variety of hardware. On a 1,000 MHZ Duron with 1.256 Gigs of memory, a simulation with $K = 1,000$, and $N = 500$ required about 15 minutes. A 600 MHZ, Pentium 4 with 256 Megs of memory increased this to about 15 minutes. By contrast, similar simulations on a 40 MHZ AMD 386 had to be run overnight. Interestingly, the VAX also required an overnight batch job.

When setting up, the process is simple and relatively efficient: simply convert the empirical data to z-scores and then create a histogram with an appropriate number of bins per sd and an appropriate range of z-scores. This depends primarily on the sample size. The simulation to create the LLR distribution uses the same bin size and range. The variable bin

TESTING BIMODALITY

widths could be used, with narrower bins near the center of the distribution and wider bins in the tails. As long as the bins form a mutually exclusive and exhaustive set for the range of interest, this is not a problem. In fact, it might be more optimal in the long run to develop algorithms that use bins that represent constant probabilities under the normal distribution.

The method can be adapted to non-normal distributions or to mixtures of non-normal distributions. There is unlimited flexibility in the choice of fitting functions. The process creates an empirical sampling distribution for whatever hypotheses are being tested.

As noted predicted bin counts could be generated using other methods, in particular the normal distribution. That route was not presented here because the use of normal distribution to predict bin counts resulted in a test with lower power. However, that method is more closely tied to the χ^2 test of fit, and the LLRT approximation to the χ^2 .

The second advantage is that the algorithm can be modified to obtain greater accuracy. Non-linear regression using a least-squares error term assumes that the theoretical error is a constant for all values of the independent variable. That is, every bin, regardless of its count, has the same contribution to the final solution. However, the error of a count (if Poisson statistics are valid) is the square root of the count. Hence, relative errors per bin increase as the count decreases. This can be used as a control in the CNLR routine. SPSS non-linear regression allows one to specify the error term. Hence, a weighted least-squares (non-linear) regression approach could be used.

In summary, the routine works; however, it must be cautioned that this algorithm is only considered an interim solution to the problem – one of many (cf., Eriksen & Eriksen, 1972; Eriksen & Yeh, 1985; Hartigan, 1974; Jones & McLachlan, 1990; Müller & Sawitzki, 1991; Roeder, 1990, 1994; Yantis, Meyer and Smith, 1991; Yellott, 1971).

Hopefully, a proper fully parametric method for assessing bimodality will be developed, one that extracts all the information contained within each individual data point rather than working through the intermediary of

a histogram. However, even if a proper parametric method is developed, it will necessarily be tied to particular parent distributions. As such, the algorithms developed herein will continue to serve some purpose with other non-normal parent distributions.

References

- Bevington, P. R., & Robinson, D. K. (1992). *Data Reduction and Error Analysis for the Physical Sciences*. Toronto: McGraw Hill, Inc.
- Do, K., & McLachlan, G. L. (1984). Estimation of mixing proportions: A case study. *Applied Statistics*, 33(2), 134-140.
- Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11(5), 583-597.
- Eriksen, C. W. & Eriksen B. A. (1972). Visual backward masking as measured by voice reaction time. *Perception and Psychophysics*, 12(1), 5-8.
- Frankland, B. W., & Zumbo, B. D. (2002). Quantifying bimodality Part I: An easily implemented method using SPSS. *Journal of Modern Applied Statistical Methods*, 1, 157-166.
- Hartigan, J. A. (1974). Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1), 117-131.
- Hayes, (1994). *Statistics*. NY: Holt, Rinehart and Winston.
- Hoffman, J. P., & Miller, A. S. (1998). Denominational influences on socially divisive issues: Polarization or continuity? *Journal for the Scientific Study of Religion*, 37(3), 528-546.
- Howell, D. (1995). *Statistical Methods for Psychology*. NY: Duxbury Press.
- Johnson, D. N., & Yantis, S. (1995). Allocating visual attention: Tests of a two-process model. *Journal of Experimental Psychology: Human Perception & Performance*, 21(6), 376-1390.
- Jones, P. N., & McLachlan, G. J. (1990). Algorithm AS 254: Maximum likelihood estimation from grouped and truncated data with finite mixture models. *Applied Statistics*, 39(2), 273-312.

Knoll, J. L. IV, et al. (1998). Heterogeneity of the psychoses: Is there a neurodegenerative psychosis? *Schizophrenia Bulletin*, 24(3), 365-379.

Müller, D. W., & Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86, 738-746.

Ottong, S. E., & Garver, D. L. (1997). A bimodal distribution of plasma HVA/MHPG in the psychoses. *Psychiatry Research*, 69(2-3), 97-103.

Reischies, F. M., Schaub, R. T., & Schlattmann, P. (1996). Normal ageing impaired cognitive functioning and senile dementia: A mixture distribution analysis. *Psychological Medicine*, 26(4), 785-790.

Roeder, K. (1990). Density estimation with confidence set exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85, 617-624.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89, 487-495.

Sussman, H. M. (1999). A neural mapping hypothesis to explain why velar stops have an allophonic split. *Brain and Language*, 70(2), 294-304.

Volbrecht, V. J., Nerger, J. L., & Harlow, C. E. (1997). The bimodality of unique green revisited. *Vision Research*, 37(4), 407-416.

Yantis, S., Meyer, D. E. & Smith, J. E. K. (1991). Analysis of multimodal mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, 110, 350-374.

Yellott, J. I. (1971) Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology*, 8(2), 159-199.

A Comparative Study of Bayesian Model Selection Criteria for Capture-Recapture Models for Closed Populations

Ross M. Gosky Sujit K. Ghosh
Appalachian State University North Carolina State University

Capture-Recapture models estimate unknown population sizes. Eight standard closed population models exist, allowing for time, behavioral, and heterogeneity effects. Bayesian versions of these models are presented and use of Akaike's Information Criterion (AIC) and the Deviance Information Criterion (DIC) are explored as model selection tools, through simulation and real dataset analysis.

Key words: AIC, Bayesian inference, capture-recapture models, closed population, DIC, Gibbs sampling, heterogeneity, MCMC, model selection, WinBUGS.

Introduction

For capture-recapture experiments involving closed populations, likelihood-based models based upon the multinomial distribution are commonly used, and a thorough treatment of these models is given by Otis, Burnham, White, and Anderson (1978). These models allow animal capture probabilities to vary based on three types of effects: time effects, heterogeneity effects, and behavioral effects. Time effects occur when capture probabilities vary by capture period. Heterogeneity effects occur when capture probabilities vary by animal. Behavioral effects occur when an animal's capture probability changes after they are captured for the first time.

This effect is called a trap-happy effect when the capture probability increases after initial capture, and is called a trap-shy effect

when the capture probability decreases after initial capture. Denoting subscripts t , h , and b to refer to time, heterogeneity, and behavioral effects, respectively, eight models have been developed, with the model subscripts indicating which effects are present in the modeling of capture probabilities. The goal of each model is to estimate the unknown population size N . The model M_0 denotes a model which has none of the three effects. Model M_t contains time effects, model M_h contains heterogeneity effects, and model M_b contains behavioral effects. Models M_{tb} , M_{bh} , M_{th} , and M_{tbh} are complex models accounting for variation in capture probabilities from each listed effect. Chao (2001) provides an overview of closed population models as well.

Pledger (2000) discussed using mixture models to fit heterogeneity effects in capture-recapture data, and discussed use of Akaike's Information Criterion (AIC) as a model selection tool. Caution in using heterogeneity models is necessary, though, as Link (2003) showed that estimates of N under M_h models are highly dependent upon the assumed distribution of capture probabilities in the population. He refers to the parameter N as non-identifiable in heterogeneity models because different, reasonable, models may fit the data equally well but give very different inferences about N . Link's results imply that distinguishing between different heterogeneity models may never be possible. However, it remains plausible that

Ross Gosky is an Assistant Professor in the Mathematical Sciences Department. His research interests are in Bayesian Statistics and Mark-Recapture models. Email: goskyrm@appstate.edu. Sujit Ghosh is a Professor in the Department of Statistics. His research interests include Bayesian Statistics, Spatial Statistics, and Survival Analysis. Email: sujit_ghosh@ncsu.edu.

estimates of N from M_h models are more accurate than those from Model M_0 , for example, in populations with heterogeneity.

Program MARK (see <http://welcome.warnercnr.colostate.edu/~gwhite/mark/mark.htm>), provides estimates of N for these closed population models as well as end-user flexibility in the specific parameterization of the models. For example, the mixture models of Pledger (2000) can be fit in Program MARK with different numbers of mixture groups specified by the user. Program MARK also provides model selection functionality based on Akaike's Information Criterion (Akaike, 1973).

Bayesian versions of closed population models have also been presented. Early approaches focused on Model M_t , such as Castledine (1981), and George and Robert (1992). Ghosh and Norris (2005) presented a Bayesian version of M_{bh} , and M_h , M_b , and M_0 as special cases of this model. Furthermore, they presented a model selection approach based upon a criterion proposed by Gelfand and Ghosh (1998). Other recent work on Bayesian models have been presented by Durban and Elston (2005) and by King and Brooks (2008). King and Brooks recommended Bayesian Model Averaging and Reversible Jump Markov Chain Monte Carlo (RJMCMC) methods for recapture/recovery data analyses, while Durban and Elston focused on Model M_{th} by adapting a log-linear modeling approach to the models of Agresti and Coull (1999). More recently, Gosky and Ghosh (2011) provide Bayesian estimation methodologies for all eight models.

Methodology

Bayesian Closed Population Capture-Recapture Models

Bayesian statistical modeling requires the development of the likelihood function of the observed data, given a set of parameters, as well as the joint prior distribution of all model parameters. A major benefit of Bayesian models for capture-recapture data is that Bayesian estimates of N , from its posterior distribution, are easily obtainable and this posterior distribution gives appropriate measures of variability for estimating N . Even when non-informative prior distributions are used for

model parameters, these estimates of variability are not based on asymptotic criteria and hold when N and the number of capture periods are relatively small (e.g., see Gosky and Ghosh, 2011). Bayesian modeling also allows for the possibility of using informative prior information about model parameters, if available.

The approach to modeling heterogeneity used is identical to that presented in Ghosh and Norris (2005), using a finite-mixture approach to heterogeneity rather than utilizing a continuous distribution to model individual capture probabilities. This basic idea was introduced by Norris and Pollock (1996) and discussed further in Pledger (2000), and has been shown to be effective in modeling heterogeneity.

Let k represent the number of capture periods in the study. Define indicator variables $X_{ij} = 1$ if animal i is captured during capture period j , for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, k$. Also denote $p_{ij} = \Pr(X_{ij} = 1)$ as the probability that animal i is captured during capture period j .

Denote the capture matrix \mathbf{X} with dimensions $N \times k$ with entry X_{ij} . Denote $X[i, \cdot]$ as the i^{th} row of \mathbf{X} , a vector with 2^k possible values because each entry in the vector is zero or one. For simplicity, these outcomes can be ordered as

Outcome 0: capture history (0, 0, 0, ..., 0, 0, 0);

Outcome 1: capture history (0, 0, 0, ..., 0, 0, 1);

Outcome 2: capture history (0, 0, 0, ..., 0, 1, 0);

Outcome 3: capture history (0, 0, 0, ..., 0, 1, 1);

through Outcome $2^k - 1$: capture history (1, 1, 1, ..., 1, 1, 1).

Each animal in the population has exactly one of the 2^k capture histories. Noting that $(X_{i1}, X_{i2}, \dots, X_{ik})$ represents the observed capture history of any animal in the population, Capture History h_i is defined corresponding to the previous ordering of outcomes, as $h_i = \sum_{j=1}^k X_{ij} 2^{k-j}$. Notice that each h_i takes values in the set $0, 1, \dots, 2^k - 1$. Denote Z_l as the number of animals with capture history l , for $l = 0, 1, \dots, 2^k$

BAYESIAN MODEL SELECTION FOR CAPTURE-RECAPTURE MODELS

- 1, i.e., $Z_l = \sum_i I(h_i = l)$ where $I(\cdot)$ denotes the indicator function which takes the value 1 if $h_i = l$ and takes the value 0 otherwise. Note that Z_0 , the number of animals with capture history $(0, 0, \dots, 0)$, cannot be observed. Also, note that $\sum_{l=0}^{2^k-1} Z_l = N$. Denote $S = N - Z_0$ as the number of animals observed during at least one capture period. Denote P_l as the probability of animal i having capture history l . Then

$$P_l = \prod_{i:h_i=l} \prod_{j=1}^k p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}} \quad (2.1)$$

Defining $L = 2^k - 1$, the joint distribution of (Z_1, Z_2, \dots, Z_L) is

$$\begin{aligned} & \Pr[Z_1 = z_1, \dots, Z_L = z_L \mid N, \mathbf{P}] \\ &= \frac{N!}{(N-S)! \prod_{l=1}^L z_l!} \prod_{l=1}^L P_l^{z_l} \left(1 - \sum_{l=1}^L P_l\right)^{N-S} \end{aligned} \quad (2.2)$$

where $\mathbf{P} = (P_1, P_2, \dots, P_L)$. Note that if N were known, this model would represent a multinomial likelihood function with counts Z_0, \dots, Z_L and probabilities P_0, \dots, P_L . However N is unknown and it is the main parameter of interest. From equation (2.2) it follows that the likelihood function of (N, \mathbf{P}) is given by

$$L(N, \mathbf{P} \mid Z) \propto \binom{N}{S} \prod_{l=1}^L P_l^{z_l} \left(1 - \sum_{l=1}^L P_l\right)^{N-S} \quad (2.3)$$

where $Z = (Z_1, \dots, Z_L)$ denotes the set of observed counts, which turns out to be the minimal sufficient statistic for this model.

It is of interest to estimate N , treating \mathbf{P} as a nuisance parameter. The capture probability vector \mathbf{P} varies depending on the specific model. A Bayesian modeling framework was adopted for each of the eight models, where

$$\Pr(N = n) \propto \frac{1}{n^\delta}, \quad n = 1, 2, \dots, N_{\max}$$

was used as the prior distribution for N , with $\delta > 0$ fixed at a specific value and N_{\max} fixed at a realistic upper bound for N . A non-informative prior distribution can be obtained with $\delta = 0.5$ (or alternatively $\delta = 1$) and a uniform prior is obtained with $\delta = 0$. The final estimate of N is obtained from the marginal posterior distribution of N by integrating out the parameters corresponding to \mathbf{P} . The most complex model, M_{tbh} , is introduced first followed by descriptions of each of the other seven models as special cases of M_{tbh} .

Model M_{tbh}

This model allows for individual heterogeneity, time, and behavior effects. For heterogeneity, a finite mixture distribution is used, representing m possibly distinct groups within the population. Behavioral effects are modeled as constant across each of the m groups and across capture periods 2 through k to minimize the number of model parameters and to allow the model to be fit to studies with a minimal number of capture periods.

Denoting $\tau_{ij} = 1$ if animal i has been captured before capture period j , then for each group, the capture probability vector is $\mathbf{p}_i = p_{i1} I(\tau_{ij} = 0) + p_{i2} I(\tau_{ij} = 1)$ and $(p_{i1}, p_{i2}) \sim F(\cdot)$, described next. A finite mixture distribution is assumed for the $2k$ -dimensional distribution function, F , specifically $dF(\mathbf{p}) = \sum_{m=1}^r \pi_m I(\mathbf{p} = \boldsymbol{\theta}_m)$, where π_m denotes the probability at support point $\boldsymbol{\theta}_m = (\theta_{11m}, \dots, \theta_{1km}, \theta_{21m}, \dots, \theta_{2km})^T$, and $\sum_{m=1}^r \pi_m = 1$. The probability

of initial capture in capture period j is represented as θ_{1jm} , where $j = 1, 2, \dots, k$ within population group m , where $m = 1, 2, \dots, r$. Similarly, θ_{2jm} is the probability of subsequent capture in capture period j within population group m . As previously stated, the behavior effect is constant across the capture periods and the m population groups. Thus $\theta_{2jm} = \theta_{1jm} + c$ for $j = 2, \dots, k$ and $m = 1, 2, \dots, r$. Furthermore, $\theta_{2jm} = 0$ for $m = 1, 2, \dots, r$ because subsequent capture is impossible in capture period one. Fixing $r = 2$ mass points representing possibly two distinct population groups implies that $1 - \pi_2$

$= \pi_1 = \pi$. Prior distributions for π and θ_{1jm} are π , $\theta_{1jm} \sim \text{Beta}(a, b)$ for $j = 1, \dots, k$ and $m = 1, 2$.

A conditional prior distribution of c given θ_{1jm} for $j = 1, \dots, k$ and $m = 1, 2$ is $\text{Uniform}(-\min_{2 \leq j \leq k; 1 \leq m \leq r} \theta_{1jm}, 1 - \max_{2 \leq j \leq k; 1 \leq m \leq r} \theta_{1jm})$. This mixture model requires restrictions for identifiability of all model parameters, so $\theta_{1j1} \leq \theta_{1j2}$ for $j = 1, 2, \dots, k$ and $\theta_{2j1} \leq \theta_{2j2}$ for $j = 2, 3, \dots, k$ is set.

Model M_{tb}

Restrict $\pi_1 = 1$ from Model M_{tbb} .

Model M_{th}

Restrict $\theta_{2jm} = \theta_{1jm}$ for $j = 2, \dots, k$ and $m = 1, \dots, r$, from Model M_{tbb} .

Model M_{bh}

Restrict $\theta_{11m} = \theta_{12m} = \dots = \theta_{1km}$ and $\theta_{22m} = \theta_{23m} = \dots = \theta_{2km}$ for $m = 1, 2, \dots, r$ from M_{tbb} . Rather than modeling $\theta_{2jm} = \theta_{1jm} + c$, choose

i.i.d. prior distributions $\theta_{11m}, \theta_{22m} \sim \text{Beta}(a, b)$ for $m = 1, \dots, r$. Fixing $r = 2$ mass points as described in M_{tbb} , restrict $\theta_{111} \leq \theta_{112}$ and $\theta_{221} \leq \theta_{222}$ for identifiability of all model parameters.

Model M_t

From Model M_{th} , restrict $\pi_1 = 1$.

Model M_h

From Model M_{bh} , restrict $\theta_{22m} = \theta_{11m}$ for $m = 1, 2, \dots, r$.

Model M_b

From Model M_{bh} , restrict $\pi_1 = 1$.

Model M_0

Restrict $\pi_1 = 1$ from Model M_h .

The number of parameters in each model as a function of r , the number of support points of the finite mixture distribution F and k , the number of capture periods, is determined from the preceding model descriptions. For example, Model M_{tbb} has parameters $N, \pi_1, \dots, \pi_{r-1}, \theta_{111}, \dots, \theta_{1k1}, \theta_{112}, \dots, \theta_{1k2}, \theta_{11r}, \dots, \theta_{1kr}$, and c . The number of parameters is thus $1 + (r-1) + kr + 1 = r(k+1) + 1$. Similarly it is established that M_{th} has $r(k+1)$ parameters, M_{bh} has $3r$ parameters, M_{tb} has $k+2$ parameters, M_h

has $2r$ parameters, M_b has 3 parameters, M_t has $k+1$ parameters, and M_0 has 2 parameters.

Posterior distributions of the model parameters for all eight models can be closely approximated using Markov Chain Monte Carlo (MCMC) methods available in the WinBUGS V1.4 software package (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>).

Model Selection Methods

Because eight possible models exist for a given closed population data set, definitive methods for model selection are necessary in such analyses. The eight models are generally (though not exclusively) nested, ranging from very simple models (M_0) to complex models (M_{tbb}). Model selection criteria allow the best model of the eight to be fit to the data. Striking a balance and finding a model that neither underfits nor over-fits the data is the motivation for model selection criteria (Burnham & Anderson, 2002).

Akaike's Information Criterion (AIC) is one such method of model selection, and seems to be the most commonly used criterion for model selection. The intent of AIC is to measure the mathematical distance between the true population and the fitted model, by using the so-called Kullback-Leibler discrepancy. To differentiate between models with different numbers of parameters, AIC adds two times the number of model parameters to the estimated Kullback-Leibler discrepancy. Thus, when two models of differing complexity fit a data set equally well AIC chooses the simpler model by penalizing the complex model for having more model parameters. The rule of parsimony says that a researcher should choose the simplest model that adequately describes the behavior of the population. Use of AIC generally supports this rule.

As the models are nested, AIC is examined as model selection tool. However, AIC is not asymptotically consistent in the sense that the probability that it chooses the correct model (given that the data has been generated from the correct model) does not converge to one as the sample size tends to infinity (Schwartz, 1978). A modified version of AIC within the Bayesian framework is used for model selection.

The Deviance Information Criterion (DIC) is strictly a Bayesian model selection criterion (Spiegelhalter, et al., 2002), which is structured similarly to AIC. The main difference between AIC and DIC is in the penalty term added to the estimated Kullback-Leibler discrepancy. DIC adds two times the effective number of parameters to the estimated Kullback-Leibler discrepancy. The effective number of parameters is a Bayesian concept. It recognizes that the number of parameters in a Bayesian model is influenced by the prior distributions of these parameters. The DIC criterion presents a methodology to measure this number of parameters. The DIC is, then, the difference between the estimated mean KL distance, and the KL distance estimated at the posterior mean of each of the model parameters.

DIC is also examined as a model selection criterion for these models. Use of DIC does not require the models to be nested. However, the modeling herein uses a mixture approach for heterogeneity models, and there are some questions about use of DIC for mixture models. Some recent suggestions have been made regarding these problems (see Celeux, et al., 2006).

Use of the Bayesian Information Criterion (BIC) was considered for model selection, but, for capture-recapture models the sample size is unclear (as N itself is a parameter and k , the number of capture periods, is usually much smaller than necessary for asymptotic properties to work). Therefore, AIC and DIC are focused on as potential model selection criteria; specifically, it is assessed whether AIC and DIC choose the correct model for a given data set.

A model selection criterion proposed by Gelfand and Ghosh (1998) is based upon minimizing the squared predictive error of the observed data, where the predictive distribution of the observed data is based partially upon the posterior distribution of the parameters, given the observed data, rather than on the prior distribution of the parameters. Ghosh and Norris (2005) discussed using this method for Model M_{bh} , and their findings were promising. This criterion is an area of future research, as it easily allows non-nested models to be directly compared and it balances between model fit and model complexity.

Results

Data Generation Process and Bayesian Analysis Method

This simulation consists of eight experiments. Experiment one contains 100 data sets generated under each modeling assumption ($M_0, M_t, M_h, \dots, M_{tbh}$). Experiments two through eight each contain 50 data sets generated under each modeling assumption. Each experiment uses Markov Chain Monte Carlo (MCMC) methods to fit each data set using each of the eight models. Thus, experiment one consists of 6,400 analyses (800 data sets each analyzed under eight models). Experiments two through eight consist of 3,200 analyses (400 data sets analyzed under eight models). Each data set is a simulated capture-recapture study with $k = 5$ capture periods. The methodology used to generate p_{ij} values is illustrated in Table 1, and detailed information regarding the data generating parameters is provided in an Appendix available at <http://www.mathsci.appstate.edu/~rmg/>.

Calculations of p_{ij} for M_{tbh} , the most complex model, are computed as $F(\mu + \beta_j + \eta\tau_{ij} + \kappa Z_i)$, where F is the Logistic distribution function

$$F(x) = [1 + e^{-x}]^{-1}, Z_i \stackrel{i.i.d.}{\sim} N(0,1),$$

where $\tau_{ij} = 1$ if the animal has been previously captured, and $\tau_{ij} = 0$ otherwise.

The approach in Table 1 resembles a 2^{5-2} fractional factorial design with factors N , Average p_{ij} , and magnitude of time, behavioral, and heterogeneity effects. Means and standard deviations of the p_{ij} for each simulation experiment are listed in the Appendix.

For each data set, and under each model, an estimate of the posterior density of N was constructed using WinBUGS Version 1.4. The median of this posterior distribution, denoted \hat{N} , was chosen to estimate N and AIC and DIC were also computed. For these simulation experiments, non-informative prior distributions were chosen for the model parameters. Specifically, $\delta = 0.5$ was chosen as the hyperparameter for the prior distribution of N ; $r = 2$ was selected for the number of support points for F , and $a = b = 0.5$ for hyperparameters

Table 1: Data Generating Assumptions for Simulation Experiments 1 to 8

Experiment Number	N	Average p_{ij}	Time Effects	Behavioral Effects	Heterogeneity
1	500	0.2	Large	Positive	Large
2	500	0.2	Small	Positive	Small
3	500	0.4	Large	Negative	Large
4	100	0.4	Large	Positive	Small
5	100	0.4	Small	Positive	Large
6	100	0.2	Large	Negative	Small
7	500	0.4	Small	Negative	Small
8	100	0.2	Small	Negative	Large

for the prior distributions of all the capture probabilities.

A burn-in period of 3,000 samples was used to allow convergence of the MCMC processes to a stable distribution. After the burn-in period, 2,000 samples were selected from each of three MCMC chains with dispersed starting values for the model parameters. Therefore, posterior distribution estimates are based upon 6,000 total samples. Convergence of the models was checked through the Gelman-Rubin statistic in WinBUGS. Table 2 shows means and percentages of times each model was selected by the MCMC estimates of AIC for Experiment one.

Analysis of AIC as a Model Selection Criterion

AIC (Akaike, 1973) has been used extensively as a model selection tool. Calculation of AIC adds a parameter penalty to the estimated Kullback-Leibler Discrepancy between the fitted model and the true model. Using θ as a general term to represent all the model parameters (e.g. $\theta = (N, \mathbf{P})$ as in Equation 2.3), X as a general term to represent the observed data (e.g. $X = (Z_1, \dots, Z_L)$ as in Equation 2.3), p' as the number of model parameters, and LogL as the log likelihood function, a form for calculation of AIC is given by

$$\text{AIC} = -2\text{LogL}(\hat{\theta} | X) + 2p' \quad (4.1)$$

where $\hat{\theta}$ is the MLE of θ under the assumed model. However, the AIC calculation used here is different from the usual form of AIC. Defining

$$D(\theta) = -2\text{LogL}(\theta | X), \quad (4.2)$$

use $\text{AIC} = E[D(\theta)|X] + 2p'$ where $E[D(\theta)|X]$ is the mean of the posterior distribution of $D(\theta)$.

Analysis of Table 2, which gives the MCMC AIC means and model selection percentages for simulation experiment one, indicates that overall the AIC is effective in determining the correct model. For the first seven columns in the table, the minimum AIC mean occurs when the fitted model matches the data generating assumptions. This suggests that AIC is capable of identifying the correct model, on average.

Perhaps more indicative of the performance of AIC is the percentage of times it chooses the correct model. For this analysis, a model with the minimum posterior mean of AIC was chosen for a given data set. When a tie occurs between two models, the simpler model is chosen. Ideally, the diagonal entries in the table should have the highest percentages of selections by AIC. The table columns represent the true model generating assumptions.

Selection of a different model from the data generating assumptions may be called a model selection error, and the percentage of AIC

BAYESIAN MODEL SELECTION FOR CAPTURE-RECAPTURE MODELS

Table 2: Simulation Experiment One Average AIC Posterior Mean and AIC Model Selection Percentages

Model Fit	True Model							
	M_0	M_h	M_t	M_b	M_{bh}	M_{tb}	M_{th}	M_{tbh}
	Avg AIC (Top Line) AIC% (Bottom Line)							
M_0	134.3 92%	197.2 0%	320.7 0%	191.8 0%	252.4 0%	533.4 0%	354.3 0%	539.0 0%
M_h	138.4 0%	159.2 95%	325.1 0%	194.4 0%	196.6 0%	537.2 0%	328.5 0%	501.4 0%
M_t	142.4 1%	205.4 0%	136.2 87%	173.2 0%	236.2 0%	164.6 2%	198.1 0%	239.7 0%
M_b	136.4 7%	197.9 0%	208.0 0%	157.9 99%	201.6 0%	227.2 0%	193.6 0%	181.3 0%
M_{bh}	142.5 0%	163.0 5%	214.8 0%	163.6 1%	165.5 99%	234.1 0%	184.5 2%	174.5 13%
M_{tb}	144.2 0%	203.8 0%	138.0 13%	165.8 0%	207.1 0%	151.1 97%	186.1 3%	174.5 11%
M_{th}	153.8 0%	175.0 0%	147.7 0%	177.0 0%	181.1 1%	161.9 2%	166.7 87%	165.1 36%
M_{tbh}	155.6 0%	176.6 0%	149.4 0%	177.1 0%	179.8 0%	162.1 0%	168.5 8%	163.5 42%
Error %	8%	5%	13%	1%	1%	3%	13%	58%

model selection errors is also listed in the last row of Table 2. In this respect, for seven of the eight models, AIC performs well. Among these seven models, for M_t and M_{th} , the percentage of selections is 87 percent, which is somewhat lower than for the other models. When M_t and M_{th} are not selected by AIC, though, AIC selects a similar model, but with more effects. This is better than the selection of an unrelated model. Model M_{tbh} does not perform as well. Data generated under the assumptions of M_{tbh} only had a 42% selection rate by AIC. When M_{tbh} was not selected in this column, the model selected was one of the sub-models containing two of the effects (M_{th} , M_{bh} , and M_{tb}).

Some of this could be due to relative weighting of the time, behavioral, and heterogeneity effects within M_{tbh} , as AIC may be picking the model based on the most significant of these effects present in any particular data set

Furthermore, with five capture periods, model M_{tbh} may be somewhat over-parameterized. Thirty-one distinct capture histories were observed, and model M_{tbh} includes 13 parameters for such data, which may lead to the estimation of effects due only to random chance. However, from an overall look at this table, it is concluded that AIC performs well as a model selection tool.

A summary of the AIC selection rates from Experiments two through eight is given in Table 3, which lists only whether AIC chose the correct model in each experiment. Thus, the 94% entry in the first row and column shows that in experiment two, AIC chose model M_0 correctly 94% of the time for the data sets generated with constant capture probability.

In column one of Table 3, a strong AIC selection rate is observed for M_0 data for all

Table 3: Selection Rates for AIC for Simulation Experiments 2 to 8 Selection Rates for Data Sets Generated Via Listed Model Assumptions

Experiment	M_0	M_b	M_t	M_h	M_{bh}	M_{tb}	M_{th}	M_{tbh}
2	94%	96%	78%	0%	86%	0%	0%	0%
3	90%	96%	90%	92%	98%	100%	94%	70%
4	92%	96%	82%	6%	62%	0%	0%	0%
5	94%	92%	36%	78%	12%	52%	2%	0%
6	84%	48%	62%	2%	8%	0%	0%	0%
7	96%	100%	88%	0%	76%	4%	0%	0%
8	92%	44%	18%	54%	6%	20%	0%	0%

experiments. Column two shows strong AIC selection rates for the M_b data sets, with the exception of experiments six and eight. These two experiments had relatively small capture probabilities, smaller population sizes of $N = 100$, and negative behavioral effects. This combination of factors makes detection of behavioral effects difficult due to small observed numbers of recaptures.

Column three shows strong selection rates for M_t data, except for experiments five and eight. However, both of those experiments have small-magnitude time effects and smaller population sizes $N = 100$, indicating a simpler model such as M_0 may be more appropriate for the data. Experiments two and seven, which also have small time effects, but a larger population size $N = 500$ show larger AIC selection percentages of 78% and 88% respectively.

Column four shows that M_h data has reasonably high selection percentages for experiments three, five, and eight, and small selection percentages for experiments two, four, six, and seven. The low selection rates occur in experiments with small heterogeneity effects in the data. It appears AIC selects a heterogeneity model when the heterogeneity effects are large, but not when they are relatively small.

Column five shows that M_{tb} data has a high AIC selection rate in experiments two, three, and seven. M_{tb} data has a moderate AIC selection rate of 62% in experiment four, and has low selection rates in experiments five, six, and eight. The low selection rates occur in

experiments with small time effects. The moderate selection rate in experiment four occurs when time effects in the data are large. Some examination of this case shows that when M_{tb} is not chosen, one of the submodels M_b or M_t is chosen by AIC.

The low selection percentages for experiments two, four, six, and seven for the heterogeneity models M_{bh} , M_{th} , and M_{tbh} occur due to the small heterogeneity effects in those experiments, and this again reflects that comparable models without heterogeneity effects can adequately fit the data. Low selection rates for M_{bh} , M_{th} , and M_{tbh} are seen in experiment eight, and for the M_{th} and M_{tbh} data for experiment five. These low rates occur for experiments where heterogeneity effects are large. However, due to the small magnitude time effects in experiments five and eight, for the M_{th} data AIC chooses M_b and M_{bh} most often as the best model, reflecting adequate fit for these data sets by simpler models. Some examination of the underlying results (not available in Table 3) shows that the penalty term for the number of parameters is the reason that M_{th} has a higher AIC value for these data sets.

For the M_{tbh} data sets, AIC chooses Model M_{bh} most commonly, followed by model M_b . The choice of M_{bh} again reflects the small magnitude of the time effects in these data sets. The choice of model M_b is surprising given that the heterogeneity in the data is strong in experiment five. However, a behavioral effect and a heterogeneity effect are not completely

BAYESIAN MODEL SELECTION FOR CAPTURE-RECAPTURE MODELS

unrelated. For capture periods two through k , the behavioral effect creates two distinct groups in the population: those which have been previously captured and those which have not been previously captured. Each group has separate capture probabilities. Although group membership is changing with each capture period, Model M_b could provide a reasonable fit to data with heterogeneity in some instances.

Finally, for experiment eight, for the M_{bh} , M_{th} , and M_{tbb} data sets, no particular model is selected overwhelmingly, and the true model is also rarely selected for these data sets. This may reflect the combination of small population size of $N = 100$, the negative behavioral effects, the average capture probabilities being 20%, and the large degree of heterogeneity in the data. For a small population, it is difficult to have one data set reflect all those sources of variation, causing problems for a selection criterion such as AIC.

Overall the performance of AIC as a model selection method for these models is encouraging. However, it is recommended that AIC be a guide to select a subset of suitable models for further analysis. Although AIC performed well in selecting the true model when the degree of underlying time, heterogeneity, or behavioral effects was large, the performance when these effects were small means that another model is selected. For this reason, it is recommended that AIC be used to narrow the set of eight models down to a smaller number of candidate models. A more detailed analysis involving other factors, such as the opinion of a subject matter expert, should be used to make the final model choice.

Analysis of DIC as a Model Selection Criterion

The DIC criterion is a recent development in model selection. DIC can be expressed similarly to AIC. Given the common use of AIC, this feature allows users to quickly understand the form and use of DIC. Additionally, DIC is easy to calculate, as it is a function of the posterior parameters and the model deviance (where deviance is related to the log-likelihood).

Using the same notation as in the definition of AIC, and again denoting $D(\theta) = -2\text{LogL}(\theta | X)$, and defining $pD = E(D(\theta)|X) - D(\hat{\theta})$,

where $E(D(\theta)|X)$ represents the posterior mean of $D(\theta)$, it is possible to calculate

$$\text{DIC} = D(\hat{\theta}) + 2pD, \quad (4.3)$$

where $\hat{\theta}$ is a posterior estimate of θ , e.g., $\hat{\theta} = E[\theta | X]$ or $\text{Median}[\theta | X]$. As stated previously, the pD term in DIC represents an effective number of parameters. The pD term measures the decrease in the deviance (increase in the likelihood) obtained by using posterior estimates of the parameters θ . Note that although DIC is structured to look like AIC, the penalty term is actually a function of the model fit, not simply a discrete number of parameters.

For computational purposes, $\text{Dev}(\theta)$ is defined as the MCMC computed deviance for any particular data set and model combination and \bar{D} as the MCMC mean of the deviance statistic, pD is computed as $pD = \bar{D} - \text{Dev}(\hat{\theta})$ and computationally, results in $\text{DIC} = \text{Dev}(\hat{\theta}) + 2pD$.

In the simulations, DIC did not perform as well as AIC in model selection. For several of the models, most notably M_{bh} , the pD penalty term in the DIC criterion was frequently negative in the simulations. Although pD is typically positive for most Bayesian statistical models, pD can be negative for a particular model and data set if the likelihood function is not log-concave. A negative pD rewards, rather than penalizes, a model for model complexity. When pD is negative, then for simple data sets (M_0 , for example), DIC selects a more complex model in the majority of cases. Of particular concern was the disproportionately large number of selections of model M_{bh} across all data sets, due to the frequency of the penalty term pD being negative. Detailed data tables regarding the performance of DIC across the eight simulation experiments are available at <http://www.mathsci.appstate.edu/~rmg/>.

Spiegelhalter, et al. (2002) stated that alternative choices for $\hat{\theta}$ could be the posterior median or posterior mode. So, pD can be calculated with these alternatives to the posterior mean of $\hat{\theta}$. Because DIC performed poorly with the posterior mean as $\hat{\theta}$, the performance of DIC

was examined when the posterior median was used for $\hat{\theta}$ instead. Ultimately, with this change, the problem of negative pD values improved, but still persisted with this change. Overall, performance of DIC in model selection for these models is inferior to that of AIC; based upon this simulation study, use of AIC as a model selection tool is recommended over DIC.

Analysis of Real Data Sets: Cottontail Rabbit Data

In Edwards and Eberhardt (1967), a capture-recapture experiment involving 135 cottontail rabbits was performed. The rabbits were released into a forty acre rabbit-proof area, and eighteen capture periods followed after a four day waiting period which gave the rabbits familiarity with their surroundings. Bayesian Models with the Program MARK models are compared. The data and a Program MARK analysis of the data are included with the MARK software package, and Pledger (2000), among others, has analyzed this data set. A total of seventy-six animals were captured at least once during the eighteen capture periods. Forty-three animals were captured once during the study, sixteen were captured twice, eight were captured three times, six were captured four times, two were captured six times, and one rabbit was captured seven times.

Using the models provided by Program MARK, Table 4 gives the estimate of N for each model, and the upper and lower limits of ninety-five percent confidence intervals for N, and the frequentist AIC statistic for each model. For Model M_{th} , the data was analyzed under two specifications, once with constant difference in capture probabilities between the two mixture groups across the time periods, and once without this restriction. Estimates and confidence limits for N are rounded to the nearest integer.

Using AIC, it is found that the M_{th} model with additive capture probability difference across the $r = 2$ groups across capture periods, and Model M_{tbh} have comparably small AIC values. The point estimator from the chosen M_{th} model is more accurate than the M_{tbh} model and the confidence interval for M_{th} is narrower.

Using WinBUGS v.1.4, the Bayesian

Table 4: Program MARK Results for Cottontail Dataset

Estimator	\hat{N}	LCL	UCL	AIC
M_0	96	87	114	379.6
M_h (2 mixture groups)	136	96	256	369.6
M_h (3 mixture groups)	157	89	593	373.5
M_b	94	82	129	381.6
M_t	95	86	112	354.6
M_{bh}	113	86	214	369.1
M_{th} (2 mixture groups; additive)	133	96	241	341.3
M_{th} (2 mixture groups; unrestricted)	98	88	117	367.0
M_{tb}	162	117	260	343.3
M_{tbh}	270	100	1698	341.9

models were fit to the cottontail data, using the non-informative prior distributions for N and the capture probabilities described in Section 4.1. For each model, Table 5 lists \hat{N} (the posterior median), the AIC posterior mean, and ninety-five percent, equal-tailed posterior interval bounds from the MCMC posterior distribution of N. Figure 1 shows the MCMC posterior density of N for Model M_{tb} .

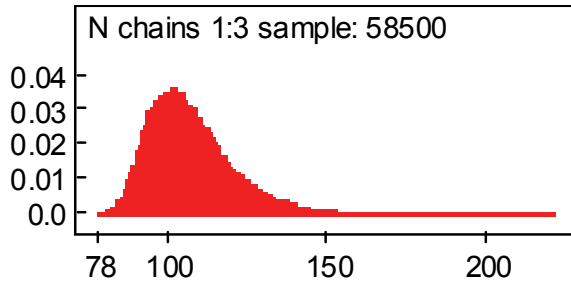
Table 5: Bayesian Model Results for Cottontail Dataset

Model	\hat{N}	2.5%	97.5%	AIC Posterior Mean
M_0	97	86	114	461.5
M_h	145	99	615	453.3
M_b	96	81	153	464.4
M_t	92	83	106	453.2
M_{tb}	104	88	138	452.8
M_{bh}	119	85	581	457.2
M_{th}	98	86	120	475.8
M_{tbh}	107	88	151	477.2

BAYESIAN MODEL SELECTION FOR CAPTURE-RECAPTURE MODELS

Note that AIC chooses Model M_{tb} . Other candidate models with comparable AIC values are M_h and M_t .

Figure 1: Posterior Density of N for Model M_{tb} for Cottontail Dataset



Model M_{tb} underestimates the true $N = 135$ but a 95% equal-tail interval from the posterior distribution of N contains the true N . The \hat{N} from the Bayesian Model M_t underestimates the true $N = 135$, as does Model M_t in Program MARK. The interval estimates produced by the two methods are similar, which is not surprising given that relatively uninformative prior distributions for the parameters were used, and the likelihood functions of the two models are the same.

The Bayesian M_h estimate is somewhat above, but relatively close to the true $N = 135$, as are both M_h estimates from Program MARK. The Bayesian posterior density of N has a higher 97.5th percentile than the upper bound for the confidence interval given for Model M_h in Program MARK. The Model M_h posterior density for N is heavily right-skewed, and the posterior interval length could be significantly shortened by choosing an interval other than an equal-tailed interval, or by lowering the confidence level.

Ultimately, the M_{tb} point estimate of N , via the posterior median of N , is comparably accurate with the M_{tb} estimator from Program MARK. Also note the Program MARK M_{th} estimator is quite accurate for N . The Bayesian M_{tb} model has a narrower confidence interval than that from Program MARK, and the Bayesian interval contains the true $N = 135$.

Because this is only one data set, general conclusions cannot be made.

Analysis of Real Data Sets: Mead's Milkweed Flower Data

Alexander, Slade, and Kettle (1997) used mark-recapture methods to estimate the number of Mead's Milkweed plants on a 4.5-ha tract of land in Kansas. The capture periods consisted of an annual search of the land area over a span of four years. Observed plants were marked with a flag so that previous captures were detectable in subsequent years. Censuses were considered impossible because these plants are perennial and do not flower every year. Presence of flowering stems makes the plants easier to observe. The authors considered the population closed over the four-year span because the plant has a long lifespan, a high survival rate, and births and deaths were considered negligible during the study. Ultimately, a total of 129 flowers were observed in the study. Twenty-two plants were observed during one capture period, fifty-six were observed during two capture periods, twenty-five were observed during three capture periods, and twenty-six were observed during all four capture periods.

Model M_{tbh} was chosen as an ideal model for the data because time effects occur due to annual variation in flowering, behavioral effects occur because the visible flags make recapture easier in subsequent years, and heterogeneity effects occur because some plants have larger underground root systems which make them more likely to flower in a given year. Alexander, et al. used Program CAPTURE for the analysis (see <http://www.mbr-pwrc.usgs.gov/software.html> for details), Model M_{tbh} was unavailable in Program CAPTURE at that time, and they ultimately found some reasonable, but non-ideal, options for simplifying the data to allow the other seven closed population models to be fit.

Both Program MARK and WinBUGS v. 1.4 were used to analyze the data set and to choose the proper model from using the AIC statistic in each case. Results are listed below in Tables 6 and 7. Note that in Program MARK, M_{tb} and M_{tbh} models were fit with behavioral

effects additive across time periods and mixture groups.

Table 6: Program MARK Results for Mead’s Milkweed Dataset

Estimator	\hat{N}	LCL	UCL	AIC
M_0	132	130	139	227.9
M_h (2 mixture groups)	135	132	144	218.7
M_b	222	162	393	115.8
M_t	129	130	135	80.2
M_{bh} (2 mixture groups)	945	230	6,769	109.7
M_{th} (2 mixture groups)	130	130	137	51.3
M_{tb}	167	137	326	48.1
M_{tbh}	1,228	233	11,769	38.4

Table 7: Bayesian Model Results for Mead’s Milkweed Dataset

Model	\hat{N}	2.5%	97.5%	AIC Posterior Mean
M_0	133	130	138	263.4
M_h	136	131	145	259.3
M_b	270	172	2823	152.4
M_t	130	130	133	119
M_{tb}	632	336	2445	79.8
M_{bh}	365	182	2823	149
M_{th}	131	130	136	82.9
M_{tbh}	131	130	632	85

Comparing the results, it is observed that in Program MARK, the AIC statistic favors Model M_{tbh} . This model has a very large upper bound for the confidence interval and thus a wide confidence interval for N. Model M_b in Program MARK has an AIC statistic that is fairly close to that of M_{tbh} and the confidence

interval for N is much narrower than that of M_{tbh} .

The Bayesian AIC statistic favors Model M_{tb} , and Models M_{th} and M_{tbh} are other possible choices. The Bayesian M_{tb} model has a large interval width and a large 97.5th percentile of the posterior distribution of N. The competing Bayesian models M_{th} and M_{tbh} have equal point estimates of N, but M_{tbh} has a much larger 97.5th percentile of the posterior distribution of N, leading to a wider posterior 95% interval.

Conclusion

In summary, useful findings for closed population capture-recapture models have been established. Eight Bayesian capture-recapture models accounting for the known sources of variability in the capture probabilities of closed animal populations were developed. Using the WinBUGS v.1.4 software, these models were easy to fit to capture-recapture data sets, and MCMC estimates of the posterior density of N are easily obtained from the output. Additionally, the modified version of AIC works well as a model selection tool for capture-recapture data sets, thus AIC is useful as a preliminary method of reducing the set of candidate models from eight down to a smaller subset worthy of further exploration to determine the best fitting model. The DIC criterion did not perform as well as AIC for capture-recapture data sets and the use of AIC over DIC is recommended.

Further areas of exploration include examining whether informative priors improve estimation of N when capture probabilities are small. Negative bias in estimating N is common for populations with heterogeneity, particularly when a significant fraction of the population has small capture probabilities. The performance of the heterogeneity models (M_h , M_{th} , M_{bh} , M_{tbh}) when the finite mixture distribution F has $r > 2$ mass points should also be examined.

Acknowledgements

The authors thank Professors Leonard A. Stefanski and Kenneth H. Pollock, both of North Carolina State University, for many helpful suggestions and contributions toward this work.

BAYESIAN MODEL SELECTION FOR CAPTURE-RECAPTURE MODELS

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, B. N. Petrov & F. Csaki (Eds.), 267-281. Akademiai Kiado, Budapest.
- Alexander, H. M., Slade, N. A., & Kettle, W. D. (1997). Application of mark-recapture models to estimation of the population size of plants. *Ecology*, 78(4), 1230-1237.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference, A practical information-theoretic approach* (2nd Ed.). NY: Springer-Verlag.
- Castledine, B. (1981). Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.
- Celeux, G., Forbes, F., Robert, C. P., & Titterton, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4), 681 – 684.
- Chao, A. (1988). Estimating animal abundance with capture frequency data. *Journal of Wildlife Management*, 52, 295-300.
- Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45, 427-438.
- Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2), 158-175.
- Coull, B.A., & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55(1), 294-301.
- Durban, J. W., & Elston, D. A. (2005). Mark-recapture with occasion and individual effects: abundance estimation through bayesian model selection in a fixed dimensional parameter space. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(3), 291-305.
- Edwards, W. R., & Eberhardt, L. (1967). Estimating cottontail abundance from live trapping data. *Journal of Wildlife Management*, 31(1), 87-96.
- Gelfand, A. E., & Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85, 1-11.
- George, E. I., & Robert, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika*, 79(4) 677-683.
- Ghosh, S. K., & Norris, J. L. (2005). Bayesian Capture-Recapture Analysis and Model Selection Allowing for Heterogeneity and Behavioral Effects. *Journal of Agricultural Biological and Environmental Statistics*, 10, 35-49.
- Gosky, R. M., & Ghosh, S. K. (2011). A Comparative Study of Bayes Estimators of Closed Population Size from Capture-Recapture Data. *Journal of Statistical Theory and Practice*, to appear.
- King, R., & Brooks, S. P. (2008). Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty. *Biometrics*, 64, 816-824.
- Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59, 1123-1130.
- Norris, J. L., & Pollock, K. H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 52, 639-649.
- Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal population. *Wildlife Monographs*, 62, 135 pp.
- Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models under mixtures. *Biometrics*, 56, 434-442.
- Schwartz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit, (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B*, 64, 583-639.

Covariate-Adjusted Constrained Bayes Predictions of Random Intercepts and Slopes

Robert H. Lyles
Emory University

René H. Moore
University of Pennsylvania

Amita K. Manatunga
Emory University

Kirk A. Easley
Emory University

Constrained Bayes methodology represents an alternative to the posterior mean (empirical Bayes) method commonly used to produce random effect predictions under mixed linear models. The general constrained Bayes methodology of Ghosh (1992) is compared to a direct implementation of constraints, and it is suggested that the former approach could feasibly be incorporated into commercial mixed model software. Simulation studies and a real-data example illustrate the main points and support the conclusions.

Key words: Mixed linear model, prediction, random effects, shrinkage.

Introduction

The standard mixed linear model (e.g., Laird & Ware, 1982) remains a popular practical tool for analyzing longitudinal, repeated measures, or otherwise correlated continuous data. In such analyses, the prediction of linear combinations of fixed and random effects can be of great interest. The typical approach implemented in commercial software is to obtain empirical best linear unbiased predictors (EBLUPs), which estimate the posterior mean of the linear combination given the response data (Littell, et al., 2006). The general acceptance of these empirical Bayes-like predictions stems from their intuitive appeal and their theoretical

underpinnings as minimal prediction mean squared error estimates (Searle, et al., 1992). They are also referred to as shrinkage estimators, given their characteristic of pulling subject-specific predictions toward a population mean.

Due to the shrinkage phenomenon, EBLUPs stemming from linear mixed models exhibit distributions that can be much narrower than those assumed to characterize the random variables being predicted. Several authors (e.g., Efron & Morris, 1971; Louis, 1984; Ghosh, 1992) have suggested potential drawbacks to this general feature and proposed methods that reduce shrinkage and/or more closely match the predictor and underlying true distributions.

One effect of overshrinkage in certain applications is that it can lead to a lack of sensitivity for identifying extreme experimental units relative to a fixed threshold (i.e., the probability that an EBLUP lies beyond a threshold given that the true random variable does can be quite small). To improve sensitivity in such a context, Lyles and Xu (1999) proposed constrained Bayes predictors of random intercepts and slopes aimed to minimize mean squared error of prediction (MSEP) given that the means and variances of the predictor distributions match those of the true random effects. Lyles, et al. (2007) introduced additional prediction criteria (e.g., regional bias and

Robert H. Lyles is an Associate Professor in the Department of Biostatistics at The Rollins School of Public Health. Email: rlyles@emory.edu. René Moore is an Assistant Professor in the Department of Biostatistics and Epidemiology in the School of Medicine. Email: rhmoore@mail.med.upenn.edu. Amita Manatunga is a Professor in the Department of Biostatistics at The Rollins School of Public Health. Email: amanatu@emory.edu. Kirk Easley is a Senior Associate in the Department of Biostatistics at The Rollins School of Public Health. Email: keasle2@emory.edu.

BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

MSEP) that are relevant when extreme subjects are of key interest and they suggested that the constrained Bayes approach can be an appealing alternative in such situations. Constrained Bayes prediction of random effects has not been widely advocated for use in the mixed linear model context.

The models considered by Lyles and Xu (1999) are extended here to use fixed and/or time-dependent covariates, and their direct constrained Bayes strategy is compared with the general paradigm advocated by Ghosh (1992). This comparison is relevant for two reasons.

First, while the criteria put forth by Lyles and Xu are specific to the mixed linear model, Ghosh's approach originates from a more general and decidedly Bayesian point of view. Ghosh provides a paradigm for minimizing a mean squared error criterion subject to matching the posterior expectation of the first two moments of a parameter distribution to corresponding moments of the histogram of the set of estimates. It is therefore useful to assess the performance of Ghosh's paradigm in the mixed model setting and to compare it against an approach that is directly rooted in that context.

Second, Ghosh's method is general, flexible, and implemented in a straightforward and consistent manner. Therefore its validation against an approach directly rooted in the mixed model setting could highlight, for practitioners and commercial mixed linear model software developers, the viability of an accessible alternative prediction method.

Methodology

Models and Posterior Mean Predictions

Two familiar normal-theory mixed linear models are used for illustration: the random intercept and random intercept/slope models, respectively.

The random intercept (or one-way random effects ANOVA) model is specified as follows (e.g., Searle, et al., 1992):

$$Y_{ij} = \mu + b_i + e_{ij} \quad (1)$$

($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$), with i indexing the subject and j indexing the observation. Typical normality assumptions dictate that $b_i \sim N(0, \sigma_b^2)$ and $e_{ij} \sim N(0, \sigma_w^2)$, with independence across subjects and between the random terms b_i and e_{ij} .

Under model (1), a common objective is to predict the i^{th} subject's random subject-specific mean, i.e., $\mu_i = \mu + b_i$ ($i=1, \dots, k$). The EBLUP, as provided by standard mixed model software, is an estimate of the posterior mean $E(\mu_i | \mathbf{Y}) = E(\mu_i | \mathbf{Y}_i)$, where \mathbf{Y} and \mathbf{Y}_i denote the complete and i^{th} subject-specific data vectors, respectively:

$$\tilde{\mu}_i = E(\mu_i | \mathbf{Y}_i = \mathbf{y}_i) = v_i \bar{y}_i + (1 - v_i) \mu \quad (2)$$

$$\text{where } \bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}, \text{ and}$$

$$v_i = \{1 + \sigma_w^2 / (n_i \sigma_b^2)\}^{-1}.$$

The parameter v_i governs the extent to which the predicted value shrinks toward the population mean μ , with more excessive shrinkage occurring when v_i is small (i.e., when $\sigma_w^2 / (n_i \sigma_b^2)$ is large). The BLUP is obtained by replacing μ in (2) by its best linear unbiased estimate (Searle, et al., 1992), whereas in practice the EBLUP also replaces the variance components in (2) by their estimates.

Next, consider the random intercept/slope model, also known as a randomized regression or linear growth curve model (e.g., Diggle, et al., 1994):

$$Y_{ij} = (\alpha + a_i) + (\beta + b_i) t_{ij} + e_{ij} \quad (3)$$

($i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$), where t_{ij} denotes the time at which Y_{ij} is measured. Typically this model assumes independence across subjects and normally distributed random effects as follows:

$$\begin{pmatrix} \mathbf{a}_i \\ \mathbf{b}_i \\ \mathbf{e}_{ij} \end{pmatrix} \sim N_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_2^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right\},$$

with $\sigma_1^2, \sigma_2^2, \sigma_{12}$, and σ^2 denoting the variances of the subject-specific intercept and slope deviations, their covariance, and the random error variance, respectively.

Under model (3), it is common to seek predictions of the i^{th} subject's random intercept ($\alpha_i = \alpha + a_i$) and slope ($\beta_i = \beta + b_i$). As with model (1) and most feasible mixed linear models, standard software provides EBLUPs for these quantities. In this case, they are estimates of the posterior means $E(\alpha_i | \mathbf{Y}) = E(\alpha_i | \mathbf{Y}_i)$ and $E(\beta_i | \mathbf{Y}) = E(\beta_i | \mathbf{Y}_i)$. The normality assumptions accompanying model (3) yield

$$\begin{aligned} \tilde{\beta}_i &= E(\beta_i | \mathbf{Y}_i = \mathbf{y}_i) \\ &= \beta + (\sigma_{12} \mathbf{1}'_{n_i} + \sigma_2^2 \mathbf{t}'_i) \Sigma_i^{-1} (\mathbf{y}_i - \alpha \mathbf{1}_{n_i} - \beta \mathbf{t}_i) \end{aligned} \quad (4)$$

where $\Sigma_i = \text{Var}(\mathbf{Y}_i) = \mathbf{Z}_i \Delta \mathbf{Z}'_i + \sigma^2 \mathbf{I}_{n_i}$, \mathbf{Z}_i is the design matrix for the simple linear regression of \mathbf{Y}_i on time (\mathbf{t}_i) for subject i , and $\Delta = \text{Var}(\mathbf{a}_i, \mathbf{b}_i)'$. Assuming $n_i \geq 2$, Lyles and Xu (1999) showed that $E(\beta_i | \mathbf{Y}_i)$ takes an appealing form:

$$\begin{aligned} \tilde{\beta}_i &= E(\beta_i | \mathbf{Y}_i) \\ &= \gamma_{i1} + \gamma_{i2} \hat{\alpha}_{i,\text{ols}} + \gamma_{i3} \hat{\beta}_{i,\text{ols}} \end{aligned} \quad (5)$$

where $\hat{\alpha}_{i,\text{ols}}$ and $\hat{\beta}_{i,\text{ols}}$ represent the ordinary least squares (OLS) intercept and slope from regressing \mathbf{Y}_i on \mathbf{t}_i . The coefficients in (5) are given by:

$$\begin{aligned} \gamma_{i2} &= (\sigma_{12} v_{\beta i} - \sigma_2^2 c_{\alpha \beta i}) / \delta_i, \\ \gamma_{i3} &= (\sigma_2^2 v_{\alpha i} - \sigma_{12} c_{\alpha \beta i}) / \delta_i, \end{aligned}$$

and

$$\gamma_{i1} = \beta(1 - \gamma_{i3}) - \alpha \gamma_{i2},$$

with

$$\delta_i = (v_{\alpha i} v_{\beta i} - c_{\alpha \beta i}^2),$$

$$v_{\alpha i} = \text{Var}(\hat{\alpha}_{i,\text{ols}}) = \sigma_1^2 + \sigma^2 [1/n_i + \bar{t}_i^2 / \{(n_i - 1)s_{t_i}^2\}],$$

$$v_{\beta i} = \text{Var}(\hat{\beta}_{i,\text{ols}}) = \sigma_2^2 + \sigma^2 / \{(n_i - 1)s_{t_i}^2\},$$

$$c_{\alpha \beta i} = \text{Cov}(\hat{\alpha}_{i,\text{ols}}, \hat{\beta}_{i,\text{ols}}) = \sigma_{12} - \bar{t}_i \sigma^2 / \{(n_i - 1)s_{t_i}^2\},$$

and where \bar{t}_i and $s_{t_i}^2$ denote the sample mean and variance of the observation times $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})'$. Similarly, it can be shown that

$$\tilde{\alpha}_i = E(\alpha_i | \mathbf{Y}_i) = \tau_{i1} + \tau_{i2} \hat{\alpha}_{i,\text{ols}} + \tau_{i3} \hat{\beta}_{i,\text{ols}} \quad (6)$$

with

$$\tau_{i2} = (\sigma_1^2 v_{\beta i} - \sigma_{12} c_{\alpha \beta i}) / \delta_i,$$

$$\tau_{i3} = (\sigma_{12} v_{\alpha i} - \sigma_1^2 c_{\alpha \beta i}) / \delta_i,$$

and

$$\tau_{i1} = \alpha(1 - \tau_{i2}) - \beta \tau_{i3}.$$

Consider the problem of predicting the unknown response under model (3) for subject i at some clinically or otherwise significant point in time (t_i^*). In other words, seeking to predict the value of

$$Y_{it}^* = E(Y_{ij} | \alpha_i, \beta_i, t_{ij} = t_i^*) = \alpha_i + \beta_i t_i^*.$$

The posterior mean of Y_{it}^* is

$$\tilde{Y}_{it}^* = E(Y_{it}^* | \mathbf{Y}_i) = \tilde{\alpha}_i + \tilde{\beta}_i t_i^* \quad (7)$$

where $\tilde{\beta}_i$ and $\tilde{\alpha}_i$ are as defined in (5) and (6), for $n_i \geq 2$. EBLUPs for $\tilde{\beta}_i$ and $\tilde{\alpha}_i$ are obtained by inserting parameter estimates into the general expressions for $E(\beta_i | \mathbf{Y}_i)$ and $E(\alpha_i | \mathbf{Y}_i)$, where $n_i = 1$ is permissible. The EBLUP for Y_{it}^* inserts the EBLUPs for $\tilde{\beta}_i$ and $\tilde{\alpha}_i$ into (7).

Constrained Bayes Predictions

The constrained Bayes (CB) approach (Louis, 1984) was extended by Ghosh (1992)

BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

into a flexible paradigm. Lyles and Xu (1999) suggested that this general idea provides a natural alternative to the EBLUP in the mixed linear models context when overshrinkage could detract from the desired application of predicted values. They applied a slight adaptation of the CB concept under models (1) and (3) by minimizing prediction mean squared error (MSEP) among unbiased candidates whose variances match that of the assumed random effects distribution. While this necessarily results in some sacrifice in overall MSEP relative to the posterior mean, it provides a set of predictions that more faithfully reproduce the underlying distribution of interest and are less likely to under-represent the extremeness of experimental units in the tails.

Under model (1), the CB predictor for μ_i recommended by Lyles and Xu is obtained directly by forcing the first two moments of the $\tilde{\mu}_i$ and μ_i distributions to match:

$$\tilde{\mu}_{i,LX} = \sqrt{v_i} \bar{y}_i + (1 - \sqrt{v_i}) \mu \quad (8)$$

The square root is indicative of the reduction in shrinkage relative to the posterior mean in (2). Under model (3), use of a Lagrangian multiplier to enforce equality of the second moments while minimizing MSEP yields a constrained Bayes alternative to the posterior mean in (5):

$$\tilde{\beta}_{i,LX} = \gamma_{i1} + \gamma_{i2} \hat{\alpha}_{i,ols} + \gamma_{i3} \hat{\beta}_{i,ols} \quad (9)$$

The coefficients in (9) are defined as

$$\gamma_{i1} = \beta(1 - \gamma_{i3}) - \alpha\gamma_{i2},$$

$$\gamma_{i2} = \pm \eta_i [\sigma_2^2 / \{v_{\beta i} + \eta_i(2c_{\alpha\beta i} + \eta_i v_{\alpha i})\}]^{1/2},$$

and

$$\gamma_{i3} = \gamma_{i2} / \eta_i,$$

where

$$\eta_i = (v_{\beta i} \sigma_{12} - \sigma_2^2 c_{\alpha\beta i})(v_{\alpha i} \sigma_2^2 - \sigma_{12} c_{\alpha\beta i})^{-1}.$$

The \pm sign in front of γ_{i2} is needed because there are two roots, although the positive root is

usually correct. The positive or negative root is taken for γ_{i2} depending on which yields the lower value of the MSEP criterion:

$$\begin{aligned} \text{MSEP} &= E(\tilde{\beta}_i - \beta_i)^2 \\ &= (\gamma_{i2}^2 v_{\alpha i} + \gamma_{i3}^2 v_{\beta i} + 2\gamma_{i2}\gamma_{i3}c_{\alpha\beta i}) - 2(\gamma_{i2}\sigma_{12} + \gamma_{i3}\sigma_2^2) + \sigma_2^2 \end{aligned} \quad (10)$$

The definitions of η_i and γ_{i2} serve to correct a subtle error in the result originally put forth by Lyles and Xu (1999). The Appendix provides analogous constrained Bayes predictors for α_i and Y_{it}^* , which are both new to the literature. Empirical constrained Bayes (ECB) predictions are obtained for practical use by replacing unknown parameters by their estimates in equations (8), (9), (A1), and (A3), and when calculating the MSEP criterion in (10).

In contrast to the preceding direct model-specific CB predictors, consider the general CB paradigm provided by Ghosh (1992). Using β_i under model (3) to illustrate, $\tilde{\beta}_{i,B}$ is first taken to indicate the posterior mean (or Bayes) predictor for subject i . An algebraic expression for $\tilde{\beta}_{i,B}$ was given in (5). Ghosh's approach defines the CB estimate ($\tilde{\beta}_{i,G}$) as follows:

$$\tilde{\beta}_{i,G} = w\tilde{\beta}_{i,B} + (1-w)\bar{\beta}_B \quad (11)$$

where

$$\bar{\beta}_B = k^{-1} \sum_{h=1}^k \tilde{\beta}_{h,B}, \quad w = (1 + H_1/H_2)^{1/2},$$

$$H_2 = \sum_{h=1}^k (\tilde{\beta}_{h,B} - \bar{\beta}_B)^2,$$

and

$$H_1 = \text{tr}\{\text{Var}(\beta - \bar{\beta}\mathbf{1}_k | \mathbf{Y})\} = (1 - k^{-1}) \sum_{h=1}^k \text{Var}(\beta_h | \mathbf{Y}_h) \quad (12)$$

with β representing the k -vector $(\beta_1, \beta_2, \dots, \beta_k)'$.

The latter equality is supplied in (12) as a result of assumed independence across experimental units for the class of mixed models under consideration here. Note that in addition to the posterior means, this paradigm requires only the corresponding posterior variances. Using the previous notation (see equation (4) and Appendix), results in:

$$\text{Var}(\beta_i | \mathbf{Y}_i) = \sigma_2^2 - \{\sigma_{12} \mathbf{1}'_{n_i} + \sigma_2^2 \mathbf{t}'_i\} \Sigma_i^{-1} \{\sigma_{12} \mathbf{1}'_{n_i} + \sigma_2^2 \mathbf{t}'_i\}', \quad (13)$$

$$\text{Var}(\alpha_i | \mathbf{Y}_i) = \sigma_1^2 - \{\sigma_1^2 \mathbf{1}'_{n_i} + \sigma_{12} \mathbf{t}'_i\} \Sigma_i^{-1} \{\sigma_1^2 \mathbf{1}'_{n_i} + \sigma_{12} \mathbf{t}'_i\}', \quad (14)$$

and

$$\begin{aligned} \text{Var}(\mathbf{Y}_{it}^* | \mathbf{Y}_i) &= \\ \text{Var}(\mathbf{Y}_{it}^*) - \{\psi_{i1} \mathbf{1}'_{n_i} + \psi_{i2} \mathbf{t}'_i\} \Sigma_i^{-1} \{\psi_{i1} \mathbf{1}'_{n_i} + \psi_{i2} \mathbf{t}'_i\}' &. \end{aligned} \quad (15)$$

ECB predictions for practical use can be obtained by replacing unknown parameters by their estimates when computing the posterior means and variances, and the building blocks for these calculations are already built into standard software for mixed linear models.

Incorporating Fixed or Time-Dependent Covariates

Consider the following extensions of models (1) and (3) to include a set of T covariates, some of which may be time-dependent:

$$Y_{ij} = \mu + b_i + \sum_{t=1}^T \theta_t c_{ijt} + e_{ij} \quad (16)$$

$$Y_{ij} = (\alpha + a_i) + (\beta + b_i) t_{ij} + \sum_{t=1}^T \theta_t c_{ijt} + e_{ij} \quad (17)$$

where c_{ijt} represents the observed value of the t^{th} covariate for subject i at time point j ($t=1, \dots, T$; $i=1, \dots, k$; $j=1, \dots, n_i$). Let $\mathbf{c}_{ij}' = (c_{ij1}, c_{ij2}, \dots, c_{ijT})$ and form the $n_i \times T$ matrix \mathbf{C}_i by stacking the row

vectors \mathbf{c}_{ij}' in order. Next, define the transformed observed data vector $\mathbf{y}_i^\bullet = \mathbf{y}_i - \mathbf{C}_i \boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_T)'$. The extension to the posterior mean formula in (2) is

$$\tilde{\mu}_i = E(\mu_i | \mathbf{Y}_i, \mathbf{C}_i) = v_i \bar{y}_i^\bullet + (1 - v_i) \mu \quad (18)$$

with μ_i and v_i defined exactly as before and $\bar{y}_i^\bullet = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}^\bullet$. In practice, predicting

$\tilde{Y}_{ij} = E(Y_{ij} | b_i, \mathbf{C}_i) = \mu_i + \mathbf{c}_{ij}' \boldsymbol{\theta}$ may be more likely. Standard mixed linear model software typically provides the EBLUP for b_i , from which EBLUPs for μ_i and \tilde{Y}_{ij} are easily obtained.

Similarly, extensions to (4) and (5) under the randomized regression model (17) are

$$\begin{aligned} \tilde{\beta}_i &= E(\beta_i | \mathbf{Y}_i, \mathbf{C}_i) \\ &= \beta + (\sigma_{12} \mathbf{1}'_{n_i} + \sigma_2^2 \mathbf{t}'_i) \Sigma_i^{-1} (\mathbf{y}_i^\bullet - \alpha \mathbf{1}_{n_i} - \beta \mathbf{t}_i) \end{aligned}$$

and

$$\begin{aligned} \tilde{\beta}_i &= E(\beta_i | \mathbf{Y}_i, \mathbf{C}_i) \\ &= \gamma_{i1} + \gamma_{i2} \hat{\alpha}_{i,ols} + \gamma_{i3} \hat{\beta}_{i,ols} \end{aligned} \quad (19)$$

where β_i , γ_{i1} , γ_{i2} , and γ_{i3} are defined as before, but with $\hat{\alpha}_{i,ols}$ and $\hat{\beta}_{i,ols}$ now representing the

OLS intercept and slope from regressing \mathbf{y}_i^\bullet on \mathbf{t}_i . The algebraic expression in (19) requires $n_i \geq 2$. Standard software typically provides EBLUPs for a_i and b_i , from which EBLUPs for α_i and β_i follow directly. In turn, the analogue to equation (7) becomes

$$\begin{aligned} \tilde{Y}_{it}^* &= E(\mathbf{Y}_{it}^* | \mathbf{Y}_i, \mathbf{C}_i) \\ &= \tilde{\alpha}_i + \tilde{\beta}_i t_{it}^* + \mathbf{c}_{i,t} \boldsymbol{\theta} \end{aligned} \quad (20)$$

which can arguably be defined only for non-time-dependent covariates unless the values of

any time dependent ones are known at time t_i^* (as indicated by the notation \mathbf{c}_{i,t^*}).

Extensions of the CB predictors $\tilde{\mu}_{i,LX}$, $\tilde{\beta}_{i,LX}$, and $\tilde{\alpha}_{i,LX}$ in equations (8), (9), and (A1) with covariate adjustment according to models (16) and (17) require no changes to the coefficients already given, once the transformation $\mathbf{y}_i^* = \mathbf{y}_i - \mathbf{C}_i\boldsymbol{\theta}$ is made. The same is true for $\tilde{Y}_{it,LX}^*$ in equation (A3), except the term $\mathbf{c}_{i,t^*}\boldsymbol{\theta}$ is added as in (20). ECB predictions for practical use follow, once estimates of the mixed linear model parameters are inserted.

In adapting the paradigm of Ghosh (1992) as in (11) and (12), ECB predictions appear straightforward for a broad class of general linear mixed models because (i) EBLUPs accounting for covariates come directly out of standard software, and (ii) the required conditional variances [e.g., (13)-(15)] are unchanged by the addition of covariates. In the case of \tilde{Y}_{it}^* , Ghosh's paradigm requires a separate application of posterior mean and variance calculations analogous to those in (11) and (12) for each unique value of t_i^* (Moore, 2006).

Example

Consider longitudinal data on CD4 cell counts collected for the Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted (P²C²) HIV Infection Study (The P²C² Study Group, 1996). This National Heart, Lung, and Blood Institute-funded study enrolled infants born to HIV-positive women during the years 1990-1993, and followed them prospectively during the first few years of life. Specifically, data was analyzed on 59 vertically infected infants who contributed a total of 539 CD4 counts over time, with the number of measurements per child ranging from 3 to 19. Initial CD4 counts were typically observed at or within a few weeks of birth. The length of follow-up on children ranged from 1 to 6 years, with a median of 3.5 years. Also recorded for

each child was the age at which he or she was determined to have reached Class A (mildly symptomatic) HIV status (Centers for Disease Control and Prevention, 1994). Across the 59 subjects, this age ranged from 0.4 to 16 months.

A mixed linear model was fit to these data, with age as the longitudinal metameter. While there was some indication of right skewness in the CD4 counts, standard transformations tended to overcorrect this and for the sake of clarity the untransformed CD4 counts were analyzed. For an illustration with covariate adjustment, the child's gender (1 for male, 0 for female) and the concurrent CD8 cell count were accounted for via the following model:

$$CD4_{ij} = (\alpha + a_i) + (\beta + b_i)AGE_{ij} + \theta_1 GENDER_i + \theta_2 CD8_{ij} + e_{ij} \quad (21)$$

The primary objective was to compare EBLUP and ECB predictions of the random intercepts ($\alpha_i = \alpha + a_i$) and random slopes ($\beta_i = \beta + b_i$). For this purpose, both the direct ECB approach patterned after Lyles and Xu (1999; 'LX ECB') and the general ECB method following Ghosh (1992) were investigated.

Next, EBLUP and Ghosh ECB predictions of Y_{it}^* were compared, where $Y_{it}^* = \alpha_i + \beta_i t_i^* + \theta_1 GENDER_i + \theta_2 CD8_i$ represents the unknown model-based CD4 count at time t_i^* . For this latter purpose, t_i^* was defined as the age at which the child was diagnosed with Class A HIV disease, and model (21) was re-fit with the initial CD8 count ($CD8_i$) in place of the time-dependent version in light of the fact that CD8 was unrecorded at the times t_i^* . Table 1 provides the coefficient and variance component estimates from fitting both versions of model (21) by maximum likelihood via SAS PROC MIXED (SAS Institute, Inc., 2004a). The table indicates a highly significant average decline of approximately 400 CD4 cells per year, little effect of gender, and a significant positive association with the CD8 count, regardless of

whether the latter was measured only initially or treated as time-dependent.

In Figure 1A, EBLUPs are plotted for the random intercepts α_i against the corresponding Ghosh ECB predictions, based on the model treating CD8 as time-dependent. The EBLUPs were obtained directly from the mixed linear model software, and the Ghosh ECBs were computed readily using the EBLUPs and posterior variance calculations with variance components replaced by their MLEs (see e.g., eqns. 11-15). The reduction in shrinkage afforded by the CB method is evidenced by the characteristic tilting in the pattern of plotted points.

Figure 1B plots the LX ECB predictions of α_i versus the Ghosh ECBs. To obtain the LX ECBs, the MLEs for variance components were inserted into the formulae provided herein, with covariate adjustment as described in Section 2.3. With a few exceptions, the two approaches produce essentially identical results. The sample means of the 59 EBLUP, Ghosh ECB, and LX ECB predicted values were 1675.5, 1675.5, and 1675.3, respectively. The corresponding sample variances were 365470, 475026, and 473752. Comparing these to $\hat{\alpha} = 1675.5$ and $\hat{\sigma}_1^2 = 468832$ (Table 1) highlights the moment matching characteristics of the CB approaches, as well as the overshrinkage of the EBLUP.

Figure 2 is the counterpart to Figure 1, for the predicted random slopes (β_i). The tilting remains prominent in Figure 2A, while Figure 2B reveals somewhat more pronounced discrepancies between the Ghosh and LX ECB point predictions than in the case of the intercepts. The sample means of the EBLUP, Ghosh ECB, and LX ECB predicted values were -388.2, -388.2, and -395.3, respectively, with sample variances of 27904, 48316, and 49401. Comparing these to $\hat{\beta} = -388.2$ and $\hat{\sigma}_2^2 = 47843$ (Table 1) again highlights the ECB moment-matching properties in action.

Figure 3 illustrates the reduction in shrinkage of the Ghosh ECB predictions (open circles) of CD4 cell counts at the time of Class A disease (Y_{it}^*), relative to the EBLUPs (closed circles). Separate plots are presented for females

and males, with overlays of the population average regression lines calculated at the overall mean of the 59 initial CD8 counts (1294.7 cells). The lines provide a relevant visual reference based on the fit of model (21) (Table 1), although the plotted points were not expected to directly follow these linear trends given that subjects with less rapidly declining CD4 counts theoretically reach Class A disease at later ages.

Results

While the close agreement of the sample means and variances of the ECB predictions to the corresponding estimated moments ($\hat{\alpha}$ and $\hat{\sigma}_1^2$, $\hat{\beta}$ and $\hat{\sigma}_2^2$) in the real-data example is indicative, simulation studies are required to further assess the quality of the variance match and to compare the performances of the Ghosh and LX ECB methods in practical settings. Several combinations of covariates and true parameter values were examined and qualitatively similar results were found. In the interest of brevity and relevance to the application presented in the previous section, simulations designed to mimic the conditions observed in the example are summarized. Simulations were carried out using matrix manipulations and standard random number generating functions available in the SAS IML package (SAS Institute, Inc., 2004b).

Performance comparison: LX vs. Ghosh CB predictors

Data was generated according to model (21) for 20,000 hypothetical subjects, with true parameter values equal to the estimates listed in the top half of Table 1. The fabricated CD4 data were unbalanced with n_i ranging randomly between 2 and 10, and measurements were unequally timed over approximate 2 month intervals. Simulated subjects were male or female with probability 0.5. For simplicity, time-varying CD8 counts were generated at each visit from a normal distribution mimicking the sample mean and variance of the initial CD8 counts in the actual example. To illustrate results for predicting Y_{it}^* , the same simulation exercise was repeated except with a time independent

BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

Table 1: Summary of mixed linear models fit to CD4 cell count data *

Model †	Coefficient	Estimate (standard error)	Variance Component	Estimate
CD8 as time- dependent	α	1675.50 (138.27)	σ_1^2	468832
	β	-388.17 (38.06)	σ_2^2	47843
	θ_1	-163.41 (146.61)	σ_{12}	-103226
	θ_2	0.26 (0.03)	σ^2	477810
CD8 as time- independent (initial value)	α	1735.88 (188.60)	σ_1^2	429957
	β	-417.51 (40.57)	σ_2^2	55206
	θ_1	-105.28 (146.61)	σ_{12}	-102537
	θ_2	0.27 (0.10)	σ^2	529062

* Data from P²C² HIV Infection Study (The P²C² Study Group, 1996)

† $CD4_{ij} = (\alpha + a_i) + (\beta + b_i) AGE_{ij} + \theta_1 GENDER_i + \theta_2 CD8 + e_{ij}$

Figure 1: EBLUP (panel A) and LX ECB (panel B) vs. Ghosh ECB predictions for random intercepts (α_i) based on the fit of model (21) with CD8 count as time-dependent

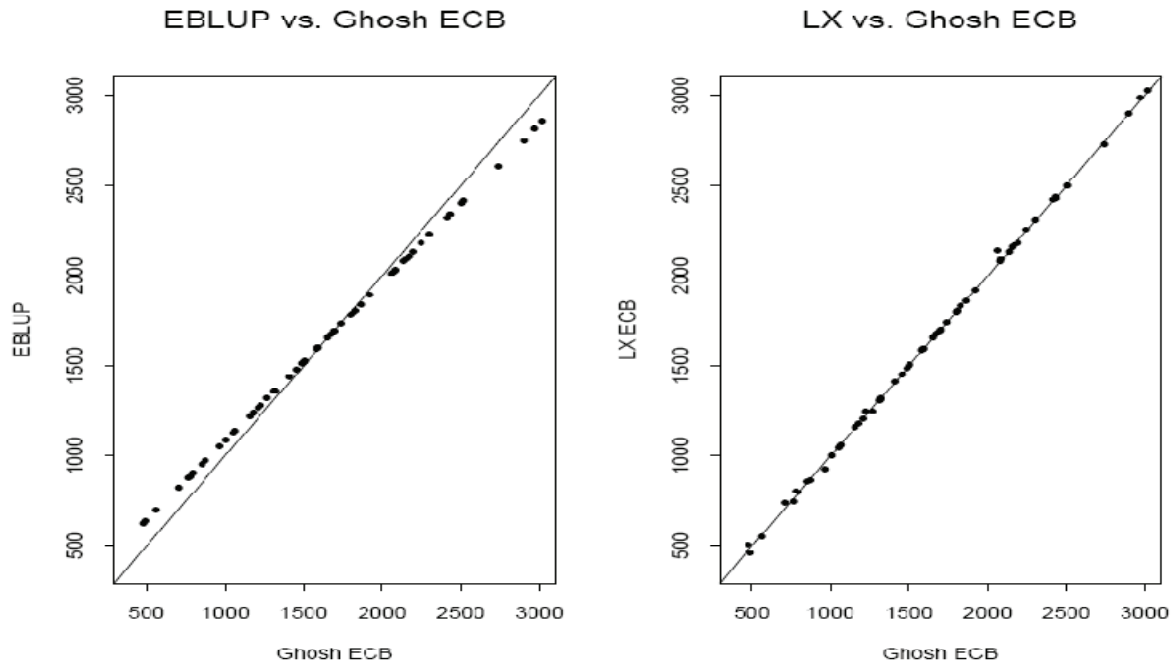


Figure 2: EBLUP (panel A) and LX ECB (panel B) vs. Ghosh ECB Predictions for Random Slopes (β_i) Based on the Fit of Model (21) with CD8 Count as Time-Dependent

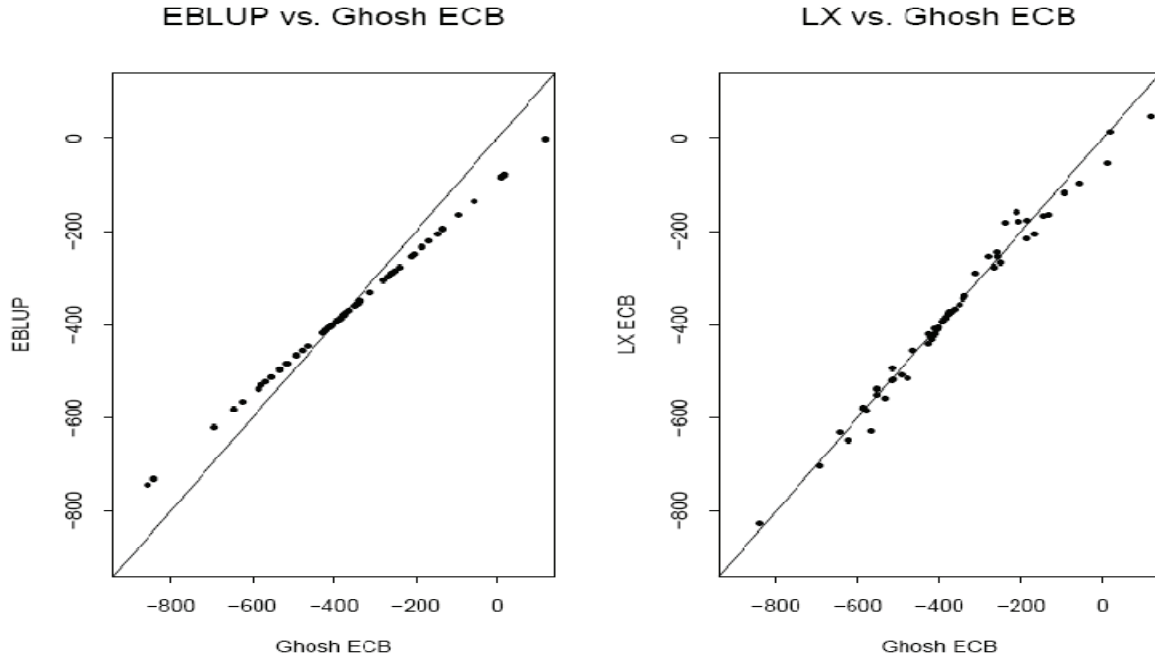
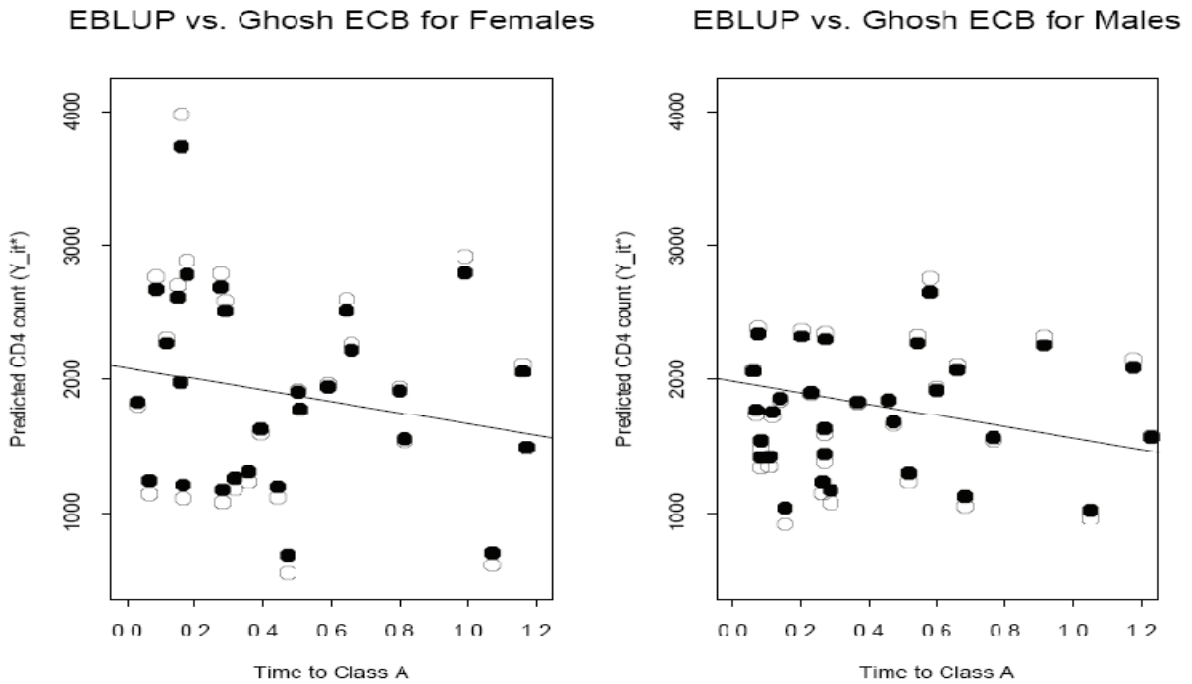


Figure 3: EBLUP (dark circle) vs. Ghosh ECB (open circle) Predictions of $Y_{it}^* = \alpha_i + \beta_i t_i^* + \theta_1 \text{GENDER}_i + \theta_2 \text{CD8}_i$ for Females (panel A) and Males (panel B), with Initial CD8 Count as a Time-Independent Covariate



BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

initial CD8 count in place of the time-varying version. The time point of interest (t_i^*) was taken to occur at 2 years for each simulated subject.

Table 2 summarizes the simulation results for predicting the α_i 's and β_i 's, and Table 3 summarizes the results for predicting Y_{it}^* . In each case, the sample means of the BLUPs and the two CB predictors closely match the true mean of the random variable being predicted.

The sample variances over 20,000 simulated subjects for both the LX and Ghosh CB methods are very close to the corresponding true variances in each case, while the overshrinkage of the BLUPs is evident by their notably tighter sampling distributions. As a final note, the empirical prediction MSEs of the LX and Ghosh methods are similar, though predictably somewhat larger than those for the corresponding BLUPs. In each case, the Ghosh method achieved a small MSE advantage relative to the LX approach.

Table 2: Simulation Results for Random Intercept and Slope Predictions^{*†}

	True α_i 's	$\tilde{\alpha}_{i, \text{BLUP}}$	$\tilde{\alpha}_{i, \text{LX}}$	$\tilde{\alpha}_{i, \text{G}}$
Mean	1675.5	1680.8	1681.2	1680.8
Variance	468832	376252	475834	474736
Prediction MSE	--	98600	105400	104469
	True β_i 's	$\tilde{\beta}_{i, \text{BLUP}}$	$\tilde{\beta}_{i, \text{LX}}$	$\tilde{\beta}_{i, \text{G}}$
Mean	-388.2	-389.4	-386.2	-389.4
Variance	47843	16115	48693	48134
Prediction MSE	--	31593	40375	40125

*Data simulated to mimic model (21) with parameters equal to estimates in Table 1 (top)

†Predictions computed assuming parameter values that generated the data

Table 3: Simulation Results for Y_{it}^* Predictions^{*†}

	True Y_{it}^* 's	$\tilde{Y}_{it, \text{BLUP}}^*$	$\tilde{Y}_{it, \text{LX}}^*$	$\tilde{Y}_{it, \text{G}}^*$
Mean	1156.4	1158.4	1158.3	1158.4
Variance	289054	177184	289249	288880
Prediction MSE	--	110636	128884	124112

*Data simulated to mimic model (21) with parameters equal to estimates in Table 1 (bottom)

†Predictions computed assuming parameter values that generated the data

Flexibility of Ghosh’s Approach under More General Covariance Structures

The LX approach, while presentable in closed form for the models considered thus far, relies upon a strict form for candidate predictors and may be cumbersome or infeasible to extend to arbitrary mixed linear models. For example, consider an extension of model (17) to incorporate serially correlated random errors, e.g., via an AR(1) structure. Rather than $\sigma^2 \mathbf{I}_{n_i}$, the covariance matrix of the i^{th} vector of random errors (\mathbf{e}_i) now takes the form

$$\text{Var}(\mathbf{e}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ & 1 & \rho & \dots & \rho^{n_i-2} \\ & & 1 & \rho & \cdot \\ & & & \cdot & \cdot \\ & & & & 1 \end{pmatrix} = \sigma^2 \mathbf{P}_{\text{AR}(1)}$$

The structured error covariance makes it less reasonable to restrict to the class of predictors that are linear combinations of $\hat{\alpha}_{i,\text{ols}}$ and $\hat{\beta}_{i,\text{ols}}$ [see eqn. (5)] in order to develop a CB predictor via the LX approach. Further, the MSEP becomes a much more difficult objective function to work with analytically.

Fortunately, the general paradigm of Ghosh (1992) encounters no difficulty with such an extension. In particular, the EBLUP remains available via common mixed linear model software, and the MVN theory-based posterior variance remains straightforward, with the only adjustment necessary to equations (13) and (14) being that the matrix $\Sigma_i = \text{Var}(\mathbf{Y}_i) = \mathbf{Z}_i \Delta \mathbf{Z}'_i + \sigma^2 \mathbf{I}_{n_i}$ becomes

$$\Sigma_i = \mathbf{Z}_i \Delta \mathbf{Z}'_i + \sigma^2 \mathbf{P}_{\text{AR}(1)}.$$

Table 4 displays the results of an additional simulation under the AR(1) error model. Data were generated under model (21) using the same true parameter values as for the simulation summarized in the top half of Table I, except with an AR(1) error structure for the covariance matrix of the random errors. The value $\rho=0.30$ was assumed. There were 5,000

simulated subjects, each with $n_i=8$ observations. The model was fit via SAS PROC MIXED and the ECB versions of $\tilde{\alpha}_{i,G}$ and $\tilde{\beta}_{i,G}$ were computed as in (11) and (12), by incorporating the EBLUPs produced by the software together with the estimated posterior variances as in (13) and (14).

As Table 4 shows, excellent matches were achieved between the sample means and variances of the ECB predictions, and the corresponding estimated population moments $(\alpha, \beta, \sigma_1^2, \sigma_2^2)$. Figure 4 displays histograms of the ECBs, which almost perfectly match the overlaid estimated theoretical normal distributions. In contrast, histograms of the EBLUPs (not shown) are characterized by markedly narrow spread as expected, thus dramatically failing to match the underlying theoretical distribution. Potential drawbacks of this overshrinkage in certain applications have been discussed at length in the literature (e.g., Louis, 1984; Ghosh, 1992; Shen & Louis, 1998; Stern & Cressie, 1999). The current example further highlights the flexibility of the Ghosh paradigm as a general approach to ECB prediction under the mixed linear model.

Conclusion

Louis (1984) and Ghosh (1992) discussed the motivation and potential benefits of constrained Bayes estimation, which seeks to optimize a traditional MSE criterion subject to matching the posterior expectation of the first two moments of a parameter distribution to the corresponding true moments. In particular, the known overall MSE advantage of the traditional posterior mean approach (which underlies the BLUP in the mixed linear model setting) is sometimes worth sacrificing to obtain a set of predictions with a histogram more closely matching a true distribution of random effects. For specific discussions of contexts in which constrained Bayes and related approaches offer tangible appeal, see Shen and Louis (1998), Lyles and Xu (1999), Stern and Cressie (1999), and Lyles, et al. (2007).

BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

Table 4: Simulation Results for Random Intercept and Slope Predictions Under AR(1) Error Model^{*†}

Parameter Estimates [‡]	ECB Sample Moments	$\tilde{\alpha}_{i,G}$	$\tilde{\beta}_{i,G}$
$\hat{\alpha} = 1683.04$ $\hat{\sigma}_1^2 = 481899$	Mean	1683.04	-389.21
$\hat{\beta} = -389.21$ $\hat{\sigma}_2^2 = 53551$	Variance	481961	53556

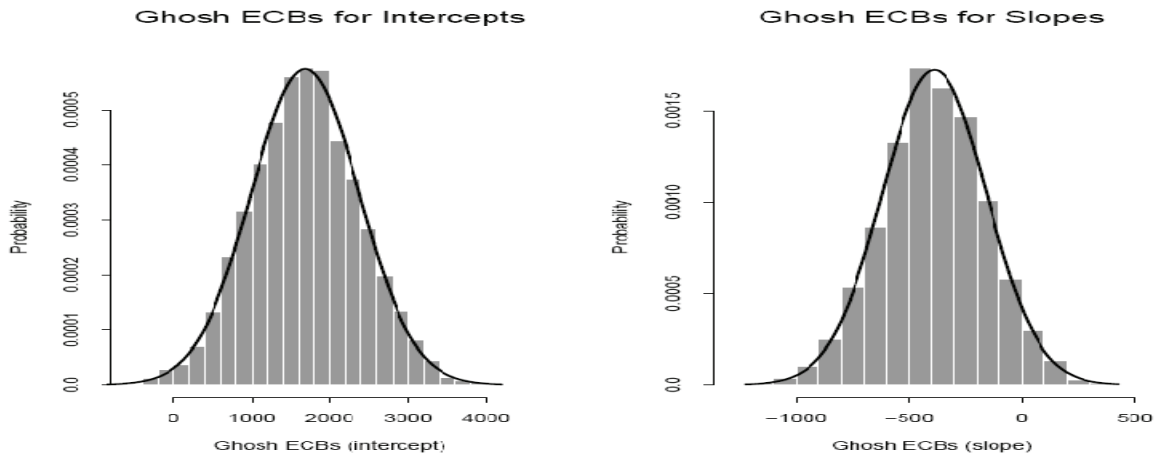
*Data simulated to mimic model (21) with $k=5000$, $n_i=8$ ($\forall i$), true parameters set equal to estimates in Table 1 (top), and $\rho=0.30$

†Ghosh ECB predictions computed by inserting MLEs of parameters

‡MLEs; Other parameter estimates: $\hat{\theta}_1 = -172.30$, $\hat{\theta}_2 = 0.23$, $\hat{\sigma}_{12} = -112073$,

$\hat{\sigma}^2 = 510618$, $\hat{\rho} = 0.29$

Figure 4: ECB Histograms Using Simulated Data from AR(1) Model (Table IV)



The purpose of this article has been to outline and compare in detail the application of a direct (LX) CB approach considered by Lyles and Xu (1999) for certain mixed linear models, as opposed to the general method of Ghosh (1992). Both approaches were explored in the presence of covariates (possibly time-dependent), and it was concluded based on simulations and a real-data example that both may be effectively applied to achieve the moment-matching goals of the CB paradigm.

The LX approach, while presentable in closed form for the models considered herein, relies upon a strict form for candidate predictors and may not be straightforward to extend to arbitrary mixed linear models. However, as highlighted previously, the general method of Ghosh (1992) appears remarkably flexible and consistent in its application. In practice, it requires only EBLUPs and estimates of the posterior variances of the random effects being predicted, with the latter readily obtainable under normal-theory mixed models. It thus seems natural to compare the performance of the Ghosh method versus the LX approach in mixed model settings where the latter is available. The simulation studies summarized (and others, unreported) consistently show the Ghosh approach to be as effective as the direct LX method at matching moments, and also suggest slight prediction MSE gains via its use for unbalanced data.

Because the primary aim was to serve as proponents of the ECB approach under the mixed linear model, the results of the current study are encouraging. The CB paradigm of Ghosh (1992) relies on building blocks that are available in commercial software for mixed linear models (e.g., SAS PROC MIXED and similar procedures in other packages such as Splus, R, SPSS, STATA or BMDP). It was shown that it performs well relative to a direct, but far less flexible, CB approach developed expressly for mixed linear models. Although further assessments will be necessary, it is hoped that these results will encourage software developers to consider the possible inclusion of options to produce the Ghosh ECB predictions in future releases. This software advance would be welcome, for the purpose of allowing practitioners the freedom to select a validated

alternative to the traditional EBLUP when overshrinkage could run counter to the objective at hand.

Acknowledgements

R. H. L. and A. K. M. were supported in part by an R01 from the National Institute of Environmental Health Sciences (ES012458). We thank the P²C²HIV study investigators and the National Heart, Lung and Blood Institute for use of their database, and appreciate support provided by the Biostatistics Core of the Emory Center for AIDS Research (P30 AI050409).

References

- Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. NY: Oxford University Press.
- Efron, B., & Morris, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators, part I: the Bayes case. *Journal of the American Statistical Association*, 66, 807-815.
- Centers for Disease Control and Prevention. (1994). 1994 revised classification for human immunodeficiency virus infection in children less than 13 years of age. *Morbidity and Mortality Weekly Report*, 43, 1-10.
- Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, 533-539.
- Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.
- Littell, R. C., et al. (2006). *SAS for mixed models, 2nd Edition*. Cary, NC: SAS Institute, Inc.
- Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 393-398.
- Lyles, R. H. & Xu, J. (1999). Classifying individuals based on predictors of random effects. *Statistics in Medicine*, 18, 35-52.
- Lyles, R. H., Manatunga, A. K., Moore, R. H., Bowman, F. D., & Cook, C. (2007). Improving point predictions of random effects for subjects at high risk. *Statistics in Medicine*, 26, 1285-1300.

BAYES PREDICTIONS OF RANDOM INTERCEPTS AND SLOPES

Moore, R.H., 2006. Prediction of random effects when data are subject to a detection limit. Unpublished Ph.D. dissertation, Department of Biostatistics, Emory University.

P²C² HIV Study Group. (1996). The pediatric pulmonary and cardiovascular complications of vertically transmitted human immunodeficiency virus (P²C² HIV) infection study: Design and methods. *Journal of Clinical Epidemiology*, 49, 1285-1294.

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

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

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). *Variance components*. NY: Wiley.

Shen, W., & Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society Series B*, 60, 455-471.

Stern, H.S., & Cressie, N. (1999). Inference for extremes in disease mapping. In: A. Lawson, et al., (Eds). *Disease mapping and risk assessment for public health*, 63-84. Chichester: Wiley.

Appendix

A constrained Bayes predictor for the i^{th} subject's random intercept (α_i) may be obtained via calculations similar to those leading to $\tilde{\beta}_{i,LX}$ in equation (8), as follows:

$$\tilde{\alpha}_{i,LX} = \tau_{i1} + \tau_{i2}\hat{\alpha}_{i,ols} + \tau_{i3}\hat{\beta}_{i,ols}, \quad (A1)$$

where

$$\tau_{i1} = \alpha (1 - \tau_{i2}) - \beta \tau_{i3},$$

$$\tau_{i2} = \pm [\sigma_1^2 / \{v_{\alpha i} + \kappa_i (2c_{\alpha\beta i} + \kappa_i v_{\beta i})\}]^{1/2},$$

and

$$\tau_{i3} = \kappa_i \tau_{i2},$$

with

$$\kappa_i = (v_{\alpha i} \sigma_{12} - \sigma_1^2 c_{\alpha\beta i}) (v_{\beta i} \sigma_1^2 - \sigma_{12} c_{\alpha\beta i})^{-1}.$$

Specifically, $\tilde{\alpha}_i$ defined in this way minimizes MSEP among predictors of the form (A1)

subject to the constraints that $E(\tilde{\alpha}_i) = E(\alpha_i) = \alpha$ and $\text{Var}(\tilde{\alpha}_i) = \text{Var}(\alpha_i) = \sigma_1^2$, where the MSEP criterion is

$$E(\tilde{\alpha}_i - \alpha_i)^2 = (\tau_{i2}^2 v_{\alpha i} + \tau_{i3}^2 v_{\beta i} + 2\tau_{i2}\tau_{i3}c_{\alpha\beta i}) - 2(\tau_{i2}\sigma_1^2 + \tau_{i3}\sigma_{12}) + \sigma_1^2 \quad (A2)$$

In an analogous manner, constrained Bayes predictor for Y_{it}^* is defined as

$$\tilde{Y}_{it,LX}^* = \phi_{i1} + \phi_{i2}\hat{\alpha}_{i,ols} + \phi_{i3}\hat{\beta}_{i,ols}, \quad (A3)$$

where

$$\phi_{i1} = \alpha (1 - \phi_{i2}) - \beta (\phi_{i3} - t_i^*),$$

$$\phi_{i2} = \pm [\psi_{i3} / \{v_{\alpha i} + \omega_i (2c_{\alpha\beta i} + \omega_i v_{\beta i})\}]^{1/2},$$

and

$$\phi_{i3} = \omega_i \phi_{i2},$$

with

$$\omega_i = (v_{\alpha i} \psi_{i2} - \psi_{i1} c_{\alpha\beta i}) (v_{\beta i} \psi_{i1} - \psi_{i2} c_{\alpha\beta i})^{-1},$$

$$\psi_{i1} = \sigma_1^2 + t_i^* \sigma_{12}, \quad \psi_{i2} = \sigma_{12} + t_i^* \sigma_2^2,$$

and

$$\psi_{i3} = \sigma_1^2 + t_i^{*2} \sigma_2^2 + 2t_i^* \sigma_{12}.$$

This minimizes MSEP for predictors of the form (A3), subject to the constraints

$$E(\tilde{Y}_{it}^*) = E(Y_{it}^*) = \alpha + \beta t_i^*$$

and

$$\text{Var}(\tilde{Y}_{it}^*) = \text{Var}(Y_{it}^*) = \phi_{i2}^2 v_{\alpha i} + \phi_{i3}^2 v_{\beta i} + 2\phi_{i2}\phi_{i3}c_{\alpha\beta i}$$

As with γ_{i2} in equation (9), technically the choice of the positive or negative root to define τ_{i2} and ϕ_{i2} should be based on which minimizes the corresponding MSEP criterion. However, it has been observed that the negative roots have never applied except in the case of γ_{i2} .

Effects of Population Distribution, Sample Size and Correlation Structure on Huberty's Effect Size R

James B. Hittner
College of Charleston

Huberty's (1994) R^2 is derived by subtracting the expected value of R^2 from an adjusted R^2 , and the square root of Huberty's R^2 is Huberty's effect size R . The present study examined the effects of population distribution, sample size and population correlation structure on the statistical power of Huberty's R .

Key words: Huberty's R ; Statistical Power; Multiple Regression; Effect Size.

Introduction

In the context of multiple regression analysis, it is often standard practice to examine whether the squared multiple correlation coefficient, R^2 , is statistically significant. The intent of such a test is to determine whether R^2 differs significantly from zero, and the null hypothesis may be stated as $H_0: \rho^2 = 0$. Although this test is widely used, it is misleading because the expected value of R^2 is not zero when $\rho = 0$. Rather, as Morrison (1990) pointed out, the expected value, or expected long-run mean, of R^2 is equal to $p / N - 1$, where p is the number of predictor variables. The implication of this equation is that R^2 should be examined in relation to the expected value of R^2 , $E(R^2)$, because the latter quantity is the value of R^2 that can be expected simply by chance.

In light of this realization, it seems more appropriate for researchers to test the null hypothesis, $H_0: \rho^2 = \rho_0^2$, where $\rho_0^2 = E(R^2)$. Darlington (1990) gave an F statistic for testing this null hypothesis and Huberty (1994) presented an adjusted R^2 index that takes into account the value of $E(R^2)$. The formula for

Huberty's adjusted R^2 index is:

$$R^2_{\text{adj}} = (R^2 - E(R^2)) / (1 - E(R^2)).$$

Huberty (1994) also presented an effect size measure for multiple regression studies that is calculated by subtracting $E(R^2)$ from Huberty's adjusted R^2 index. This effect size measure seems more appropriate than either R^2 or the adjusted R^2 given that it simultaneously accounts for both shrinkage and the sample size-to-predictor ratio.

Despite the apparent appropriateness of Huberty's effect size measure, standard statistical software packages, such as SPSS and Minitab, report only R^2 and adjusted R^2 values. Furthermore, although fourteen years have passed since the article was first published, very little, if any, quantitative research has been conducted on Huberty's proposed effect size measure. Due to this omission from the statistical literature, the present study generated simulated data and examined Huberty's effect size measure under different population distributions, sample sizes, and population correlation structures.

Methodology

Random variables were generated from the following three population distributions: Normal ($\mu = 0$, $\sigma = 1$), Weibull ($\lambda = 0.5$, $k = 1.2$), and Poisson ($\mu = \lambda = 0.5$). These distributions differ

James Hittner is a Professor of psychology. His quantitative research interests include linear regression and correlation, confirmatory factor analysis, and statistical software development. Email: hittnerj@cofc.edu.

HUBERTY'S EFFECT SIZE

in shape and are representative of the types of data distributions often encountered in applied research. The Weibull distribution, for example, is commonly used to model failure characteristics such as infant mortality, random failures, product wear-out, and the breaking strength of materials; it is also appropriate for lifetime modeling/survival analyses. Although similar in form to the exponential distribution the Weibull distribution can accommodate hazard changes over time, unlike the exponential which assumes a constant hazard rate (Heo, Faith, & Allison, 1998). The Poisson distribution is a discrete distribution that is often used to model counts, such as the number of arrivals, deaths, or failures in a given time period, and it can also be used to model the number of times a random event occurs over a given distance or across a particular spatial area. Such modeling of frequency count data per unit time, distance or area is tantamount to modeling rate data.

For each of the three population distributions, four random variables were generated for three different sample sizes (N 's of 50, 100, 200) and three different population correlation structures (ρ 's of 0.15, 0.30 and 0.65, representing low, moderate and high levels of correlation, respectively). This data generation process resulted in a total of 27 sets of four random variables (i.e., 3 distributions x 3 sample sizes x 3 correlation structures). For each set of four random variables, the specified correlation structure was induced by adding a multiple of a random variable, U , from the same population distribution to each randomly generated variable ($X1$, $X2$, $X3$, and Y). For each variable set, the value of the multiplicative constant, c , was chosen to produce the desired correlation. The specific algorithm was as follows:

$$\begin{aligned}X1_{\text{new}} &= (X1 + cU) / (1 + c^2) \\X2_{\text{new}} &= (X2 + cU) / (1 + c^2) \\X3_{\text{new}} &= (X3 + cU) / (1 + c^2) \\Y_{\text{new}} &= (Y + cU) / (1 + c^2)\end{aligned}$$

In generating the new, correlated, variables the choice as to which variable constituted Y was arbitrary. For consistency, the fourth correlated variable was always designated as Y . An important point concerning this

methodology is that the algorithm produces variables that correlate, on average, at the specified level of correlation. By generating variables that demonstrate approximate rather than exact and unvarying levels of correlation, the above algorithm produces sets of correlated variables that more closely mirror real-world datasets. For example, in the case of the Weibull distribution with $N = 200$ and a population correlation structure of 0.30, the mean empirical correlation for the four variables was 0.303 and the 95% confidence interval for the mean r ranged from 0.272 to 0.334. All of the variables in the present study were generated using the Statistical Package for the Social Sciences (SPSS, version 14).

For each of the 27 simulated datasets, a simultaneous multiple regression analysis was conducted (using SPSS) whereby Y was regressed onto the three predictor variables ($X1$, $X2$ and $X3$). The resulting R^2 value, along with the sample size, N , and the number of predictors, p , was then entered into a SAS data step program to calculate the expected value of R^2 , Huberty's adjusted R^2 index, and Huberty's effect size measure (the SAS data step program is available from the author upon request). For each of the 27 datasets, the square root of Huberty's R^2 effect size measure - hereafter referred to as Huberty's effect size R - was examined to determine whether, given a specified sample size (50, 100, 200), number of predictors (3), level of statistical power (0.80) and alpha level (0.05), the value of R would be large enough to attain statistical significance at $p \leq 0.05$. The relevant power calculations were carried out using a FORTRAN program written by Dunlap, Xin, and Myers (2004). This program calculates power using the random, or unconditional, approach recommended by Gatsonis and Sampson (1989). Monte Carlo simulation results reported by Dunlap, et al. (2004) indicate that the random approach is more accurate than the more commonly used fixed approach. For each generated dataset, Huberty's effect size R was evaluated against the minimally detectable population R given the specified sample size, power = 0.80, alpha = 0.05, and $p = 3$ predictors. Based on Dunlap et al.'s power program, the minimally detectable population R values under these conditions for

N 's of 200, 100, and 50 are 0.231, 0.323, and 0.448, respectively. Considering the above R values as comparative benchmarks, the objective of this study was to examine the effects of population distribution, sample size, and population correlation structure on the power of Huberty's effect size R , where power is defined as being adequate (≥ 0.80) when Huberty's R exceeds the minimally detectable population R .

Results

For all cases with a correlation structure of $\rho = 0.65$, Huberty's effect size R estimates exceeded the minimally detectable population R , thereby demonstrating adequate levels of statistical power. For cases with a correlation structure of $\rho = 0.30$, six of the nine Huberty R estimates demonstrated adequate power, two demonstrated inadequate power, and one could not be calculated. The two underpowered cases were the Weibull distribution at $N = 100$ and the Poisson distribution at $N = 100$. The incalculable estimate was for the Weibull distribution at $N = 50$. Huberty's effect size R could not be computed for this case because the value of Huberty's adjusted R^2 index (0.030) was less than the expected value of R^2 (0.061). The difference between these two values equals Huberty's effect size measure, R^2 , which in this case amounted to -0.031 (i.e., $0.030 - 0.061$). Because the square root of a negative number cannot be computed, the value of Huberty's effect size R for this case is incalculable. For cases with a correlation structure of $\rho = 0.15$, six of the nine Huberty R estimates were underpowered and the remaining three could not be calculated (for the same reasons as noted above). The Huberty effect size R estimates and other relevant data for each case examined in this study are presented in Table 1.

One finding of interest concerns the two underpowered cases with a correlation structure of $\rho = 0.30$ (the Weibull and Poisson distributions at $N = 100$). In an effort to explain these findings, the mean empirical correlations and the coefficients of variation (CV; standard deviation of the empirical correlations divided by the mean correlation) for the Weibull and Poisson cases were compared against the corresponding, adequately powered, Normal

distribution case. All of the mean correlation comparisons were statistically nonsignificant (all Fisher Z -tests < 0.50 , all p -values > 0.60). By contrast, all pairwise likelihood ratio tests on the CV's were statistically significant (p 's < 0.005), with the Weibull and Poisson CV's being significantly larger than the Normal distribution CV. These data suggest that, relative to the Normal case, the greater noise-to-signal ratio in the empirically generated correlations for the Weibull and Poisson cases may have contributed to their compromised levels of statistical power.

Another finding of interest was the negative value for Huberty's R^2 (and corresponding incalculable value for Huberty's effect size R) for the Weibull distribution at $N = 50$ and correlation structure of $\rho = 0.30$. Although the reason for this finding is not entirely clear, one possible explanation is that the small sample size (50) and relatively large CV (0.326) interacted with the shape (i.e., moments) of the Weibull distribution to produce an insufficiently large R^2 value. With respect to the cases with a correlation structure of $\rho = 0.15$, the fact that all six of the calculated Huberty R estimates were underpowered (three were incalculable) suggests that such a low level of intercorrelation among predictors and criterion generated a regression model that lacks adequate statistical power. It is important to note, however, that the data generation algorithm used in this study produced empirical correlations for the $\rho = 0.15$ cases that were noticeably more variable, as evidenced by the CV's, than were the correlations for the 0.30 and 0.65 cases. This heightened level of variability could have contributed to the underpowered estimates for the $\rho = 0.15$ cases. These same two factors (low level of intercorrelation, greater variability in estimated correlations), more so than sample size and distribution type, are the likely reasons underlying the incalculable Huberty R estimates.

One point worth mentioning about statistical power analysis in the context of multiple regression is that the algorithms used to compute integrals from the distribution of R^2 assume that the joint distribution of predictors and criterion is multivariate normal (Dunlap et al., 2004; Gatsonis & Sampson, 1989). When the multivariate distribution deviates from

HUBERTY'S EFFECT SIZE

Table 1: Huberty's Effect Size R and Related Statistics

Distrib	N	rstruct	ExpRsqr	HuberRsqr	HuberES	HuberESR	Meanr	Sdr	CVr
Weib	200	0.15	0.015	0.038	0.023	0.152	0.145	0.031	0.214
Weib	200	0.30	0.015	0.164	0.149	0.386	0.303	0.030	0.099
Weib	200	0.65	0.015	0.554	0.539	0.734	0.655	0.022	0.034
Poiss	200	0.15	0.015	0.033	0.018	0.134	0.139	0.048	0.345
Poiss	200	0.30	0.015	0.168	0.153	0.391	0.321	0.040	0.125
Poiss	200	0.65	0.015	0.557	0.542	0.736	0.665	0.022	0.033
Norm	200	0.15	0.015	0.023	0.008	0.089	0.156	0.064	0.410
Norm	200	0.30	0.015	0.140	0.125	0.354	0.321	0.058	0.181
Norm	200	0.65	0.015	0.519	0.504	0.710	0.659	0.034	0.052
Weib	100	0.15	0.030	0.001	-0.030		0.146	0.092	0.630
Weib	100	0.30	0.030	0.099	0.068	0.261	0.295	0.078	0.264
Weib	100	0.65	0.030	0.496	0.465	0.682	0.643	0.043	0.067
Poiss	100	0.15	0.030	0.012	-0.018		0.153	0.063	0.412
Poiss	100	0.30	0.030	0.123	0.093	0.305	0.312	0.060	0.192
Poiss	100	0.65	0.030	0.508	0.478	0.691	0.651	0.034	0.052
Norm	100	0.15	0.030	0.056	0.026	0.161	0.145	0.061	0.421
Norm	100	0.30	0.030	0.244	0.214	0.463	0.357	0.051	0.143
Norm	100	0.65	0.030	0.594	0.563	0.750	0.672	0.030	0.045
Weib	50	0.15	0.061	-0.054	-0.115		0.165	0.118	0.715
Weib	50	0.30	0.061	0.030	-0.031		0.304	0.099	0.326
Weib	50	0.65	0.061	0.491	0.430	0.656	0.668	0.043	0.064
Poiss	50	0.15	0.061	0.113	0.051	0.226	0.152	0.171	1.125
Poiss	50	0.30	0.061	0.315	0.254	0.504	0.390	0.128	0.328
Poiss	50	0.65	0.061	0.651	0.589	0.767	0.693	0.074	0.107
Norm	50	0.15	0.061	0.076	0.015	0.122	0.161	0.104	0.646
Norm	50	0.30	0.061	0.274	0.212	0.460	0.369	0.090	0.244
Norm	50	0.65	0.061	0.608	0.547	0.740	0.688	0.042	0.061

Notes: Distrib = Population distribution (Weibull, Poisson, Normal); N = Population sample size; rstruct = Population correlation structure; ExpRsqr = Expected value of R^2 ; HuberRsqr = Huberty's adjusted R^2 ; HuberES = Huberty's adjusted R^2 minus the expected value of R^2 ; HuberESR = The square root of HuberES; Meanr = Arithmetic average of empirically generated correlations (i.e., correlations among X_1 , X_2 , X_3 , and Y); Sdr = Standard deviation of empirically generated correlations; CVr = Coefficient of variation for empirically generated correlations (i.e., Sdr / Meanr). Blank entries for HuberESR indicate incalculable values (see text for details).

normality, then power estimates may become biased. However, the extent of bias is difficult to quantify and represents an important topic for future research. Another point worth noting is that the present investigation focused solely on factors affecting the power of Huberty's overall multiple regression coefficient. Factors affecting the power of individual predictors within the context of a larger regression model were not considered (for a treatment of this topic, the reader is referred to Maxwell, 2000). Though it is commonplace in multiple regression to test the partial contribution of a single predictor in the context of other predictor variables, such a practice is not without interpretive problems (Dunlap & Landis, 1998). A final point is that, in the present study, datasets with known *a priori* properties, in terms of population distribution, sample size and correlation structure, were generated and the obtained power of Huberty's effect size R was examined for each generated dataset. The present investigation was not a Monte Carlo simulation study in which the empirical properties of one or more statistical tests were examined. Such Monte Carlo work designed to investigate the power and efficiency (Type I error rate) of a significance test of Huberty's R represents an important direction for future research.

Conclusion

This study examined the power of Huberty's effect size R under three different population distributions (Weibull, Poisson, Normal), sample sizes (N 's of 50, 100, 200), and population correlation structures (ρ 's of 0.15, 0.30, and 0.65). For all conditions with a correlation structure of 0.65, Huberty's R demonstrated adequate statistical power. For cases with a correlation structure of 0.30, six of the eight estimated Huberty R values maintained adequate power (one value could not be calculated). For cases with a correlation structure of 0.15, the Huberty R values were either underpowered (six cases) or incalculable (three cases).

These results suggest that - in the context of multiple regression research - Huberty's effect size R maintains adequate statistical power under a variety of distributional

shapes, samples sizes and correlation structures. The notable exception to this rule concerns cases with a correlation structure of 0.15, in which all of the estimated Huberty R values (six of nine cases) were underpowered. Such low power estimates suggest that practitioners of multiple regression analysis should restrict their attention to variables that correlate above 0.15 if they hope to maintain adequate statistical power for Huberty's effect size R (at least for models with 3 predictors and sample sizes ≤ 200). The precise *magnitude* of correlation needed to maintain adequate power for Huberty's R under various distributional shapes and sample size conditions is a topic for future research. It is hoped that the present study fosters a greater appreciation of Huberty's R and that the findings motivate additional research into factors that influence the statistical power of Huberty's effect size R .

References

- Darlington, R. B. (1990). *Regression and linear models*. NY: McGraw-Hill.
- Dunlap, W. P., & Landis, R. S. (1998). Interpretations of multiple regression borrowed from factor analysis and canonical correlation. *Journal of General Psychology, 125*, 397-407.
- Dunlap, W. P., Xin, X., & Myers, L. (2004). Computing aspects of power for multiple regression. *Behavior Research Methods, Instruments, & Computers, 36*, 695-701.
- Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin, 106*, 516-524.
- Heo, M., Faith, M. S., & Allison, D. B. (1998). Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest. *Mechanisms of Ageing and Development, 102*, 45-53.
- Huberty, C. J. (1994). A note on interpreting an R^2 value. *Journal of Educational and Behavioral Statistics, 19*, 351-356.
- Maxwell, S.E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5*, 434-458.
- Morrison, D. F. (1990). *Multivariate statistical methods*. NY: McGraw-Hill.

A Note on Hypothesis Tests after Correction for Autocorrelation: Solace for the Cochrane-Orcutt Method?

Terry E. Dielman
Texas Christian University

The behavior of the t test in small samples for coefficient significance in time-series regressions is examined after using the Prais-Winsten (PW) and Cochrane-Orcutt (CO) corrections for autocorrelation. Results are compared to ordinary least squares and generalized least squares.

Key words: First-order autocorrelation generalized least squares, ordinary least squares, Prais-Winsten, time series regression.

Introduction

The Prais-Winsten (PW) and Cochrane-Orcutt (CO) methods are popular procedures for correcting for autocorrelation in time-series regression models. Both methods transform the data using a differencing transformation to remove autocorrelation. Ordinary least squares (OLS) applied to the transformed observations will yield estimators that are asymptotically more efficient than OLS applied to the original data.

The PW and CO methods are essentially equivalent except for the treatment of the first observation in the data set. The CO method simply omits the first observation, while the PW method transforms the observation and retains it. Asymptotically, there is no difference in the efficiency of estimators produced by the two methods. In previous studies of small sample behavior, however, the superior performance of the PW procedure has been documented. Using the CO procedure results in estimators that are less efficient in small samples. Under certain conditions, the CO estimator can even be less efficient than OLS applied to the original data.

Due to the inefficiency of the CO estimator, comparisons of hypothesis testing results from models estimated by PW and CO have not been considered. This article examines the behavior of the t test in small samples for coefficient significance in time-series regressions. Tests are compared using four estimation procedures: OLS, CO, PW and generalized least squares estimation (GLS) using the true value of the autocorrelation coefficient.

The results suggest that the PW and CO methods perform similarly when testing hypotheses, but in certain cases, CO outperforms PW. This does not, however, mean that either method performed particularly well. Both had levels of significance that were much higher than desirable in certain circumstances. The poor performance of these procedures in situations when they are intended to correct for autocorrelation suggests the need for either better estimates of the autocorrelation coefficient, better procedures for correcting for autocorrelation, or alternative approaches that will result in improved hypothesis tests.

Methodology

The following simple regression model is considered:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \text{ with } \varepsilon_t = \rho \varepsilon_{t-1} + \eta_t \quad (1)$$

for $t=1,2,\dots,T$. In equation (1), y_t and x_t are the t^{th} observations on the dependent and

Terry E. Dielman is a Professor in the M.J. Neeley School of Business, Department of Information Systems and Supply Chain Management. Email: t.dielman@tcu.edu.

explanatory variables, respectively, and ε_t is a random disturbance for the t^{th} observation and may be subject to autocorrelation. The η_t represents disturbance components that are assumed to be independent and identically distributed. The parameters β_0 and β_1 are unknown and must be estimated. The parameter ρ is the autocorrelation coefficient, with $|\rho| < 1$. Using matrix notation, the model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_T \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_T \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_T \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad (3)$$

Two procedures to correct for autocorrelation are examined. These are the Prais-Winsten (1954) and Cochrane-Orcutt (1949) procedures. Both procedures transform the data using the autocorrelation coefficient, ρ , after which the transformed data are used in estimation. The procedures differ in their treatment of the first observation, (x_1, y_1) . The PW transformation matrix is:

$$\mathbf{M}_{PW} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ -\rho & 1 & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & -\rho & 1 \end{bmatrix} \quad (4)$$

Pre-multiplying the model in (2) by \mathbf{M}_{PW} yields

$$\mathbf{M}_{PW}\mathbf{Y} = \mathbf{M}_{PW}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}_{PW}\boldsymbol{\varepsilon} \quad (5)$$

or

$$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\eta} \quad (6)$$

where \mathbf{Y}^* contains the transformed dependent variable values and \mathbf{X}^* is the matrix of transformed independent variable values, so

$$\mathbf{Y}^* = \left[\sqrt{1-\rho^2} y_1, y_2 - \rho y_1, \dots, y_T - \rho y_{T-1} \right] \quad (7)$$

and

$$\mathbf{X}^* = \begin{bmatrix} \sqrt{1-\rho^2} & \sqrt{1-\rho^2} x_1 \\ 1-\rho & x_2 - \rho x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1-\rho & x_T - \rho x_{T-1} \end{bmatrix} \quad (8)$$

In (6), $\boldsymbol{\eta}$ is the vector of serially uncorrelated η_t errors.

The CO transformation matrix is the $(T-1) \times 1$ matrix obtained by removing the first row of the \mathbf{M}_{PW} transformation matrix. The use of the CO transformation means that $(T-1)$ observations, rather than T , are used to estimate the model. In the CO transformation, the first observation is omitted, whereas it is transformed and included in the estimation in the PW transformation. Asymptotically, the loss of this single observation is probably of minimal concern. However, for small samples, omitting the first observation has been shown to result in an estimator inferior to that obtained when the first observation is retained and transformed. See Dielman & Pfaffenberger (1984), Maeshiro (1979), and Park & Mitchell (1980) for simulation studies demonstrating the efficiency gains of PW, and Doran (1981), Magee (1987), Taylor (1981), and Thornton (1987) for analytical results.

In practice the value of ρ will be unknown and it must be estimated from sample data. The estimators of ρ used will be as follows:

$$\hat{\rho}_{PW} = \frac{\sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=2}^T \hat{\varepsilon}_t^2} \quad (9)$$

HYPOTHESIS TESTS AFTER CORRECTION FOR AUTOCORRELATION

when all T observations are used, and

$$\hat{\rho}_{CO} = \frac{\sum_{t=2}^T \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^{T-1} \hat{\varepsilon}_t^2} \quad (10)$$

when $T-1$ observations are used, where the $\hat{\varepsilon}_t$ represent OLS residuals. Park and Mitchell (1980) showed that these two estimators minimize the error sum of squares conditional on β when T and $T-1$ observations are used, respectively, in the estimation process.

The actual estimation procedures for both PW and CO are iterative procedures. OLS is run to obtain estimates of the regression coefficients and, subsequently, the $\hat{\varepsilon}_t$. The estimator of the autocorrelation coefficient, ρ , is computed, the data are transformed, and new estimates of the regression coefficients are obtained. The autocorrelation coefficient estimate is recomputed and compared to the previous estimate. In the results, if these estimates differ by less than 0.000001, the iterative procedure stops. The procedure also stops when it reaches 25 iterations. If boundary conditions are encountered the estimate of ρ is set at ± 0.999999 .

The model considered in this article is described in equation (1). The explanatory variable values are generated as follows:

1. $x_t = \lambda x_{t-1} + u_t$ for $t = 1, 2, \dots, T$ with the u_t chosen from the $N(0,2)$ distribution. The values of λ used were 0.0, 0.4 and 0.8.
2. A stochastic time trend is used. In this case $x_t = \lambda t + u_t$ for $t = 1, 2, \dots, T$ and u_t is chosen from the $N(0,2)$ distribution for $\lambda = 0.4$ and 0.8.

Once generated, these values are held fixed throughout the experiment for each sample size. The disturbances, η_t , are chosen from the $N(0,1)$ distribution. After generating the η_t , the ε_t values are created as $\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$ where $\varepsilon_0 = \frac{\eta_0}{1-\rho^2}$ and η_0 is an initial draw from the disturbance

distribution. The explanatory variable values were generated independently of the disturbances.

The parameter β_0 was set equal to zero (without loss of generality). The parameter β_1 was set equal to zero to examine the level of significance. For each factor combination in the experimental design, ten thousand Monte Carlo trials were used to assess levels of significance. A sample size of $T = 20$ was used. The values of ρ were 0.0, 0.2, 0.4, 0.6, 0.8, and 0.95. The null hypothesis $H_0: \beta_1 = 0$ is tested using the t test and the number of rejections of the null hypothesis was recorded to assess the level of significance.

The hypothesis tests were also conducted with $\beta_1 = 0.2, 0.3, 0.4, 0.5,$ and 1.0. When H_0 is rejected, the proportion of correct rejections can be used to construct empirical power functions. Power comparisons based on the original simulations are complicated by the differences in the observed significance levels. A valid power comparison can be made only if the true significance levels of the tests are similar, which is clearly not the case based on our results.

The power comparison was accomplished using a procedure suggested by Zhang and Boos (1994). From the original 10,000 simulations with $\beta_1 = 0$, the test statistics were sorted and the critical values producing a 5% level of significance were chosen for each design point. These values represent estimates of the critical values under the null hypothesis that produce an exact 5% level of significance. The simulation was repeated with the non-zero values of β_1 , using the empirically determined critical values. The test statistics from the second set of simulations will have similar levels of significance, making their powers comparable. Zhang and Boos (1994) suggested using a larger number of Monte Carlo trials to estimate the correct critical value under the null hypothesis if possible. In this experiment 10,000 trials under the null and 5,000 under the alternative hypotheses were used.

Results are reported for four estimation procedures: OLS (assuming $\rho = 0$), PW and CO and GLS (which is the PW procedure using the true value of ρ). All random numbers were

generated using IMSL subroutines and the simulation was written in FORTRAN.

Results

Consider Tables 1 and 2. These tables show the number of rejections of the true null hypothesis that the slope is zero for all factor combinations in the Monte Carlo simulation. Table 1 shows the results for the autoregressive independent variable; Table 2 for the stochastic trend variable. The most striking results are for the autoregressive case when λ is 0.8 and the stochastic trend case for λ equal to both 0.4 and 0.8. As the level of autocorrelation increases, the observed levels of significance become very high for OLS, but this is not unexpected. OLS is not expected to perform well when disturbances are autocorrelated.

However, the two methods that correct for autocorrelation do not perform well either. PW has very high rejection rates with some cases approaching 50%. The rejection rates for CO are high as well, but often not as high as PW. This is particularly evident when the independent variable is autoregressive. These results suggest that correcting for autocorrelation does not guarantee reliable inferences about the slope coefficient.

Selected power comparisons using 5,000 Monte Carlo trials are shown in Table 3 for the autoregressive independent variable with $\lambda = 0.8$ and in Table 4 for the stochastic trend variable with $\lambda = 0.8$. When the independent variable is autoregressive, CO generally has power equal to or slightly higher than PW. Figures 1 and 2 plot the empirical power curves for $\rho = 0.0$ and $\rho = 0.95$ from Table 3. When $\rho = 0.0$ there is little difference in adjusted power; when $\rho = 0.95$ CO has higher power than PW.

When the independent variable is a stochastic trend, there is little difference between PW and CO as evidenced in the empirical power curves in Figures 3 and 4 for $\rho = 0.0$ and $\rho = 0.95$, respectively. In this case, when $\rho = 0.95$, it is especially troublesome that OLS has higher adjusted power than either PW or CO, which supposedly adjust for autocorrelation. This result is driven by the very high levels of significance for OLS of course.

Conclusion

Previous studies have shown that the PW method is superior to CO as a correction for autocorrelation in terms of estimator efficiency. However, these results do not hold up in an examination of inference results. CO generally performs as well and in many cases better than PW in terms of observed level of significance and adjusted power. This should not be taken as a suggestion that the PW method should be abandoned and CO resorted to, however. Perhaps both methods should be abandoned and a better approach sought for handling autocorrelation in regression models. In terms of inference, a bootstrap approach as Rayner (1991) suggested might be preferred to either the PW or CO method. Alternatively, as suggested by Mizon (1995), perhaps another approach to correcting for autocorrelation should be considered. Bayesian estimators (see Ohtani, 1990, and Kennedy & Simons, 1991) also hold promise for improvements.

References

- Cochrane, D., & Orcutt, G. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32-61.
- Dielman, T., & Pfaffenberger, R. (1984). Small sample properties of estimators in the autocorrelated error model: A review and some additional simulations. *Statistical Papers/Statistische Hefte*, 30, 163-183.
- Doran, H. (1981). Omission of an observation from a regression analysis: A discussion on efficiency loss with applications. *Journal of Econometrics*, 16, 367-374.
- Kennedy, P., & Simons, D. (1991). Fighting the teflon factor: Comparing classical and Bayesian estimators for autocorrelated errors. *Journal of Econometrics*, 48, 15-27.
- Maeshiro, A. (1979). On the retention of the first observation in serial correlation adjustment of regression models. *International Economic Review*, 20, 259-265.
- Magee, L. (1987). A note on Cochrane-Orcutt estimation. *Journal of Econometrics*, 35, 211-218.

HYPOTHESIS TESTS AFTER CORRECTION FOR AUTOCORRELATION

Table 1: Empirical Significance Level: Number of Rejections of True Null Hypothesis $H_0: \beta_1 = 0$ Using Autoregressive Independent Variable (10,000 Trials)

Lambda = 0.0	Rho =					
	0.00	0.20	0.40	0.60	0.80	0.95
OLS	508	601	623	531	342	226
PW	754	757	716	637	523	462
GLS	508	500	516	514	526	530
CO	699	716	697	617	471	424
Lambda = 0.4	Rho =					
	0.00	0.20	0.40	0.60	0.80	0.95
OLS	516	712	870	852	700	491
PW	808	819	826	755	640	546
GLS	516	516	501	503	508	493
CO	744	771	764	694	566	447
Lambda = 0.8	Rho =					
	0.00	0.20	0.40	0.60	0.80	0.95
OLS	521	865	1307	1848	2658	3613
PW	806	942	1048	1232	1442	1679
GLS	521	512	512	506	505	496
CO	726	877	995	1057	949	822

Table 2: Empirical Significance Level: Number of Rejections of True Null Hypothesis $H_0: \beta_1 = 0$ Using Stochastic Trend Independent Variable (10,000 Trials)

Lambda = 0.4	Rho =					
	0.00	0.20	0.40	0.60	0.80	0.95
OLS	453	936	1685	2848	4537	6111
PW	842	1026	1291	1815	2965	4496
GLS	453	458	475	482	477	494
CO	805	1008	1307	1860	2955	4419
Lambda = 0.8	Rho =					
	0.00	0.20	0.40	0.60	0.80	0.95
OLS	467	1013	1822	3087	4812	6352
PW	831	1039	1347	1941	3181	4711
GLS	467	456	453	459	484	487
CO	819	1036	1399	2038	3288	4706

DIELMAN

Table 3: Adjusted Power Comparisons Using Autoregressive Independent Variable with
Lambda = 0.8 (5,000 trials)

Rho = 0.0	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	3006	4572	4961	4998	5000
PW	250	3107	4535	4922	4988	5000
GLS	250	3006	4572	4961	4998	5000
CO	250	3108	4550	4920	4992	5000
Rho = 0.2	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	2476	4079	4819	4976	5000
PW	250	2332	3856	4695	4944	5000
GLS	250	2361	4043	4817	4978	5000
CO	250	2457	3991	4731	4953	5000
Rho = 0.4	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	2039	3407	4429	4845	5000
PW	250	1737	3003	4175	4749	5000
GLS	250	1860	3420	4507	4907	5000
CO	250	1932	3254	4346	4795	5000
Rho = 0.6	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	1684	2711	3659	4364	5000
PW	250	1286	2313	3359	4252	5000
GLS	250	1526	2925	4120	4729	5000
CO	250	1655	2771	3749	4490	5000
Rho = 0.8	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	1478	2153	2849	3475	4906
PW	250	1089	1808	2656	3497	4967
GLS	250	1424	2771	3935	4637	5000
CO	250	1651	2430	3287	3970	4977
Rho = 0.95	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	1430	1874	2361	2852	4473
PW	250	973	1526	2235	3048	4821
GLS	250	1532	2911	4072	4686	5000
CO	250	1726	2307	3025	3626	4835

HYPOTHESIS TESTS AFTER CORRECTION FOR AUTOCORRELATION

Figure 1: Power Curve for Testing Slope Equal Zero:
Autoregressive With Lambda = 0.8; rho = 0.0

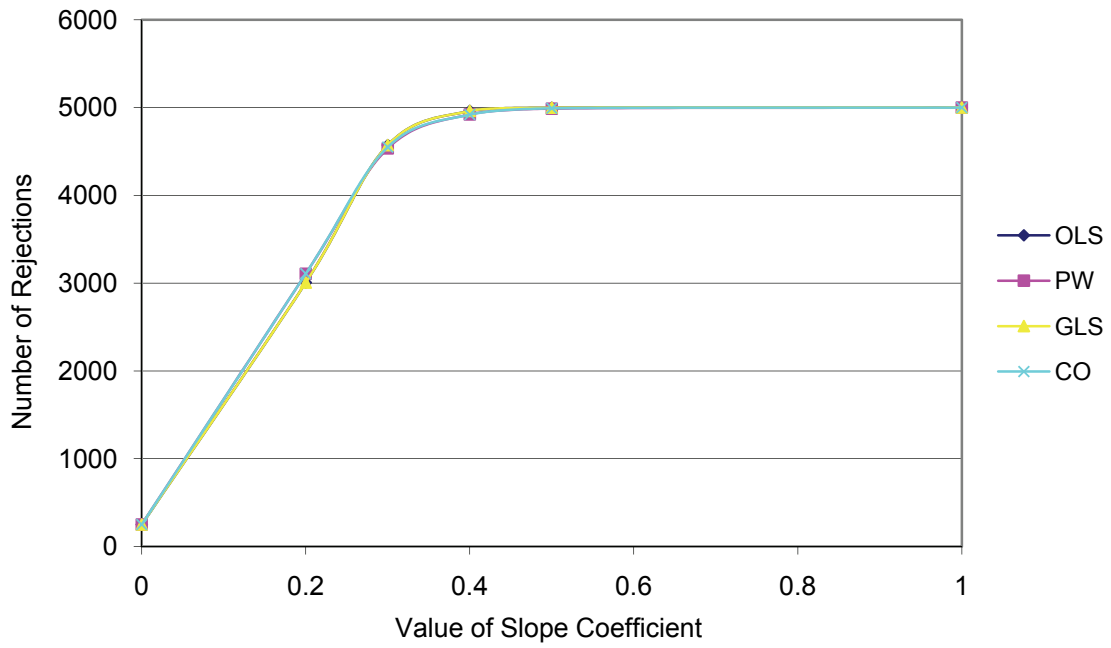
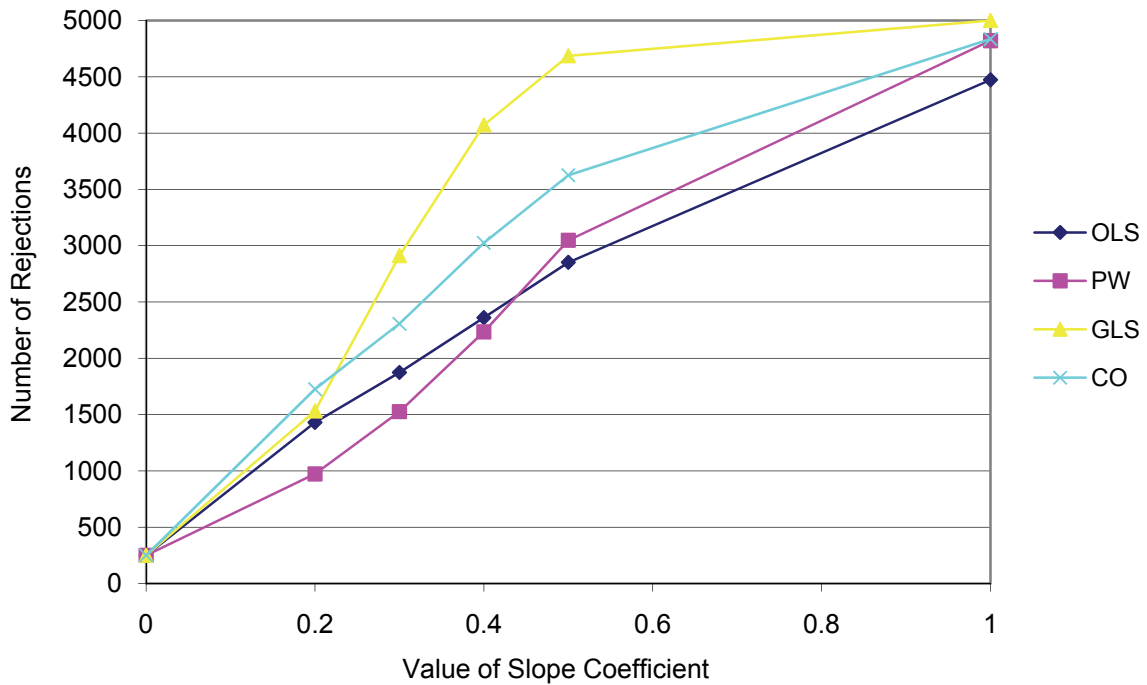


Figure 2: Power Curve for Testing Slope Equal Zero:
Autoregressive With Lambda=0.8, rho=0.95



DIELMAN

Table 4: Adjusted Power Comparisons Using Stochastic Trend Independent Variable with
Lambda = 0.8 (5,000 trials)

Rho = 0.00	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	4866	4999	5000	5000	5000
PW	250	4785	4983	4998	5000	5000
GLS	250	4866	4999	5000	5000	5000
CO	250	4666	4959	4993	4996	5000
Rho = 0.20	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	4475	4990	5000	5000	5000
PW	250	4124	4873	4988	5000	5000
GLS	250	4494	4994	5000	5000	5000
CO	250	3900	4788	4959	4989	5000
Rho = 0.40	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	3591	4817	4992	5000	5000
PW	250	2860	4389	4869	4984	5000
GLS	250	3768	4882	4998	5000	5000
CO	250	2656	4207	4752	4931	5000
Rho = 0.60	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	2268	3856	4712	4959	5000
PW	250	1520	2963	4176	4714	5000
GLS	250	2673	4294	4916	4996	5000
CO	250	1440	2840	3964	4540	4997
Rho = 0.80	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	1133	2124	3209	4022	5000
PW	250	715	1406	2295	3193	4962
GLS	250	1713	3143	4291	4841	5000
CO	250	701	1371	2232	3049	4880
Rho = 0.95	Beta =					
	0.0	0.2	0.3	0.4	0.5	1.0
OLS	250	685	1161	1766	2446	4661
PW	250	415	713	1112	1608	4074
GLS	250	1294	2457	3634	4450	5000
CO	250	442	733	1164	1660	4024

HYPOTHESIS TESTS AFTER CORRECTION FOR AUTOCORRELATION

Figure 3: Power Curve for Testing Slope Equal Zero:
Stochastic Trend With Lambda = 0.8; rho = 0.0

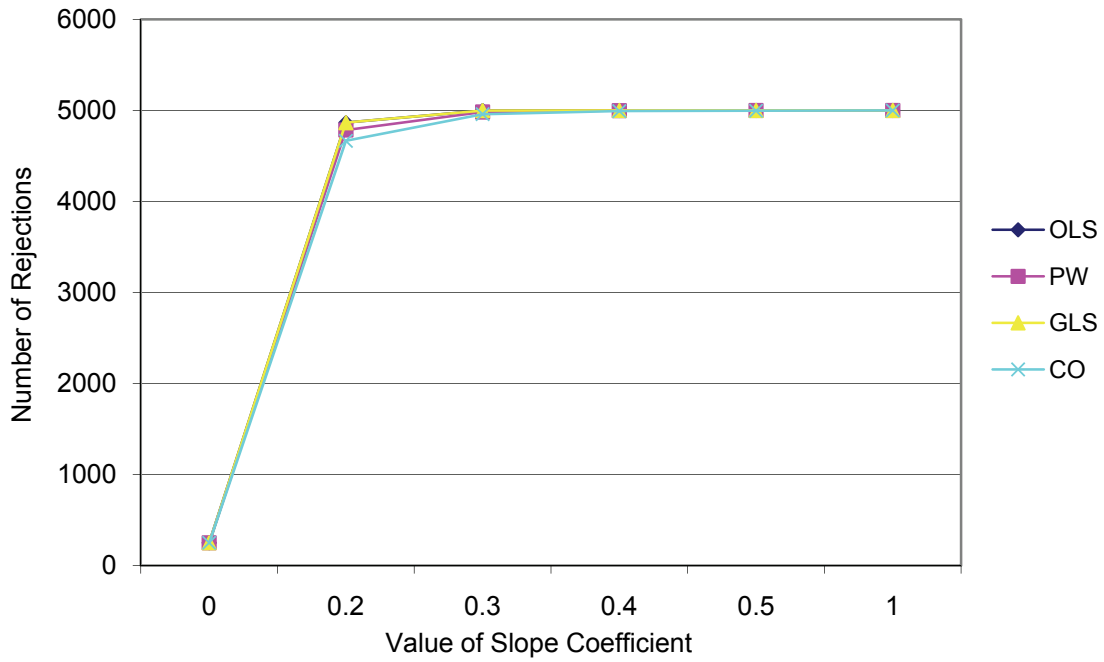
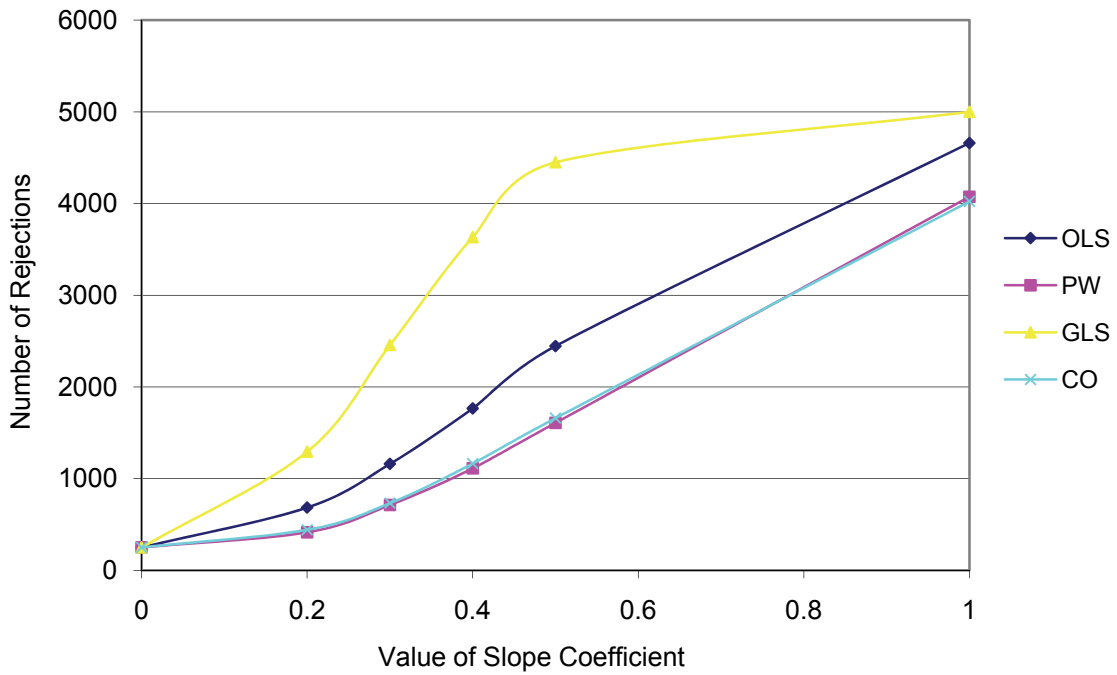


Figure 4: Power Curve for Testing Slope Equal Zero:
Stochastic Trend With Lambda = 0.8; rho = 0.95



Mizon, G. (1995). A simple message for autocorrelation correctors: Don't. *Journal of Econometrics*, 69, 267-288.

Ohtani, K. (1990). On estimating and testing in a linear regression model with autocorrelated errors. *Journal of Econometrics*, 44, 333-346.

Park, R., & Mitchell, G. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics*, 13, 185-201.

Prais, S., & Winsten, C. (1954). *Trend estimators and serial correlation*. Cowles Commission Discussion Paper: Stat. No. 383, Chicago, IL.

Rayner, R. (1991). Resampling methods for tests in regression models with autocorrelated errors. *Economics Letters*, 36, 281-284.

Taylor, W. (1981). On the efficiency of the Cochrane-Orcutt estimator. *Journal of Econometrics*, 17, 67-82.

Thornton, D. (1987). A note on the efficiency of the Cochrane-Orcutt estimator of the AR(1) regression model. *Journal of Econometrics*, 36, 369-376.

Zhang, J., & Boos, D. (1994). Adjusted power estimates in Monte Carlo experiments. *Communication in Statistics - Simulation and Computation*, 23, 165-173.

Beyond Kappa: Estimating Inter-Rater Agreement with Nominal Classifications

Nol Bendermacher Pierre Souren

Radboud University
Nijmegen, The Netherlands

Cohen's Kappa and a number of related measures can all be criticized for their definition of correction for chance agreement. A measure is introduced that derives the corrected proportion of agreement directly from the data, thereby overcoming objections to Kappa and its related measures.

Key words: Interrater agreement, Cohen's Kappa, nominal data, reliability.

Introduction

The most popular measure of inter-rater agreement in the case of nominal classification is Cohen's kappa (Cohen, 1960). Kappa is a member of a family of measures that are all defined by the same basic formula (Zwick, 1988):

$$A = \frac{f - p_c(A)}{1 - p_c(A)} \quad (1.1)$$

where f = the observed proportion of agreement and $p_c(A)$ = the definition of chance agreement according to measure A . The measures of this family differ only in their definitions of chance agreement $p_c(A)$.

Methodology

A General Model

Starting with n cases classified by two raters into c exhaustive and mutually exclusive categories, the population distribution of the c categories is given by the vector V . The joint distribution of the ratings is given by the c by c population matrix X . The model distinguishes

three types of classifications: (1) a correct observation, (2) a correct guess, and (3) a wrong guess. The second type is a correct classification, but not a correct observation. The model assumes a fixed probability p_r that rater r makes a correct observation, i.e., a classification of type (1). Fixed means that p_i is independent of the true category V_i of the case and of its classification by the other rater. Rater agreement, as far as it is not based on chance, arises if both raters make a correct observation. Assuming that raters act independently, the probability of such non-chance-agreement is $p_1 p_2$. Therefore a measure of inter-rater agreement is defined as: $s = p_1 p_2$.

If rater r performs a correct observation, the probabilities of the categories are given by the population distribution V . However, if the rater does not, the classifications follow an error distribution W_r . The error distributions may differ from V and from each other. It is assumed that W_r is independent of the true category of the case. The model parameters are p_1 , p_2 , V , W_1 and W_2 as defined above. In order to simplify the formulas $q_r = 1 - p_r$ and $D_r = W_r - V$ are also defined. This article will show that s and V can be estimated directly from the observed sample of classifications by the raters, without any assumptions regarding the error distributions W_1 and W_2 .

Nol Bendermacher is a retired member of the Research Technical Support Group of the Faculty of the Social Sciences. Pierre Souren is a current member of this group. Email them at Bendermacher@hotmail.com and p.souren@socsci.ru.nl

Some Measures for Inter-rater Agreement

In formula (1.1), f is the proportion of cases classified in the same way by both raters, and $p_c(A)$ is the correction for chance agreement according to measure A. The denominator is a scaling factor restricting the measure to a maximum of 1.

Bennett, Alpert and Goldstein (1954) assumed that a rater who does not recognize the true category of a case draws from a uniform distribution, thus giving each category an equal chance. In terms of the general model, they

assume $W_1 = W_2 = \left(\frac{1}{c}, \frac{1}{c}, \dots, \frac{1}{c}, \frac{1}{c}\right)^T$. If both

raters draw from this common error distribution, the probability of chance agreement on one specific category is $\frac{1}{c^2}$ and the overall expected

proportion of chance agreement is $\frac{c}{c^2} = \frac{1}{c}$.

Therefore, Bennett, Alpert and Goldstein (1954) defined the correction for chance agreement as

$$p_c(A) = \frac{1}{c} \text{ for their measure S.}$$

At least two objections to this choice exist:

1. In many situations it is plausible that the true distribution V of the cases deviates from uniformity and that the raters, knowing so, adjust their guessing distributions accordingly.
2. Scott (1955) objected that if W_1 and/or W_2 deviate from uniformity, the proportion of agreement by chance will always be greater than $\frac{1}{c}$. In other words, $\frac{1}{c}$ is a lower limit for the proportion of agreement by chance, meaning that S is an upper bound for inter-rater agreement.

S has been presented several times under different names and different notations. For the case of two categories S is equal to the random error RE (Maxwell, 1977). With only two categories, this measure is equal to the difference between the proportion of agreement

and the proportion of disagreement: $\frac{f}{1-f}$. For

the general case, Brennan and Prediger (1981)

reported the measure as κ_n , Zwick (1988) mentioned Guilford's G, for the two categories case, and Janson and Vegelius' C for the general case.

Scott (1955) tried to overcome the second objection by introducing the assumption that both raters, when guessing, follow the true distribution. In terms of the general model, Scott assumed that $W_1 = W_2 = V$. Therefore he estimated the distribution by the average of the two marginal distributions. His measure is called

$$\pi \text{ and } p_c(\pi) \text{ is defined as } \sum_{i=1}^c \left(\frac{M_{1i} + M_{2i}}{2}\right)^2,$$

where M_{1i} and M_{2i} are the two observed marginal proportions of category i .

Cohen (1960) objected to Scott that one source of disagreement is precisely the tendency of the raters to spread their ratings differently over the categories: "one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgments differently over the categories." Therefore, Cohen dropped the assumption of equal marginal distributions and defined the proportion of chance agreement as

$$\sum_{i=1}^c M_{1i}M_{2i}.$$

It can be seen, however, that the marginal distributions are a mix of the true distribution V and the error distributions W_1 and W_2 , more precisely, $M_r = p_r V + q_r W_r$, so Cohen's estimation of chance agreement is only correct under the null hypothesis that p_1 and p_2 are both zero, or under the assumption that $W_1 = W_2 = V$. The latter assumption would mean that the two marginal distributions are equal, so Scott's π could be used as well. As Brennan and Prediger (1981) stated: "For *descriptive* purposes, therefore, when marginals are free it seems questionable to reduce observed agreement by $\sum P_i P_{.i}$, which is directly dependent on agreement in the marginals" (p. 692). Other objections and alternatives to Kappa have also been brought forward. For details, readers are referred to Perreault and Leigh (1989) and Brennan and Prediger (1981).

The next section will elaborate on the formal model and investigate possibilities to identify and estimate the model parameters.

What is special to this approach is that the inter-rater agreement is estimated without any assumptions regarding the rater distributions W_1 and W_2 . In addition, a short outline of an algorithm that performs the required calculations is provided and an extension for the case of three simultaneous raters is introduced. Two computer programs, called Raters2 and Raters3, that implement these ideas are available at <http://www.ru.nl/socialewetenschappen/rtog/software/statistische/kunst/>.

Table 1 shows the two-way frequency distribution and the corresponding proportions, Cohen (1960, p. 45) used as an illustration. The proportion of joint judgments is the sum of the diagonal cells, here called f . In this example $f = 0.70$. Cohen defined chance agreement as $\sum_{i=1}^c M_{1i}M_{2i}$. In the example the correction is $0.30 + 0.09 + 0.02 = 0.41$, so the corrected proportion of joint judgments is:

$$f - \sum_{i=1}^c M_{1i}M_{2i} = 0.29.$$

If this value is rescaled by dividing it by its maximum, Cohen's Kappa results:

$$\text{Kappa} = \frac{f - \sum_{i=1}^c M_{1i}M_{2i}}{1 - \sum_{i=1}^c M_{1i}M_{2i}} = 0.4915 \quad (1.2)$$

Table 1: Cohen's Example Data

Frequencies				Proportions			
88	14	18	120	0.44	0.07	0.09	0.60
10	40	10	60	0.05	0.20	0.05	0.30
2	6	12	20	0.01	0.03	0.06	0.10
100	60	40	200	0.50	0.30	0.20	1.00

The General Model in Detail

From the model parameters, the population distribution X of the simultaneous classifications can be derived. Any cell $X(i,j)$ of X defines the probability of a joint classification in category i by rater 1 and category j by rater 2. X can be estimated from the two-way frequency matrix of the ratings in the sample, which will be indicated as \hat{X} . X can be interpreted as a

weighted sum of four c by c matrices, corresponding to the behavior of the raters:

X_1 : Both raters perform a correct observation. The probability of a score in a diagonal cell X_{1ii} is the product of: (a) the probability V_i that the case belongs to category i , (b) the probability p_1 that rater 1 performs a correct observation and (c) the probability p_2 that rater 2 performs a correct observation. Thus, $X_{1ii} = p_1 p_2 V_i$. The probability of a score in an off-diagonal cell is zero, so X_1 is a diagonal matrix.

X_2 : Only rater 1 performs a correct observation. The probability of a score in a cell X_{2ij} is the product of: (a) the probability V_i that the case belongs to category i , (b) the probability p_1 that rater 1 performs a correct observation, (c) the probability q_2 that rater 2 guesses and (d) the probability W_{2j} that rater 2 guesses category j . Thus, $X_{2ij} = p_1 V_i q_2 W_{2j}$.

X_3 : Only rater 2 performs a correct observation. The probability of a score in a cell X_{3ij} is the product of: (a) the probability V_j that the case belongs to category j , (b) the probability p_2 that rater 2 performs a correct observation, (c) the probability q_1 that rater 1 guesses and (d) the probability W_{1i} that rater 1 guesses category i . Thus, $X_{3ij} = p_2 V_j q_1 W_{1i}$.

X_4 : Both raters are guessing. The probability of a score in a cell X_{4ij} is the product of: (a) the probability q_1 that rater 1 is guessing, (b) the probability q_2 that rater 2 is guessing, (c) the probability W_{1i} that rater 1 guesses category i and (d) the probability W_{2j} that rater 2 guesses category j . Thus, $X_{4ij} = q_1 q_2 W_{1i} W_{2j}$.

The matrix X is the sum of these 4 matrices and its content can be summarized as follows:

For $i \neq j$:

$$\begin{aligned} X_{ij} &= p_1 q_2 V_i W_{2j} + q_1 p_2 W_{1i} V_j + q_1 q_2 W_{1i} W_{2j} \\ &= (1 - p_1 p_2) V_i V_j + q_1 V_j D_{1i} + q_2 V_i D_{2j} + q_1 q_2 D_{1i} D_{2j}, \end{aligned} \quad (2)$$

and, for $i = j$:

$$X_{ii} = p_1 p_2 V_i + p_1 q_2 V_i W_{2i} + q_1 p_2 V_i W_{1i} + q_1 q_2 W_{1i} W_{2i} \\ = p_1 p_2 V_i + (1 - p_1 p_2) V_i^2 + q_1 V_i D_{1i} + q_2 V_i D_{2i} + q_1 q_2 D_{1i} D_{2i}. \quad (3)$$

The marginal distributions M_1 and M_2 of X are given by:

$$M_r = p_r \cdot V + q_r W_r = V + q_r D_r, \text{ for } r = 1, 2. \quad (4)$$

A similar model is given by Klauer and Batchelder (1996).

Comparing s to Cohen's Kappa

The measure s and Cohen's Kappa can be compared based on the following derivation: from (3) it is evident that

$$X_{ii} = s V_i (1 - V_i) + V_i^2 + q_1 V_i D_{1i} + q_2 V_i D_{2i} + q_1 q_2 D_{1i} D_{2i}$$

and, from (4),

$$M_{1i} M_{2i} = V_i^2 + q_1 V_i D_{1i} + q_2 V_i D_{2i} + q_1 q_2 D_{1i} D_{2i}$$

thus,

$$X_{ii} - M_{1i} M_{2i} = s V_i (1 - V_i), \quad (5)$$

and,

$$s = \frac{X_{ii} - M_{1i} M_{2i}}{V_i (1 - V_i)} \\ = \frac{f - \sum_{i=1}^c M_{1i} M_{2i}}{\sum_{i=1}^c V_i (1 - V_i)} \\ = \frac{f - \sum_{i=1}^c M_{1i} M_{2i}}{1 - \sum_{i=1}^c V_i^2}.$$

Comparing this result with the formula for Kappa in (1.2) it follows that Kappa and s are only equivalent if $\sum_{i=1}^c M_{1i} M_{2i} = \sum_{i=1}^c V_i^2$. From (4) it becomes clear that such is the case only if $p_1 = p_2 = 1$, or if $W_1 = W_2 = V$. The $p_1 = p_2 = 1$ assumption is very unrealistic. The assumption that both W -vectors equal the true distribution

implies that the two marginal distributions M_1 and M_2 are equal. This is a severe and unnecessary restriction that Cohen rejected when he introduced Kappa. In his example, as shown in Table 1, the two marginal distributions differ significantly ($\chi^2 = 34.6959$, $df = 3$, $p = 0.0000$). Table 2 shows Kappa as well as the results of an analysis of Cohen's example according to the model presented herein.

Table 2: Parameter Estimates According to Proposed Model for Cohen's Example

V	W ₁	W ₂	Parameter Estimates
0.6861	0.0000	0.0000	$s = 0.6280$
0.2347	0.7620	0.4683	$p_1 = 0.8696$
0.0792	0.2380	0.5317	$p_2 = 0.7221$
			$kappa = 0.4915$

$$\text{model fit: } \chi^2 = 2.0325, \text{ df} = 1, p = 0.1540$$

Identifiability of Model Parameters

By the identifiability of the parameters is meant that their values can be uniquely derived from the joint distribution matrix X . If $B_i = X_{ii} - M_{1i} M_{2i}$, then from (5):

$$B_i = s V_i (1 - V_i) \quad (6)$$

With at least 3 non-zero entries in V , the largest entry is the one closest to 0.5. Therefore, it corresponds to the largest entry in B . In other words: if B_m is (one of) the largest entry(s) in B , V_m is (one of) the largest entry(s) in V . From (6):

$$\frac{V_j (1 - V_j)}{V_m (1 - V_m)} = \frac{B_j}{B_m}$$

and, as a consequence, for all $j \neq m$,

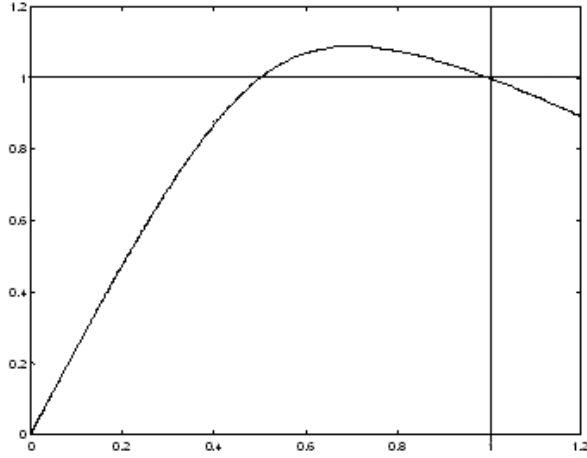
$$V_j = 0.5 \pm \sqrt{0.25 - V_m (1 - V_m) \frac{B_j}{B_m}}$$

Because there can be only one entry in V greater than 0.5, the sign before the square root must be negative for all $j \neq m$:

$$V_j = 0.5 - \sqrt{0.25 - V_m(1 - V_m) \frac{B_j}{B_m}} \quad (7)$$

It can be proved that there is only one value V_m for which the sum of elements in V according to (7) becomes 1, provided that: X obeys to the model, $c > 2$ (and consequently $V_m < 1$), $s > 0$, and, by definition, the sum of the elements in V equals 1. Figure 1 shows an example of the sum $g(V_m) = V_m + \sum_{j \neq m} V_j$ as a function of V_m and with V_j defined by (7).

Figure 1: Example of the Function $g(V_m)$



Thus, there is only one vector V for which equation (7) holds and for which the elements of V add up to 1. So V can be identified.

Once V has been identified, s can also be derived from (6):

$$s = \frac{B_i}{V_i(1 - V_i)} \quad (8)$$

for any i , except those for which $V_i = 0$.

Although the product $p_1 p_2$ (i.e., s) can be identified, it is generally impossible to identify its components p_1 and p_2 . From (4) it is known that $q_r D_r = M_r - V$, but looking at formulas (2) and (3) for the cells in X a multiplication of D_r by a constant h can be compensated by dividing q_r by the same h . Thus, neither W_1 and W_2 , nor p_1 and p_2 can be identified.

The good news is that boundaries can be identified, within which these parameters are enclosed. The boundaries follow from the facts that: all cells of V , W_1 and W_2 represent probabilities and therefore must be in the range $[0,1]$, and that V , W_1 and W_2 must add up to 1. Therefore, the following series of restrictions can be derived:

1. $s \leq p_1 \leq 1$ and $s \leq p_2 \leq 1$.

2. From (4) it is known that

$$s \frac{V_i}{M_{2i}} = p_1 \frac{p_2 V_i}{p_2 V_i + q_2 W_{2i}}, \text{ thus,}$$

$$s \frac{V_i}{M_{2i}} \leq p_1 \text{ and } p_2 \leq \frac{M_{2i}}{V_i}.$$

3. Similarly, it is known that:

$$s \frac{V_i}{M_{1i}} \leq p_2 \text{ and } p_1 \leq \frac{M_{1i}}{V_i}$$

4. Since all values of W_1 are between 0 and 1, it is known that:

$$q_1(1 - W_{1i}) \geq 0$$

$$\Rightarrow q_1(1 - V_i) - q_1(W_{1i} - V_i) \geq 0$$

$$\Rightarrow p_1(1 - V_i) + q_1(1 - V_i) - q_1(W_{1i} - V_i) \geq p_1(1 - V_i)$$

$$\Rightarrow 1 - V_i - q_1(W_{1i} - V_i) \geq p_1(1 - V_i)$$

$$\Rightarrow \frac{1 - V_i - q_1(W_{1i} - V_i)}{1 - V_i} \geq p_1$$

$$\Rightarrow \frac{1 - M_{1i}}{1 - V_i} \geq p_1, \text{ and consequently:}$$

$$s \frac{1 - V_i}{1 - M_{1i}} \leq p_1.$$

5. In the same way the following may be derived:

$$\frac{1 - M_{1i}}{1 - V_i} \geq p_1 \text{ and } s \frac{1 - V_i}{1 - M_{1i}} \leq p_2.$$

These restrictions can be summarized by the following boundaries for all i and k :

$$p_1 \cdot p_2 \leq p_1 \leq 1 \quad (9)$$

$$p_1 \cdot p_2 \leq p_2 \leq 1 \quad (10)$$

If $M_{2i} < 1$ (and $V_i < 1$), then:

$$s \frac{V_i}{M_{2i}} \leq p_1 \leq \frac{M_{1k}}{V_k} \quad (11)$$

If $M_{1i} < 1$ (and $V_i < 1$), then:

$$s \frac{V_k}{M_{1k}} \leq p_2 \leq \frac{M_{2i}}{V_i} \quad (12)$$

If $M_{2i} < 1$ (and $V_i < 1$), then:

$$s \frac{1 - V_i}{1 - M_{2i}} \leq p_1 \leq \frac{1 - M_{1k}}{1 - V_k} \quad (13)$$

If $M_{1i} < 1$ (and $V_i < 1$), then:

$$s \frac{1 - V_k}{1 - M_{1k}} \leq p_2 \leq \frac{1 - M_{2i}}{1 - V_i} \quad (14)$$

These formulas are cross-linked: the minimum for p_1 in (11) goes together with the maximum p_2 in (12) and the maximum for p_1 in (11) corresponds to the minimum for p_2 in (12). The link comes from the fact that their product must be s . The formulas in (13) and (14) are connected in a similar way.

The limits from (11) through (14) all hinge upon the differences between the true distribution V and the rater error distributions W_1 and W_2 . If $W_2 = V$ the lower limit for p_1 is s and the maximum for p_2 is 1. If $W_1 = V$ the lower limit for p_2 is s and the maximum for p_1 is 1. From these formulas it is also observed that, for categories with V -values close to 0 or 1, even small differences between W_1 or W_2 and V will impose strong restrictions.

It must be noted that this model cannot be applied if the number of categories is only 2.

Reparametrization

The parameters, as defined to this point, are neither all identifiable, nor are they independent. W_r and p_r cannot be identified, only the combination $q_r D_r = (1 - p_r)(W_r - V)$. Therefore, a reparametrization from the original set of parameters to the set $[s, V, q_1 D_1$ and $q_2 D_2]$ is used in the estimations. Moreover, the vector

V adds up to 1 and the vectors $q_1 D_1$ and $q_2 D_2$ add up to 0, which means that their elements cannot be estimated independent. Therefore, following Klauer and Batchelder (1996) the model is reparametrized again as follows:

$$A_i^* = \frac{A_i}{1.1 - \sum_{j=1}^{i-1} A_j} \quad \text{for } i = 1, c,$$

where $A = V, q_1 D_1$ and $q_2 D_2$ respectively. The last element A_c is dropped. The back-translation to the original parameters is performed by the formula:

$$A_i = A_i^* \left(1.1 - \sum_{j=1}^{i-1} A_j \right), \quad \text{for } i = 1, c$$

Initial Parameter Estimations: V and s

For the parameter estimations from an observed matrix \hat{X} one may proceed in two steps. The first step is a procedure directly derived from the model and uses only information from the diagonal and the marginal frequencies of the observed matrix. In the second step a general minimization algorithm is applied to minimize a criterion (for instance, the negative of the likelihood) based on all cells of \hat{X} . This algorithm starts from the estimations produced by the first step.

For the first step define:

$$g(x) = x + \sum_{j \neq m}^c \left(0.5 - \sqrt{0.25 - x(1-x) \frac{\hat{B}_j}{\hat{B}_m}} \right)$$

with $\hat{B}_i = \hat{X}_{ii} - \hat{M}_{1i} \hat{M}_{2i}$ and $\hat{B}_m =$ the largest value in \hat{B} . (Figure 1 shows an example of this function.) From (7) it is clear that V_m can be estimated by the value of x for which $g(x) = 1$ with $0 < x < 1$. Starting with evaluations of g at $1/c$ and a suitable maximum (for instance f), an estimate of V_m can be found by a simple iteration process using, for example, the bisection method. The remaining elements of V can be estimated by:

$$\hat{V}_j = 0.5 - \sqrt{0.25 - \hat{V}_m(1 - \hat{V}_m) \frac{\hat{B}_j}{\hat{B}_m}} \quad (15)$$

Once estimates of V are obtained, s can be estimated on the base of (8) as:

$$\hat{s} = \frac{\hat{B}_i}{\hat{V}_i(1 - \hat{V}_i)} \text{ for any } i \text{ (unless } \hat{V}_i = 0),$$

or for a combination of the estimates for different i .

However, with sampled data this direct method may easily fail. Therefore a numerically more robust algorithm to find the same initial estimates of V and s was designed, called the *ping-pong algorithm*, a detailed description of which is provided later.

Initial Estimation of W_1 and W_2

Although the parameters W_1 , W_2 , p_1 and p_2 are not identifiable, the inequalities (9) through (14) offer the possibility to set boundaries around them. These boundaries may define a very narrow area, especially for categories that are very frequent or very rare. But unfortunately it is the infrequent categories for which the analysis produces the least reliable estimations. The problem becomes most serious if there are many categories and relatively few observations, i.e., if n/c is small. If the limits given by (9) through (14) restrict the estimates \hat{p}_1 and \hat{p}_2 to single values, as occurred when Cohen's example was analyzed, W_1 and W_2 can also be estimated using (4) as:

$$\hat{W}_r = \hat{V} + \frac{1}{\hat{q}_r} (M_r - \hat{V})$$

Final Estimations and Model Test

The initial parameter estimates based on the considerations above are based completely on the diagonal and marginal distributions of \hat{X} disregarding any information in the off-diagonal cells. In the final estimation procedure information from all cells will be used. A criterion is defined for the dissimilarity between the reconstruction X^* of X from the parameter estimates and the observed matrix \hat{X} , and a

powerful minimization technique like the Davidon-Fletcher-Powell algorithm is used to improve the initial parameter estimates. An attractive criterion is based on the negative of the likelihood ratio with a small adjustment, defined as:

$$e_{ij} = \text{Max}(X^*_{ij}, \epsilon)$$

$$\text{Crit} = \sum_{i=1}^c \sum_{j=1}^c \hat{X}_{ij} \cdot \text{LN} \left(\frac{\hat{X}_{ij}}{e_{ij}} \right) + \text{penalty}$$

The term ϵ is a small value to prevent division by zero and to avoid too exotic values of $\frac{\hat{X}_{ij}}{e_{ij}}$,

for instance $\epsilon = 1.0e-20$. The *penalty* serves to force the parameters within the restrictions of the model (for instance $0 \leq s \leq 1$).

The estimation procedure as designed starts with the ping-pong algorithm resulting in estimates \hat{V} and \hat{s} , after which the reparametrizations and the minimization procedure are applied. When the final parameter estimates are obtained, a model test can be performed based on the test statistic for the likelihood ratio:

$$\chi^2 = 2 \cdot n \cdot \left(\sum_{i=1}^c \sum_{j=1}^c \hat{X}_{ij} \cdot \text{LN} \left(\frac{\hat{X}_{ij}}{e_{ij}} \right) \right)$$

The associated number of degrees of freedom is $c^2 - 3 \cdot c + 1$.

The whole model as described above is based on the assumption that s is greater than zero. If $p_1 = 0$ or $p_2 = 0$, the value of any cell X_{ij} is equal to the product of the corresponding marginal probabilities M_{1i} and M_{2j} , even if X_{ij} is a diagonal cell. This assumption that $s > 0$ may

be tested by the statistic $t = f - \sum_{i=1}^c M_{1i} \cdot M_{2i}$,

which is (approximately) distributed as Student's t with 1 degree of freedom. Confidence intervals for the parameters may be constructed by the use of the information matrix or, if the Hessian matrix is singular, by bootstrapping methods.

The Ping-Pong Algorithm

The ping-pong algorithm is designed to simultaneously estimate $s = p_1.p_2$, and the largest element V_m in V . Once V_m is estimated, the entire vector V can be estimated according to (15). In order to grasp the basic idea of the algorithm, assume that the exact values for B and f are known. Then the logic is as follows. Define: $t_i =$ upper boundary for s in the i^{th} iteration, and $u_i =$ lower boundary for V_m in the i^{th} iteration.

1. From (3) $s \leq f = \sum_{i=1}^c X_{ii}$, so choose $t_0 = f$.

2. From (7):

$$1 = \sum_{i=1}^c V_i = \frac{1}{2}(c-1) + V_m - \sum_{j \neq m} \sqrt{0.25 - V_m(1 - V_m) \frac{B_j}{B_m}}$$

so, using (8)

$$V_m = 1 - \frac{1}{2}(c-1) + \sum_{j \neq m} \sqrt{0.25 - \frac{B_j}{s}}$$

and as a consequence:

$$V_m \leq 1 - \frac{1}{2}(c-1) + \sum_{j \neq m} \sqrt{0.25 - \frac{B_j}{t_i}} \text{ for any}$$

step i

3. From (8): $s = \frac{B_m}{V_m(1 - V_m)}$, thus

$$s \leq \frac{B_m}{u_i(1 - u_i)}$$

Now the following procedure is applied:

1) $t_0 = f$

$$2) u_i = 1 - \frac{1}{2}(c-1) + \sum_{j \neq m} \sqrt{0.25 - \frac{B_j}{t_i}}$$

$$3) t_i = \frac{B_m}{u_{i-1}(1 - u_{i-1})}$$

4) Repeat from 2) until convergence is reached.

This algorithm converges to $t_i = \hat{s}$ and $u_i = \hat{V}_m$.

When working with the sample estimators \hat{B} and \hat{f} it may be necessary to make some corrections during the iteration process:

1. In the iteration process t_i may exceed the value $t_0 = \hat{f}$. In that case, force \hat{B}_m to $\hat{f}.u_{i-1}(1 - u_{i-1})$ and set t_i equal to \hat{f} . In order to keep the sum of \hat{B} unchanged, replace the other elements of \hat{B} according to the following rule:

$$\hat{B}_i \leftarrow \bar{B} + (\hat{B}_i - \bar{B}) \frac{\hat{B}_m^* - \bar{B}}{\hat{B}_m - \bar{B}}$$

where \bar{B} is the mean of the \hat{B} -values, \hat{B}_m the original estimate of B_m and \hat{B}_m^* the corrected estimate.

2. If the estimate u_i becomes less than $1/c$ force it back to $1/c$ and adjust the B -values accordingly:

$$u_i < \frac{1}{c},$$

so

$$1 - \frac{1}{2}(c-1) + \sum_{j \neq m} \sqrt{0.25 - \frac{B_j}{t_i}} < \frac{1}{c}$$

Adjust the B -vector by a vector B^* , such that

$$1 - \frac{1}{2}(c-1) + \sum_{j \neq m} \sqrt{0.25 - \frac{B_j^*}{t_i}} = \frac{1}{c},$$

which means that

$$\sum_{j \neq m} \sqrt{0.25 - \frac{B_j^*}{t_i}} = 0.5c + \frac{1}{c} - 1.5$$

Make the adjustment by taking B^* such that each term in the summation, except B_m , is multiplied by:

$$a = \frac{0.5c + \frac{1}{c} - 1.5}{\sum_{j \neq m} \sqrt{0.25 - \frac{B_j}{t_i}}} = \frac{0.5c - 1.5 + \frac{1}{c}}{0.5c - 1.5 + u_i}$$

BEYOND KAPPA

This is realized by replacing each B_j , except B_m , by $B_j^* = 0.25t_i(1-a^2) + a^2B_j$.

Three Raters

Under the given model, the expansion to three simultaneous raters is straightforward. Moreover, with three simultaneous raters, all parameters are identifiable if there are at least three categories. The notation must be extended to three p-values p_1, p_2 and p_3 , three q-values q_1, q_2 and q_3 , three W-vectors W_1, W_2 and W_3 , and three marginal distributions M_1, M_2 and M_3 . In addition the matrix X will now have three dimensions. The formulas for the probabilities in the cells of X are more complicated: X_{ijk} is the sum of the corresponding cells in eight submatrices as shown in Tables 3a through 3c.

Table 3a: Formulas for Two Parts of the Matrix X in Case of Three Raters

Raters i, j, k	123	i = j = k	i = k ≠ j
		X_{ijk}	X_{ijk}
X_1	ccc	$p_1p_2p_3V_i$	0
X_2	cci	$p_1p_2V_{iq_3W_3k}$	0
X_3	cic	$p_1p_3V_{iq_2W_2j}$	$p_1p_3V_{iq_2W_2j}$
X_4	cii	$p_1V_{iq_2W_2j}q_3W_3k$	$p_1V_{iq_2W_2j}q_3W_3k$
X_5	icc	$p_2p_3V_{jq_1W_1i}$	0
X_6	ici	$p_2V_{jq_1W_1i}q_3W_3k$	$p_2V_{jq_1W_1i}q_3W_3k$
X_7	iic	$p_3V_{kq_1W_1i}q_2W_2j$	$p_3V_{kq_1W_1i}q_2W_2j$
X_8	iii	$q_1W_1i q_2W_2j q_3W_3k$	$q_1W_1i q_2W_2j q_3W_3k$

Table 3b: Formulas for Two Parts of the Matrix X in Case of Three Raters

Raters i, j, k	123	i = j ≠ k	i ≠ j = k
		X_{ijk}	X_{ijk}
X_1	ccc	0	0
X_2	cci	$p_1.p_2V_{iq_3W_3k}$	0
X_3	cic	0	0
X_4	cii	$p_1V_{iq_2W_2j}q_3W_3k$	$p_1V_{iq_2W_2j}q_3W_3k$
X_5	icc	0	$p_2p_3V_{jq_1W_1i}$
X_6	ici	$p_2V_{jq_1W_1i}q_3W_3k$	$p_2V_{jq_1W_1i}q_3W_3k$
X_7	iic	$p_3V_{kq_1W_1i}q_2W_2j$	$p_3V_{kq_1W_1i}q_2W_2j$
X_8	iii	$q_1W_1i q_2W_2j q_3W_3k$	$q_1W_1i q_2W_2j q_3W_3k$

Table 3c: Formulas for One Part of the Matrix X in Case of Three Raters

Raters i, j, k	123	i ≠ j ≠ k
		X_{ijk}
X_1	ccc	0
X_2	cci	0
X_3	cic	0
X_4	cii	$p_1V_{iq_2W_2j}q_3W_3k$
X_5	icc	0
X_6	ici	$p_2V_{jq_1W_1i}q_3W_3k$
X_7	iic	$p_3V_{kq_1W_1i}q_2W_2j$
X_8	iii	$q_1W_1i q_2W_2j q_3W_3k$

Submatrix X_1 contains those ratings for which all three raters make a correct observation, as indicated by the code ccc, which means correct-correct-correct. The value in cell i, j, k depends on the equality of the three indices as indicated by the column headings. The other submatrices are organized in the same way: X_2 contains ratings where raters 1 and 2 made correct observations, but rater three did not (he guessed, correctly or not), indicated by the label cci (correct-correct-incorrect).

Table 4: Frequency Matrix with Three Categories and Three Raters

		Rater 3 = 1			Rater 3 = 2			Rater 3 = 3		
		Rater 2			Rater 2			Rater 2		
Rater 1		37	16	19	32	21	13	0	2	7
		19	11	7	30	103	38	9	11	16
		5	7	2	10	22	11	11	13	28

Table 4 shows an example of the three-way distribution in a sample with size 500. The data were generated by random sampling from a theoretical distribution based on the probabilities given in tables 3a-3c, with the following parameters: $p_1 = 0.5$, $p_2 = 0.4$, $p_3 = 0.6$, $V^T = (0.3, 0.5, 0.2)$, $W_1^T = (0.2, 0.5, 0.3)$, $W_2^T = (0.3, 0.4, 0.3)$, and $W_3^T = (0.1, 0.7, 0.2)$. Initial estimations for p_1 , p_2 , p_3 , V , W_1 , W_2 and W_3 can be derived from the three marginal planes, which can be computed from \hat{X} by summation over the categories of one rater:

$$X^{12}_{ij} = \sum_{k=1}^c X_{ijk}$$

$$X^{13}_{ij} = \sum_{k=1}^c X_{ikj}$$

$$X^{23}_{ij} = \sum_{k=1}^c X_{kij}$$

Tables 5, 6 and 7 show these planes for the example in table 4.

Table 5: The Marginal Planes for Raters 1 and 2 in the Example

\hat{X}^{12} :	Rater 2			\hat{M}_1
Rater 1	0.138	0.078	0.078	0.294
	0.116	0.250	0.122	0.488
	0.052	0.084	0.082	0.218
\hat{M}_2	0.306	0.412	0.282	1.000

Table 6: The Marginal Planes for Raters 1 and 3 in the Example

\hat{X}^{13} :	Rater 3			\hat{M}_1
Rater 1	0.144	0.132	0.018	0.294
	0.074	0.342	0.072	0.488
	0.028	0.086	0.104	0.218
\hat{M}_3	0.246	0.560	0.194	1.000

Table 7: The Marginal Planes for Raters 2 and 3 in the Example

\hat{X}^{23} :	Rater 3			\hat{M}_2
Rater 2	0.122	0.144	0.040	0.306
	0.068	0.292	0.052	0.412
	0.056	0.124	0.102	0.282
\hat{M}_3	0.246	0.560	0.194	1.000

Define:

$$B^{12}_i = X^{12}_{ii} - M_{1i} \cdot M_{2i}$$

$$B^{13}_i = X^{13}_{ii} - M_{1i} \cdot M_{3i}$$

$$B^{23}_i = X^{23}_{ii} - M_{2i} \cdot M_{3i}$$

In the example above these values are estimated by:

$$\hat{B}^{12T} = [0.038036, 0.048944, 0.020524]$$

$$\hat{B}^{13T} = [0.071676, 0.068720, 0.061708]$$

$$\hat{B}^{23T} = [0.046724, 0.061280, 0.047292]$$

Because $B^{12}_i = p_1 p_2 V_i (1 - V_i)$ and analogously $B^{13}_i = p_1 p_3 V_i (1 - V_i)$ and $B^{23}_i = p_2 p_3 V_i (1 - V_i)$:

$$\frac{V_j(1 - V_j)}{V_i(1 - V_i)} = \frac{B^{12}_j}{B^{12}_i} = \frac{B^{13}_j}{B^{13}_i}$$

$$= \frac{B^{23}_j}{B^{23}_i} = \frac{B^{12}_j + B^{13}_j + B^{23}_j}{B^{12}_i + B^{13}_i + B^{23}_i} \quad (16)$$

$$= \frac{1}{3} \left(\frac{B^{12}_j}{B^{12}_i} + \frac{B^{13}_j}{B^{13}_i} + \frac{B^{23}_j}{B^{23}_i} \right)$$

The largest value V_m in V can be estimated by setting it to the value of x for which the function $g(x) = 1$, where g is defined as:

$$g(x) = x + \sum_{j \neq m}^c \left(0.5 - \sqrt{0.25 - x(1-x)} \frac{\left(\hat{B}^{12}_j + \hat{B}^{13}_j + \hat{B}^{23}_j \right)}{\left(\hat{B}^{12}_m + \hat{B}^{13}_m + \hat{B}^{23}_m \right)} \right)$$

or as

$$g(x) = x + \sum_{j \neq m}^c \left(0.5 - \sqrt{0.25 - x(1-x)} \frac{1}{3} \left(\frac{\hat{B}^{12}_j}{\hat{B}^{12}_m} + \frac{\hat{B}^{13}_j}{\hat{B}^{13}_m} + \frac{\hat{B}^{23}_j}{\hat{B}^{23}_m} \right) \right)$$

In this function the index m refers to the largest value (or one of the largest values) in the B-vectors. Because the three estimated B-vectors in a sample may have different orders, choose m as the index for which

$$\hat{B}_m^{12} + \hat{B}_m^{13} + \hat{B}_m^{23} \geq \hat{B}_i^{12} + \hat{B}_i^{13} + \hat{B}_i^{23}$$

for all i. In the example, from (15), $\hat{V}_m = 0.422659$.

From (14) follows that, for the other elements of V,

$$V_j = 0.5 - \sqrt{0.25 - V_m(1 - V_m) \left(\frac{\hat{B}_j^{12} + \hat{B}_j^{13} + \hat{B}_j^{23}}{\hat{B}_m^{12} + \hat{B}_m^{13} + \hat{B}_m^{23}} \right)}$$

and it is found that:

$$\hat{V}^T = [0.348216, 0.422659, 0.229124]$$

Once the initial estimate of V is made, the parameters p_1 , p_2 and p_3 can be estimated in the following way: from (8) it is known that, for all i,

$$s^{12} = p_1 p_2 = \frac{B_i^{12}}{V_i(1 - V_i)},$$

so the product can be estimated by averaging over i-values:

$$\hat{s}^{12} = \frac{1}{c} \sum_{i=1}^c \frac{\hat{B}_i^{12}}{\hat{V}_i(1 - \hat{V}_i)}.$$

In the same way s^{13} and s^{23} can be estimated and estimates of the parameters p_1 , p_2 and p_3 can be found by combining the three estimated s-values. For any triple (i, j, k) raters:

$$\frac{s^{ij} s^{ik}}{s^{jk}} = \frac{p_i p_j p_i p_k}{p_j p_k} = p_i^2,$$

so the p-values can be estimated from their estimated products:

$$\hat{p}_i = \sqrt{\frac{\hat{s}^{ij} \hat{s}^{ik}}{\hat{s}^{jk}}}.$$

In the example: $\hat{s}^{12} = 0.176141$, $\hat{s}^{13} = 0.315598$, $\hat{s}^{23} = 0.241583$ and $\hat{p}_1 = 0.479694$, $\hat{p}_2 = 0.367195$, $\hat{p}_3 = 0.657915$. Once initial estimates for V and the p-parameters are obtained, the estimation of the W-vectors is straightforward. From (4), it is known that, for rater r, $M_r = p_r V + (1-p_r)W_r$, so W_r can be estimated by:

$$W_r = \frac{1}{1 - \hat{p}_r} (M_r - \hat{p}_r \hat{V}).$$

In the example this results in the following initial estimates:

$$\hat{W}_1^T = [0.244016, 0.548241, 0.207744],$$

$$\hat{W}_2^T = [0.281504, 0.405815, 0.312682],$$

$$\hat{W}_3^T = [0.049413, 0.824141, 0.126448],$$

but with sample data these formulas may lead to negative entries in the estimated W-vectors. If that occurs the initial estimate for the W-vector at hand can be set equal to the estimated V.

Final estimates, using information from all cells in \hat{X} , can be computed by methods analogous to those described, minimizing the adjusted likelihood ratio.

Conclusion

When Cohen (1960) introduced his measure Kappa, he provided a good index to estimate inter-rater agreement in the case of a nominal category system that could be easily computed by hand. Cohen argued that differences in the marginal distributions must be taken into account, but, as shown, his measure Kappa does so correctly only if the marginal distributions are equal. For practical reasons, especially the fact that computers were mostly unavailable in 1960, Kappa could be considered the best available instrument at the time, but with modern computers advancements can be made. A model based on Cohen's ideas and a procedure to correctly estimate its parameters was presented herein. The model allows - to a certain extent - to separately estimate the qualities of two raters by giving two measures p_1 and p_2 . It also breaks

apart the rater characteristics (W_1 and W_2) on one hand and the true distribution of the categories (V) on the other.

If the estimates p_r and W_r are truly independent from the distribution V , it becomes possible first to assess these statistics for one rater (using a second rater) in a pilot study, and then to use them in order to find boundaries for the V -values in the main study without the need for a second rater. The formula to be used

follows from (4): $\hat{V}_i = \frac{M_{ri} - \hat{q}_r \hat{W}_{ri}}{\hat{p}_r}$.

References

- Bennett, E.M., Alpert, R., & Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18, 303-308
- Brennan, R. L., & Dale J. Prediger, D. J. (1981). Coefficient kappa: some uses, misuses and alternatives, *Educational and Psychological Measurement*, 41, 687-699.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46
- Klauer, K.C., & Batchelder, W.H. (1996). Structural analysis of subjective categorical data. *Psychometrika*, 61, 199-240.
- Maxwell, A.E. (1977). Coefficients of agreement between Observers and their interpretation. *British Journal of Psychiatry*, 130, 79-83.
- Perreault, W.D., & Leigh, E. (1989). Reliability of nominal data based on qualitative judgements. *Journal of Marketing Research*, 26, 135-148.
- Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.
- Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374-378.

Multiple Regression in Pair Correlation Solution

Stan Lipovetsky
GfK Custom Research North America

Behavior of the coefficients of ordinary least squares (OLS) regression with the coefficients regularized by the one-parameter ridge (Ridge-1) and two-parameter ridge (Ridge-2) regressions are compared. The ridge models are not prone to multicollinearity. The fit quality of Ridge-2 does not decrease with the profile parameter increase, but the Ridge-2 model converges to a solution proportional to the coefficients of pair correlation between the dependent variable and predictors. The Correlation-Regression (CORE) model suggests meaningful coefficients and net effects for the individual impact of the predictors, high quality model fit, and convenient analysis and interpretation of the regression. Simulation with three correlations show in which areas the OLS regression coefficients have the same signs with pair correlations, and where the signs are opposite. The CORE technique should be used to keep the expected direction of the predictor's impact on the dependent variable.

Key words: multiple regression, ridge regression, multicollinearity, net effects, simulation modeling.

Introduction

Regression analysis is one of the main tools of statistical modeling. It is efficient for prediction but often produces poor results in the analysis of the individual predictors importance due to multicollinearity (Dillon & Goldstein, 1984; Weisberg, 1985; Grapentine, 1997). Multicollinearity among predictors makes parameter estimates fluctuate uncontrollably with only a minor change in the sample, produces signs of coefficients in regression opposite to the signs of pair correlations, and yields theoretically important variables with insignificant coefficients. Multicollinearity also causes a reduction in statistical power that leads to wider confidence intervals for the coefficients, leaving some to be incorrectly identified as insignificant, while the ability to determine the difference between parameters is also degraded (Mason & Perreault, 1991). To

overcome the deficiencies of multicollinearity, a ridge regression technique was developed (Hoerl & Kennard, 1970, 1988, 2000; Brown, 1994). However, compared to the ordinary least squares (OLS) regression, the quality of fit of the one-parameter ridge, or Ridge-1, is worse. This quality decreases with an increase of the ridge parameter used to attain interpretable signs of the regression coefficients.

Other approaches include regularization methods based on the principal components, on the quadratic L_2 -metric, lasso regression based on the linear L_1 -metric, and other L_p -metrics used for modeling (Frank & Friedman, 1993; Wildt, 1993; Tibshirani, 1996; Hawkins & Yin, 2002; Efron, et al., 2004; Lipovetsky, 2007). A useful two-parameter ridge model is considered in (Lipovetsky, 2006) where it is shown that the quality of fit of the Ridge-2 model is much better than that of the regular Ridge-1 regression and is close to the OLS model. With an increase of the profile parameter, the quality of the Ridge-2 model stays high, and its solution becomes proportional to the coefficients of pair correlations of the dependent variable with the predictors. The quality of fit can be very similar for the models with rather different coefficients (Ehrenberg, 1982; Weisberg, 1985).

Stan Lipovetsky is Senior Research Director at GfK Custom Research North America in Minneapolis, MN. Email him at: stan.lipovetsky@gfk.com.

Methodology

$$R^2 = 1 - S^2 = \beta'(2r - C\beta) \quad (4)$$

Ordinary Least Squares Regression and Ridge-1 Regression

Consider the ordinary least squares (OLS) regression and some of its features. For the standardized (centered and normalized by standard deviation) variables, a multiple linear regression is $y_i = \beta_1 x_{i1} + \dots + \beta_n x_{in} + \varepsilon_i$, or in the matrix form:

$$y = X\beta + \varepsilon \quad (1)$$

where X is N by n matrix with elements x_{ij} of i^{th} observations ($i=1, \dots, N$) by j^{th} independent variables ($j=1, \dots, n$), y is the vector of observations for the dependent variable, β is the n^{th} order vector of beta-coefficients for the standardized regression, $X\beta = \tilde{y}$ is the theoretical predicted by the model vector of the dependent variable, and ε is a vector of deviations from the theoretical relationship. The Least Squares (LS) objective for the regression corresponds to minimizing the sum of squared deviations:

$$\begin{aligned} S^2 &= \|\varepsilon\|^2 \\ &= (y - X\beta)'(y - X\beta) \\ &= 1 - 2\beta'r + \beta' C \beta \end{aligned} \quad (2)$$

where prime denotes transposition, variance of the standardized y equals one, $y'y = 1$, and notations C and r correspond to the correlation matrix $C = X'X$ and vector of the correlations with the dependent variable $r = X'y$. The first order condition of minimization $\partial S^2 / \partial \beta = 0$ yields a system of equations with the corresponding solution:

$$C\beta = r, \quad \beta = C^{-1}r \quad (3)$$

The vector of standardized coefficients of regression β in the OLS solution (3) is defined via the inverse correlation matrix C^{-1} . The quality of the model is estimated by the residual sum of squares (2), or by the coefficient of multiple determination:

The Pythagorean connection between the unit of the original standardized empirical sum of squares with the sum of squares explained (R^2) and non-explained (S^2) by the regression, is $R^2 + S^2 = 1$.

The minimum of the objective (2), when the equation $C\beta = r$ (3) is satisfied, corresponds to the maximum of the coefficient of multiple determination which reduces to:

$$R^2 = \beta' C \beta = \beta' r \quad (5)$$

The items $(\beta'r)_j$ of the scalar product in (5) define the net effects, $NetEff_j$, which can be used to estimate the individual contribution of each j^{th} regressor:

$$R^2 = \beta'r = \sum_{j=1}^n \beta_j r_{yj} \equiv \sum_{j=1}^n NetEff_j \quad (6)$$

where r_{yj} are the pair correlations of y with the regressors x_j .

If any regressors are highly correlated or multicollinear, correlation matrix C (3) becomes ill-conditioned, its determinant is close to zero, and the inverse matrix in (3) produces a solution with highly inflated values of the coefficients of regression. The values of these coefficients often have signs opposite to the corresponding pair correlations of regressors with the dependent variable, so the net effects (6) become negative. Such a model can be used for prediction, but it is useless for analyzing and interpreting the predictors' role in the model.

The one-parameter ridge model (Ridge-1) is widely used for overcoming the difficulties of multicollinearity. Adding a regularization of the squared norm for the vector of regression coefficients (that prevents their inflation) to LS objective (2) yields a conditional objective:

$$\begin{aligned} S^2 &= \|\varepsilon\|^2 + k \|\beta_{rd}\|^2 \\ &= 1 - 2\beta'_{rd} r + \beta'_{rd} C \beta_{rd} + k \beta'_{rd} \beta_{rd} \end{aligned} \quad (7)$$

MULTIPLE REGRESSION IN PAIR CORRELATION SOLUTION

where β_{rd} denotes a vector of the ridge regression estimates for the coefficients in (1), and k is a positive profile parameter. Minimizing the objective (7) by vector β_{rd} yields a system of equations and its corresponding solution as:

$$(C + kI)\beta_{rd} = r, \quad \beta_{rd} = (C + kI)^{-1}r \quad (8)$$

where I is the identity matrix of n^{th} order. The solution (8) exists even for a singular matrix C . If $k = 0$ the Ridge-1 model (7)-(8) reduces to the OLS regression model (2)-(3).

The eigenproblem for the matrix of correlations among the regressors is $Ca = \lambda a$, (9) so the matrix can be presented as $C = A \text{diag}(\lambda_j) A'$, where A is the matrix of eigenvectors a_j in its columns, and $\text{diag}(\lambda_j)$ is a diagonal matrix of the eigenvalues λ_j . By the eigenproblem results, the Ridge-1 solution (8) can be represented as follows:

$$\beta_{rd} = A \text{diag}((\lambda_j + k)^{-1}) A' r \quad (10)$$

Increasing the profile parameter k drives the Ridge-1 solution (10) to zero at a rate of $1/k$. The coefficient of multiple determination (4) for the Ridge-1 model can be presented as:

$$\begin{aligned} R_{rd}^2 &= r' A \text{diag} \left(\frac{2}{\lambda_j + k} - \frac{\lambda_j}{(\lambda_j + k)^2} \right) A' r \\ &= r' A \text{diag} \left(\frac{\lambda_j + 2k}{(\lambda_j + k)^2} \right) A' r \end{aligned} \quad (11)$$

So the quality of fit for the Ridge-1 model also reaches zero in a proportion reciprocal to k . This means that increasing the profile parameter k could yield coefficients with interpretable signs, but small values, and poor quality of fit for the model.

Two-Parameter Ridge and Correlation-Regression Model

Consider a generalization of the regularization (7) with several positive parameters k :

$$\begin{aligned} S^2 &= \|y - Xb\|^2 + k_1 \|b\|^2 + k_2 \|X'y - b\|^2 + k_3 \|y'(y - Xb)\|^2 \\ &= \left[\begin{array}{l} (1 - 2b'r + b'Cb) + k_1(b'b) + \\ k_2(r'r - 2b'r + b'b) + k_3(1 - 2b'r + b'rr'b) \end{array} \right]. \end{aligned} \quad (12)$$

The vector b is an estimator of the coefficients of regression (1) by the multiple objective (12), where the first two items coincide with those in the Ridge-1 objective (7). The next item with k_2 pushes the estimates b to be closer to the pair correlations r with the dependent variable, which helps us obtain a solution with interpretable coefficients. The last item with k_3 expresses the relation $y'\mathcal{E} = 1 - R^2$, so its minimum corresponds to the maximum coefficient of multiple determination (more details are given in Lipovetsky, 2006). Minimization (12) yields a matrix equation

$$Cb + k_1 b + k_2 b + k_3 r r' b = r + k_2 r + k_3 r.$$

The scalar product $r'b$ can be considered as another constant and combined with the parameter k_3 , so this item at the left-hand side is proportional to vector r and can be transferred to the right-hand side of this equation. By combining constants at each side of this equation, it is easy to reduce it to the following system with the corresponding solution:

$$(C + kI)b = qr, \quad b = q(C + kI)^{-1}r \quad (13)$$

where k and q are two new constant parameters. It is the Ridge-2 model that is proportional to the Ridge-1 (8) with the term q .

For a current profile ridge parameter k , the value of the second parameter q can be found by a criterion of maximum quality of fit. Substituting solution (13) into the coefficient of multiple determination (4) yields:

$$\begin{aligned} \tilde{R}^2 &= 2q[r'(C + kI)^{-1}r] - \\ & q^2[r'(C + kI)^{-1}C(C + kI)^{-1}r] \end{aligned} \quad (14)$$

The coefficient of multiple determination \tilde{R}^2 for the Ridge-2 model is a concave quadratic by q function, and it reaches its maximum at the value:

$$q = \frac{r'(C + kI)^{-1}r}{r'(C + kI)^{-1}C(C + kI)^{-1}r}, \quad (15)$$

so the parameter q is uniquely defined as a quotient of two quadratic forms dependent on the profile parameter k . While the term k serves for regularization of an ill-conditioned matrix, the term q is used for tuning the quality of the model fit.

Using the term (15) in (13) presents the Ridge-2 solution in the explicit form:

$$b = \frac{r'(C + kI)^{-1}r}{r'(C + kI)^{-1}C(C + kI)^{-1}r}(C + kI)^{-1}r \quad (16)$$

Substituting q (15) into (14) yields the maximum coefficient of multiple determination in two following equivalent forms:

$$\tilde{R}^2 = \frac{[r'(C + kI)^{-1}r]^2}{r'(C + kI)^{-1}C(C + kI)^{-1}r} = b'Cb = r'b \quad (17)$$

Both Ridge-2 (17) and OLS (5) coefficients of multiple determination can be presented similarly as scalar products of the vectors of regression coefficients and pair correlations. The coefficient of multiple determination for Ridge-2 (17) is smaller than that of the OLS (5) but larger than that of Ridge-1 (11).

Consider the behavior of the Ridge-2 solution with the parameter k increasing. In the limit of large k , the matrix $C + kI$ gets a dominant diagonal, so the inverse matrix $(C + kI)^{-1}$ reduces to the scalar matrix $k^{-1}I$, and the term (15) becomes:

$$q = \frac{k^{-1}r'r}{k^{-2}r'Cr} = k\gamma, \quad \gamma = \frac{r'r}{r'Cr}, \quad (18)$$

so q is linearly proportional to k with a constant γ defined by the positive ratio of two quadratic forms. Similarly, in the limit of large k , the Ridge-2 solution (16) eventually converges to the independent of k asymptote:

$$b = \frac{k^{-2}r'r}{k^{-2}r'Cr}r = \left(\frac{r'r}{r'Cr}\right)r \equiv \gamma r \quad (19)$$

where γ is a constant from (18). Thus, in contrast to diminishing to zero Ridge-1 coefficients (8), the coefficients of the Ridge-2 solution (19) become proportional to the vector r of the pair correlations of y with each regressor. It is a model which can be called Correlation-Regression (CORE) model. It can also be described in terms of the pair-wise regressions of y by each x_j separately, where a beta-coefficient equals the pair correlation r_{yj} of y with the variable x_j .

The signs of CORE coefficients b (19) coincide with the signs of the pair correlations r . It guarantees the clear interpretability of this solution, and the positive net effect contributions $b_j r_j = \gamma r_j^2$ (6) of the regressors into the coefficient of multiple determination (17). With k increasing, the coefficient of multiple determination (17) reaches the limit:

$$\tilde{R}^2 = \frac{k^{-2}(r'r)^2}{k^{-2}r'Cr} = \frac{(r'r)^2}{r'Cr} = \gamma(r'r) \quad (20)$$

Thus, eventually, while k increases, the coefficient of multiple determination becomes a constant independent of k .

Numerical runs support the features of the eventual ridge regression. With increasing parameter k , the Ridge-2 coefficient of multiple determination \tilde{R}^2 (17) stays consistently close to the maximum R^2 (5) of the OLS model, while

MULTIPLE REGRESSION IN PAIR CORRELATION SOLUTION

the Ridge-1 coefficient R_{rd}^2 (11) quickly diminishes to zero. In Ridge-2 modeling k can be increased without losing the quality of regression fit and prediction, until reaching the asymptotic solution (19) of interpretable coefficients of multiple regression proportional to pair correlations of y with the x -s, with the coefficients of multiple determination (20).

The constant γ (18) used in the CORE solution (19)-(20) can be obtained in a simpler approach. If the vector r of the pair correlations of y with regressors is taken for the coefficients β in a multiple regression (1), then the vector of theoretical values of the dependent variable is $\tilde{y} = Xr$. Consider a pair regression of the observed values y on the theoretical aggregate \tilde{y} , so a model $y = \gamma\tilde{y}$, with a slope coefficient γ . As in any pair regression, this coefficient is defined as follows:

$$\gamma = \frac{\text{cov}(y, \tilde{y})}{\text{cov}(\tilde{y}, \tilde{y})} = \frac{y'Xr}{r'X'Xr} = \frac{r'r}{r'Cr} \quad (21)$$

where C and r are defined as in (2)-(3). The slope (21) coincides with the coefficient γ in (18). Then the model is $y = \gamma\tilde{y} = X(\gamma r) = Xb$, with the same coefficients b as in (19). Also, the coefficient of pair correlation between y and the aggregate \tilde{y} is:

$$\begin{aligned} \text{cor}(y, \tilde{y}) &= \frac{\text{cov}(y, \tilde{y})}{\sqrt{\text{cov}(y, y) \cdot \text{cov}(\tilde{y}, \tilde{y})}} \\ &= \frac{r'r}{\sqrt{r'Cr}} \end{aligned} \quad (22)$$

with $y'y = 1$ as in (2). This coefficient (22) squared yields the same expression (20), as a regular pair correlation squared equals the coefficient of multiple determination in the model by only one predictor.

A simple solution for the coefficients of regression can also be based on the relation between the coefficients of multiple determination R^2 (5) and $R_{y,(-j)}^2$ in the regressions of y by all n and by $n-1$ predictors

without x_j variable, respectively. The increment U_j from $R_{y,(-j)}^2$ to R^2 is defined by the j^{th} coefficient of regression (1) and by the multiple determination $R_{j,(-j)}^2$ of the regression x_j by all the other $n-1$ predictors:

$$U_j = \beta_j^2(1 - R_{j,(-j)}^2) = \beta_j^2 / VIF_j \quad (23)$$

where VIF_j is the so-called variance inflation factor (Weisberg, 1985). The VIF value for each regressor equals the diagonal elements of the inverse correlation matrix of predictors, $VIF_j = (1 - R_{j,(-j)}^2)^{-1} = (C^{-1})_{jj}$. The measure (23) of predictor importance is considered in (Darlington, 1968; Harris, 1975; Lipovetsky & Conklin, 2005).

A criterion of proportionality $U_j = g \text{NetEff}_j$ (where g is a constant) between the increments (23) and net effects (6) for each predictor can be used to estimate the coefficients of regression by the relation $\beta_j^2 / VIF_j = g\beta_j r_{yj}$, which yields the solution: $\beta_j = g(r_{yj} VIF_j)$. The constant g is estimated by the same expression (21) up to using the vector with elements $r_{yj} VIF_j$ in place of the vector r with the elements r_{yj} . However, the numerical simulations show that the results based on this approach are very close to those obtained in a simple pair correlation CORE solution (19)-(20). This means that in the eventual ridge solution (19) the coefficients of regression yield the increments in (23) approximately proportional to the net effects (6) in the coefficient of multiple determination.

Another way to obtain CORE-type model consists in the rearranging the OLS objective by opening parentheses and squaring the items in (2) explicitly:

$$\begin{aligned}
 S^2 &= \sum_{i=1}^N (y_i - \beta_1 x_{i1} - \dots - \beta_n x_{in})^2 \\
 &= \sum_{i=1}^N \left[\left(\frac{1}{n} y_i - \beta_1 x_{i1} \right) + \dots + \left(\frac{1}{n} y_i - \beta_n x_{in} \right) \right]^2 \\
 &= \sum_{i=1}^N \left[\sum_{j=1}^n \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 + \right. \\
 &\quad \left. 2 \sum_{j>k}^n \left(\frac{1}{n} y_i - \beta_j x_{ij} \right) \left(\frac{1}{n} y_i - \beta_k x_{ik} \right) \right] \\
 &= \left[\sum_{j=1}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 + \right. \\
 &\quad \left. 2 \sum_{j>k}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right) \left(\frac{1}{n} y_i - \beta_k x_{ik} \right) \right].
 \end{aligned} \tag{24}$$

so the LS objective (2) can be presented as the total of squared deviations $y_i/n - \beta_j x_{ij}$ in the pair-wise regressions of $1/n^{\text{th}}$ portion of y by each x_j separately, plus double cross-products of such deviations from each two pair-wise regressions by variables x_j and x_k . If the cross-products of deviations are small in comparison with squared deviations, the result (24) reduces to the total of least squares objectives by each variable separately:

$$S_{pair}^2 = \sum_{j=1}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 \equiv \sum_{j=1}^n S_j^2 \tag{25}$$

Minimizing (25) yields coefficients $\beta_j = r_{yj}/n$ equal to the pair correlations of y/n with the variables x_j . This multiple regression's coefficients are proportional to the pair correlations (similarly to the solution (19) of CORE model), and each predictor explains $1/n^{\text{th}}$ portion of the dependent variable. The constant γ in (19) is also used for sharing the regressors influence on the dependent variable, and it approximately equals $1/n$ as well.

In place of skipping cross-products in reducing LS objective to (25), it is possible to use them with a diminished influence by

inserting a varying parameter g into the result (24):

$$\begin{aligned}
 S_{multi}^2 &= \sum_{j=1}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 + \\
 &\quad g \cdot 2 \sum_{j>k}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right) \left(\frac{1}{n} y_i - \beta_k x_{ik} \right)
 \end{aligned} \tag{26}$$

For $g = 0$ the multi-objective S_{multi}^2 reduces to the pair objective (25), for $g=1$ (26) coincides with the regular LS objective (2), and for intermediate g values from 0 to 1 it corresponds to a model between the pair-wise CORE and regular OLS regressions. The objective (26) is identical to the expression:

$$\begin{aligned}
 S_{multi}^2 &= g \cdot \sum_{j=1}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 + \\
 &\quad g \cdot 2 \sum_{j>k}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right) \left(\frac{1}{n} y_i - \beta_k x_{ik} \right) \\
 &\quad + (1-g) \cdot \sum_{j=1}^n \sum_{i=1}^N \left(\frac{1}{n} y_i - \beta_j x_{ij} \right)^2 \\
 &= g \cdot S^2 + (1-g) \cdot S_{pair}^2,
 \end{aligned} \tag{27}$$

where S^2 and S_{pair}^2 are OLS and CORE objectives defined in (24)-(25). Minimizing the objective (27) yields a system of equations and its corresponding solution as in (13), with the parameters $k = (1-g)/g$ and $q = 1 + k/n$. Further results can be derived as in the relations (14)-(20).

Numerical Simulation

All pair correlations in vector r can be positive, or the scales of the predictors with negative correlations with y can be reversed to make all correlations positive. The positive regression solution (or of the same signs as pair correlations) can be obtained if the system $C\beta = r$ of normal equations (3) satisfies the conditions of the Farkas lemma (Craven, 1978). In practice, it is convenient to use more explicit

MULTIPLE REGRESSION IN PAIR CORRELATION SOLUTION

criteria, for instance, a criterion proposed by Redheffer (2000), which can be written in terms of correlations: for the satisfied conditions

$$\sum_{j \neq i} \max(r_{ij}, 0) < r_{yj} \leq r_{ii} + \sum_{j \neq i} \min(r_{ij}, 0) \quad (28)$$

the system $C\beta = r$ has a positive solution $\beta > 0$. In (28) r_{ij} and r_{yj} are the elements of correlation matrix C and vector r , respectively, and the diagonal elements $r_{ii} = 1$. The criterion (28) is a sufficient but not a necessary condition for positive regression coefficients.

Consider an example of a regression by two predictors, $y = \beta_1 x_1 + \beta_2 x_2$, when the normal system and its solution (3) explicitly are:

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}, \quad \beta_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \quad (29)$$

If r_{12} reaches one, the OLS coefficients (29) are becoming inflated, and of different signs, although r_{y1} becomes close to r_{y2} , so it could be more reasonable to have both coefficients (29) of the same impact on the dependent variable y . At the same time, the eventual ridge regression solution (19) in this case of two predictors is:

$$b_1 = \frac{r_{y1}^2 + r_{y2}^2}{r_{y1}^2 + r_{y2}^2 + 2r_{y1}r_{y2}r_{12}} r_{y1}, \quad (30)$$

$$b_2 = \frac{r_{y1}^2 + r_{y2}^2}{r_{y1}^2 + r_{y2}^2 + 2r_{y1}r_{y2}r_{12}} r_{y2}$$

so the coefficients of regression have the same signs as the pair correlations of the predictors with the dependent variable, and their values are not inflated.

For the model $y = \beta_1 x_1 + \beta_2 x_2$, the correlation matrix of all three variables is a non-negatively definite matrix, so its determinant can be presented in the following inequality:

$$\begin{vmatrix} 1 & r_{12} & r_{y1} \\ r_{12} & 1 & r_{y2} \\ r_{y1} & r_{y2} & 1 \end{vmatrix} = (1 - r_{y1}^2)(1 - r_{y2}^2) - (r_{12} - r_{y1}r_{y2})^2 \geq 0 \quad (31)$$

so for any two given correlations, r_{y1} and r_{y2} , the third one r_{12} can have values within the range satisfying the inequality (31).

Numerical simulation results of the OLS solution (29) for the set of r_{y1} and r_{y2} in the wide range of their values, and several values of r_{12} are given in Tables 1-6. Each table presents the coefficients β_1 , and the other coefficient β_2 can be obtained in the transposed across the second diagonal of the matrix location. Table 1 shows the results for $r_{12} = 0$, Table 2 – the results for $r_{12} = 0.2$, etc., through the last Table 6 for $r_{12} = 0.99$. The tables for the negative values of r_{12} can be obtained from the given tables by their reflection across the vertical axis of the central column for $r_{y2} = 0$ in Tables 1-6.

Tables 1-6 have filled cells only at the locations where the condition (31) is satisfied. The bold font in the tables marks those cells where the OLS coefficients (29) have the signs of pair correlations, $sign(\beta_1) = sign(r_{y1})$ and $sign(\beta_2) = sign(r_{y2})$. The tables show that with the parameter r_{12} increasing from zero to one the shape of the feasible solutions area changes from anisotropic circular to a straight line direction, corresponding to the regression as the expectation of the dependent variable conditioned on the independent variables in their tri-variate normal distribution. What is more interesting – the proportion of the cells where one or two coefficients β_1 and β_2 have signs opposite to the signs of the pair correlations r_{y1} and r_{y2} is rather high (the solutions non-marked by bold font). The frequency to obtain hardly interpretable regression coefficients is substantial, and there is no way to reduce the

occurrence of such a solution in regular regression modeling. However, the CORE solution (30) yields coefficients of regression that are always of the same signs as the pair relations: $sign(b_1) = sign(r_{y_1})$ and $sign(b_2) = sign(r_{y_2})$, in each feasible cell of Tables 1-6 where the condition (31) is satisfied.

Conclusion

The two-parameter ridge regression model and its solution proportional to the pair correlation coefficients are considered. The results of the eventual ridge regression are robust, not prone to multicollinearity effects, and are easily interpretable. The suggested approach is useful for theoretical consideration of regression models and for the practical needs of regression analysis.

References

- Brown, P. J. (1994). *Measurement, regression and calibration*. Oxford: Oxford University Press.
- Craven, B. D. (1978). *Mathematical programming and control theory*. London: Chapman & Hall.
- Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 79, 161-182.
- Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis, methods and applications*. NY: Wiley.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407-489.
- Ehrenberg, A. S. C. (1982). How good is best. *J. of Royal Statistical Society*, A, 145, 364-366.
- Frank, I., & Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.
- Grapentine, A. (1997). Managing multicollinearity. *Marketing Research*, 9, 11-21.
- Hawkins, D. M., & Yin, X. (2002). A faster algorithm for ridge regression of reduced rank data. *Computational Statistics & Data Analysis*, 40, 253-262.
- Harris, R. (1975). *A primer of multivariate statistics*. NY: Academic Press.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Hoerl, A. E., & Kennard, R. W. (1988). Ridge regression, in: *Encyclopedia of Statistical Sciences*, Kotz, S., & Johnson, N. L (Eds.), 8, 129-136.
- Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42, 80-86.
- Lipovetsky, S., & Conklin, M. (2005). Incremental net effects in multiple regression. *International Journal of Mathematical Education in Science and Technology*, 36, 361-373.
- Lipovetsky, S. (2006). Two-parameter ridge regression and its convergence to the eventual pairwise model. *Mathematical and Computer Modelling*, 44, 304-318.
- Lipovetsky, S. (2007). Optimal Lp-metric for minimizing powered deviations in regression. *Journal of Modern Applied Statistical Methods*, 6, 219-227.
- Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28, 268-280.
- Redheffer, R. (2000). All solutions in the unit cube. *The American Mathematical Monthly*, problem # 10764, 107(9), 868-869.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58, 267-288.
- Weisberg, S. (1985). *Applied linear regression*. NY: Wiley.
- Wildt, A. R. (1993). Equity estimation and assessing market response. *Journal of Marketing Research*, 30, 437-451.

MULTIPLE REGRESSION IN PAIR CORRELATION SOLUTION

Table 1: OLS solutions for β_1 , when $r_{12} = 0$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99						0.99					
0.80				0.80	0.80	0.80	0.80	0.80			
0.60			0.60	0.60	0.60	0.60	0.60	0.60	0.60		
0.40		0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	
0.20		0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
-0.20		-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	
-0.40		-0.40	-0.40	-0.40	-0.40	-0.40	-0.40	-0.40	-0.40	-0.40	
-0.60			-0.60	-0.60	-0.60	-0.60	-0.60	-0.60	-0.60		
-0.80				-0.80	-0.80	-0.80	-0.80	-0.80			
-0.99						-0.99					

Table 2: OLS solutions for β_1 , when $r_{12} = 0.2$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99							0.99				
0.80				0.92	0.88	0.83	0.79	0.75	0.71		
0.60			0.75	0.71	0.67	0.63	0.58	0.54	0.50	0.46	
0.40		0.58	0.54	0.50	0.46	0.42	0.38	0.33	0.29	0.25	
0.20		0.38	0.33	0.29	0.25	0.21	0.17	0.13	0.08	0.04	0.00
0.00		0.17	0.13	0.08	0.04	0.00	-0.04	-0.08	-0.13	-0.17	
-0.20	0.00	-0.04	-0.08	-0.13	-0.17	-0.21	-0.25	-0.29	-0.33	-0.38	
-0.40		-0.25	-0.29	-0.33	-0.38	-0.42	-0.46	-0.50	-0.54	-0.58	
-0.60		-0.46	-0.50	-0.54	-0.58	-0.63	-0.67	-0.71	-0.75		
-0.80			-0.71	-0.75	-0.79	-0.83	-0.88	-0.92			
-0.99					-0.99						

Table 3: OLS solutions for β_1 , when $r_{12} = 0.4$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99								0.99			
0.80					1.05	0.95	0.86	0.76	0.67	0.57	
0.60				0.90	0.81	0.71	0.62	0.52	0.43	0.33	
0.40			0.76	0.67	0.57	0.48	0.38	0.29	0.19	0.10	0.00
0.20		0.62	0.52	0.43	0.33	0.24	0.14	0.05	-0.05	-0.14	
0.00		0.38	0.29	0.19	0.10	0.00	-0.10	-0.19	-0.29	-0.38	
-0.20		0.14	0.05	-0.05	-0.14	-0.24	-0.33	-0.43	-0.52	-0.62	
-0.40	0.00	-0.10	-0.19	-0.29	-0.38	-0.48	-0.57	-0.67	-0.76		
-0.60		-0.33	-0.43	-0.52	-0.62	-0.71	-0.81	-0.90			
-0.80		-0.57	-0.67	-0.76	-0.86	-0.95	-1.05				
-0.99				-0.99							

LIPOVETSKY

Table 4: OLS solutions for β_1 , when $r_{12} = 0.6$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99									0.98		
0.80							1.06	0.88	0.69	0.50	
0.60					1.13	0.94	0.75	0.56	0.38	0.19	0.01
0.40				1.00	0.81	0.63	0.44	0.25	0.06	-0.13	
0.20			0.88	0.69	0.50	0.31	0.13	-0.06	-0.25	-0.44	
0.00			0.56	0.38	0.19	0.00	-0.19	-0.38	-0.56		
-0.20		0.44	0.25	0.06	-0.13	-0.31	-0.50	-0.69	-0.88		
-0.40		0.13	-0.06	-0.25	-0.44	-0.63	-0.81	-1.00			
-0.60	-0.01	-0.19	-0.38	-0.56	-0.75	-0.94	-1.13				
-0.80		-0.50	-0.69	-0.88	-1.06						
-0.99			-0.98								

Table 5: OLS solutions for β_1 , when $r_{12} = 0.8$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99										0.97	
0.80								1.33	0.89	0.44	0.02
0.60							1.22	0.78	0.33	-0.11	
0.40					1.56	1.11	0.67	0.22	-0.22	-0.67	
0.20				1.44	1.00	0.56	0.11	-0.33	-0.78		
0.00				0.89	0.44	0.00	-0.44	-0.89			
-0.20			0.78	0.33	-0.11	-0.56	-1.00	-1.44			
-0.40		0.67	0.22	-0.22	-0.67	-1.11	-1.56				
-0.60		0.11	-0.33	-0.78	-1.22						
-0.80	-0.02	-0.44	-0.89	-1.33							
-0.99		-0.97									

Table 6: OLS solutions for β_1 , when $r_{12} = 0.99$.

$r_{y1} \backslash r_{y2}$	-0.99	-0.80	-0.60	-0.40	-0.20	0.00	0.20	0.40	0.60	0.80	0.99
0.99											0.50
0.80										0.40	
0.60									0.30		
0.40								0.20			
0.20							0.10				
0.00						0.00					
-0.20					-0.10						
-0.40				-0.20							
-0.60			-0.30								
-0.80		-0.40									
-0.99	-0.50										

Quel Test for Two Linear Restrictions in the Nonlinear Models

Krishna K. Saha
Central Connecticut State University

An alternative Wald type test called the quel test is developed for two linear restrictions by finding the critical region based on the quel utilizing the repeated values of estimated parameters of interest under the null. Simulation shows evidence that the full quel test performs best in that it holds nominal level well and shows monotonic increasing power properties.

Key words: Bootstrap technique, nonlinear models, percentile confidence contour, power, quel, size.

Introduction

Considerable interest exists in testing linear restrictions in the nonlinear models such as logit, Tobit and exponential models. For testing such hypotheses in the context of nonlinear models, an asymptotic test such as the Wald test or the likelihood ratio test is usually employed. The Wald test has an advantage over the likelihood ratio test since the Wald test requires the maximum likelihood estimates of the parameters only under the alternate hypothesis.

Unfortunately, for small samples the Wald test does not perform well in terms of size and power property. In some situations, the power of the Wald test first increases then eventually starts decreasing when alternative hypothesis parameters increase in distance from the null hypothesis. The Wald test behaves this way because, for certain parameter values, the estimated covariance matrix of the maximum likelihood estimator increases faster than the square of the distance between the parameter estimate and null value (see, for example, Hauck & Donner, 1977; Vaeth, 1985; Mantel, 1987; Nelson & Savin, 1988, 1990). Moreover, the biased estimates of the parameters being tested can cause the power of the Wald test to drop below its size at local alternatives (for example

see, Goh, 1998). These two types of behavior discussed above are usually known as non-monotonicity in the power function and local biasedness respectively.

Other important situations exist in which the estimated covariance matrix cannot be assessed and may not have an explicit form. For example, testing for the presence of first-order moving average disturbances in a linear regression model the information matrix is not well defined if the parameter of the moving average process is 1 or -1 (see Goh, 1998).

This article introduces to construct the alternative Wald type tests that do not depend on the estimated covariance matrix, and use nonparametric ideas and computer simulation to judge whether the estimates observed are likely to have come from a null hypothesis data generating process. Applying the above concepts, we construct the bivariate generalizations of the boxplot based on a generalized quel introduced by Goldberg and Iglewicz (1992) which is defined as four separate quarter ellipses matched on their major and minor axes so that the quel is continuous and smooth. Previously, many authors including Turkey (1947), Scott (1985) and Beckett and Gould (1987) attempted to estimate the confidence contours of a bivariate density, but those approaches had serious shortcomings.

The primary aim of this article is to construct new tests that solve the problem of non-monotonicity in the power function, but do not face the limitations discussed above in

Krishna K. Saha is an Associate Professor in the Department of Mathematical Sciences. Email: sahakrk@mail.ccsu.edu.

practice. These new tests only require simulated estimates of the parameters of interest under the null hypothesis but do not involve estimating the covariance matrix. Moreover, these new tests require defining the rejection region based on the contour points of the percentile confidence limits of quels (half and full) of the values of estimated parameters of interest under the null hypothesis. Furthermore, the null hypothesis is rejected if the sample data estimates fall outside the percentile confidence limit of a quel (half or full).

Methodology

The Wald and LR Tests

Let y_1, \dots, y_n be n independent observations distributed with density function $f(y_t | x_t, \theta)$, $t=1, \dots, n$, where x_t is a vector of covariates and θ is an unknown $k \times 1$ parameter vector. Let $\theta = (\beta', \eta')$, where $\beta = (\beta_1, \beta_2)'$ are two parameters of interest and η is a $(k-2) \times 1$ vector of nuisance parameters. The log-likelihood function is given by $l(\theta) = \sum_{t=1}^n \ln f(y_t | x_t, \theta)$. The interest lies in testing the composite hypotheses

$$H_0 : \beta = \beta_0 \text{ against } H_1 : \beta \neq \beta_0, \quad (1)$$

where β_0 is a 2×1 vector of known constants. Let $\hat{\theta} = (\hat{\beta}', \hat{\eta}')$ be the maximum likelihood estimators of θ under the alternative hypotheses. Then the Wald test statistic is

$$W = (\hat{\beta} - \beta_0)(R\hat{V}(\hat{\theta})R')^{-1}(\hat{\beta} - \beta_0), \quad (2)$$

where $R = (I_2 : 0)$, I_2 is the 2×2 identity matrix, $\hat{V}(\hat{\theta})$ is a constant estimator of $V(\hat{\theta})$ with replace θ by $\hat{\theta}$ and

$$V(\hat{\theta}) = \left(E \left[- \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right] \right)$$

is the covariance matrix of $\hat{\theta}$. Under the standard regularity conditions (see, for example, Godfrey, 1988), W asymptotically follows a χ^2 distribution with 2 degrees of freedom under the null hypothesis. The null hypothesis is rejected for large values of W (for details see Goh, 1998).

Neyman and Pearson (1928) first proposed the likelihood ratio (LR) test for testing a composite hypothesis. Note that the Wald and LR tests have the same first-order asymptotic properties and they are asymptotically equivalent (see Rao, 1973). Several authors have studied the asymptotic relationship between these two tests (see, for example, Gourieroux and Monfort 1995, Chapter 17; Hendry 1995, Chapter 13). Let $\hat{\theta}_0$ be the maximum likelihood estimators of θ under the null hypotheses. Then the LR test for the hypothesis in (1) involves rejecting the null hypothesis for large values of

$$LR = 2 \left[l(\hat{\theta}) - l(\hat{\theta}_0) \right], \quad (3)$$

which, under the standard regularity conditions, follows a χ^2 distribution with 2 degrees of freedom asymptotically under the null hypothesis.

The Quel (Full or Half) Test

As observed for some nonlinear models, the estimated covariance matrix is not always available. Thus, some new test procedures namely, full quel and half quel tests, for two linear restrictions, are outlined which do not require an expression of this matrix. As only the quel for a two-dimensional case can be constructed (see Goldberg & Iglewicz, 1992), attention is limited to testing problems involving only two restrictions.

The Percentile Confidence Contour Points of a Quel (Full or Half)

Let (u_i, v_i) , $i = 1, 2, \dots, N$, be a set of simulated maximum likelihood estimates of (β_1, β_2) for the i^{th} sample under the null hypothesis. Specifically, $(u_1, v_1), (u_2, v_2), \dots,$

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

(u_N, v_N) , are bivariate observations of size N for (U, V) . Following the method of an asymmetric plot provided by Goldberg and Iglewicz (1992), the two-dimensional confidence contour based on the standardized errors of each point of (U, V) can be found. To obtain these errors, requires first finding the location, scale, and correlation estimators for (U, V) as well as the two additional parameters represented by the proportions of the total standard deviation due to residuals in the positive direction of the major and minor axes of the asymmetric plot. Goldberg and Iglewicz (1992) introduced the estimation of those parameters by an extended biweight bivariate estimator (BIWT) and one-step biweight estimator (BIWT-1) as being efficient. Based on the extended BIWT and BIWT-1 method provided by Goldberg and Iglewicz (1992), the location, scale, and correlation estimators for (U, V) can be easily obtained, as well as the two additional parameters.

Let μ_{ub}^{*q} , μ_{vb}^{*q} , σ_{ub}^q , σ_{vb}^q and r_{uvb}^q be the location, scale, and correlation estimates of U and V respectively, and let γ_1 and γ_2 be the estimates of the two additional parameters. In order to find the boundary points of the confidence contour of a quel, regardless of size, define $G_1 = U_s^q + V_s^q$ and $G_2 = U_s^q - V_s^q$, as the major and minor axes in this order, where U_s^q and V_s^q are the standardized values of U and V based on the location and scale estimators for a quel as $u_{si}^q = (u_i - \mu_{ub}^{*q}) / \sigma_{ub}^q$ and $v_{si}^q = (v_i - \mu_{vb}^{*q}) / \sigma_{vb}^q$, respectively. Note that $sign(G_1 - G_2) = sign(r_{uvb}^q)$. Therefore, the major and minor axes must be redefined with respect to the correlation estimator r_{uvb}^q as $G_{1i}^* = (u_{si}^q + v_{si}^q) / r_{1uvb}^{*q}$ and $G_{2i}^* = (u_{si}^q - v_{si}^q) / r_{2uvb}^{*q}$, where $r_{1uvb}^{*q} = \sqrt{2(1 + r_{uvb}^q)}$ and $r_{2uvb}^{*q} = \sqrt{2(1 - r_{uvb}^q)}$, respectively. In addition, the standardized errors for the construction of a quel whose percentile point approximately determines the percentile confidence contour of the rejection region for

the quel (see Figure 1) must be computed. Compute the standardized errors, ξ_i^q , based on G_{1i}^* and G_{2i}^* using the additional parameter estimates γ_1 and γ_2 as

$$\xi_i^q = \sqrt{\nabla_{1i}^2 + \nabla_{2i}^2}, \quad \text{for } i = 1, 2, \dots, N, \quad (4)$$

where for $l = 1, 2$,

$$\nabla_{li} = \begin{cases} G_{li}^* / 2\gamma_l & \text{if } G_{li}^* > 0 \\ G_{li}^* / 2(1 - \gamma_l) & \text{otherwise.} \end{cases}$$

Note that these errors assess the distances of each point obtained from the observations of U and V to the center $(\mu_{ub}^{*q}, \mu_{vb}^{*q})$. Let $\xi_{percentile}^q$ be the percentile of the standardized errors ξ_i^q ($i = 1, 2, \dots, N$) in equation (4).

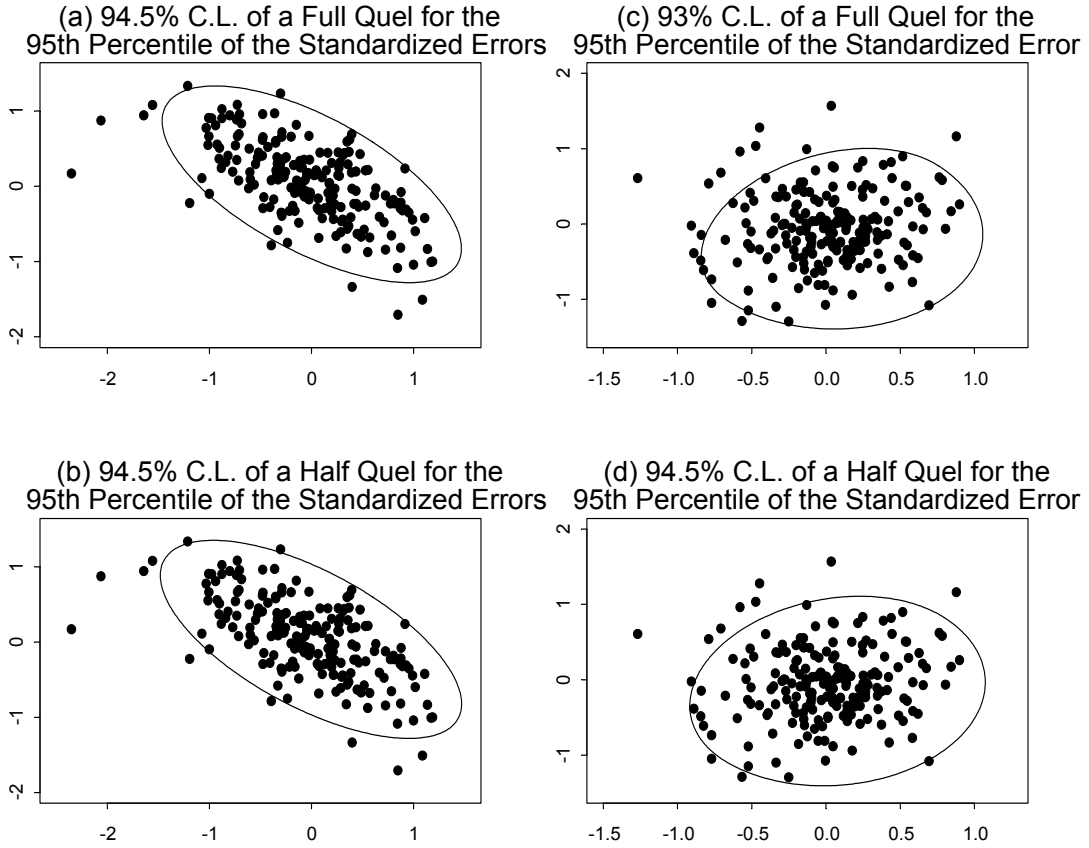
The construction of a quel depends on two things, the percentile of the errors $\xi_{percentile}^q$ and the estimators of two additional parameters γ_1 and γ_2 . As a result, options of different values of $\xi_{percentile}^q$, γ_1 and γ_2 create different kinds of quel. In this case, these values are chosen from two different options, which assess two different quels called full and half quels. These two options for constructing full and half quels are discussed in the Appendix.

Upon acquiring the percentile of the standardized errors as well as the two additional estimators from either options for a full or half quel as shown in the Appendix, it is easy to find the boundary points of the percentile confidence limit for the full or half quel. In doing so, based on $\xi_{percentile}^q$ as well as γ_l^{fq} for $l = 1, 2$ from option-I in the Appendix, the lengths of the vertices in all four quadrants from the origin for the full quel are:

$$\phi_1^{(-1)} = \xi_{percentile}^{fq} [2(1 - \gamma_1^{fq})] \sqrt{(1 + r_{uvb}^q) / 2}$$

$$\phi_1^{(+1)} = \xi_{percentile}^{fq} [2(\gamma_1^{fq})] \sqrt{(1 + r_{uvb}^q) / 2}$$

Figure 1: Exact Percentile Confidence Contour of a Full or Half Quel for the 95th Percentile of the Standardized Errors ξ_i^q ($i = 1, 2, \dots, N$) for a Full or Half Quel when $N = 200$.*



*The scatter points represent the simulated ML estimates of (β_1, β_2) of size N . The underlying distribution was the two-regressor binary logit model of size $n = 30$ using design matrix X_3 [(a), (b)] and the three-regressor binary logit model of size $n = 30$ using design matrix X_2 [(c), (d)].

$$\phi_1^{(+1)} = \xi_{\text{percentile}}^{fq} \left[2\gamma_1^{fq} \sqrt{(1+r_{uvb}^q)/2} \right]$$

$$\phi_2^{(+1)} = \xi_{\text{percentile}}^{fq} \left[2\gamma_2^{fq} \sqrt{(1-r_{uvb}^q)/2} \right].$$

Next, based on the parametric equations of an ellipse in terms of angle θ^{fq} with range 0 to 360 degrees using $\phi_1^{(-1)}$, $\phi_1^{(+1)}$, $\phi_2^{(-1)}$, and $\phi_2^{(+1)}$ as

$$\Phi_1 = \phi_1^{\text{sign}(\cos\theta^{fq})} \cos\theta^{fq}$$

and

$$\Phi_2 = \phi_2^{\text{sign}(\sin\theta^{fq})} \sin\theta^{fq}, \quad (5)$$

the boundary points of the percentile confidence contour for the full quel are given by

$$X^{fq} = \mu_{ub}^q + (\Phi_1 + \Phi_2)\sigma_{ub}^q$$

and

$$Y^{fq} = \mu_{vb}^q + (\Phi_1 - \Phi_2)\sigma_{vb}^q. \quad (6)$$

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

In a manner similar to the case of a full quel discussed above, the boundary points of the percentile confidence limit for a half quel can also be obtained as in equation (6) by using equation (5) with $\xi_{percentile}^{fq}$ and γ_1^{fq} for $l = 1, 2$ replaced by $\xi_{percentile}^{hq}$ and γ_1^{hq} for $l = 1, 2$ from option-II in the Appendix.

A Point on the Percentile Confidence Limit of a Quel for a Fixed Angle

Consider the case of using a full quel to find the boundary point P_{H_0} . In doing so, it is necessary to find a solution of an angle θ^{fq} , based on the angle A_{H_1} so that the x and y coordinates of a point P_{H_0} using equations (5) and (6) for a particular solution of this θ^{fq} can easily be obtained. To solve this θ^{fq} based on the angle A_{H_1} , define

$$\begin{aligned} X^{fq} &= D_{H_0} \cos A_{H_1} \\ Y^{fq} &= D_{H_0} \sin A_{H_1} \end{aligned} \quad (7)$$

Using equation (5), results in an updated equation for the solution of an angle θ^{fq} from equations (6) and (7) as given by

$$\begin{aligned} &\mu_{vb}^q + \left[\phi_1^{sign(\cos \theta^{fq})} \cos \theta^{fq} - \phi_2^{sign(\sin \theta^{fq})} \sin \theta^{fq} \right] \sigma_{vb}^q \\ &= \left\{ \mu_{ub}^q + \left[\begin{array}{l} \phi_1^{sign(\cos \theta^{fq})} \cos \theta^{fq} + \\ \phi_2^{sign(\sin \theta^{fq})} \sin \theta^{fq} \end{array} \right] \sigma_{ub}^q \right\} \tan A_{H_1} \end{aligned} \quad (8)$$

To obtain the solution of θ^{fq} from (8), start with the combination of ϕ_1 and ϕ_2 from the previous section, which depends on the sign of $\cos \theta^{fq}$ and $\sin \theta^{fq}$. Based on the values of $\cos \theta^{fq}$ and $\sin \theta^{fq}$, the solution for an angle θ^{fq} in four different cases is obtained as follows.

Case I: Values of $\cos \theta^{fq}$ and $\sin \theta^{fq}$ are both positive

$$\begin{aligned} \phi_1^{sign(\cos \theta^{fq})} &= \phi_1^{(+1)} \\ \phi_2^{sign(\sin \theta^{fq})} &= \phi_2^{(+1)}. \end{aligned} \quad (9)$$

Using (9), solve for the angle θ^{fq} from equation (8), which is

$$\theta^{fq} = \arcsin \frac{2b_1 \pm \sqrt{4a_1^2 (a_1^2 - 1 + b_1^2)}}{2(a_1^2 + b_1^2)}, \quad (10)$$

where

$$a_1 = \frac{\phi_1^{(+1)} (\sigma_{vb}^q - \sigma_{ub}^q \tan A_{H_1})}{\mu_{ub}^q \tan A_{H_1} - \mu_{vb}^q}$$

and

$$b_1 = \frac{\phi_2^{(+1)} (\sigma_{vb}^q + \sigma_{ub}^q \tan A_{H_1})}{\mu_{ub}^q \tan A_{H_1} - \mu_{vb}^q}.$$

Case II: Values of $\cos \theta^{fq}$ and $\sin \theta^{fq}$ are positive and negative respectively

$$\begin{aligned} \phi_1^{sign(\cos \theta^{fq})} &= \phi_1^{(+1)} \\ \phi_2^{sign(\sin \theta^{fq})} &= \phi_2^{(-1)} \end{aligned} \quad (11)$$

Similar to Case-I, find the solution of an angle θ^{fq} from (8) using equation (11) as

$$\theta^{fq} = \arcsin \frac{2b_2 \pm \sqrt{4a_1^2 (a_1^2 - 1 + b_1^2)}}{2(a_1^2 + b_1^2)}, \quad (12)$$

where

$$b_2 = \frac{\phi_2^{(-1)} (\sigma_{vb}^q + \sigma_{ub}^q \tan A_{H_1})}{\mu_{ub}^q \tan A_{H_1} - \mu_{vb}^q}.$$

Case III: Values of $\cos \theta^{fq}$ and $\sin \theta^{fq}$ are, respectively, negative and positive

$$\begin{aligned} \phi_1^{sign(\cos \theta^{fq})} &= \phi_1^{(-1)} \\ \phi_2^{sign(\sin \theta^{fq})} &= \phi_2^{(+1)} \end{aligned} \quad (13)$$

Using equation (13), obtain the angle θ^{fq} from equation (8) as

$$\theta^{fq} = \arcsin \frac{2b_1 \pm \sqrt{4a_2^2(a_2^2 - 1 + b_1^2)}}{2(a_2^2 + b_1^2)}, \quad (14)$$

where

$$a_2 = \frac{\phi_1^{(-1)}(\sigma_{vb}^q - \sigma_{ub}^q \tan A_{H_1})}{\mu_{ub}^q \tan A_{H_1} - \mu_{vb}^q}.$$

Case IV: Values of $\cos \theta^{fq}$ and $\sin \theta^{fq}$ are both negative

$$\begin{aligned} \phi_1^{\text{sign}(\cos \theta^{fq})} &= \phi_1^{(-1)} \\ \phi_2^{\text{sign}(\sin \theta^{fq})} &= \phi_2^{(-1)}. \end{aligned} \quad (15)$$

In this final case, evaluate the angle θ^{fq} from equation (8) by using equation (15), to get

$$\theta^{fq} = \arcsin \frac{2b_2 \pm \sqrt{4a_2^2(a_2^2 - 1 + b_2^2)}}{2(a_2^2 + b_2^2)}. \quad (16)$$

Upon achieving the solutions of the angle θ^{fq} from all four cases in equations (10), (12), (14) and (16), these solutions need to be adjusted by considering all four quadrants as

$$\theta^* = \begin{cases} \theta^* \text{ and } 180^\circ - \theta^{fq}, & \text{if } \theta^{fq} > 0 \\ 180^\circ - \theta^{fq} \text{ and } 360^\circ - \theta^{fq}, & \text{otherwise.} \end{cases} \quad (17)$$

Angle θ^{fq} has two solutions for each case but imposing equation (17) means there are four solutions for θ^{fq} for each case, which in turn gives sixteen solutions from all four cases. In practice, only one of the solutions of θ^{fq} is accountable for the angle A_{H_1} . In order to obtain this particular solution, use all sixteen solutions in equation (5) to find the corresponding x and y coordinates for the boundary points based on equation (6). Consequently, find an angle for this boundary point for each solution of θ^{fq} deeming all four quadrants. Assume $P_{H_0}^{fq}(X^{fq}, Y^{fq})$ is a

boundary point having an angle $A_{H_0}^{fq}$ for a specific value of θ^{fq} . Now, finding this particular value of θ^{fq} , which applies to A_{H_1} requires finding which $A_{H_0}^{fq}$ is such that $|A_{H_1} - A_{H_0}^{fq}| \approx 0$, that is, $A_{H_0}^{fq}$ is equal or close to A_{H_1} . As a result, $P_{H_0}^{fq}(X^{fq}, Y^{fq})$ would be a boundary point of a full quel for angle A_{H_1} (see, Figures 2a and 2c). In a manner similar to the case for the full quel, the angle θ^{fq} may be solved for, which is responsible for an angle A_{H_1} and boundary point $P_{H_0}^{fq}(X^{fq}, Y^{fq})$ of a half quel for angle A_{H_1} found (see, Figures 2b and 2d) for the particular solution of θ^{fq} .

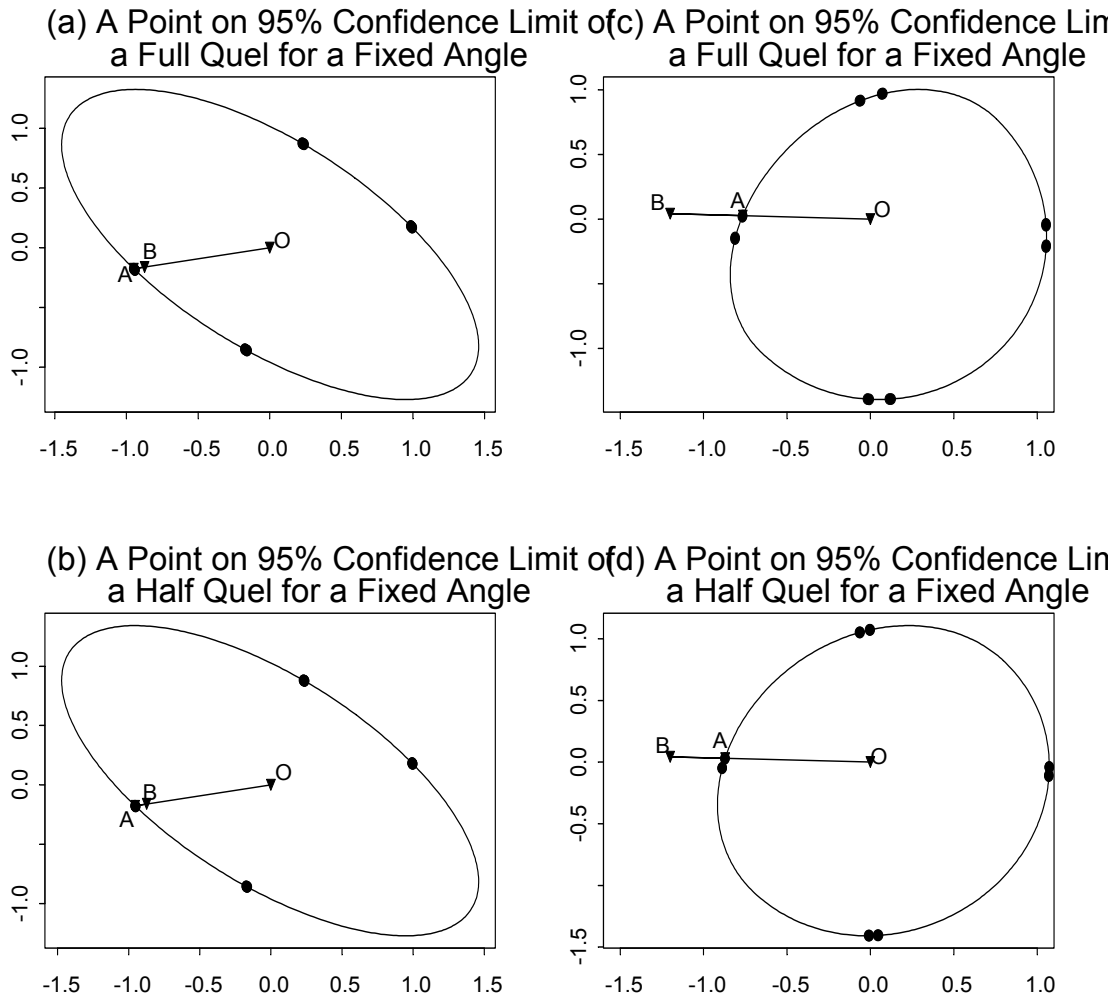
Outline of the New Tests

An outline of the new test procedure is as follows:

1. Estimate the parameter vector θ for the given data set, $\hat{\theta} = (\hat{\beta}', \hat{\eta}')$. Assume P_{H_1} is the sample point for $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ and compute $D_{H_1} = \sqrt{\hat{\beta}_1 + \hat{\beta}_2}$ and $A_{H_1} = \arctan(\hat{\beta}_1 + \hat{\beta}_2)$.
2. Utilizing the estimate of η in step 1 and the null values of β , construct $\hat{\theta}_0 = (\beta_0', \hat{\eta}')$. Generate a sample of size n under the null from the density function $f(y_i | x_i, \theta)$ by setting $\theta = \hat{\theta}_0$ and estimate $\beta = (\beta_1, \beta_2)'$ for this sample. Repeat this process N times and let (u_i, v_i) , $i = 1, 2, \dots, N$, be the estimates of $\beta = (\beta_1, \beta_2)'$ for the i^{th} sample under the null.
3. Based on the values (u_i, v_i) , $i = 1, 2, \dots, N$, in step 2, obtain the contour points of the $100(1 - \alpha)\%$ confidence limit of the quel

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

Figure 2: Point A on 95% Confidence Contour of a Full Quel for a Fixed Angle A_{H_1} Obtained from a Point B when $N = 200$, where the Points A and B Are the Points P_{H_0} and P_{H_1} as Defined so that $D_{H_0} = AO$ and $D_{H_1} = BO$.*



*The underlying distribution was the two-regressor binary logit model of size $n = 30$ using design matrix X_3 [(a), (b)] and the three-regressor binary logit model of size $n = 30$ using design matrix X_2 [(c), (d)]. H_0 is not rejected for [(a), (b)] and is rejected for [(c), (d)].

(half or full) by considering $\xi_{percentile}^{fq} = \xi_{1-\alpha}^{fq}$ or $\xi_{percentile}^{hq} = \xi_{1-\alpha}^{hq}$ for a full or half quel, respectively.

4. Corresponding to the angle A_{H_1} in step 1, obtain a point $P_{H_1}(X, Y)$ on the contour of the $100(1 - \alpha)\%$ confidence limit of the quel (full or half) following subsection 3.2 and compute $D_{H_0} = \sqrt{X^2 + Y^2}$. Reject H_0 if $D_{H_1} > D_{H_0}$.

Examples

Consider the problems of testing two linear restrictions associated with the two-regressor binary logit and the three-regressor binary logit models.

A binary logit model associated with two regressors $x_t = (x_{1t}, x_{2t})'$ and errors ζ_t is given by

$$\begin{aligned} \Pr(y_t = 1) &= \nu(x_t'\theta) \\ &= \frac{1}{[1 + \exp(-x_t'\theta)]}, t = 1, 2, \dots, n \end{aligned} \quad (18)$$

where

$$y_t = \begin{cases} 1 & \text{if } x_t'\theta + \xi_t > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$\theta = \beta$ and ξ_t is a standard logistic distribution with $\nu(u) = [1 + \exp(-u)]^{-1}$ being the standard logistic distribution function. Let $\hat{\theta}$ be the ML estimate of θ . The covariance matrix of $\hat{\theta} = \hat{\beta}$ for this model is given by $V(\hat{\theta}) = V(\hat{\beta}) = [\sum_t \nu_t(1 - \nu_t)x_t x_t']^{-1}$, where $\nu_t = \nu(x_t'\hat{\theta})$. Thus, the Wald test statistic is similar to (2) with $V(\hat{\theta})$ and $R = I_2$. The LR test statistic for this model is given by (3), where $\hat{\theta} = \hat{\beta}$, $\hat{\theta}_0 = \beta_0$ and

$$l(\theta) = \sum_t [y_t \ln \nu_t + (1 - y_t) \ln(1 - \nu_t)]$$

with

$$v_t = \nu(x_t'\theta).$$

Also, consider the three regressors binary logit model having the same form as (18) but with $x_t = (x_{1t}, x_{2t}, x_{3t})'$ and $\theta = (\beta', \eta)'$. Here η is a scalar nuisance parameter. Let $\hat{\theta} = (\hat{\beta}', \hat{\eta})'$ be the ML estimate of θ under the alternative hypothesis. In this three-regressor binary logit model, the Wald test statistic is defined in (2) with the covariance matrix of $\hat{\theta}$ as,

$$V(\hat{\theta}) = V(\hat{\beta}) = [\sum_t \nu_t(1 - \nu_t)x_t x_t']^{-1},$$

where $\hat{\nu}_t = \nu(x_t'\hat{\theta})$ and $R = (I_2 : 0)$. The LR test statistic of this model is given by (3), where $\hat{\theta}_0 = (\beta_0', \hat{\eta})'$ is the ML estimate of θ under the null hypothesis.

In these two linear restrictions testing problems, all the test statistics defined follow the asymptotic Chi-squared distribution with two degrees of freedom under the null and standard regularity conditions.

Simulation Study

The object of the simulation study is to investigate the small-sample properties of the Wald, LR, and quel (full or half) tests for hypothesis testing problems involving two linear restrictions in both models discussed in terms of size and power. For testing two linear restrictions, $\beta_0 = (0, 0)'$ was used so that the null hypothesis is $H_0 : \beta_1 = \beta_2 = 0$ and the alternative hypothesis is $H_1 : \text{at least one of } \beta_1 \text{ or } \beta_2 \neq 0$. In this case, four design matrices for x_t were used as follows: X_1 : Two independent series of independent $N(0, 1)$ random drawings (x_{1t} and x_{2t}), X_2 : Three independent series of independent $N(0, 1)$ random drawings (x_{1t}, x_{2t} and x_{3t}), X_3 : Quarterly Australian private capital movements (\$'000 million) (x_{1t}) and government capital movements (\$'00 million)

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

(x_{2t}) beginning the first quarter of 1968, and X_4 : Quarterly Australian private capital movements (\$'000 million) (x_{1t}), the same capital movements lagged one quarter (x_{2t}) and government capital movements (\$'00 million) (x_{3t}) beginning the first quarter of 1968. All tests were performed at the 5% nominal level using the sample sizes, $n = 30$ and 80 . In the three-regressor binary logit model, value of the nuisance parameter was set at $\eta = 0.1$. Each experiment was based on 1,000 replications. Empirical sizes and powers of the Wald and LR tests were estimated using asymptotic critical values. The empirical sizes and powers of the quel (full and half) test were also computed for both models. In each replication of the experiment, $N = 200$ samples were drawn from the data generating process under the null hypothesis to find the contour points of the percentile confidence limit of a quel (full or half).

Empirical sizes and powers of the Wald, LR and quel (full or half) tests are reported in Tables 1-4 for the selected small-sample and experiments noted above. In the analysis of size, the rejection probabilities of the tests under the null, which are outside the range $[0.0322, 0.0678]$, were significantly different from 5% at the 0.01 level for 1,000 replications. Based on these rejection probabilities, Tables 1 and 2 show that estimated sizes for the Wald and LR tests are significantly different from 5% at the 0.01 level in both models for sample size $n = 30$ and Table 3 shows that estimated sizes for the Wald test were significantly different from 5% at the 0.01 level in the two-regressor binary logit model for sample size $n = 80$. Of these tests, the LR test in general shows extreme liberal behavior, whereas the Wald test shows extreme conservative behavior in most data situations. Both proposed tests reported in Tables 1-4 perform extremely well and hold nominal level reasonably well in all instances. However, the performance of the full quel test is uniformly best in that it holds nominal level well in all data distribution situations with no apparent anti-conservative behavior.

Empirical powers of all tests were computed for the parameter space around the null that divided into five different regions based on the signs of the parameter values. In both models, the powers of the Wald test are non-monotonic at non-local alternatives in most of the regions (see, for example, regions 1 and 5 in Table 1). The Wald power function becomes almost flat at zero, for example, region 5 in Table 2. In the most serious case of region 4 in Table 2, the power for the Wald test is below 26% whereas at the same point in the parameter space, our proposed tests attain a power of 100%. The LR test, as well as the proposed new tests, has monotonic power functions in all the five regions of the parameter space in all cases for both of the models considered here. In some regions, the LR test perform better than all other tests, for example, region 2 in Table 2. However, power estimates of the LR test are erroneous because this test is liberal. The proposed quel tests show excellent power properties in most data situations. Of these new tests, power of the full quel test is better than that of the half quel test in all five regions except for a few points of some regions in the parameter space. Moreover, the full quel test has more balanced power compared to that of the half quel test at the same local alternatives in most of the regions.

Overall, the full quel test has consistently higher power and holds its level quite well. In some situations, the half quel test showed good power property and well controlled level. Among the LR, half quel and full quel tests, the full quel test can be recommended for testing two linear restrictions in these nonlinear models.

Conclusion

The Wald test requires an analytical form of the variance-covariance matrix of the ML estimators of the parameters, and it shows extreme conservative and non-monotonic power behavior caused by inaccuracy of the estimated covariance matrix of the estimator. In this article an alternative Wald type test was proposed to resolve this problem of small-sample local biasedness and non-monotonic power behavior of the Wald test for two linear restrictions. The proposed new tests have desirable size with

Table 1: Estimates of Size and Power for the Wald, LR, Full Quel, and Half Quel Tests at the $\alpha = 5\%$ Level of Significance for Testing $H_0 : \beta_1 = \beta_2 = 0$ in the Two-Regressor Binary Logit Model Using Design Matrix X_1 when $n = 30$.

Region	β_1	β_2	Asymptotic Tests		New tests	
			Wald	LR	Full Quel	Half Quel
	0.00	0.000	0.015*	0.058	0.048	0.046
1	-0.30	0.000	0.068	0.114	0.135	0.131
	-0.60	0.000	0.221	0.313	0.354	0.356
	-0.90	0.000	0.504	0.633	0.690	0.688
	-1.50	0.000	0.875	0.949	0.965	0.965
	-2.00	0.000	0.890	1.000	0.994	0.994
	-3.50	0.000	0.467	1.000	1.000	1.000
2	0.30	0.000	0.067	0.146	0.154	0.149
	0.60	0.000	0.232	0.401	0.410	0.405
	0.90	0.000	0.504	0.691	0.704	0.697
	1.50	0.000	0.877	0.960	0.965	0.963
	2.00	0.000	0.859	0.994	0.998	0.998
	3.50	0.000	0.462	1.000	1.000	1.000
3	-0.30	-0.300	0.087	0.200	0.193	0.183
	-0.50	-0.500	0.231	0.405	0.415	0.416
	-0.90	-0.900	0.692	0.863	0.870	0.869
	-1.40	-1.400	0.802	0.986	0.990	0.991
	-1.90	-1.900	0.580	0.999	1.000	1.000
	-3.35	-3.350	0.118	1.000	1.000	1.000
4	0.30	0.300	0.097	0.202	0.209	0.204
	0.50	0.500	0.253	0.418	0.439	0.435
	0.90	0.900	0.659	0.847	0.863	0.859
	1.40	1.400	0.788	0.991	0.996	0.995
	1.90	1.900	0.561	0.999	1.000	1.000
	3.35	3.350	0.117	1.000	1.000	1.000
5	-0.30	0.300	0.097	0.176	0.183	0.178
	-0.50	0.500	0.268	0.418	0.425	0.425
	-0.95	0.950	0.779	0.915	0.924	0.917
	-1.55	1.550	0.815	1.000	0.998	0.998
	-2.25	2.250	0.535	1.000	1.000	1.000
	-3.35	3.350	0.230	1.000	1.000	1.000

Note: * Size is significantly different from 5% at the 1% level.

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

Table 2: Estimates of Size and Power for the Wald, LR, Full Quel, and Half Quel Tests at the $\alpha = 5\%$ Level of Significance for Testing $H_0 : \beta_1 = \beta_2 = 0$ in the Two-Regressor Binary Logit Model Using Design Matrix X_4 when $n = 30$.

Region			Asymptotic tests		New tests	
	β_1	β_2	Wald	LR	Full Quel	Half Quel
	0.00	0.00	0.019*	0.076*	0.053	0.044
1	-0.40	0.00	0.022	0.088	0.092	0.089
	-0.75	0.00	0.030	0.116	0.134	0.121
	-0.95	0.00	0.039	0.131	0.156	0.141
	-1.25	0.00	0.060	0.182	0.197	0.185
	-1.90	0.00	0.160	0.311	0.349	0.328
	-3.25	0.00	0.522	0.709	0.769	0.759
2	0.15	0.00	0.023	0.077	0.075	0.066
	0.55	0.00	0.037	0.109	0.097	0.082
	0.90	0.00	0.063	0.150	0.127	0.116
	1.55	0.00	0.125	0.253	0.237	0.229
	2.25	0.00	0.289	0.455	0.416	0.409
	3.10	0.00	0.502	0.689	0.667	0.645
3	-0.10	-0.10	0.017	0.077	0.060	0.051
	-0.35	-0.35	0.026	0.098	0.081	0.077
	-0.75	-0.75	0.085	0.201	0.170	0.153
	-1.15	-1.15	0.215	0.384	0.305	0.295
	-1.45	-1.45	0.355	0.538	0.460	0.442
	-2.25	-2.25	0.756	0.876	0.771	0.791
4	0.10	0.10	0.022	0.079	0.094	0.095
	0.35	0.35	0.046	0.114	0.774	0.775
	0.75	0.75	0.107	0.215	0.998	0.999
	1.15	1.15	0.259	0.415	1.000	1.000
	1.45	1.45	0.393	0.570	1.000	1.000
	2.25	2.25	0.751	0.881	1.000	1.000
5	-0.10	0.10	0.019	0.077	0.087	0.078
	-0.35	0.35	0.021	0.085	0.092	0.095
	-0.75	0.75	0.019	0.099	0.113	0.106
	-1.15	1.15	0.022	0.122	0.143	0.134
	-1.45	2.45	0.029	0.147	0.184	0.165
	-2.25	2.25	0.040	0.247	0.296	0.265

Note: * Size is significantly different from 5% at the 1% level.

Table 3: Estimates of Size and Power for the Wald, LR, Full Quel, and Half Quel Tests at the $\alpha = 5\%$ Level of Significance for Testing $H_0 : \beta_1 = \beta_2 = 0$ in the Two-Regressor Binary Logit Model Using Design Matrix X_3 when $n = 80$.

Region			Asymptotic tests		New tests	
	β_1	β_2	Wald	LR	Full Quel	Half Quel
	0.00	0.00	0.015*	0.061	0.058	0.061
1	-0.40	0.00	0.427	0.573	0.600	0.590
	-0.75	0.00	0.905	0.983	0.976	0.975
	-0.95	0.00	0.983	0.997	0.999	0.999
	-1.25	0.00	1.000	1.000	1.000	1.000
	-1.90	0.00	1.000	1.000	1.000	1.000
	-3.25	0.00	1.000	1.000	1.000	1.000
2	0.15	0.00	0.060	0.144	0.154	0.142
	0.55	0.00	0.732	0.814	0.820	0.821
	0.90	0.00	0.986	0.995	0.997	0.996
	1.55	0.00	1.000	1.000	1.000	1.000
	2.25	0.00	1.000	1.000	1.000	1.000
	3.10	0.00	1.000	1.000	1.000	1.000
3	-0.10	-0.10	0.395	0.710	0.779	0.768
	-0.35	-0.35	0.998	1.000	1.000	1.000
	-0.75	-0.75	1.000	1.000	1.000	1.000
	-1.15	-1.15	1.000	1.000	1.000	1.000
	-1.45	-1.45	1.000	1.000	1.000	1.000
	-2.25	-2.25	1.000	1.000	1.000	1.000
4	0.10	0.10	0.369	0.691	0.694	0.694
	0.35	0.35	1.000	1.000	1.000	1.000
	0.75	0.75	1.000	1.000	1.000	1.000
	1.15	1.15	1.000	1.000	1.000	1.000
	1.45	1.45	1.000	1.000	1.000	1.000
	2.25	2.25	1.000	1.000	1.000	1.000
5	-0.10	0.10	0.200	0.454	0.492	0.492
	-0.35	0.35	0.985	1.000	0.999	0.999
	-0.75	0.75	1.000	1.000	1.000	1.000
	-1.15	1.15	1.000	1.000	1.000	1.000
	-1.45	2.45	1.000	1.000	1.000	1.000
	-2.25	2.25	1.000	1.000	1.000	1.000

Note: * Size is significantly different from 5% at the 1% level.

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

Table 4: Estimates of Size and Power for the Wald, LR, Full Quel, and Half Quel Tests at the $\alpha = 5\%$ Level of Significance for Testing $H_0 : \beta_1 = \beta_2 = 0$ in the Two-Regressor Binary Logit Model Using Design Matrix X_2 when $n = 80$.

Region			Asymptotic Tests		New tests	
	β_1	β_2	Wald	LR	Full Quel	Half Quel
	0	0	0.035	0.058	0.050	0.051
1	-0.30	0.00	0.192	0.244	0.275	0.271
	-0.60	0.00	0.668	0.732	0.754	0.749
	-0.90	0.00	0.948	0.968	0.973	0.973
	-1.50	0.00	0.999	1.000	1.000	1.000
	-2.00	0.00	1.000	1.000	1.000	1.000
	-3.50	0.00	1.000	1.000	1.000	1.000
2	0.30	0.00	0.215	0.258	0.263	0.257
	0.60	0.00	0.664	0.735	0.764	0.745
	0.90	0.00	0.936	0.956	0.968	0.960
	1.50	0.00	1.000	1.000	1.000	0.999
	2.00	0.00	1.000	1.000	1.000	1.000
	3.50	0.00	1.000	1.000	1.000	1.000
3	-0.30	0.30	0.358	0.437	0.460	0.453
	-0.50	-0.50	0.833	0.875	0.875	0.873
	-0.90	-0.90	1.000	1.000	1.000	1.000
	-1.40	-1.40	1.000	1.000	1.000	1.000
	-1.90	-1.90	1.000	1.000	1.000	1.000
	-3.35	-3.35	0.991	1.000	1.000	1.000
4	0.30	0.30	0.424	0.460	0.468	0.464
	0.50	0.50	0.842	0.842	0.859	0.865
	0.90	0.90	0.999	0.999	1.000	1.000
	1.40	1.40	1.000	1.000	1.000	1.000
	1.90	1.90	1.000	1.000	1.000	1.000
	3.35	3.35	0.988	1.000	1.000	1.000
5	-0.30	0.30	0.284	0.332	0.345	0.343
	-0.50	0.50	0.675	0.733	0.756	0.747
	-0.95	0.95	0.992	0.997	0.992	0.995
	-1.55	1.55	1.000	1.000	1.000	1.000
	-2.25	2.25	0.999	1.000	1.000	1.000
	-3.35	3.35	0.995	1.000	1.000	1.000

good power properties, which are developed defining the critical region based on constructing a quel (full or half) utilizing the bootstrap that are known as full and half quel tests. These new test procedures do not suffer from the non-monotonic power and local biasedness behavior of the Wald test. More importantly, the full quel test performs uniformly best in that it holds nominal level quite well and shows comparable power in most instances. In addition, this full quel test can occasionally surpass the LR and half quel test in terms of power over most of the regions of the parameter space. Furthermore, in contrast to the Wald test, this new test does not require an analytical form of the variance-covariance matrix of the ML estimators of the parameters. This adds to its practical advantage when this matrix causes the non-monotonic power of the Wald test or is difficult to obtain. In light of this, the full quel test is best applied with the use of the quel critical region via bootstrap.

References

- Beckett, S., & Gould, W. (1987). Rangefinder box plots: A note. *The American Statistician*, 41, 149.
- Godfrey, L. G. (1988). *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*. Cambridge: Cambridge University Press.
- Goh, K. L. (1998). Some solutions to small-sample problems of Wald tests in econometrics. Unpublished Ph.D. thesis, Monash University.
- Goldberg, K. M., & Iglewicz, B. (1992). Bivariate extensions of the boxplot. *Technometrics*, 34, 307-320.
- Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models*. Cambridge: Cambridge University Press.
- Hauck, Jr., W.W. & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72, pp. 851-853.
- Hendry, D. F. (1995). *Dynamic econometrics*. NY: Oxford University Press.
- Mantel, N. (1987). Understanding Wald's test for exponential families. *American Statistician*, 41, 147-148.
- Nelson, F. D., & Savin, N. E. (1988). The non-monotonicity of the power function of the Wald test in non-linear models. *Working Paper Series No. 88-7, Department of Economics*, University of Iowa.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A, 175-240, 263-294.
- Rao, C. R. (1973). *Linear statistical inference and its applications (2nd ed.)*. NY: John Wiley & Sons).
- Scott, D. W. (1985). Frequency polygons: Theory and application. *Journal of the American Statistical Association*, 30, 348-354.
- Turkey, J. W. (1947). Non-parametric estimation II: Statistically equivalent blocks and tolerance regions-the continuous case. *The Annals of Mathematical Statistics*, 18, 529-539.
- Vaeth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review*, 53, 199-214.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.

Appendix: Options for a Quel

The confidence contour points for full or half quel are computed by considering the following two options of the percentile of the standardized errors $\xi_{\text{percentile}}^{fq}$ and the estimators γ_1 and γ_2 .

Option I - Full quell:

Constant ratio: In this case, the percentile of the standardized errors $\xi_{\text{percentile}}^{fq}$ and the two additional estimators γ_1^{fq} and γ_2^{fq} are taken into account to construct the full quel, respectively, the same as $\xi_{\text{percentile}}^q$, γ_1 and γ_2 discussed in subsection 3.2, that is, $\xi_{\text{percentile}}^{fq} = \xi_{\text{percentile}}^q$ and $\gamma_l^{fq} = \gamma_l$ for $l = 1, 2$.

Option II: Half quel

Constant difference in the direction of the angle of the parametric equations: In this

QUEL TEST FOR TWO LINEAR RESTRICTIONS IN THE NONLINEAR MODELS

option to find those values for the construction of the half quel, first compute the standardized errors as ξ_i^{*q} based on equation (4) in subsection 3.2 with γ_1 and γ_2 replaced by γ_1^a and γ_2^a . Then define $\vartheta = \xi_{percentile}^{*q} / \xi_{median}^{*q}$ and $\gamma_l^a = (2\gamma_l - 1 + \vartheta) / 2\vartheta$ for $l = 1, 2$, where $\xi_{percentile}^{*q}$ and ξ_{median}^{*q} are the percentile and median of the errors ξ_i^{*q} . In this case, $\xi_{percentile}^{*q}$ and γ_l^a are functions of each other. Thus, the solution of γ_l^a for $l = 1, 2$ can be simply found by the following iterative procedure:

- Start with the initial value of ϑ , ϑ_0 .
- Using this initial value ϑ_0 , compute γ_l^a using the above formula for $l = 1, 2$.
- Based on the above values of γ_l^a , compute the standardized errors ξ_i^{*q} from equation (4) in subsection 3.2.
- Compute the percentile and median of the errors ξ_i^{*q} and compute ϑ from the above equation.
- Stop if the convergence condition holds, that is, $|\vartheta - \vartheta_0| < 0.01$, ϑ and ϑ_0 are the current and previous iteration's values, respectively. Store the final solution of γ_l^a as γ_l^{hq} for $l = 1, 2$ and compute the percentile of the standardized errors obtained from equation (4) with these values, γ_l^{hq} for $l = 1, 2$, as $\xi_{percentile}^{hq}$ for the construction of a half quel.

Least Absolute Value vs. Least Squares Estimation and Inference Procedures in Regression Models with Asymmetric Error Distributions

Terry E. Dielman
Texas Christian University

A Monte Carlo simulation is used to compare estimation and inference procedures in least absolute value (LAV) and least squares (LS) regression models with asymmetric error distributions. Mean square errors (MSE) of coefficient estimates are used to assess the relative efficiency of the estimators. Hypothesis tests for coefficients are compared on the basis of empirical level of significance and power.

Key words: L_1 regression, least absolute deviations, robust regression, simulation.

Introduction

The use of regression analysis relies on the choice of a criterion in order to estimate the coefficients of the explanatory variables. Traditionally, the least squares (LS) criterion has been the method of choice; however, the least absolute value (LAV) criterion provides an alternative. LAV regression coefficients are chosen to minimize the sum of the absolute values of the residuals. By minimizing sums of absolute values rather than sums of squares, the effect of outliers on the coefficient estimates is diminished.

In most previous studies comparing the performance of LAV and LS estimation, the distributions examined have been symmetric. Fat-tailed distributions that introduce outliers have been used, but these have typically been symmetric fat-tailed distributions (Laplace, Cauchy, etc). This study examined the performance of LAV and LS coefficient estimators when the regression disturbances come from asymmetric distributions.

Also, hypothesis tests for coefficient significance are examined. For the LAV

regression, the tests compared include the likelihood ratio (LR) test, the Lagrange multiplier (LM) test suggested by Koenker and Bassett (1982) and a bootstrap test. The tests are compared in terms of both observed significance level and empirical power. Four alternative variance estimates are considered for the LR and bootstrap tests. The LAV tests are also compared with the traditional t-test for LS regression.

Methodology

Least Absolute Value Estimation and Testing

The model considered in this article is the linear regression model:

$$y_i = \beta_0 + \sum_{k=1}^K \beta_k x_{ik} + \varepsilon_i$$
$$i = 1, 2, \dots, n \quad (1)$$

where y_i is the i^{th} observation on the dependent variable, x_{ik} is the i^{th} observation on the k^{th} explanatory variable, and ε_i is a random disturbance for the i^{th} observation. The distribution of the disturbances may not be normal or even symmetric in this examination. The parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_K$ are unknown and must be estimated. For a discussion of algorithms to produce LAV coefficient estimates, see Dielman (1992, 2005).

Terry E. Dielman is Professor of Decision Sciences in the M.J. Neeley School of Business. Email: t.dielman@tcu.edu.

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

In matrix notation, the model in (1) can be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where \mathbf{Y} is an $n \times 1$ vector of values of the dependent variable, \mathbf{X} is an $n \times (K+1)$ matrix of values of the explanatory variables, including a column of ones for the constant, $\boldsymbol{\beta}$ is a $(K+1) \times 1$ vector of the regression coefficients to be estimated and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of disturbances. Bassett and Koenker (1978) showed that, under reasonable conditions, the LAV coefficient estimator has an asymptotic distribution that converges to

$N(\boldsymbol{\beta}, \lambda^2(\mathbf{X}'\mathbf{X})^{-1})$ where $\frac{\lambda^2}{n}$ is the asymptotic variance of the sample median for a sample of size n from the disturbance distribution.

Equation (2) can be rewritten in the following form:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (3)$$

The coefficient vector $\boldsymbol{\beta}$ and the data matrix \mathbf{X} from equation (2) have been partitioned: $\boldsymbol{\beta}_1$ is a $k_1 \times 1$ vector of coefficients to remain in the model and \mathbf{X}_1 is the associated part of the original data matrix, \mathbf{X} ; $\boldsymbol{\beta}_2$ represents the $k_2 \times 1$ vector of coefficients to be included in a hypothesis test, and \mathbf{X}_2 is the associated part of the original data matrix, \mathbf{X} . The test considered is the basic test for coefficient significance, i.e., $H_0: \boldsymbol{\beta}_2 = \mathbf{0}$. In the simulation $\boldsymbol{\beta}_2$ consists of a single coefficient.

Koenker and Bassett (1982) proposed three procedures for conducting hypothesis tests on the LAV regression model coefficients. The three tests are based on Wald, likelihood ratio (LR), and Lagrange multiplier (LM) test statistics, each of which has the same limiting Chi-square distribution. The LR and LM statistics will be examined in the Monte Carlo simulation. In previous studies, the Wald test has been shown to be inferior to the LR and LM statistics in small samples, so it is not included in this study (See, for example, Dielman and Pfaffenberger, 1988, 1990, 1992; Dielman, 2006).

The Lagrange Multiplier (LM) test statistic for the test of the null hypothesis $H_0: \boldsymbol{\beta}_2 = \mathbf{0}$ is given by

$$LM = \mathbf{g}'_2 \mathbf{D} \mathbf{g}_2, \quad (4)$$

where \mathbf{g}_2 is the appropriate portion of the normalized gradient of the unrestricted LAV objective function, evaluated at the restricted estimate, and \mathbf{D} is the appropriate block of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix to be used in the test.

The Likelihood Ratio (LR) test statistic (assuming the disturbances follow a Laplace distribution) is

$$LR = \frac{2(SAD_1 - SAD_2)}{\lambda} \quad (5)$$

where SAD_1 is the sum of the absolute deviations of the residuals in the restricted or reduced model (i.e., $\boldsymbol{\beta}_2 = \mathbf{0}$) and SAD_2 is the sum of the absolute deviations of the residuals in the unrestricted model.

The LR test statistic requires the estimation of the scale parameter λ , whereas the LM test statistic does not. One often-suggested estimator for λ can be computed as follows:

$$\hat{\lambda} = \frac{\sqrt{n'} [e_{(n'-m-1)} - e_{(m)}]}{z_{\alpha/2}},$$

where,

$$m = \frac{n' + 1}{2} - z_{\alpha/2} \sqrt{\frac{n'}{4}} \quad (6)$$

where the $e_{(i)}$ are ordered residuals from the LAV-fitted model, and $n' = n - r$ where r is the number of zero residuals. A value of $\alpha = 0.05$ is usually suggested. This estimator will be referred to as the SECI estimator. See McKean and Schrader (1984), McKean and Schrader (1987), Sheather (1987), Dielman and Pfaffenberger (1990, 1992) and Dielman and Rose (1995, 1996) for discussions and uses of this estimator.

When computing the variance of the slope coefficient in a LAV regression, the estimator of λ in equation (6) will be used. However, four different options in constructing this estimator will be considered. These options are as follows:

- SECI1: $\hat{\lambda}_1$ uses $z = 1.96$ ($\alpha = 0.05$ value) and n' = total number of observations (n).
- SECI2: $\hat{\lambda}_2$ uses $t_{0.025}$ with n degrees of freedom rather than the z value and $n' =$ total number of observations (n).
- SECI3: $\hat{\lambda}_3$ uses $z = 1.96$ ($\alpha = 0.05$ value) and $n' = n - r$ where r is the number of zero residuals.
- SECI4: $\hat{\lambda}_4$ uses $t_{0.025}$ with $n - r$ degrees of freedom rather than the z value and $n' = n - r$ where r is the number of zero residuals.

The notation L1, L2, L3 and L4 will be used to indicate the LR test using variance estimator 1, 2, 3, or 4. Much of the literature in this area recommends using the estimator SECI3. However, Dielman (2006) performed a simulation study that suggested using SECI2. These results were for symmetric distributions only. Results for asymmetric distributions will be examined in this paper. In addition, the bootstrap tests were not included in the previous study.

The bootstrapping methodology provides an alternative to the LR and LM tests. In a LAV simple regression, for example, a bootstrap test statistic for $H_0: \beta_1 = 0$ can be computed in several ways (see Li & Maddala, 1996). The following procedure will be used in this study: The model shown as equation (1) is estimated (when $K = 1$ for simple regression) using LAV estimation procedures and residuals

are obtained. The test statistic, $\frac{|\hat{\beta}_1 - 0|}{se(\hat{\beta}_1)}$, is

computed from the regression on the original data, where $se(\hat{\beta}_1)$ represents the standard error of the coefficient estimate. The residuals, e_i ($i = 1, 2, \dots, n$), from this regression are saved, centered, and resampled (with replacement, excluding zero residuals), to obtain a new sample of disturbances, e_i^* . The e_i^* values are used to create pseudo-data as follows:

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i^* \quad (7)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the initial LAV estimates of the intercept and slope. The coefficients in equation (7) are then re-estimated to obtain new parameter estimates, $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, and the test

statistic $T = \frac{|\hat{\beta}_1^* - \hat{\beta}_1|}{se(\hat{\beta}_1^*)}$ is computed and saved.

The process of computing T is repeated a large number of times. For a test to be performed at a particular level of significance, α , the critical value is the $(1 - \alpha)^{th}$ percentile from the ordered test statistic values. If the original test statistic is larger than this critical value, then the null hypothesis that $\beta_1 = 0$ is rejected. The extension to a single coefficient in a multiple regression is easily accomplished.

Although Li and Maddala (1996) suggested that the pseudo-data generating process can proceed in other ways, the method outlined here is fairly typical. Research by van Giersbergen and Kiviet (2002) and Dielman and Rose (2002) suggest that the aspect of primary importance is that the resampling scheme should mimic the null distribution of the test statistic to be bootstrapped. This suggestion is followed in the bootstrap approach used in this paper. Results from the traditional LS t-test are compared to those from the LAV-based tests.

Description of the Simulation Experiment

The simulation is based on the model in equation (1). The sample sizes used are $n = 20, 30, 40$ and 100 . The disturbances are generated using stable distributions with the following combinations of characteristic exponent (alpha) and skewness parameter (beta):

Beta = 0.0, 0.4 and 0.8 with Alpha = 1.2

Beta = 0.0, 0.4 and 0.8 with Alpha = 1.8

In addition the normal (beta = 0.0 with alpha = 1.2) and Cauchy distributions (beta = 0.0 with alpha = 1.0) were used. The normal and Cauchy distributions serve as extremes. Stable distributions are infinite variance distributions when the characteristic exponent is less than 2.0, so the LAV estimator would be expected to outperform LS in these cases. When the

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

characteristic exponent equals 2.0 (and beta is zero), the distribution is normal and LS will be optimal. For a characteristic exponent close to 2.0 (and a symmetric distribution), we would expect LS to perform relatively better than for an exponent near 1.0 (Cauchy disturbances). As alpha approaches 1.0, LAV is expected to perform better than LS.

The independent variables are generated as independent standard normal random variables, independent of the disturbances. Bootstrap tests used 199 bootstrap replications. The value of β_0 is set equal to zero (without loss of generality). In the simple regression, the value of β_1 is set equal to 0.0 to assess the level of significance and is set equal to 0.2, 0.4, 0.6, 0.8, 1.0 and 2.0 to examine power. In the multiple regressions, all slope coefficients are set equal to zero (without loss of generality), except for one coefficient which is set equal to 0.0 to assess the level of significance and is set equal to 0.2, 0.4, 0.6, 0.8, 1.0 and 2.0 to examine power. For each factor combination in the experimental design, 5,000 Monte Carlo simulations are used, and the number of rejections of the null hypothesis of whether the selected slope coefficient is equal to zero is counted for each setting. All testing is done using a nominal 5% level of significance.

Results

Estimation

Table 1 contains ratios of mean square errors (MSEs) for estimates of the intercept and slope coefficients in simple regressions ($K = 1$) and for the intercept and one of the coefficients in the multiple regressions ($K = 3$ and 5) for sample size $n = 20$ in Panel A, $n = 30$ in Panel B, $n = 40$ in Panel C, and $n = 100$ in Panel D. The extremes of alpha = 0.0 (Cauchy) and alpha = 2.0 (normal) show the range of possibilities when distributions are symmetric. LS is always preferred to LAV when disturbances are normal. The ratio of MSEs is consistently 0.8 except when $n = 20$ and $K = 5$, in which case the preference for LS is even stronger.

The LAV estimator is preferred over LS for alpha of 1.8 and 1.2, although the advantage decreases as alpha approaches two (normal distribution) as would be expected. The only

exception to this rule is when $n = 20$ and $K = 5$ when LS is preferred for beta = 0.0 or 0.4, that is, when the skewness is less extreme. LAV is preferred in all cases when beta = 0.8.

When alpha = 1.8, the preference for LAV over LS increases in all cases as skewness increases. When alpha = 1.2 and $K = 1$, the preference for LAV over LS decreases (although LAV is still better than LS by a wide margin). With alpha = 1.2 and $K = 3$ or 5 , the results are mixed in terms of the increase or decrease of the preference for LAV over LS based on skewness. This may be a result of looking at an efficiency measure for only a single coefficient. Regardless, LAV is still preferable to LS by a wide margin when alpha = 1.2.

Hypothesis Tests

Tables 2 through 5 contain the median percentage of trials in which H_0 : coefficient = 0 is rejected for various combinations of test and coefficient values for $n = 20, 30, 40$ and 100 , respectively, when $K = 1$. The medians are taken over the disturbance distributions. Thus, the results for the symmetric distributions (beta = 0.0) include Stable distributions with alpha = 1.0 (Cauchy) 1.2, 1.8 and 2.0 (normal). The asymmetric distributions include Stable with alpha = 1.2 and 1.8 when beta is either 0.4 or 0.8. When the coefficient value is zero, the empirical significance levels can be assessed; when it is non-zero, power for the tests can be compared. Tables 6 through 9 contain the same information for $K = 3$ while tables 10 through 13 contain results for $K = 5$.

The empirical level of significance for the LS t-test never exceeds 0.06 in any of the experimental settings (nominal level = 0.05). However, the test lacks power when compared to the LAV tests. For example, consider Table 5 with $K = 1$ and $n = 100$. All tests have empirical level of significance 0.05, but LST has considerably lower power.

There is little difference in performance for skewed and symmetric error distributions. When LAV is preferred to LS, the preference is due to the presence of outliers from the fat-tailed distribution rather than from any lack of symmetry in the distributions.

Among the LAV tests, the bootstrap tests and the LM test tend to maintain a median

DIELMAN

Table 1: Ratios of mean square error of estimates of intercept and slope (or one of the slope coefficients if $K = 3$ or 5): LS/LAV. Numbers greater than one favor LAV, numbers less than one favor LS. Alpha is the characteristic exponent of the Stable distribution; beta is the skewness parameter. (Alpha = 2.0 is the normal distribution, Alpha = 0.0 is the Cauchy).

Panel A: Intercept ($n = 20$)				
Alpha	K	Beta		
		0.0	0.4	0.8
0.0	1	102.0		
	3	55.9		
	5	82.3		
1.2	1	83.1	68.8	38.1
	3	17.8	57.0	21.4
	5	25.6	22.2	14.7
1.8	1	1.3	1.3	1.4
	3	1.2	1.2	1.2
	5	0.7	0.7	2.5
2.0	1	0.8		
	3	0.8		
	5	0.4		

Panel A: Slope ($n = 20$)				
Alpha	K	Beta		
		0.0	0.4	0.8
0.0	1	69.5		
	3	46.3		
	5	28.6		
1.2	1	99.1	83.9	52.8
	3	13.7	27.6	11.9
	5	10.0	10.7	12.8
1.8	1	1.2	1.2	1.2
	3	1.1	1.1	1.1
	5	0.7	0.7	1.4
2.0	1	0.8		
	3	0.8		
	5	0.5		

Panel B: Intercept ($n = 30$)				
Alpha	K	Beta		
		0.0	0.4	0.8
0.0	1	130.8		
	3	104.7		
	5	1016.7		
1.2	1	76.8	67.2	37.0
	3	64.6	53.3	29.8
	5	44.9	37.7	25.6
1.8	1	1.3	1.3	1.4
	3	1.2	1.3	1.4
	5	1.1	1.2	1.3
2.0	1	0.8		
	3	0.8		
	5	0.8		

Panel B: Slope ($n = 30$)				
Alpha	K	Beta		
		0.0	0.4	0.8
0.0	1	90.4		
	3	56.0		
	5	933.9		
1.2	1	71.4	83.4	56.1
	3	50.4	43.2	29.7
	5	58.0	51.1	38.8
1.8	1	1.3	1.4	1.5
	3	1.2	1.3	1.3
	5	1.3	1.3	1.4
2.0	1	0.8		
	3	0.8		
	5	0.8		

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

Table 1: continued

Panel C: Intercept ($n = 40$)					Panel C: Slope ($n = 40$)				
		Beta					Beta		
Alpha	K	0.0	0.4	0.8	Alpha	K	0.0	0.4	0.8
0.0	1	176.2			0.0	1	123.8		
	3	136.3				3	74.3		
	5	130.0				5	109.7		
1.2	1	53.7	46.0	29.2	1.2	1	25.7	24.0	21.0
	3	39.0	35.3	26.2		3	28.8	28.9	29.7
	5	41.2	36.8	26.3		5	22.6	23.0	24.0
1.8	1	1.4	1.5	1.6	1.8	1	1.2	1.3	1.3
	3	1.4	1.5	1.6		3	1.4	1.4	1.5
	5	1.4	1.5	1.6		5	1.3	1.4	1.4
2.0	1	0.8			2.0	1	0.8		
	3	0.8				3	0.8		
	5	0.8				5	0.8		

Panel D: Intercept ($n = 100$)					Panel D: Slope ($n = 100$)				
		Beta					Beta		
Alpha	K	0.0	0.4	0.8	Alpha	K	0.0	0.4	0.8
0.0	1	1467.5			0.0	1	1017.6		
	3	1555.1				3	751.2		
	5	1513.6				5	278.5		
1.2	1	96.4	96.4	76.1	1.2	1	57.3	50.7	38.5
	3	119.2	121.6	96.4		3	117.3	132.0	149.6
	5	117.9	121.1	95.4		5	99.4	107.1	115.9
1.8	1	2.0	2.3	2.6	1.8	1	1.2	1.3	1.3
	3	2.4	2.8	3.0		3	2.5	2.9	3.3
	5	2.4	2.8	3.1		5	2.3	2.6	3.0
2.0	1	0.8			2.0	1	0.8		
	3	0.8				3	0.8		
	5	0.8				5	0.8		

significance level close to nominal. The LR tests often deviate considerably from nominal. However, LR2 has median significance level closer to nominal than the other LR tests in most cases. Performance is similar for the LR tests for skewed and symmetric distributions. Among the bootstrap tests, there is little difference in performance for any of the experimental settings.

In choosing among the LAV tests, it appears that the LR2 test maintains relatively high power - even when the level of significance is lower compared to the other tests. Also, the LM test is consistently lower in power. This negates some of the advantage the LM test might have due to the fact that it does not need an estimate of the nuisance parameter. As noted, the bootstrap tests have levels of significance that tend to be close to the nominal level. Power for the bootstrap tests can be slightly lower than that for LR2, even when the level of significance is equal or lower for LR2. Increasing the number of bootstrap iterations might improve the power of these tests. When sample size is large ($n = 100$), there is little difference among any of the LAV based tests. These tests still improve on the LS t-test even in large samples.

The variance estimate used to obtain LR2 uses n in the computations rather than $n - r$ (where r is the number of zero residuals). This adjustment for zero residuals does not appear to be necessary. The variance estimates used to obtain LR1 and LR2 differ in that LR1 uses the z value while LR2 uses the appropriate t value in the computations. This provides some improvement in test performance for LR2 in small samples but the advantage vanishes for a sample size of 100.

Conclusion

Previous research examining small sample performance of some of the test statistics discussed in this article based on symmetric error distributions include Dielman (2006), Dielman and Pfaffenberger (1988, 1990, 1992), Dielman and Rose (1995, 1996, 2002), Koenker (1987) and Stangenhuis (1987). The results of these studies suggest that, in small samples, the LR and LM tests generally outperform the Wald

test (not considered in the present study) in terms of both power and observed significance level.

The LR and LM tests differ in that the LR test requires an estimate of the λ parameter discussed previously, while the LM test does not. However, using a fairly simple estimate of this scale parameter, the LR test has generally performed as well as, or better than, the LM test. In addition to the Wald, LR, and LM tests, bootstrap approaches have also been examined for inference in LAV regression. Dielman and Pfaffenberger (1988) used a bootstrap approach to estimate the scale parameter, λ , but the significance tests based on these bootstrap estimates did not perform particularly well.

Dielman and Rose (1995) compared a true bootstrap test statistic with the LR and LM tests, and found that the bootstrap performed well in small samples. Dielman and Rose (2002) compared the LR, LM and three versions of the bootstrap suggested by Li and Maddala (1996) along with the LS t-test. They found that the LR test performed at least as well as, and often better than, the competing tests. Prior results for symmetric error distributions are consistent with the results from this study for both symmetric and asymmetric error distributions.

If error distributions are suspected to be fat-tailed, improvements in estimation and inference are possible using LAV estimation rather than LS. This is true regardless of whether the distributions are symmetric or skewed. When choosing a test procedure for LAV estimated models, the bootstrap approaches perform reasonably well for all cases examined here. If a likelihood ratio test is to be used, LR2 seems to perform better than the other choices examined here. In addition, the LM test performs reasonably well in most settings examined although the power may be somewhat lower than the LR2 test. Differences in performance between the LAV based tests are small once the sample size reaches 100.

References

Bassett, G. W., & Koenker, R. W. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73, 618-622.

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

Table 2: Median percentages of rejections of $H_0: \beta_1 = 0$, normal explanatory variable, $n = 20$, $K = 1$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.03	0.07	0.08	0.06	0.06	0.06	0.06	0.06	0.04
0.2	0.10	0.06	0.10	0.12	0.08	0.08	0.08	0.08	0.09	0.07
0.4	0.20	0.14	0.21	0.23	0.15	0.16	0.15	0.15	0.18	0.15
0.6	0.37	0.28	0.38	0.41	0.28	0.29	0.27	0.27	0.31	0.27
0.8	0.56	0.47	0.57	0.60	0.43	0.46	0.43	0.43	0.47	0.41
1.0	0.72	0.64	0.73	0.75	0.59	0.62	0.59	0.59	0.61	0.54
2.0	0.99	0.98	0.99	0.99	0.97	0.98	0.97	0.97	0.94	0.81

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.04	0.03	0.07	0.05	0.05	0.06	0.05	0.05	0.06	0.04
0.2	0.07	0.06	0.10	0.08	0.08	0.08	0.07	0.07	0.09	0.07
0.4	0.16	0.13	0.20	0.18	0.15	0.15	0.14	0.15	0.17	0.13
0.6	0.29	0.26	0.36	0.32	0.27	0.28	0.24	0.25	0.29	0.25
0.8	0.46	0.44	0.54	0.50	0.41	0.42	0.38	0.40	0.43	0.38
1.0	0.63	0.60	0.70	0.66	0.57	0.58	0.52	0.55	0.56	0.53
2.0	0.98	0.98	0.99	0.98	0.97	0.97	0.95	0.96	0.92	0.82

Table 3: Median percentages of rejections of $H_0: \beta_1 = 0$, normal explanatory variable, $n = 30$, $K = 1$ for symmetric (with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.04	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05
0.2	0.11	0.08	0.12	0.13	0.09	0.09	0.10	0.10	0.10	0.09
0.4	0.26	0.20	0.27	0.28	0.21	0.21	0.21	0.21	0.22	0.21
0.6	0.48	0.40	0.49	0.50	0.37	0.39	0.38	0.38	0.40	0.41
0.8	0.69	0.62	0.70	0.71	0.57	0.59	0.57	0.57	0.58	0.61
1.0	0.84	0.79	0.84	0.85	0.74	0.76	0.74	0.74	0.73	0.77
2.0	1.00	0.99	1.00	1.00	0.99	1.00	1.00	1.00	0.97	0.97

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.04	0.06	0.07	0.06	0.06	0.06	0.06	0.05	0.05
0.2	0.11	0.08	0.11	0.12	0.09	0.09	0.09	0.09	0.09	0.08
0.4	0.25	0.20	0.26	0.27	0.19	0.20	0.19	0.19	0.21	0.18
0.6	0.47	0.40	0.48	0.49	0.36	0.37	0.36	0.36	0.41	0.33
0.8	0.69	0.62	0.69	0.71	0.55	0.58	0.56	0.56	0.60	0.49
1.0	0.84	0.80	0.85	0.86	0.73	0.75	0.73	0.73	0.76	0.60
2.0	1.00	1.00	1.00	1.00	0.99	1.00	0.99	0.99	0.98	0.83

DIELMAN

Table 4: Median percentages of rejections of $H_0: \beta_1 = 0$, normal explanatory variable, $n = 40$, $K = 1$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8$, and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.05	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
0.2	0.13	0.14	0.14	0.14	0.11	0.11	0.11	0.11	0.12	0.09
0.4	0.37	0.38	0.37	0.38	0.30	0.30	0.30	0.30	0.32	0.25
0.6	0.65	0.66	0.66	0.67	0.56	0.56	0.56	0.56	0.58	0.44
0.8	0.85	0.86	0.86	0.86	0.78	0.78	0.78	0.78	0.78	0.59
1.0	0.96	0.96	0.96	0.96	0.92	0.92	0.92	0.92	0.90	0.68
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.06	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05
0.2	0.13	0.14	0.13	0.14	0.11	0.11	0.11	0.11	0.12	0.09
0.4	0.35	0.36	0.36	0.37	0.28	0.28	0.27	0.27	0.32	0.25
0.6	0.64	0.65	0.65	0.66	0.53	0.53	0.53	0.53	0.56	0.44
0.8	0.85	0.86	0.85	0.86	0.76	0.76	0.77	0.77	0.77	0.59
1.0	0.95	0.96	0.95	0.96	0.91	0.91	0.91	0.91	0.90	0.68
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.86

Table 5: Median percentages of rejections of $H_0: \beta_1 = 0$, normal explanatory variable, $n = 100$, $K = 1$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8$, and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.2	0.24	0.24	0.24	0.25	0.22	0.22	0.22	0.22	0.23	0.14
0.4	0.68	0.68	0.68	0.69	0.63	0.63	0.63	0.63	0.64	0.39
0.6	0.94	0.94	0.94	0.94	0.92	0.92	0.91	0.91	0.92	0.57
0.8	0.99	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.2	0.22	0.22	0.22	0.22	0.20	0.20	0.19	0.19	0.20	0.14
0.4	0.63	0.64	0.64	0.64	0.57	0.57	0.56	0.56	0.62	0.39
0.6	0.94	0.94	0.94	0.94	0.88	0.88	0.88	0.88	0.91	0.57
0.8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

Table 6: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 20$, $K = 3$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.09	0.04	0.09	0.11	0.06	0.06	0.06	0.06	0.06	0.04
0.2	0.14	0.08	0.14	0.15	0.08	0.08	0.08	0.08	0.11	0.07
0.4	0.27	0.18	0.27	0.29	0.15	0.16	0.15	0.15	0.20	0.16
0.6	0.45	0.33	0.45	0.47	0.27	0.29	0.27	0.27	0.34	0.29
0.8	0.63	0.52	0.64	0.66	0.41	0.45	0.41	0.41	0.49	0.44
1.0	0.78	0.69	0.78	0.80	0.56	0.62	0.56	0.56	0.62	0.56
2.0	0.99	0.98	0.99	0.99	0.96	0.97	0.96	0.96	0.92	0.82

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.09	0.05	0.08	0.10	0.06	0.06	0.05	0.05	0.06	0.04
0.2	0.14	0.08	0.13	0.14	0.08	0.08	0.07	0.07	0.10	0.07
0.4	0.26	0.17	0.25	0.28	0.15	0.15	0.14	0.14	0.20	0.16
0.6	0.44	0.33	0.42	0.45	0.26	0.29	0.25	0.25	0.33	0.29
0.8	0.62	0.51	0.61	0.63	0.41	0.44	0.39	0.39	0.48	0.43
1.0	0.77	0.68	0.76	0.78	0.55	0.61	0.54	0.54	0.62	0.56
2.0	0.99	0.98	0.99	0.99	0.95	0.97	0.95	0.95	0.93	0.82

Table 7: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 30$, $K = 3$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.09	0.06	0.10	0.10	0.06	0.06	0.06	0.06	0.07	0.05
0.2	0.14	0.09	0.14	0.15	0.08	0.09	0.08	0.08	0.11	0.08
0.4	0.26	0.20	0.27	0.29	0.16	0.16	0.16	0.16	0.22	0.15
0.6	0.44	0.36	0.45	0.47	0.28	0.30	0.28	0.28	0.36	0.27
0.8	0.63	0.55	0.65	0.66	0.43	0.47	0.44	0.44	0.53	0.42
1.0	0.79	0.72	0.80	0.81	0.59	0.63	0.60	0.60	0.67	0.54
2.0	0.99	0.99	0.99	0.99	0.97	0.98	0.98	0.98	0.96	0.80

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.09	0.06	0.10	0.10	0.06	0.06	0.05	0.05	0.07	0.05
0.2	0.13	0.09	0.14	0.15	0.08	0.08	0.08	0.08	0.11	0.08
0.4	0.26	0.20	0.26	0.27	0.15	0.17	0.15	0.15	0.22	0.15
0.6	0.44	0.36	0.44	0.46	0.27	0.29	0.27	0.27	0.36	0.27
0.8	0.63	0.55	0.63	0.65	0.43	0.46	0.42	0.42	0.52	0.42
1.0	0.80	0.72	0.79	0.80	0.59	0.62	0.59	0.59	0.66	0.54
2.0	0.99	0.99	0.99	0.99	0.97	0.98	0.97	0.97	0.96	0.80

DIELMAN

Table 8: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 40$, $K = 3$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.08	0.08	0.08	0.09	0.05	0.05	0.05	0.05	0.06	0.05
0.2	0.14	0.14	0.14	0.15	0.09	0.09	0.08	0.08	0.12	0.08
0.4	0.31	0.32	0.32	0.33	0.19	0.19	0.19	0.19	0.26	0.19
0.6	0.54	0.55	0.55	0.56	0.36	0.36	0.37	0.37	0.45	0.34
0.8	0.75	0.76	0.76	0.77	0.57	0.57	0.57	0.57	0.65	0.49
1.0	0.88	0.89	0.89	0.89	0.75	0.75	0.75	0.75	0.79	0.60
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.82

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.08	0.09	0.08	0.09	0.05	0.05	0.05	0.05	0.06	0.05
0.2	0.13	0.14	0.14	0.14	0.08	0.08	0.08	0.08	0.12	0.08
0.4	0.30	0.31	0.31	0.32	0.19	0.19	0.19	0.19	0.25	0.18
0.6	0.53	0.54	0.54	0.55	0.36	0.36	0.36	0.36	0.44	0.34
0.8	0.75	0.76	0.75	0.76	0.57	0.57	0.56	0.56	0.63	0.49
1.0	0.88	0.89	0.88	0.89	0.74	0.74	0.74	0.74	0.78	0.60
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.82

Table 9: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 100$, $K = 3$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8,$ and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05
0.2	0.25	0.26	0.26	0.27	0.21	0.21	0.21	0.21	0.23	0.14
0.4	0.68	0.68	0.69	0.69	0.60	0.60	0.60	0.60	0.63	0.38
0.6	0.94	0.94	0.94	0.94	0.90	0.90	0.90	0.90	0.91	0.57
0.8	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05
0.2	0.24	0.24	0.24	0.25	0.20	0.20	0.20	0.20	0.22	0.14
0.4	0.67	0.67	0.67	0.68	0.55	0.55	0.56	0.56	0.62	0.38
0.6	0.93	0.93	0.93	0.93	0.87	0.87	0.88	0.88	0.88	0.57
0.8	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

Table 10: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 20$, $K = 5$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8$, and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.18	0.09	0.11	0.13	0.05	0.05	0.05	0.05	0.15	0.06
0.2	0.21	0.10	0.13	0.15	0.06	0.06	0.05	0.05	0.17	0.07
0.4	0.27	0.14	0.18	0.20	0.07	0.08	0.07	0.07	0.20	0.09
0.6	0.35	0.21	0.25	0.28	0.10	0.11	0.09	0.09	0.24	0.14
0.8	0.45	0.31	0.34	0.37	0.13	0.15	0.14	0.14	0.29	0.20
1.0	0.56	0.40	0.44	0.47	0.18	0.21	0.19	0.19	0.33	0.28
2.0	0.90	0.83	0.84	0.86	0.51	0.63	0.56	0.56	0.53	0.65

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.19	0.09	0.10	0.12	0.05	0.05	0.04	0.04	0.15	0.06
0.2	0.21	0.10	0.12	0.14	0.05	0.05	0.05	0.05	0.17	0.07
0.4	0.27	0.14	0.16	0.18	0.07	0.07	0.06	0.06	0.20	0.09
0.6	0.35	0.21	0.23	0.26	0.09	0.10	0.09	0.09	0.23	0.14
0.8	0.44	0.30	0.32	0.34	0.13	0.15	0.13	0.13	0.28	0.20
1.0	0.55	0.40	0.41	0.44	0.18	0.20	0.17	0.17	0.32	0.28
2.0	0.89	0.82	0.82	0.84	0.51	0.62	0.53	0.53	0.51	0.65

Table 11: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 30$, $K = 5$ for symmetric (stable with $\beta = 0.0$ for $\alpha = 1.0, 1.2, 1.8$, and 2.0) and skewed (stable with $\beta = 0.4$ and 0.8 for $\alpha = 1.2$ and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.14	0.08	0.13	0.14	0.05	0.05	0.05	0.05	0.08	0.04
0.2	0.20	0.13	0.18	0.20	0.08	0.08	0.07	0.07	0.11	0.07
0.4	0.35	0.26	0.34	0.36	0.15	0.16	0.15	0.15	0.19	0.16
0.6	0.55	0.46	0.55	0.57	0.26	0.30	0.28	0.28	0.34	0.30
0.8	0.74	0.66	0.74	0.75	0.42	0.47	0.45	0.45	0.50	0.45
1.0	0.86	0.81	0.86	0.87	0.58	0.64	0.61	0.61	0.64	0.58
2.0	1.00	0.99	1.00	1.00	0.96	0.98	0.97	0.97	0.95	0.82

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.14	0.08	0.12	0.13	0.05	0.05	0.05	0.05	0.07	0.04
0.2	0.19	0.13	0.17	0.19	0.07	0.08	0.08	0.08	0.10	0.07
0.4	0.35	0.26	0.33	0.34	0.15	0.16	0.15	0.15	0.20	0.16
0.6	0.55	0.46	0.53	0.55	0.27	0.30	0.27	0.27	0.34	0.30
0.8	0.73	0.65	0.72	0.73	0.41	0.46	0.43	0.43	0.50	0.45
1.0	0.86	0.81	0.85	0.86	0.57	0.63	0.59	0.59	0.64	0.58
2.0	1.00	0.99	1.00	1.00	0.96	0.98	0.97	0.97	0.95	0.82

DIELMAN

Table 12: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 40$, $K = 5$ for symmetric (stable with beta = 0.0 for alpha = 1.0, 1.2, 1.8, and 2.0) and skewed (stable with beta = 0.4 and 0.8 for alpha = 1.2 and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.11	0.12	0.12	0.13	0.06	0.06	0.06	0.06	0.07	0.05
0.2	0.20	0.21	0.21	0.21	0.10	0.10	0.10	0.10	0.13	0.09
0.4	0.44	0.45	0.44	0.45	0.24	0.24	0.24	0.24	0.30	0.23
0.6	0.69	0.70	0.70	0.71	0.45	0.45	0.45	0.45	0.53	0.42
0.8	0.87	0.87	0.88	0.88	0.66	0.66	0.67	0.67	0.73	0.57
1.0	0.95	0.96	0.96	0.96	0.82	0.82	0.83	0.83	0.86	0.67
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.85

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.12	0.12	0.11	0.12	0.06	0.06	0.06	0.06	0.06	0.05
0.2	0.20	0.21	0.20	0.21	0.10	0.10	0.10	0.10	0.13	0.09
0.4	0.43	0.44	0.43	0.44	0.23	0.23	0.23	0.23	0.30	0.23
0.6	0.69	0.70	0.69	0.70	0.44	0.44	0.44	0.44	0.52	0.42
0.8	0.87	0.87	0.87	0.88	0.65	0.65	0.65	0.65	0.72	0.57
1.0	0.95	0.96	0.96	0.96	0.82	0.82	0.82	0.82	0.85	0.67
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.85

Table 13: Median percentages of rejections of H_0 : coefficient = 0, normal explanatory variable, $n = 100$, $K = 5$ for symmetric (stable with beta = 0.0 for alpha = 1.0, 1.2, 1.8, and 2.0) and skewed (stable with beta = 0.4 and 0.8 for alpha = 1.2 and 1.8) distributions.

Panel A: Symmetric Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.08	0.08	0.08	0.08	0.05	0.05	0.05	0.05	0.05	0.05
0.2	0.28	0.28	0.29	0.29	0.19	0.19	0.19	0.19	0.21	0.14
0.4	0.72	0.72	0.72	0.73	0.58	0.58	0.58	0.58	0.61	0.39
0.6	0.95	0.95	0.95	0.95	0.88	0.88	0.88	0.88	0.89	0.57
0.8	0.99	0.99	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88

Panel B: Skewed Distributions										
Beta	LR1	LR2	LR3	LR4	B1	B2	B3	B4	LM	LST
0.0	0.08	0.08	0.08	0.08	0.05	0.05	0.05	0.05	0.05	0.06
0.2	0.27	0.27	0.27	0.28	0.18	0.18	0.19	0.19	0.20	0.15
0.4	0.70	0.70	0.70	0.71	0.55	0.55	0.56	0.56	0.61	0.39
0.6	0.94	0.94	0.94	0.95	0.87	0.87	0.87	0.87	0.89	0.58
0.8	1.00	1.00	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.67
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.73
2.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88

COMPARISON OF LAV AND LS ESTIMATION AND INFERENCE PROCEDURES

Dielman, T. E. (2006). Variance estimates and hypothesis tests in least absolute value regression. *Journal of Statistical Computation and Simulation*, 76, 103-114.

Dielman, T. E. (2005). Least absolute value regression: Recent contributions. *Journal of Statistical Computation and Simulation*, 75, 263-286.

Dielman, T. E. (1992). Computational algorithms for least absolute value regression. In Dodge, Y., (Ed.): *L1-statistical analysis and related methods*, 311-326. Amsterdam: Elsevier Science Publishers.

Dielman, T. E., & Pfaffenberger, R. (1992). A further comparison of tests of hypotheses in LAV regression. *Computational Statistics and Data Analysis*, 14, 375-384.

Dielman, T. E., & Pfaffenberger, R. (1990). Tests of linear hypotheses in LAV regression. *Communications in Statistics – Simulation and Computation*, 19, 1179-1199.

Dielman, T. E., & Pfaffenberger, R. (1988). Bootstrapping in least absolute value regression: An application to hypothesis testing. *Communications in Statistics – Simulation and Computation*, 17, 843-856.

Dielman, T. E., & Rose, E. L. (2002). Bootstrap versus traditional hypothesis testing procedures for coefficients in least absolute value regression. *Journal of Statistical Computation and Simulation*, 72, 665-675.

Dielman, T. E., & Rose, E. L. (1996). A note on hypothesis testing in LAV multiple regression: A small sample comparison. *Computational Statistics and Data Analysis*, 21, 463-470.

Dielman, T. E. & Rose, E. L. (1995). A bootstrap approach to hypothesis testing in least absolute value regression. *Computational Statistics and Data Analysis*, 20, 119-130.

Koenker, R. (1987). A comparison of asymptotic testing methods for L1-regression. In Dodge, Y., (Ed.): *Statistical data analysis based on the L1-norm and related methods*, 287-295. Amsterdam: Elsevier Science Publishers.

Koenker, R., & Bassett, G. (1982). Tests of linear hypotheses and L1 estimation. *Econometrica*, 50, 1577-1583.

Li, H., & Maddala, G. S. (1996). Bootstrapping time series models. *Econometric Reviews*, 15, 115-158.

McKean, J., & Schrader, R. (1987). Least absolute errors analysis of variance. In: Dodge, Y., (Ed.): *Statistical data analysis based on the L1-norm and related methods*, 297-305. Amsterdam: Elsevier Science Publishers.

McKean, J., & Schrader, R. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics – Simulation and Computation*, 13, 751-773.

Sheather, S. J. (1987). Assessing the accuracy of the sample median: Estimated standard errors versus interpolated confidence intervals. In: Dodge, Y., (Ed.): *Statistical data analysis based on the L1-Norm and related methods*, 203-215. Amsterdam: Elsevier Science Publishers.

Stangenhuis, G. (1987). Bootstrap and inference procedures for L1- regression. In Dodge, Y., (Ed.): *Statistical data analysis based on the L1-norm and related methods*, 323-332. Amsterdam: Elsevier Science Publishers.

van Giersbergen, N. P. A., & Kiviet, J. F. (2002). How to implement the bootstrap in static or stable dynamic regression models: test statistic versus confidence region approach. *Journal of Econometrics*, 108, 133-156.

A Monte Carlo Comparison of Regression Estimators When the Error Distribution is Long-Tailed Symmetric

Oya Can Mutan
ODTU, Turkey

Birdal Şenoğlu
Ankara University, Turkey

The performances of the ordinary least squares (OLS), modified maximum likelihood (MML), least absolute deviations (LAD), Winsorized least squares (WIN), trimmed least squares (TLS), Theil's (Theil) and weighted Theil's (Weighted Theil) estimators are compared under the simple linear regression model in terms of their bias and efficiency when the distribution of error terms is long-tailed symmetric.

Key words: Long-tailed symmetric, ordinary least squares, modified maximum likelihood, least absolute deviations, Winsorized least squares, trimmed least squares, Theil's method, Weighted Theil's method.

Introduction

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad (1)$$

where ($i = 1, 2, \dots, n$), y_i is the response variable, x_i is a nonstochastic explanatory variable and β_0 and β_1 are the unknown parameters. Traditionally, error terms e_i ($1 \leq i \leq n$) are assumed to be independently and identically distributed (iid) normal $N(0, \sigma^2)$ and the regression coefficients β_0 and β_1 are estimated by using the OLS estimators given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2)$$

respectively.

The OLS estimators are optimal only if the error distribution is normal. However, in most real life applications, nonnormal distributions are more prevalent; see, Pearson (1932), Geary (1947), Huber (1981), Şenoğlu (2005) and Şenoğlu (2007). Additionally, the occurrence of outliers in a data set is another indication of nonnormality. Due to these weaknesses of the OLS estimators, statisticians prefer to use the alternative regression estimators which are more efficient and robust under nonnormality

However, the choice of which method to use is not defined clearly for different types of error distributions. In the literature, there exists a very limited number of researches comparing alternative regression methods, see Tam (1996) and Nevitt and Tam (1998). In this study, our main concern is to identify the most efficient method when the error distribution is long-tailed symmetric and also to see the effect of nonnormality on the efficiencies and robustness of the regression estimators.

Oya Can Mutan is a Statistician in the Capital Markets Board of Turkey. She received her B.S. (statistics), M.S. (statistics, economics) and Ph.D. (statistics) degrees from Middle East Technical University in Turkey. Email: oya.canmutan@spk.gov.tr. Birdal Şenoğlu is an Associate Professor in the Department of Statistics at Ankara University, Turkey. He received his B.S. and Ph.D. degrees (statistics) from Middle East Technical University in Turkey and his M.S. (statistics) degree from Iowa State University, USA. Email: senoglu@science.ankara.edu.tr.

Long-tailed Symmetric (LTS) Distribution

The LTS distribution has the probability density function:

$$LTS(p, \sigma): f(e) \propto \frac{1}{\sigma} \left\{ 1 + \frac{e^2}{k\sigma^2} \right\}^{-p},$$

$$-\infty < e < \infty;$$

with $k = 2p - 3$ and $p \geq 2$. The mean and variance of the random variable e is 0 and σ^2 , respectively. See also the following table for the Pearson coefficient of kurtosis, i.e., $\beta_2 = \mu_4 / \mu_2^2$ of the $LTS(p, \sigma)$ distribution:

$p =$	2.5	3.5	5.0	10	∞
$\beta_2 =$	∞	9	4.2	3.4	3.0

This reduces to the normal distribution when p is equal to ∞ .

Methodology

OLS is the most popular method for estimating the parameters of the simple linear regression model. This is partly due to the relative simplicity of its computations. However, the OLS method is very sensitive to outliers and to nonnormality. To remedy these problems, alternative regression methods have been developed that are not sensitive to the violations of the assumptions of the simple linear regression model. The only disadvantage of these alternative methods is their computational difficulty. Today, however, computational difficulties are unimportant issue because of the improvements in computer technology (see Birkes & Dodge, 1993; Rousseeuw & Leroy, 1987).

The Modified Maximum Likelihood Method

The maximum likelihood (ML) estimators are the solutions of the equations

$$\partial \ln L / \partial \beta_0 = 0,$$

$$\partial \ln L / \partial \beta_1 = 0,$$

and

$$\partial \ln L / \partial \sigma = 0. \tag{3}$$

These equations do not have explicit solutions. Tiku, et al. (2001) express likelihood equations in terms of order statistics (for a given β_1), since complete sums are invariant to ordering.

$$z_{(i)} = \frac{y_{[i]} - \beta_0 - \beta_1 x_{[i]}}{\sigma}, \quad (1 \leq i \leq n)$$

where $(y_{[i]}, x_{[i]})$ is that pair of observations which correspond to $z_{(i)}$ ($1 \leq i \leq n$); $(y_{[i]}, x_{[i]})$ are called the concomitants of $z_{(i)}$. They linearize the intractable functions $g(z_{(i)}) = z_{(i)} / \left\{ 1 + (1/k)z_{(i)}^2 \right\}$ by using the first two terms of a Taylor series expansion by using the following linear approximation

$$g(z_{(i)}) \cong \alpha_i + \beta_i z_{(i)}, \quad 1 \leq i \leq n \tag{4}$$

where

$$\alpha_i = \frac{(2/k)t_{(i)}^3}{\left(1 + (1/k)t_{(i)}^2\right)^2}$$

and

$$\beta_i = \frac{1 - (1/k)t_{(i)}^2}{\left(1 + (1/k)t_{(i)}^2\right)^2}$$

$t_{(i)}$'s ($i = 1, 2, \dots, n$) are the expected values of the order statistics $z_{(i)}$, i. e., $t_{(i)} = E(z_{(i)})$.

Incorporating (4) in (3), results in modified likelihood equations:

$$\partial \ln L^* / \partial \beta_0 = 0,$$

$$\partial \ln L^* / \partial \beta_1 = 0,$$

and

$$\partial \ln L^* / \partial \sigma = 0.$$

These equations have explicit solutions called as MML estimators:

$$\hat{\beta}_0 = \bar{y}_{[.]} - \hat{\beta}_1 \bar{x}_{[.]}$$

$$\hat{\beta}_1 = K + D\hat{\sigma}$$

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nC}}{2\sqrt{n(n-2)}}$$

where

$$\bar{y}_{[.]} = \sum_{i=1}^n \beta_i y_{[i]} / m, \quad \bar{x}_{[.]} = \sum_{i=1}^n \beta_i x_{[i]} / m,$$

$$\left(m = \sum_{i=1}^n \beta_i \right),$$

$$K = \sum_{i=1}^n \beta_i (x_{[i]} - \bar{x}_{[.]}) y_{[i]} / \sum_{i=1}^n \beta_i (x_{[i]} - \bar{x}_{[.]})^2,$$

$$D = \sum_{i=1}^n \alpha_i x_{[i]} / \sum_{i=1}^n \beta_i (x_{[i]} - \bar{x}_{[.]})^2,$$

$$B = (2p/k) \sum_{i=1}^n \alpha_i \{y_{[i]} - \bar{y}_{[.]} - K(x_{[i]} - \bar{x}_{[.]})\}$$

and

$$C = (2p/k) \sum_{i=1}^n \beta_i \{y_{[i]} - \bar{y}_{[.]} - K(x_{[i]} - \bar{x}_{[.]})\}^2.$$

Least Absolute Deviations (LAD)

The LAD regression method was developed by Roger Joseph Boscovich in 1757, see Birkes and Dodge (1993). The LAD estimators of regression coefficients, β_0 and β_1 , are found by minimizing the function:

$$F = \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_i)|. \quad (5)$$

Although the logic behind LAD is not more difficult than the concept of OLS, calculation of the LAD estimates is more troublesome. An algorithmic method is used for the calculation of the LAD estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, since there are no exact formulas.

This algorithm starts with one of the data points (x, y) , say (x_1, y_1) , and tries to find the best line passing through it. The line passing through (x_1, y_1) also passes through another data point denoted by (x_2, y_2) . Next we find

the best line passing through (x_2, y_2) . As the algorithm continues, we obtain increasingly better lines and finally the most recent line obtained will be the same as the previous line. This line is the best line and it is called as LAD regression line, see Birkes and Dodge (1993) for more detailed information.

Winsorized Least Squares

The WLS which is an iterative method is another alternative to OLS method; see Yale and Forsythe (1976). Smoothing techniques based on the OLS estimation are applied to reduce the effect of the outliers in the sample. The basic idea is to replace the most extreme residual with the next closest residual in the sample in an iterative way. In the literature, the studies show that Winsorization does not worsen a good linear relationship on non-contaminated data. On the contrary, it improves the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, when the sample is contaminated with outliers.

Trimmed Least Squares

The fourth method is the TLS introduced by Rousseeuw in 1984. The TLS estimation procedure is similar to the OLS estimation, but in TLS procedure, the fit is not so much affected from the outliers, because the data points corresponding to a specified percentage of the highest residuals based on an initial OLS estimation are removed. The OLS estimates of slope and intercept for the remaining data are called TLS estimates, see Rousseeuw and Leroy (1987) and Nevitt and Tam (1998). The aim is to minimize

$$\sum_{i=1}^h (y_i - \beta_0 - \beta_1 x_i)^2 \quad (6)$$

As it is seen in equation (6), rather than smoothing the data as in Winsorized regression, the outlying cases are deleted, therefore the $n-h$ observations do not affect the estimators.

Theil's Method

Theil's nonparametric regression method using the median as robust measures (see Theil, 1950) is presented. In Theil's

method, the only assumption is that the error terms are identically and independently distributed (i.i.d); this is different than the robust methods.

Sprent (1993) stated that for a simple linear regression model to obtain the slope of a line that fits the data points, the set of all slopes

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i}$$

of lines joining pairs of data points $(x_i, y_i), (x_j, y_j), x_j \neq x_i$, for $1 \leq i < j \leq n$ should be calculated.

Hussain and Sprent (1983) say that no generality is lost if $1 \leq i < j \leq n$ is taken, assuming that the x_i 's are arranged in ascending order (note that $b_{ij} = b_{ji}$). According to these results the Theil's slope estimator is:

$$\hat{\beta}_1 = med\{b_{ij} | x_j \neq x_i\}$$

where $x_1 \leq x_2 \leq \dots \leq x_n$.

It is known that median estimators are less affected from the outlying values in the data set as compared to the mean estimators, i.e., they are resistant estimators. The corresponding intercept term is defined as the median of the $y_i - \hat{\beta}_1 x_i$ terms (see Birkes & Dodge, 1993).

Weighted Theil's Method

A modified version of the Theil's method is called a Weighted Theil's Regression Method. In this method, different than the Theil's original method, each of the pairwise slopes are weighted using a weighting scheme. The weighted Theil slope estimator for the n observations in the sample data is the weighted median of these b_{ij} 's. w_{ij} , as the weighting procedure, can be taken as

$$x_j - x_i, j - i \text{ or } |x_j - x_i|,$$

see, for example Jaeckel (1972) and Scholz [16] and Birkes and Dodge (1993). In this study, the

weights $w_{ij} = \frac{(x_i - x_j)^2}{\sum (x_i - x_j)^2}$ were used to

calculate the slope estimator $\hat{\beta}_1 = \sum w_{ij} b_{ij}$. The intercept estimator is calculated in a similar fashion as in Theil's original method.

Results

The design points x_i ($1 \leq i \leq n$) follow an equally spaced, sequential additive series ($x_i = 1, 2, \dots, n$) (see Hussain & Sprent, 1983) and are common to all random samples (y_1, y_2, \dots, y_n) for the $N = [100,000/n]$ (integer) Monte Carlo runs. The error terms, e_i , are generated from the long-tailed symmetric distribution given above, and β_0, β_1 and σ are taken to be 0, 1 (1 in the remainder of this article) without loss of generality. The simulated means, variances and mean square errors (MSE) of the estimators are computed for some selected values of p (2.0, 2.5, 3.0, 3.5 and 5.0) and the results are given in Table 1.

From the simulation results presented in Table 1, all of the methods of estimation produced negligible bias therefore comparisons may be made in terms of MSE for both $\hat{\beta}_0$ and $\hat{\beta}_1$. In view of MSE, the following conclusions are put forth for the intercept estimator $\hat{\beta}_0$:

- WIN20 and WIN10 outperformed other estimators at all sample sizes for $p < 3$. For moderate ($n = 20$) and large sample sizes ($n=50$) they had the smallest MSE when $p = 3.0$. For values of the shape parameter p greater than 3, WIN20 and WIN10 were the preferred estimators for large sample sizes ($n=50$).
- The performance of the MML is best for small sample sizes ($n=10$) when $p=3$. When $p = 3.5$ and 5, the highest performance was achieved by MML for small ($n=10$) and moderate ($n=20$) samples.
- LAD and TLS performed poorly at all sample sizes for all values of the shape parameter p . As expected, the performance of OLS was the worst for $p = 2.5$, however, it consistently increased with the value of

MUTAN & ŞENOĞLU

Table 1: Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, $n=10$

Method	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
p=2.0						
OLS	0.003516	0.442733	0.442745	0.998563	0.011514	0.011516
MML	0.004207	0.341639	0.341656	0.998604	0.009006	0.009008
LAD	-0.002318	0.362488	0.362493	0.999963	0.009799	0.009799
WIN10	0.000232	0.361181	0.361181	0.999536	0.009411	0.009411
WIN20	0.001934	0.300163	0.300167	0.999102	0.007824	0.007825
TLS	0.006592	0.329992	0.330035	0.998576	0.008706	0.008708
Theil	0.000764	0.314738	0.314738	0.999397	0.008095	0.008096
Wtd.Theil	0.001506	0.312057	0.312059	0.999328	0.008060	0.008060
P=2.5						
OLS	-0.003356	0.461119	0.461130	1.000817	0.012041	0.012042
MML	-0.003358	0.413896	0.413908	1.000816	0.010877	0.010878
LAD	-0.001238	0.494956	0.494957	1.000694	0.013322	0.013322
WIN10	-0.003894	0.459236	0.459251	1.000988	0.012092	0.012093
WIN20	-0.001129	0.385763	0.385764	1.000446	0.010191	0.010191
TLS	-0.002565	0.445692	0.445699	1.000634	0.011855	0.011855
Theil	-0.002769	0.413026	0.413033	1.000909	0.010785	0.010786
Wtd.Theil	-0.000713	0.407067	0.407068	1.000667	0.010531	0.010531
P=3.0						
OLS	-0.002395	0.459847	0.459853	1.000911	0.012078	0.012079
MML	-0.001450	0.410860	0.410862	1.000782	0.010912	0.010913
LAD	0.003457	0.556958	0.556970	0.999881	0.015020	0.015020
WIN10	0.002749	0.475428	0.475435	0.999938	0.012637	0.012637
WIN20	0.001543	0.415174	0.415177	1.000308	0.010967	0.010968
TLS	-0.002892	0.485833	0.485841	1.000915	0.012907	0.012908
Theil	0.000275	0.448417	0.448417	1.000647	0.011503	0.011503
Wtd.Theil	0.000458	0.438228	0.438228	1.000618	0.011210	0.011210
P=3.5						
OLS	-0.013050	0.470511	0.470681	1.000804	0.012082	0.012082
MML	-0.010891	0.434622	0.434741	1.000796	0.011308	0.011309
LAD	-0.012993	0.594436	0.594605	1.001073	0.016032	0.016034
WIN10	-0.014704	0.510295	0.510512	1.001517	0.013519	0.013521
WIN20	-0.010134	0.446861	0.446964	1.000764	0.011649	0.011650
TLS	-0.009950	0.524629	0.524728	1.000705	0.013743	0.013743
Theil	-0.009799	0.472920	0.473016	1.000470	0.011964	0.011964
Wtd.Theil	-0.009878	0.470252	0.470350	1.000552	0.011806	0.011806
P=5.0						
OLS	0.006726	0.473619	0.473664	0.999226	0.012242	0.012243
MML	0.006238	0.459306	0.459345	0.999366	0.011917	0.011917
LAD	0.005333	0.653941	0.653969	0.999511	0.017332	0.017332
WIN10	0.004859	0.542576	0.542600	0.999847	0.014320	0.014320
WIN20	0.006534	0.482789	0.482832	0.999342	0.012526	0.012526
TLS	0.005403	0.587715	0.587744	0.999960	0.015314	0.015314
Theil	0.005450	0.523069	0.523098	0.999733	0.013058	0.013058
Wtd.Theil	0.007827	0.507404	0.507465	0.999458	0.012595	0.012596

LONG- TAILED SYMMETRIC DISTRIBUTION REGRESSION ESTIMATORS

Table 1 (continued): Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, $n=20$

<i>m</i>	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
p=2.0						
OLS	-0.002366	0.214414	0.214420	1.000324	0.001462	0.001463
MML	-0.001976	0.141376	0.141380	1.000244	0.000964	0.000964
LAD	-0.016728	0.152118	0.152398	1.001465	0.001097	0.001099
WIN10	-0.000748	0.168078	0.168079	1.000103	0.001165	0.001165
WIN20	-0.000829	0.128047	0.128048	1.000164	0.000879	0.000879
TLS	-0.000038	0.144824	0.144824	1.000100	0.001000	0.001000
Theil	0.007148	0.132135	0.132187	0.999243	0.000896	0.000897
Wtd.Theil	-0.000949	0.130729	0.130730	1.000045	0.000880	0.000880
P=2.5						
OLS	-0.004576	0.210169	0.210190	1.000161	0.001458	0.001458
MML	-0.005110	0.173648	0.173674	1.000131	0.001211	0.001211
LAD	-0.024790	0.207893	0.208508	1.001937	0.001483	0.001487
WIN10	0.000021	0.205904	0.205904	0.999652	0.001469	0.001469
WIN20	-0.006372	0.161651	0.161691	1.000210	0.001144	0.001144
TLS	-0.006068	0.186094	0.186131	1.000211	0.001325	0.001325
Theil	0.005634	0.173694	0.173725	0.999042	0.001185	0.001186
Wtd.Theil	-0.005509	0.171549	0.171580	1.000126	0.001166	0.001167
P=3.0						
OLS	-0.000997	0.217897	0.217898	1.000303	0.001517	0.001518
MML	-0.001199	0.190935	0.190936	1.000301	0.001320	0.001320
LAD	-0.015553	0.236484	0.236726	1.001847	0.001681	0.001684
WIN10	-0.001128	0.227378	0.227379	1.000144	0.001614	0.001614
WIN20	0.000029	0.181401	0.181401	1.000151	0.001256	0.001256
TLS	0.002355	0.211460	0.211466	1.000057	0.001474	0.001474
Theil	0.014359	0.195260	0.195466	0.999013	0.001304	0.001305
Wtd.Theil	0.001850	0.192893	0.192896	1.000188	0.001278	0.001278
P=3.5						
OLS	-0.005599	0.215732	0.215764	1.001062	0.001529	0.001530
MML	-0.002278	0.193426	0.193431	1.000902	0.001370	0.001371
LAD	-0.019423	0.262883	0.263260	1.002378	0.001877	0.001882
WIN10	-0.002735	0.242673	0.242680	1.000908	0.001750	0.001751
WIN20	-0.001386	0.195807	0.195809	1.000829	0.001384	0.001385
TLS	0.003151	0.232698	0.232707	1.000321	0.001637	0.001637
Theil	0.008258	0.211309	0.211377	0.999694	0.001439	0.001439
Wtd.Theil	-0.003741	0.209351	0.209365	1.000870	0.001413	0.001414
P=5.0						
OLS	-0.001472	0.206327	0.206329	1.000286	0.001458	0.001458
MML	-0.001661	0.196991	0.196994	1.000312	0.001395	0.001395
LAD	-0.019671	0.282823	0.283210	1.002131	0.002007	0.002011
WIN10	0.002690	0.250279	0.250286	0.999782	0.001833	0.001833
WIN20	-0.002567	0.202167	0.202173	1.000406	0.001418	0.001419
TLS	-0.003974	0.243164	0.243180	1.000674	0.001704	0.001704
Theil	0.013557	0.220453	0.220637	0.999055	0.001461	0.001462
Wtd.Theil	0.001284	0.217649	0.217651	1.000161	0.001438	0.001438

MUTAN & ŞENOĞLU

Table 1 (continued): Means, Variances and MSE's for the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, $n=50$

Method	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
p=2.0						
OLS	-0.000594	0.084085	0.084085	1.000087	0.000104	0.000104
MML	0.001371	0.047691	0.047692	1.000004	0.000055	0.000055
LAD	0.000378	0.052987	0.052987	0.999985	0.000063	0.000063
WIN10	0.006258	0.065587	0.065626	0.999828	0.000079	0.000079
WIN20	0.000867	0.046586	0.046586	1.000034	0.000055	0.000055
TLS	0.001880	0.050630	0.050634	1.000005	0.000060	0.000060
Theil	-0.000014	0.047390	0.047390	1.000006	0.000054	0.000054
Wtd.Theil	-0.000386	0.047201	0.047202	1.000025	0.000053	0.000053
P=2.5						
OLS	0.002424	0.086628	0.086634	0.999785	0.000099	0.000099
MML	-0.000641	0.066412	0.066413	0.999911	0.000076	0.000076
LAD	0.001515	0.080364	0.080366	0.999884	0.000094	0.000094
WIN10	0.004181	0.091114	0.091131	0.999756	0.000108	0.000108
WIN20	-0.000303	0.064850	0.064850	0.999878	0.000075	0.000075
TLS	-0.004784	0.076472	0.076495	1.000056	0.000087	0.000087
Theil	0.002601	0.068287	0.068294	0.999896	0.000075	0.000075
Wtd.Theil	0.002145	0.068267	0.068272	0.999915	0.000075	0.000075
P=3.0						
OLS	-0.012133	0.085280	0.085428	1.000378	0.000100	0.000100
MML	-0.011707	0.073272	0.073409	1.000390	0.000085	0.000085
LAD	-0.012364	0.089730	0.089883	1.000416	0.000106	0.000106
WIN10	-0.007459	0.096523	0.096579	1.000233	0.000115	0.000115
WIN20	-0.009552	0.071668	0.071759	1.000327	0.000084	0.000084
TLS	-0.010372	0.078219	0.078326	1.000291	0.000094	0.000094
Theil	-0.009797	0.075452	0.075548	1.000351	0.000084	0.000084
Wtd.Theil	-0.009190	0.074861	0.074945	1.000330	0.000083	0.000083
P=3.5						
OLS	-0.013384	0.081143	0.081322	1.000466	0.000093	0.000093
MML	-0.012900	0.070895	0.071062	1.000445	0.000082	0.000082
LAD	-0.009534	0.092041	0.092131	1.000356	0.000108	0.000108
WIN10	-0.012675	0.089156	0.089317	1.000384	0.000108	0.000108
WIN20	-0.012857	0.069653	0.069818	1.000447	0.000081	0.000081
TLS	-0.012912	0.079559	0.079725	1.000442	0.000093	0.000093
Theil	-0.012624	0.077350	0.077510	1.000469	0.000083	0.000083
Wtd.Theil	-0.012442	0.076734	0.076889	1.000476	0.000082	0.000082
P=5.0						
OLS	0.000349	0.080924	0.080924	1.000022	0.000093	0.000093
MML	-0.002554	0.075887	0.075893	1.000110	0.000088	0.000088
LAD	-0.004909	0.110364	0.110388	1.000214	0.000129	0.000129
WIN10	0.000915	0.100494	0.100495	1.000063	0.000122	0.000122
WIN20	-0.001840	0.076396	0.076399	1.000070	0.000088	0.000088
TLS	-0.002449	0.093636	0.093642	1.000074	0.000108	0.000108
Theil	-0.003242	0.083709	0.083720	1.000146	0.000090	0.000090
Wtd.Theil	-0.002844	0.083042	0.083050	1.000120	0.000089	0.000089

LONG- TAILED SYMMETRIC DISTRIBUTION REGRESSION ESTIMATORS

the shape parameter p since OLS is the optimal method under normality and the $LTS(p, \sigma)$ distribution approaches normal as $p \rightarrow \infty$. Results were not reproduced for the sake of brevity, however.

For the slope estimator $\hat{\beta}_1$:

- For $p = 2$ and 2.5 , the performances of the WIN20 and WIN10 were the best at sample sizes 10 and 20 and Wtd.Theil and Theil provide the smallest MSE for the large sample sizes ($n = 50$).
- For $p = 3.0$, WIN20 demonstrated the strongest performance with lowest MSE at all sample sizes except for $n = 10$, in which case MML provides the smallest MSE.
- MML, WIN10 and Wtd.Theil were the preferred methods for $p = 3.5$. When $p = 5.0$, MML, WIN10 and WIN20 have the smallest MSE.
- The LAD and TLS slope estimators showed very poor performance with the largest MSE values at all sample sizes for all values of the shape parameter, p .
- The performance of the OLS slope estimator is similar to the OLS intercept estimator.

Robustness

In practice, a model is identified by Q-Q plots or goodness of fit tests. Neither of these methods, nor in fact any other method, identifies a model exactly or uniquely. In other words, the value of the shape parameter p in $LTS(p, \sigma)$ might be misspecified. Assume, for illustration, that the true distribution is the $LTS(3.5, \sigma)$. To represent a large number of plausible alternatives, consider the following sample models:

- Model (1): $LTS(2.0, \sigma)$
- Model (2): $LTS(5.0, \sigma)$
- Model (3): Outlier Model; $(n-r)$ observations from $LTS(3.5, \sigma)$ and r observations from $LTS(3.5, 4\sigma)$ where $r = [0.5 + 0.1n]$
- Model (4): Mixture Model; $0.90LTS(3.5, \sigma) + 0.10LTS(3.5, 4\sigma)$

- Model (5): Contamination Model; $0.90LTS(3.5, \sigma) + 0.10 Normal(0, 4)$

The simulated means, variances and MSE of the regression estimators for the alternative models are shown in Table 2. It should be noted that an estimator $\hat{\theta}$ of θ is called robust if it is fully efficient (or nearly so) for an assumed model but maintains high efficiencies for plausible alternatives to the assumed model. Based on the information in Table 2, the following conclusions are put forth for the intercept estimator $\hat{\beta}_0$:

- WIN10 and WIN20 showed the strongest performance with lowest MSE for Models (1), (3), (4) and (5) at all sample sizes except for a sample of size 50 in Models (1) and (5) in which case the Wtd. Theil provides the smallest MSE.
- MML demonstrated the strongest performance with lowest MSE as compared to other methods in Model (2).
- OLS and LAD showed very poor estimator performance at all sample sizes with largest MSE values for Models (1), (3), (4), (5) and Model (2), respectively.

For the slope estimator $\hat{\beta}_1$:

- WIN10 and WIN20 provided the smallest MSE for Models (1), (3), (4) and (5) at sample sizes 10 and 20, however, for the sample size $n = 50$, the Wtd. Theil's slope estimator had the strongest efficiency.
- The highest performance for Model (2), similar to intercept estimator $\hat{\beta}_0$, is achieved by MML.
- OLS and LAD have the highest MSE values for Models (1), (3), (4), (5) and Model (2), respectively. Therefore, they are not preferred estimators under these sample models.

Conclusion

The OLS estimation procedure provides good results when the error terms have a normal distribution. However, in real life, it is nearly impossible to find a data set that satisfies all of

Table 2: Means, Variances and MSE's for the sample models (1)-(5), $n=10$

Method	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
Model (1)						
OLS	-0.012290	0.485174	0.485325	1.001940	0.012996	0.013000
MML	-0.010016	0.352918	0.353018	1.001675	0.009657	0.009660
LAD	-0.009981	0.357653	0.357753	1.001655	0.009603	0.009605
WIN10	-0.007079	0.365626	0.365676	1.001175	0.009806	0.009807
WIN20	-0.010102	0.317492	0.317594	1.001656	0.007995	0.007998
TLS	-0.009358	0.325208	0.325295	1.001784	0.008635	0.008638
Theil	-0.010226	0.307593	0.307697	1.001484	0.008033	0.008035
Wtd.Theil	-0.012949	0.308182	0.308350	1.001986	0.008008	0.008012
Model (2)						
OLS	-0.006355	0.470296	0.470337	1.000850	0.012108	0.012109
MML	-0.006313	0.456893	0.456933	1.000915	0.011789	0.011790
LAD	-0.008034	0.656990	0.657055	1.001427	0.017477	0.017479
WIN10	-0.008389	0.546347	0.546417	1.001590	0.014378	0.014380
WIN20	-0.006451	0.480393	0.480435	1.001023	0.012422	0.012423
TLS	-0.005099	0.577213	0.577239	1.000909	0.015085	0.015086
Theil	-0.005327	0.520019	0.520047	1.000805	0.012966	0.012966
Wtd.Theil	-0.006527	0.507479	0.507522	1.001039	0.012562	0.012563
Model (3)						
OLS	0.012384	1.223769	1.223923	0.998354	0.032164	0.032167
MML	0.009457	0.753716	0.753806	0.998816	0.020260	0.020262
LAD	0.016420	0.751355	0.751625	0.997478	0.019980	0.019987
WIN10	-0.010434	0.754763	0.754872	1.000511	0.019574	0.019574
WIN20	0.008255	0.630553	0.630621	0.998781	0.016398	0.016400
TLS	0.012560	0.667084	0.667242	0.998366	0.017401	0.017404
Theil	0.007114	0.660818	0.660869	0.999191	0.016871	0.016872
Wtd.Theil	0.006276	0.658813	0.658853	0.999105	0.016838	0.016839
Model (4)						
OLS	-0.015771	1.169783	1.170031	1.003291	0.030152	0.030163
MML	-0.015086	0.776937	0.777164	1.002934	0.020509	0.020518
LAD	-0.022904	0.798735	0.799260	1.003815	0.021329	0.021343
WIN10	-0.026484	0.830451	0.831153	1.003516	0.021298	0.021310
WIN20	-0.013370	0.661862	0.662040	1.002650	0.017077	0.017084
TLS	-0.011041	0.710763	0.710885	1.002151	0.018798	0.018803
Theil	-0.016685	0.694106	0.694385	1.002835	0.017787	0.017795
Wtd.Theil	-0.015797	0.690942	0.691192	1.002849	0.017729	0.017737
Model (5)						
OLS	-0.004107	1.179549	1.179566	1.001699	0.030212	0.030215
MML	-0.001795	0.797778	0.797782	1.001125	0.021203	0.021204
LAD	0.004313	0.797272	0.797291	0.999694	0.021461	0.021461
WIN10	-0.011044	0.839572	0.839694	1.001791	0.022213	0.022217
WIN20	0.000062	0.684177	0.684177	1.000728	0.017888	0.017889
TLS	0.002536	0.742882	0.742889	1.000399	0.019719	0.019719
Theil	-0.000683	0.701909	0.701910	1.000896	0.018376	0.018377
Wtd.Theil	-0.001727	0.715698	0.715701	1.000978	0.018841	0.018842

LONG- TAILED SYMMETRIC DISTRIBUTION REGRESSION ESTIMATORS

Table 2 (continued): Means, Variances and MSE's for the sample models (1)-(5), $n=20$

Method	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
Model (1)						
OLS	-0.003752	0.222670	0.222685	0.999951	0.001546	0.001546
MML	-0.004145	0.138954	0.138971	1.000165	0.000966	0.000966
LAD	-0.014743	0.152306	0.152524	1.001444	0.001072	0.001074
WIN10	-0.004923	0.163155	0.163179	1.000060	0.001149	0.001149
WIN20	-0.001735	0.123830	0.123833	1.000051	0.000863	0.000863
TLS	-0.002840	0.142264	0.142272	1.000216	0.000988	0.000988
Theil	0.011090	0.129912	0.130035	0.998996	0.000873	0.000874
Wtd.Theil	0.001468	0.128770	0.128772	0.999909	0.000861	0.000861
Model (2)						
OLS	-0.009421	0.220500	0.220589	1.000323	0.001527	0.001527
MML	-0.007871	0.208822	0.208884	1.000299	0.001453	0.001453
LAD	-0.015340	0.296896	0.297132	1.001461	0.002084	0.002086
WIN10	-0.008323	0.263561	0.263631	1.000316	0.001861	0.001861
WIN20	-0.006477	0.212390	0.212432	1.000258	0.001475	0.001475
TLS	-0.000358	0.260313	0.260313	0.999871	0.001816	0.001816
Theil	0.010322	0.231483	0.231589	0.998944	0.001514	0.001515
Wtd.Theil	-0.002871	0.228255	0.228263	1.000186	0.001491	0.001491
Model (3)						
OLS	0.008763	0.534048	0.534125	0.998650	0.003708	0.003710
MML	0.009852	0.271805	0.271903	0.998716	0.001954	0.001955
LAD	-0.004580	0.312706	0.312727	0.999984	0.002225	0.002225
WIN10	0.007319	0.356939	0.356993	0.998841	0.002524	0.002525
WIN20	0.010996	0.254113	0.254234	0.998615	0.001813	0.001815
TLS	0.012679	0.292134	0.292295	0.998599	0.002067	0.002069
Theil	0.025097	0.270968	0.271598	0.997220	0.001856	0.001864
Wtd.Theil	0.010753	0.268535	0.268651	0.998603	0.001824	0.001826
Model (4)						
OLS	-0.011834	0.530361	0.530501	1.000266	0.003641	0.003641
MML	-0.007635	0.285413	0.285471	1.000330	0.002019	0.002019
LAD	-0.021156	0.320834	0.321282	1.001696	0.002262	0.002265
WIN10	-0.002971	0.383144	0.383153	0.999619	0.002664	0.002664
WIN20	-0.004989	0.263167	0.263192	1.000165	0.001853	0.001853
TLS	-0.002033	0.301227	0.301231	0.999793	0.002084	0.002084
Theil	0.007851	0.274877	0.274938	0.998851	0.001875	0.001876
Wtd.Theil	-0.005549	0.272119	0.272150	1.000128	0.001839	0.001839
Model (5)						
OLS	-0.014204	0.546967	0.547169	1.000832	0.003830	0.003830
MML	-0.007622	0.291418	0.291476	1.000401	0.002046	0.002046
LAD	-0.018763	0.323247	0.323599	1.001347	0.002247	0.002249
WIN10	-0.012408	0.388890	0.389044	1.000889	0.002683	0.002684
WIN20	-0.007292	0.271799	0.271852	1.000305	0.001893	0.001893
TLS	-0.000146	0.296508	0.296508	0.999684	0.002040	0.002040
Theil	0.006440	0.283805	0.283846	0.999057	0.001913	0.001914
Wtd.Theil	-0.007353	0.281584	0.281638	1.000388	0.001892	0.001892

Table 2 (continued): Means, Variances and MSE's for the sample models (1)-(5), $n=50$

Method	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Mean	Variance	MSE	Mean	Variance	MSE
Model (1)						
OLS	-0.000805	0.074541	0.074542	0.999982	0.000084	0.000084
MML	-0.000114	0.044790	0.044790	0.999976	0.000050	0.000050
LAD	0.004056	0.051563	0.051579	0.999809	0.000059	0.000059
WIN10	-0.002538	0.065713	0.065719	1.000100	0.000077	0.000077
WIN20	0.000232	0.045412	0.045412	0.999965	0.000052	0.000052
TLS	0.001835	0.048722	0.048726	0.999970	0.000055	0.000055
Theil	-0.000018	0.044688	0.044688	0.999963	0.000050	0.000050
Wtd.Theil	-0.000159	0.044719	0.044719	0.999970	0.000050	0.000050
Model (2)						
OLS	-0.005070	0.082616	0.082642	1.000184	0.000095	0.000095
MML	-0.003343	0.078478	0.078489	1.000146	0.000091	0.000091
LAD	-0.004388	0.107137	0.107156	1.000242	0.000124	0.000124
WIN10	-0.002386	0.102003	0.102008	1.000153	0.000125	0.000125
WIN20	-0.002738	0.079004	0.079012	1.000119	0.000092	0.000092
TLS	-0.001691	0.099586	0.099589	1.000081	0.000116	0.000116
Theil	-0.001652	0.086823	0.086826	1.000123	0.000093	0.000093
Wtd.Theil	-0.001529	0.086498	0.086501	1.000126	0.000092	0.000092
Model (3)						
OLS	0.001170	0.218986	0.218987	1.000070	0.000247	0.000247
MML	0.005980	0.138048	0.138083	0.999923	0.000149	0.000149
LAD	0.007935	0.116472	0.116535	0.999940	0.000131	0.000131
WIN10	0.006277	0.158747	0.158786	0.999961	0.000176	0.000176
WIN20	0.007735	0.102137	0.102197	0.999891	0.000112	0.000112
TLS	0.005534	0.118133	0.118164	1.000001	0.000130	0.000130
Theil	0.008831	0.099889	0.099967	0.999887	0.000108	0.000109
Wtd.Theil	0.009458	0.100008	0.100097	0.999849	0.000108	0.000108
Model (4)						
OLS	0.007550	0.213060	0.213117	0.999828	0.000249	0.000249
MML	0.008803	0.134974	0.135051	0.999741	0.000155	0.000155
LAD	0.009354	0.118182	0.118269	0.999582	0.000142	0.000142
WIN10	0.014525	0.166463	0.166674	0.999435	0.000197	0.000197
WIN20	0.007484	0.104324	0.104380	0.999757	0.000123	0.000123
TLS	0.003825	0.115579	0.115593	0.999892	0.000138	0.000138
Theil	0.006978	0.103516	0.103565	0.999747	0.000119	0.000119
Wtd.Theil	0.007271	0.103704	0.103757	0.999727	0.000118	0.000119
Model (5)						
OLS	0.000823	0.213641	0.213642	1.000111	0.000251	0.000251
MML	0.001214	0.139224	0.139226	1.000019	0.000158	0.000158
LAD	-0.006313	0.123031	0.123071	1.000148	0.000146	0.000146
WIN10	0.002004	0.175000	0.175004	1.000008	0.000198	0.000198
WIN20	0.000914	0.109873	0.109874	0.999948	0.000129	0.000129
TLS	0.001631	0.116706	0.116709	0.999897	0.000135	0.000135
Theil	-0.000120	0.107528	0.107528	0.999936	0.000122	0.000122
Wtd.Theil	-0.000712	0.107393	0.107394	0.999947	0.000122	0.000122

the normality assumptions, therefore, alternative regression methods are needed. In this study, efficiency and robustness properties of some prominent robust and nonparametric regression estimators have been compared via Monte Carlo simulation when the error terms come from long-tailed symmetric $LTS(p, \sigma)$ distributions.

The methods giving the smallest MSE for various shape parameters and sample models were defined clearly for different sample sizes. If the distribution of error terms is $LTS(p, \sigma)$ in a simple linear regression model, it is therefore suggested that the selection procedure for the most efficient and robust method of estimation should be accomplished according to the results given above.

References

- Birkes, D., & Dodge, Y. (1993). *Alternative methods of regression*. New York, NY: Wiley.
- Geary, R. C. (1947). Testing for normality. *Biometrika*, 34, 209-242.
- Huber, P. J. (1981). *Robust statistics*. NY: John Wiley.
- Hussain, S. S., & Sprent, P. (1983). Nonparametric regression. *Journal of the Royal Statistical Society*, A146, 182-191.
- Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Annals of Mathematical Statistics*, 43, 1449-1458.
- Nevitt, T., & Tam, H. P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model. *Multiple Linear Regression Viewpoints*, 25, 54-69.
- Pearson, E. S. (1932). The analysis of variance in cases of nonnormal variation. *Biometrika*, 23, 114-133.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: Wiley.
- Scholz, F.W. (1978). Weighted median regression estimates. *The Annals of Statistics*, 6(3), 603-609.
- Sprent, P. (1993). *Applied nonparametric statistical methods*. NY: Chapman and Hall.
- Şenoğlu, B. (2005). Robust 2^k factorial design with Weibull error distributions. *Journal of Applied Statistics*, 32(10), 1051-1066.
- Şenoğlu, B. (2007). Estimating parameters in one-way analysis of covariance model with short-tailed symmetric error distributions. *Journal of Computational and Applied Mathematics*, 201, 275-283.
- Tam, H. P. (1996). *A review of nonparametric regression techniques*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, 12, 85-91.
- Tiku, M. L., Islam, M. Q., & Selçuk, A. (2001). Non-normal regression II: Symmetric distributions. *Communications in Statistics Theory and Methods*, 30, 1021-1045.
- Yule, C., & Forsythe, A. B. (1976). Winsorized regression. *Technometrics*, 18, 291-300.

The Comparison of Model Selection Criteria When Selecting Among Competing Hierarchical Linear Models

Tiffany A. Whittaker
The University of Texas at Austin

Carolyn F. Furlow
Georgia State University

Little is known about the use and accuracy of model selection criteria when selecting among a set of competing multilevel models. The practices of applied researchers and the performance of five model selection criteria are examined when selecting the correct multilevel model using simulation techniques.

Key words: Hierarchical linear modeling, multilevel modeling, model selection criteria.

Introduction

Researchers are typically interested in comparing the fit of various theoretically plausible models to data. Hierarchical Linear Modeling (HLM), or multilevel modeling, has become a widely used tool to aid in the explanation of predictive theoretical models within the social and behavioral sciences. As is common with other statistical techniques (e.g., multiple linear regression, structural equation modeling), there exist various criteria for model comparison and selection within the HLM arena. Little is known, however, about the accuracy of various selection criteria within the HLM arena.

The purpose of this article is twofold: (1) to examine the current practices of researchers in the field when comparing and selecting hierarchical linear models; and (2) to examine the performance of various model selection techniques with respect to selecting the correct hierarchical linear model from a group of competing models.

Tiffany Whittaker is an Assistant Professor in the Department of Educational Psychology in the College of Education. Email: tiffany.whittaker@mail.utexas.edu. Carolyn Furlow is a statistician with the Science Applications International Corporation. Email: carolyn.f.furlow@saic.com.

Model Comparison and Selection Criteria in the HLM Arena

One method for the comparison of nested hierarchical linear models is the Chi-square difference test. The Chi-square difference test incorporates the deviance statistic in its calculation. The deviance statistic is given as:

$$-2[LL_{current\ model} - LL_{saturated\ model}] \quad (1)$$

where $LL_{current\ model}$ is the Log Likelihood (LL) value obtained from fitting the proposed model to the data. $LL_{saturated\ model}$ is the Log Likelihood value of fitting the best possible fitting model, the saturated model, to the data which results in a LL value of zero. Consequently, the deviance statistic reduces to $-2LL$.

The difference between two nested models' deviance statistics, which is asymptotically distributed as a Chi-square (χ^2) statistic, may be used to determine if a significant difference between the two models exists when adding or eliminating model parameters:

$$\chi^2_{-2LL\ difference} = -2LL_{restricted} - 2LL_{unrestricted}, \quad (2)$$

where $-2LL_{restricted}$ is the deviance statistic for the nested, less parameterized (restricted) model and $-2LL_{unrestricted}$ is the deviance statistic for

MODEL SELECTION CRITERIA WITH HLM

the more parameterized, less restricted (unrestricted) model, with corresponding degrees of freedom equal to the difference in the number of parameters estimated (q) in each model:

$$df_{-2LL\text{difference}} = q_{\text{restricted}} - q_{\text{unrestricted}} \quad (3)$$

When the $\chi^2_{-2LL\text{difference}}$ indicates a significant difference between two hierarchically related models, the nested model with less parameters has been oversimplified. That is, the less parameterized (nested) model has significantly decreased the overall fit of the model when compared to the model with more parameters. In this situation, then, the more parameterized model would be selected over the less parameterized model. On the other hand, when the $\chi^2_{-2LL\text{difference}}$ test is not significant, the two models are comparable in terms of overall model fit. In this situation, the less parameterized would most likely be selected over the more parameterized model in support of parsimony.

When hierarchical linear models are non-nested, the $\chi^2_{-2LL\text{difference}}$ test is an inappropriate method to assess significant model fit differences because neither of the two models can serve as a baseline comparison model. Still, there are instances in which different theoretical models posited to support the data are non-nested. In this situation, information criteria may be used for model comparison and selection. The benefit of using information criteria in the model selection process is that they may be used to compare and select among a set of nested and/or non-nested models.

The most popular information criterion is Akaike's (1973) information criterion (AIC) which compensates for the number of parameters in the model to encourage parsimony:

$$AIC = -2LL + 2q, \quad (4)$$

where $-2LL$ is the deviance statistic for a given model and q is the number of parameters estimated in the given model. When comparing two competing models, the model with the

lowest AIC value would be selected as the model demonstrating better fit than its comparison model.

The AIC is asymptotically efficient, meaning that it will select the best finite dimensional model (closest to the correct/true model) if the correct/true model is infinite dimensional. The AIC, however, has often been criticized for lack of consistency (Bozdogan, 1987; Hannon & Quinn, 1979; Hurvich & Tsai, 1989; Schwarz, 1978). Consistent model selection criteria select the correct/true model reliably (probabilities close to or at 1) when the correct/true model exists among the set of competing models. In addition, the AIC has been shown to incorrectly select more highly parameterized models, particularly when the ratio of estimated parameters to sample size is large (Hurvich & Tsai, 1989). Consequently, additional information criteria, which have extended the AIC to account for both model complexity and sample size, have been proposed.

Although various information criteria exist, this paper will focus on the information criteria readily available in current versions of SAS's PROC MIXED (version 9.2; SAS Institute Inc., 2007) and/or SPSS when using the Mixed Models command (version 16.0; SPSS Inc., 2007). SAS's PROC MIXED is a commonly used multilevel software program and with the recent addition of the Mixed Models command in SPSS, it too should become increasingly used when conducting multilevel analyses. Both software programs are able to provide more than one information criterion in the output. These include the Bayesian information criterion (BIC; Schwarz, 1978):

$$BIC = -2LL + \ln(N)q; \quad (5)$$

Hannon and Quinn's (1979) information criterion, *which is only available in SAS* (HQIC):

$$HQIC = -2LL + 2q\ln(\ln(N)); \quad (6)$$

and Bozdogan's (1987) consistent AIC (CAIC):

$$CAIC(k) = -2LL + [\ln(N) + 1]q; \quad (7)$$

where \ln is the natural log and N is the sample size. While the BIC, HQIC, and CAIC were proposed to be more consistent model selection criteria, Hurvich and Tsai (1989) proposed a criterion that extends the AIC to correct for its tendency to overfit models (select highly parameterized models) which is still asymptotically efficient, called the finite sample corrected AIC (AICC):

$$\text{AICC} = -2LL + 2qN/(N-q-1). \quad (8)$$

There remains debate concerning which model selection feature is best (efficiency versus consistency). Some may argue that models are simply approximations of the truth and that researchers will never know if the true model exists in their set of competing models, supporting the use of efficient model selection criteria. Others, however, may argue that they are able to measure all relevant variables and thus have the correct model in their set of competing models, supporting the use of consistent model selection criteria.

The point of this article is not to argue in favor of either efficiency or consistency as it depends upon the context and the discipline (Burnham & Anderson, 2002; McQuarrie & Tsai, 1998; Shi & Tsai, 2002). Nonetheless, the current paper will assess the performance of these five model selection criteria (both efficient and consistent ones) in terms of selecting the correct multilevel model from among a set of competing incorrect models. This performance standard does support the definition of consistency; unfortunately, it is difficult to assess the performance of these model selection criteria otherwise.

To our knowledge, there is no study that has compared the performance of all five of these information criteria with respect to selecting the correct model among a set of competing models in the HLM arena. The most recent and relevant study was conducted by Gurka (2006) who examined the performance of the AIC, AICC, BIC, and CAIC in terms of selecting the correct multilevel growth curve model under various conditions, including different sample sizes, total variances, ICC values, model misspecification, criteria calculation, and estimation methods.

The model selection criteria were assessed in three different scenarios: 1) the ability to select the correct fixed effects given a compound symmetric covariance structure; 2) the ability to select the correct random effects given the fixed effects in the model; and 3) the ability to select the correct fixed and random effects in the model. Overall, the results indicated that the BIC and CAIC tended to outperform both the AIC and AICC. In addition, the AICC tended to outperform the AIC when selecting the correct model. None of the criteria performed well under the small sample size condition (with 25 cases at level-2 and 3 observations within each case). All four criteria performed well when selecting the correct random effects model (in more than 90% of the replications), regardless of total variance and ICC conditions. When selecting the correct fixed effects only and the correct fixed and random effects models, the criteria performed worse as the ICC values increased with the larger total variance conditions.

The impetus behind Gurka's (2006) study was the interest in comparing these criteria under different estimation methods available in multilevel software packages. The five model selection criteria presented in Equations 4 through 8 above are calculated under full information maximum likelihood (FIML) estimation as opposed to restricted maximum likelihood (REML) estimation in which the calculations change a bit with respect to N and q . When using FIML, the likelihood function contains both the fixed effects and the random effects (Raudenbush & Bryk, 2002). REML, however, rests on the assumption that fixed effects are uncertain and should be estimated separately from the random effects. It has been argued that deviance statistics, as well as the information criteria, of different models can be compared when the models differ only in their random effects under REML estimation while the deviance statistics, as well as the information criteria, of different models can be compared when the models differ in their fixed effects or their random effects under FIML estimation (Verbeke & Molenberghs, 2000).

Gurka (2006) questioned why the information criteria calculated under REML estimation could not be used in the model

MODEL SELECTION CRITERIA WITH HLM

selection process when comparing models containing fixed effects. As a result, he compared the performance of the four model selection criteria under FIML and REML estimation conditions. The findings indicated that the selection criteria performed better or equally well under REML estimation compared to FIML estimation when selecting the fixed effects model. As Gurka (2006) noted, the question as to whether the information criteria of fixed effects models may be compared under REML estimation should be examined further. In spite of this, the current paper does not examine this question. Instead, FIML estimation will solely be used as the models compared in the current paper differ with respect to their fixed and random effects.

SAS and SPSS differ with respect to the calculations of the BIC and the CAIC. More specifically, sample size in SAS is equal to the number of observations at level-2 (m) whereas sample size in SPSS is equal to the total number of observations at level-1 (N) when calculating the BIC and the CAIC under FIML estimation. The AICC, however, is calculated identically in both SAS and SPSS, using the total number of observations at level-1 (N) in the calculation. In cross-sectional designs, it seems reasonable to use the number of observations at level-1 as N in the calculation of these criteria.

In contrast, it seems more reasonable to use the number of observations at level-2 (m) in the calculation of these criteria in growth curve modeling designs. Additionally, Raudenbush and Liu (2000) reported that in their research on power with HLM designs, the sample size at level-2 was typically more important for power than the sample size at level-1. This research would also seem to indicate the utility of using m in the calculation of these criteria. Gurka (2006) also examined the performance of the model selection criteria (AICC, BIC, and CAIC) when using N versus m in their calculation. The results indicated that the criteria tended to perform better in terms of selecting the correct model when they were calculated using the number of observations at level-2 (m) as opposed to the number of observations at level-1 (N) under FIML estimation.

To summarize, there is no study, to our knowledge, that has examined all five model

selection criteria (AIC, AICC, BIC, CAIC, and HQIC) simultaneously within the HLM arena. Gurka (2006) recently examined all of these criteria, with the exception of the HQIC. Hence, it is unknown how the HQIC will compare with the remaining model selection criteria examined in his study under different conditions. In addition, Gurka used a fairly simple correct model, both in terms of fixed and random effects, with only two predictors included in the model, and the single slope coefficient from level-1 was not allowed to randomly vary at level-2. This is unfortunate as researchers commonly allow slopes to vary randomly and the capability to model random slopes is a major advantage of multilevel modeling.

Researchers are also typically interested in examining more complex models that include more than just two predictors. Consequently, it is unclear how the model selection criteria will perform when comparing a set of simple models versus more complicated models. In addition, because Gurka was interested in how the criteria perform under a growth curve modeling context, the sample sizes used in his study were not reflective of those found in typical HLM designs where individuals are nested within groups. Thus, the purpose of this article is to examine the performance of all five model selection criteria in terms of selecting the correct multilevel model (with slopes allowed to randomly vary) under various conditions, including criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and ICCs.

Content Analysis

In order to evaluate the use of model selection criteria within the HLM arena, a content analysis was conducted. When conducting the content analysis, several different characteristics were assessed. More specifically, interest was placed on 1) the frequency with which model selection criteria are used by applied researchers when selecting among competing hierarchical models; 2) the types of model selection criteria used by applied researchers in the model comparison/selection process; and 3) if model selection criteria were

used, what multilevel software package was used when conducting the analyses.

Content Analysis Procedure

To assess these characteristics, a search in PsycInfo was conducted using the following search terms: “HLM,” “Hierarchical Linear Modeling,” “Multilevel Modeling,” and “Random Effects Modeling.” All applied articles using HLM techniques published between January 2002 and March 2007 were collected. Two hundred twenty articles were collected as a result of this search. These 220 articles were examined in order to collect information concerning the three characteristics mentioned above (see Table 1 for complete information on the content analysis characteristics).

Content Analysis Results

Of the 220 articles reviewed, the authors of 45 articles reported using some form of model selection criteria whereas the authors of 175 articles did not report using model selection criteria. The most commonly used model selection criteria was the Chi-square difference test followed by both the AIC and the BIC used together. Neither the AICC, the CAIC, nor the HQIC was used in any of these reviewed articles. The articles in which model selection criteria were used were also reviewed to determine what type of multilevel software package was used to conduct the analyses. The most popular software used was HLM (Raudenbush, Bryk, & Congdon, 2007) followed by MLwiN (Rasbash, Charlton, Browne, Healy, & Cameron, 2005) and SAS’s PROC MIXED (SAS Institute Inc., 2003). LISREL (Jöreskog & Sörbom, 2006), MIXREG (Hedeker & Gibbons, 1999), and Mplus (Muthén & Muthén, 2007) were used less frequently. The authors of the remaining 20 articles did not report which software package was used. For a list of all the articles collected in this study, please contact the first author.

Methodology

Simulation Study

A Monte Carlo simulation study was conducted in order to examine the performance of five different model selection criteria when

selecting the correct multilevel model from a group of competing multilevel models. The performance of these criteria was examined under varied conditions, including criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and the intraclass correlation (ICC) value.

Model Selection Criteria Calculation

The five model selection criteria (AIC, AICC, BIC, CAIC, HQIC) were examined under all conditions. To compare whether m or N is best in the calculation of the criteria (except the AIC as sample size is not used in its calculation), the AICC, BIC, CAIC, and HQIC were calculated in all conditions using the number of observations at level-2 (m ; as calculated in SAS) and the number of observations at level-1 (N ; as calculated in SPSS), resulting in the following nine model selection criteria: the AIC, the AICC $_m$, the AICC $_N$, the BIC $_m$, the BIC $_N$, the CAIC $_m$, the CAIC $_N$, the HQIC $_m$, and the HQIC $_N$. Although the Chi-square difference test was more commonly used by applied researchers, as demonstrated by the content analysis, a number of the misspecified models (described below) examined in this study were non-nested, rendering the Chi-square difference test ineffectual across all possible model comparisons. Therefore, the Chi-square difference test was not used as one of the model selection criteria.

Model Complexity

To examine whether the model selection criteria would perform differently when selecting among a simple set of multilevel models versus a more complex set of multilevel models, a simple generating model and a complex generating model were used. The simple generating model (Simple Model 1) consisted of a two-level model in which one predictor is included at both the participant-level (level-1) and the group-level (level-2) and is as follows:

MODEL SELECTION CRITERIA WITH HLM

$$\begin{aligned}
 & \text{Level - 1:} \\
 & Y_{ij} = \beta_{0j} + \beta_{1j}X1 + r_{ij} \\
 & \text{Level - 2:} \\
 & \beta_{0j} = \gamma_{00} + \gamma_{01}W1 + u_{0j} \\
 & \beta_{1j} = \gamma_{10} + u_{1j}
 \end{aligned} \tag{9}$$

In the simple generating model, the parameters from level-1 were all allowed to randomly vary at level-2. However, there was no cross-level interaction between X1 and W1. The variance/covariance matrix at level-2 associated with this model is:

$$\text{Var} \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{pmatrix}. \tag{10}$$

The complex generating model (Complex Model 1) consisted of the same variance/covariance structure as the simple model but included a more complex fixed effects structure in which two predictors were included at both the participant-level (level-1) and the group-level (level-2):

$$\begin{aligned}
 & \text{Level - 1:} \\
 & Y_{ij} = \beta_{0j} + \beta_{1j}X1 + \beta_{2j}X2 + r_{ij} \\
 & \text{Level - 2:} \\
 & \beta_{0j} = \gamma_{00} + \gamma_{01}W1 + \gamma_{02}W2 + u_{0j} \\
 & \beta_{1j} = \gamma_{10} + \gamma_{11}W1 + \gamma_{12}W2 + u_{1j} \\
 & \beta_{2j} = \gamma_{20} + \gamma_{21}W1 + \gamma_{22}W2
 \end{aligned} \tag{11}$$

While the simple model did not include a cross-level interaction, there were four cross-level interaction terms estimated in the complex model.

Model Misspecification

The simple and complex hierarchical linear model sets consisted of eight different nested and non-nested models, including the correct simple and complex generating model, respectively. The models examined were misspecified by incorrectly adding a parameter, incorrectly removing a parameter, or incorrectly

adding and removing a parameter from the correct model. For the simple model set, the seven models were misspecified as follows:

- a) by including a cross-level interaction between X1 and W1, γ_{02} (Simple Model 2);
- b) by dropping X1 from the model which results in the loss of the level-2 equation for the prediction of β_{1j} (Simple Model 3);
- c) by dropping W1, γ_{01} , from the model (Simple Model 4);
- d) by dropping u_{1j} from the model, thus the corresponding variance, τ_{11} , and covariance, τ_{10} were not estimated (Simple Model 5);
- e) by dropping u_{0j} from the model, thus the corresponding variance, τ_{00} , and covariance, τ_{10} were not estimated (Simple Model 6);
- f) by dropping u_{1j} from the model and including the cross-level interaction between X1 and W1, γ_{02} (Simple Model 7); and
- g) by dropping u_{0j} from the model and including the cross-level interaction between X1 and W1, γ_{02} (Simple Model 8).

Of all of these misspecified models, Model 2 is the more parameterized, incorrect nested model, Models 3 – 6 are less parameterized, incorrect nested models, and Models 7 – 8 are non-nested, incorrect models.

For the complex model set, the seven models were misspecified as follows:

- a) by including u_{2j} , thus estimating the corresponding variance, τ_{22} , and covariance, τ_{20} (Complex Model 2);
- b) by including an interaction between X1 and X2 that was fixed at level-2, γ_{30} (Complex Model 3);

WHITTAKER & FURLOW

Table 1: Characteristics of the Applied HLM Articles Reviewed
(January 2002 – March 2007)

Characteristic	Frequency
Reported Use of Model Selection Criteria	45
Model Selection Criteria Used	
Chi-Square Difference Test	35
AIC	2
BIC	1
AIC with BIC	3
Chi-Square Difference Test with AIC	2
Chi-Square Difference Test with BIC	1
Chi-Square Difference Test with AIC & BIC	1
HLM Software Used	
HLM	10
MLwiN	7
SAS PROC MIXED	4
LISREL	2
MIXREG	1
Mplus	1
Did Not Specify	20
Did Not Report Use of Model Selection Criteria	175

- c) by including an interaction between W1 and W2 in the intercept equation, γ_{03} (Complex Model 4);
- d) by dropping u_{1j} , from the model, thus the corresponding variance, τ_{11} , and covariance, τ_{10} , were not estimated (Complex Model 5);
- e) by dropping the cross-level interaction between X1 and W2, γ_{12} (Complex Model 6);
- f) by dropping u_{1j} from the model and including u_{2j} (Complex Model 7); and
- g) by dropping W2 from the intercept equation, γ_{02} , and including u_{2j} (Complex Model 8).

Of all of these misspecified models, Models 2 – 4 are more parameterized, incorrect nested models, Models 5 – 6 are less parameterized,

incorrect nested models, and Models 7 – 8 are non-nested, incorrect models.

Number of Groups at Level-2 and Participants per Group

The number of groups modeled at level-2 was varied to be either 20 or 40 to represent small to moderate sizes. Within each group, the sample size was varied to be either 15 or 30 participants to represent fairly small to moderate to large total sample sizes (300, 600, and 1,200, respectively).

Parameter Magnitude

The magnitude of all of the slope coefficients was varied to equal .5 or .7 to represent moderate to large magnitudes. The overall intercept (γ_{00}) remained constant at a value of 1 and the intercept values for the slope

MODEL SELECTION CRITERIA WITH HLM

equations (γ_{10} and γ_{20} in the complex model and only γ_{10} in the simple model) remained constant at a value of .5.

Intraclass Correlation (ICC) Value

The conditional intraclass correlation (ICC), which represents the proportion of the residual variance between groups remaining after including explanatory variables, was varied to equal either .1 or .3. The level-1 residual variance was set to equal .5. The level-2 variance components, τ_{00} and τ_{11} , were set to be equal to one another and their values were dictated by the ICC and the level-1 variance. This resulted in level-2 variances equal to 0.05555556 with an ICC of .1 and 0.214285714 with an ICC of .3. The level-2 covariance term, τ_{01} , was assumed to be equal to 0.

Simulation Study Procedure

SAS (version 9.1) was used to generate raw data according to the correct simple and complex generating models (see Equations 9 and 11) under the 16 combinations of different number of groups, participants per group, parameter magnitude, and ICC value conditions, resulting in 32 conditions. For each of the 32 conditions, 1,000 sets of raw data were generated. Each variable was generated to be standard normal. Once each data set was generated, all eight models (one correct and seven misspecified) were fit to the data using full information maximum likelihood (FIML) estimation in SAS's PROC MIXED procedure. The nine model selection criteria under examination were calculated for each of the models. The number of times each criteria selected each of the models was then documented.

Results

The selection rates of the nine criteria are presented in Tables 2 – 9. The simple model selection rates are presented in Tables 2 – 5 and the complex model selection rates are presented in Tables 6 – 9. None of the criteria performed well in the smallest total sample size (20 groups

x 15 participants per group = 300) and low ICC value (.1) conditions, regardless of parameter magnitude (see Tables 2 and 6). Overall, however, the accuracy of the selection criteria with respect to selecting the correct hierarchical linear model tended to increase as total sample size and ICC values increased. Further, the criteria generally performed better when selecting the correct model from the simple multilevel model set than when selecting the correct model from the complex multilevel model set.

Parameter magnitude did not have an effect on all of the selection criteria in all of the conditions. In general, the criteria tended to perform similarly in both low and high parameter magnitude conditions. Still, it did have an effect on the performance of the $AICC_m$ in two conditions. More specifically, the $AICC_m$ selected the correct model more frequently in the high parameter condition when group size was equal to 20 and the ICC value was high in the complex model set (see Tables 6 – 7).

The AIC and the $AICC_N$ were the least accurate selection criteria. These criteria never correctly selected the Simple or Complex Model 1 in more than 84% or 62% of the replications in any one condition, respectively. When the AIC or the $AICC_N$ did not select the correct multilevel model, they tended to select the more parameterized, misspecified models.

The next least accurate criterion was the $HQIC_m$, which never selected the Simple or Complex Model 1 in more than 89% or 73% of the replications in a condition, respectively. The $HQIC_N$ outperformed its m -calculated counterpart in all but four conditions (see Tables 2 and 4). Still, while the $HQIC_N$ selected the Simple Model 1 in more than 90% of the replications in more than half of the conditions, it never selected the Complex Model 1 in more than 90% of the replications in any condition. When the $HQIC_m$ or $HQIC_N$ did not select the correct multilevel model, they tended to incorrectly select the more parameterized, misspecified models.

The next least accurate criterion was the BIC_m . It correctly selected Simple Model 1 in more than 90% of the replications in half of the conditions but never correctly selected Complex Model 1 in more than 90% of the replications in

WHITTAKER & FURLOW

Table 2: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 300 Total Participants (20 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	57.8	31.1	57.8	50.0	20.1	37.4	14.4	57.3	43.3
M2	11.8	0.0	10.8	3.7	0.0	0.9	0.0	9.9	1.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	10.5	27.3	11.0	17.4	33.7	24.1	36.9	11.4	21.5
M6	10.4	32.6	10.9	18.6	39.6	28.5	43.7	11.7	23.7
M7	7.0	7.6	7.0	8.4	5.9	7.7	4.7	7.2	8.2
M8	2.5	1.4	2.5	1.9	0.7	1.4	0.3	2.5	1.5
Parameter Magnitude = .5, ICC = .3									
M1	81.5	92.0	82.8	88.4	89.9	91.1	85.8	83.7	90.3
M2	17.6	1.9	16.2	9.5	1.9	5.3	0.7	15.2	6.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.9	4.9	1.0	1.7	6.2	2.8	9.8	1.1	2.4
M5	0.0	0.3	0.0	0.0	0.7	0.1	1.1	0.0	0.0
M6	0.0	0.4	0.0	0.2	0.9	0.4	1.9	0.0	0.2
M7	0.0	0.3	0.0	0.1	0.3	0.2	0.6	0.0	0.2
M8	0.0	0.2	0.0	0.1	0.1	0.1	0.1	0.0	0.1
Parameter Magnitude = .7, ICC = .1									
M1	59.3	30.9	59.2	49.5	18.4	35.7	12.3	58.0	42.9
M2	10.8	0.2	9.8	3.9	0.2	1.3	0.1	9.2	2.3
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	10.9	28.1	11.2	17.5	35.6	25.7	39.9	12.0	21.4
M6	10.5	31.8	10.9	19.2	39.7	28.7	43.2	11.9	24.2
M7	6.3	7.8	6.6	8.1	5.4	7.3	4.0	6.8	7.8
M8	2.2	1.2	2.3	1.8	0.7	1.3	0.5	2.1	1.4
Parameter Magnitude = .7, ICC = .3									
M1	82.8	97.3	83.6	90.1	96.2	94.1	94.9	83.9	92.3
M2	17.0	1.4	16.2	9.4	1.4	5.1	0.8	15.7	7.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.5	0.0	0.1	0.8	0.3	1.3	0.0	0.3
M5	0.1	0.3	0.1	0.2	0.9	0.3	1.6	0.1	0.2
M6	0.0	0.1	0.0	0.0	0.3	0.0	1.1	0.0	0.0
M7	0.1	0.4	0.1	0.2	0.4	0.2	0.3	0.3	0.2
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

MODEL SELECTION CRITERIA WITH HLM

Table 3: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (20 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	79.8	82.1	80.2	83.2	70.9	82.6	64.6	80.9	83.1
M2	16.5	1.3	16.1	8.2	0.6	4.0	0.3	14.8	4.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.1	0.0	0.0	0.1	0.1	0.2	0.0	0.0
M5	0.8	4.0	0.8	2.2	8.8	3.3	11.4	1.1	3.1
M6	1.5	9.0	1.5	4.3	16.4	7.5	20.4	1.9	6.5
M7	1.2	2.9	1.2	1.8	2.8	2.3	2.8	1.2	2.3
M8	0.2	0.6	0.2	0.3	0.4	0.2	0.3	0.1	0.2
Parameter Magnitude = .5, ICC = .3									
M1	81.7	93.5	82.2	88.5	91.3	91.4	87.7	83.4	90.7
M2	17.7	2.5	17.2	10.1	1.5	5.8	1.0	16.0	6.7
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.6	4.0	0.6	1.4	7.0	2.8	11.1	0.6	2.6
M5	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	78.0	82.3	78.6	81.5	70.7	82.9	65.2	79.4	82.8
M2	18.0	1.6	17.3	10.7	1.0	4.8	0.5	16.0	6.1
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.1	4.9	1.1	2.5	10.8	3.9	14.1	1.4	3.4
M6	1.1	7.4	1.1	3.0	14.3	5.1	17.6	1.5	4.6
M7	1.3	3.2	1.4	1.9	2.8	2.7	2.4	1.3	2.5
M8	0.5	0.6	0.5	0.4	0.4	0.6	0.2	0.4	0.6
Parameter Magnitude = .7, ICC = .3									
M1	82.6	97.9	83.2	91.4	97.9	94.9	97.7	84.3	94.0
M2	17.4	2.0	16.8	8.6	1.5	5.1	1.1	15.7	6.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.1	0.0	0.0	0.6	0.0	1.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

WHITTAKER & FURLOW

Table 4: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (40 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	79.8	83.0	80.4	78.9	59.0	73.1	51.2	81.0	78.9
M2	16.0	6.6	15.3	4.2	1.1	2.2	0.4	10.3	4.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.7	3.8	1.7	6.1	17.4	10.3	21.4	3.4	6.2
M6	1.2	4.1	1.2	7.9	20.7	11.4	25.2	3.0	7.9
M7	0.8	1.8	0.9	2.2	1.5	2.4	1.5	1.4	2.3
M8	0.5	0.7	0.5	0.7	0.3	0.6	0.3	0.9	0.7
Parameter Magnitude = .5, ICC = .3									
M1	81.9	91.6	82.5	94.0	98.8	96.7	99.1	88.1	94.0
M2	18.1	8.4	17.5	6.0	1.2	3.3	0.7	11.9	6.0
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	80.1	84.0	80.5	82.0	61.0	76.9	51.4	82.2	82.0
M2	16.5	8.8	15.7	5.1	0.4	2.0	0.1	11.7	4.9
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	1.2	2.9	1.2	5.6	15.5	8.1	20.4	2.7	5.6
M6	1.3	2.9	1.5	6.1	21.2	11.1	26.3	2.4	6.3
M7	0.6	1.1	0.8	1.2	1.6	1.5	1.7	0.8	1.2
M8	0.3	0.3	0.3	0.0	0.3	0.4	0.1	0.2	0.0
Parameter Magnitude = .7, ICC = .3									
M1	82.0	90.5	82.4	93.2	98.1	95.7	98.6	87.0	93.2
M2	18.0	9.5	17.6	6.8	1.9	4.3	1.4	13.0	6.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

MODEL SELECTION CRITERIA WITH HLM

Table 5: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 1200 Total Participants (40 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Simple Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	83.5	91.4	83.8	93.8	97.9	96.1	97.7	88.4	94.7
M2	16.4	8.4	16.1	5.8	0.8	3.3	0.4	11.4	4.9
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.1	0.4	0.1	0.5	0.0	0.1
M6	0.1	0.1	0.1	0.2	0.6	0.3	1.1	0.1	0.2
M7	0.0	0.1	0.0	0.1	0.3	0.2	0.3	0.1	0.1
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	83.2	91.5	83.6	93.6	98.9	96.2	99.3	88.1	94.2
M2	16.8	8.5	16.4	6.4	1.0	3.8	0.5	11.9	5.8
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	82.3	89.9	82.5	92.8	98.2	95.1	97.8	87.0	93.5
M2	17.5	9.9	17.3	6.9	1.0	4.5	0.5	12.8	6.2
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.1	0.1	0.1	0.1	0.3	0.1	0.4	0.1	0.1
M6	0.1	0.1	0.1	0.2	0.4	0.2	1.2	0.1	0.2
M7	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	82.7	90.2	83.0	92.1	98.8	95.3	99.2	87.3	92.9
M2	17.3	9.8	17.0	7.9	1.2	4.7	0.8	12.7	7.1
M3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 is the more parameterized, incorrect nested model. M3 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

WHITTAKER & FURLOW

Table 6: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 300 Total Participants (20 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	41.8	1.0	45.0	45.9	30.1	43.3	24.3	44.6	45.7
M2	8.9	0.0	6.2	2.2	0.2	0.5	0.0	6.4	0.7
M3	11.0	0.0	9.3	5.9	1.0	2.6	0.6	9.4	3.8
M4	19.8	0.0	18.2	12.4	3.2	7.5	2.0	18.2	9.5
M5	14.9	96.9	18.0	31.6	65.0	44.5	72.8	18.1	38.6
M6	0.0	1.9	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M7	3.6	0.2	3.3	2.0	0.4	1.6	0.2	3.3	1.7
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	56.1	16.8	60.0	72.9	84.1	80.3	82.1	59.7	78.1
M2	9.5	0.0	7.8	4.1	0.2	1.4	0.0	8.0	2.5
M3	14.3	0.0	12.8	7.6	1.7	5.2	1.1	12.8	6.1
M4	18.4	0.0	17.4	12.5	3.6	7.8	2.4	17.5	9.5
M5	0.1	21.1	0.2	0.6	2.3	0.8	3.6	0.2	0.6
M6	1.4	62.1	1.6	2.2	8.1	4.4	10.8	1.6	3.1
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.2	0.0	0.2	0.1	0.0	0.1	0.0	0.2	0.1
Parameter Magnitude = .7, ICC = .1									
M1	40.8	2.0	43.5	45.8	35.9	43.5	30.2	43.3	45.4
M2	8.2	0.0	6.6	2.0	0.1	0.4	0.0	6.7	0.8
M3	10.9	0.0	9.4	5.4	0.5	2.7	0.5	9.4	4.0
M4	19.7	0.2	18.0	11.5	3.0	7.5	2.0	18.1	9.1
M5	17.6	97.4	19.9	33.1	60.1	44.3	67.3	20.0	38.9
M6	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	2.8	0.1	2.6	2.2	0.4	1.6	0.0	2.5	1.8
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	55.6	48.0	60.4	72.5	90.7	83.0	91.9	60.0	77.7
M2	9.9	0.0	7.5	2.6	0.1	0.7	0.0	7.7	1.2
M3	14.3	0.0	13.3	10.0	2.6	5.6	1.6	13.3	8.0
M4	19.9	0.0	18.5	14.4	4.4	9.8	3.1	18.6	12.3
M5	0.2	22.4	0.2	0.4	1.6	0.6	2.2	0.3	0.6
M6	0.1	29.6	0.1	0.1	0.5	0.3	1.1	0.1	0.2
M7	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

MODEL SELECTION CRITERIA WITH HLM

any one condition. When the BIC_m did not select the correct model, it tended to select both the less parameterized and more parameterized, misspecified models, depending upon the condition. More specifically, it tended to select the less parameterized, misspecified models when ICC values were low with smaller total sample sizes and the more parameterized, misspecified models when ICC values were high with larger total sample sizes.

Interestingly, the $AICC_m$ tended to outperform the $AICC_N$ in all but two conditions in the simple model set, but only in approximately half of the conditions in the complex model set. More specifically, the $AICC_m$ did not outperform the $AICC_N$ until total sample size reached a moderate size (20 groups x 30 participants per group = 600), high parameter magnitude (.7), and high ICC value (.3) (see Tables 7 – 9). The $AICC_m$ correctly selected Simple Model 1 in more than 90% of the replications in more than half of the conditions and correctly selected Complex Model 1 in more than 90% of the replications in a little less than half of the conditions. When the $AICC_m$ did not select the correct model, it tended to incorrectly select the less parameterized, misspecified models.

The $CAIC_m$ performed fairly comparably to the $AICC_m$, selecting the Simple Model 1 in more than 90% of the replications in a little more than half of the conditions but correctly selected Complex Model 1 in more than 90% of the replications in a little less than half of the conditions. When the $CAIC_m$ did not select the correct model, it tended to select the more parameterized and less parameterized, misspecified models depending upon the condition. For example, it tended to select the less parameterized, misspecified models when ICC values were low with smaller sample sizes and the more parameterized, misspecified models when ICC values were high with larger sample sizes.

The BIC_N and $CAIC_N$ performed the most accurately and fairly similarly. While the BIC_N did perform slightly better than the $CAIC_N$, these differences were generally small. The BIC_N correctly selected Simple Model 1 in more than 90% of the replications in a little more than half of the conditions and correctly selected

Complex Model 1 in more than 90% of the replications in half of the conditions. The BIC_N outperformed its m -calculated counterpart in more than half of the conditions. Nonetheless, when the BIC_m outperformed the BIC_N in the remaining conditions, the ICC value was low (see Tables 2 – 4, 6, and 8). When the BIC_N did not select the correct model, it generally tended to incorrectly select the less parameterized models.

The $CAIC_N$ correctly selected Simple Model 1 in more than 90% of the replications in half of the conditions and correctly selected Complex Model 1 in more than 90% of the replications in a little more than half of the conditions. The $CAIC_N$ outperformed the $CAIC_m$ in a little more than half of the conditions. Similar to the BIC, when the $CAIC_m$ outperformed the $CAIC_N$ in the remaining conditions, the ICC value tended to be low (see Tables 2 – 4, and 6 - 8), with the exception of two conditions (see Tables 2 and 3). When the $CAIC_N$ did not select the correct model, it tended to incorrectly select the less parameterized, misspecified models. It should be mentioned that when the m -calculated BIC and CAIC outperformed their N -calculated counterparts, the differences were quite large, particularly within the Simple Model set. In contrast, when the N -calculated BIC and CAIC outperformed their m -calculated counterparts, the differences were not as large.

It must be noted that the results presented are based on 1,000 replications in which all of the eight simple and complex models did not encounter any estimation problems. Hence, replications in which any model encountered a problem involving a non-positive definite variance component matrix or a convergence problem were discarded.

Additional replications were conducted until 1,000 replications in which problems did not exist were reached (see Table 10 for a summary of replications needed and percentage of usable replications in each generating condition). Less estimation problems were encountered when running the simple models than when running the complex models. Overall, fewer problems were encountered as total sample size and ICC values increased. Non-positive definite covariance matrix and

WHITTAKER & FURLOW

Table 7: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (20 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	54.8	27.3	57.1	70.8	75.4	77.8	72.3	58.8	76.3
M2	12.2	0.0	10.9	4.8	0.3	2.0	0.0	9.8	2.7
M3	12.1	0.0	11.7	6.9	0.9	3.7	0.2	11.0	4.5
M4	18.7	0.0	17.9	12.3	2.4	7.7	1.5	17.7	8.9
M5	2.0	68.5	2.2	4.9	20.9	8.5	25.9	2.5	7.3
M6	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.2	0.0	0.2	0.3	0.1	0.3	0.1	0.2	0.3
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	53.4	33.1	55.5	72.2	88.8	82.3	86.9	57.5	80.1
M2	14.4	0.0	13.3	4.6	0.4	2.0	0.0	12.0	2.6
M3	13.1	0.0	12.6	8.4	1.2	5.2	0.9	12.2	5.6
M4	18.4	0.0	17.8	12.6	2.0	7.8	1.4	17.3	9.1
M5	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.7	66.3	0.8	2.0	7.6	2.5	10.8	1.0	2.4
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.2
Parameter Magnitude = .7, ICC = .1									
M1	51.4	33.3	53.5	68.0	75.9	76.4	72.6	55.4	75.6
M2	12.8	0.0	11.8	5.7	0.1	1.5	0.0	11.2	2.1
M3	12.7	0.0	11.9	7.8	0.8	4.4	0.4	11.2	5.1
M4	20.1	0.0	19.7	13.7	2.2	8.4	1.2	19.0	9.6
M5	2.2	66.4	2.3	4.0	20.6	8.5	25.7	2.4	6.6
M6	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.8	0.0	0.8	0.8	0.4	0.8	0.1	0.8	1.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	53.4	73.6	56.7	72.6	95.3	84.6	96.6	58.7	81.4
M2	13.6	0.0	12.1	5.2	0.4	2.0	0.1	11.0	2.6
M3	12.5	0.0	11.7	7.7	0.8	3.4	0.2	11.2	4.7
M4	20.5	0.0	19.5	14.5	2.9	9.9	1.9	19.1	11.2
M5	0.0	1.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0
M6	0.0	24.9	0.0	0.0	0.6	0.1	1.1	0.0	0.1
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

MODEL SELECTION CRITERIA WITH HLM

Table 8: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 600 Total Participants (40 Groups X 15 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	56.3	79.1	57.8	76.5	70.7	75.7	63.1	67.6	76.7
M2	12.1	0.6	11.5	2.4	0.1	0.9	0.0	6.5	2.4
M3	12.9	1.7	12.5	3.4	0.4	1.7	0.3	8.6	3.3
M4	15.4	3.2	14.6	5.7	1.1	3.5	0.7	11.0	5.6
M5	2.1	14.4	2.2	10.6	26.9	17.3	35.3	5.0	10.6
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	1.2	1.0	1.4	1.4	0.8	0.9	0.6	1.3	1.4
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	55.9	92.4	57.9	83.5	97.2	91.5	98.4	69.2	83.8
M2	13.2	0.8	12.1	3.8	0.2	1.4	0.0	7.6	3.6
M3	14.8	3.1	14.2	5.1	1.0	3.2	0.5	10.9	5.1
M4	16.1	3.7	15.8	7.6	1.5	3.9	0.8	12.3	7.5
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.1	0.0	0.3	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	57.0	81.1	59.5	77.7	71.0	77.6	64.2	67.5	77.7
M2	12.0	0.4	10.8	1.6	0.0	0.7	0.0	6.8	1.6
M3	13.3	3.1	12.6	5.0	1.4	3.1	0.8	9.6	4.9
M4	14.7	2.4	14.0	5.5	1.0	2.5	0.7	11.0	5.5
M5	2.6	12.2	2.6	9.7	26.4	15.8	34.1	4.4	9.8
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.4	0.8	0.5	0.5	0.2	0.3	0.2	0.7	0.5
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	56.1	91.0	58.8	83.9	96.7	90.2	98.3	70.0	83.9
M2	15.5	0.9	13.8	3.0	0.1	1.1	0.0	8.4	3.0
M3	11.8	3.1	11.3	5.2	1.6	3.1	0.8	8.6	5.2
M4	16.6	5.0	16.1	7.9	1.6	5.6	0.9	13.0	7.9
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

WHITTAKER & FURLOW

Table 9: Percentage of Times Out of 1,000 Replications Each Model Selection Criteria Selected Each Hierarchical Linear Model with 1,200 Total Participants (40 Groups X 30 Participants Per Group) as a Function of Parameter Magnitude and ICC Value – Complex Model Set

	AIC	AICC _m	AICC _N	BIC _m	BIC _N	CAIC _m	CAIC _N	HQIC _m	HQIC _N
Parameter Magnitude = .5, ICC = .1									
M1	59.4	92.1	60.3	85.4	97.6	91.7	97.8	71.4	86.8
M2	13.2	0.8	12.8	3.2	0.0	1.0	0.0	7.9	2.9
M3	12.3	3.0	12.1	4.7	0.6	3.0	0.3	9.6	4.1
M4	15.1	3.9	14.8	6.6	1.1	4.1	0.4	11.1	6.0
M5	0.0	0.2	0.0	0.1	0.7	0.2	1.5	0.0	0.2
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .5, ICC = .3									
M1	58.6	92.1	59.6	85.3	97.2	91.6	97.7	71.8	87.2
M2	13.4	0.9	13.1	2.6	0.2	1.1	0.1	7.5	2.1
M3	13.0	2.7	12.4	5.3	0.7	2.9	0.6	9.3	4.9
M4	15.0	4.3	14.9	6.8	1.4	4.3	0.9	11.4	5.8
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.5	0.1	0.7	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .1									
M1	55.3	91.2	56.6	82.7	96.6	90.1	96.7	68.5	84.7
M2	14.7	1.3	14.2	3.7	0.1	1.7	0.0	9.0	3.3
M3	13.0	2.8	12.8	5.9	0.9	3.2	0.3	9.7	5.1
M4	17.0	4.4	16.4	7.5	1.2	4.7	1.1	12.6	6.7
M5	0.0	0.3	0.0	0.2	1.2	0.3	1.9	0.1	0.2
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Parameter Magnitude = .7, ICC = .3									
M1	59.3	93.7	61.1	87.6	97.6	92.9	98.7	72.7	89.2
M2	12.5	0.5	11.7	1.8	0.2	1.0	0.0	6.6	1.4
M3	14.5	2.3	13.9	5.0	0.6	2.5	0.3	10.0	4.2
M4	13.7	3.5	13.3	5.6	1.6	3.6	1.0	10.7	5.2
M5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Note: M1 is the correct model. M2 – M4 are more parameterized, incorrect nested models. M5 – M6 are less parameterized, incorrect nested models. M7 – M8 are non-nested, incorrect models.

MODEL SELECTION CRITERIA WITH HLM

convergence problems could both be encountered for different models within the same replication. Convergence problems were encountered more frequently when trying to fit Complex Model 2 (which incorrectly included a random effect), Complex Model 7 (which was missing a random effect and incorrectly included a random effect), and Complex Model 8 (which was missing a fixed effect and incorrectly included a random effect).

Conclusion

The current study examined the performance of the AIC, AICC, BIC, CAIC, and the HQIC when selecting the correct multilevel model under different criteria calculation, model complexity, model misspecification, number of groups at level-2, number of participants per group, parameter magnitude, and intraclass correlation (ICC) value conditions. Several of the study factors, either in isolation or in conjunction with another factor or factors, did affect the performance of the model selection criteria. For example, none of the model selection criteria performed well with respect to selecting the correct model when total sample size and ICC values were small and the performance of the model selection criteria improved as total sample size and ICC value increased.

The criteria generally performed more accurately when selecting the correct model from the simple model set than the complex model set. This seems reasonable given that adding or dropping parameters from a simple model would result in a more highly misspecified model than when adding or dropping parameters from a complex model. For example, dropping a random slope from a model in which there is only 1 random slope would result in a more highly misspecified model than dropping a random slope from a model in which there are 2 or more random slopes. Thus, the criteria would be more likely to select the correct model from among a set of severely misspecified models in the simple model set than a set of moderately misspecified models in the complex model set.

Although parameter magnitude did not appear to have a great impact on the

performance of the model selection criteria, it did impact the $AICC_m$ in two conditions. That is, the $AICC_m$ performed more accurately in the high parameter magnitude condition when group size was 20 with a high ICC value in the complex model set. Again, this appears to be an isolated occurrence as parameter magnitude did not generally affect the remaining criteria.

The efficient model selection criteria, the AIC and the $AICC_N$, did not perform as well as the remaining, consistent criteria. This is to be expected when the definition of the performance standard, such as the one used in this study, is consistency (i.e., the selection of the correct model from among a set of competing models). These results corroborate the findings in Gurka's (2006) study.

To date, the HQIC, to the best of our knowledge, has not been examined in the relevant literature. Thus, the performance of the HQIC under various conditions and in comparison to the remaining criteria was of interest in the current study. The results indicated that while the HQIC performed more accurately than the AIC and the $AICC_N$, it did not perform more accurately than the BIC, CAIC, or the $AICC_m$ when selecting the correct model.

The $AICC_m$ proved to be a contender, not only outperforming its N -calculated counterpart in almost all conditions, but also performing comparably to the $CAIC_m$, next to the most accurately performing criteria (BIC_N and $CAIC_N$). Gurka (2006) also found that the $AICC_m$ performed adequately. Gurka (2006) recommended the use of the BIC_m and the $CAIC_m$ based on his findings, however, the BIC_N and $CAIC_N$ outperformed their m -calculated counterparts in several conditions in the current study. When the BIC_m and the $CAIC_m$ did outperform their N -calculated counterparts, the ICC value was low. Also, the differences in the rates of choosing the correct model were appreciably higher for the m -calculated criteria in these conditions, particularly within the Simple Model set.

The results of the current study did not determine which one model selection criterion will perform optimally in every situation encountered. It is clear, however, that the BIC, the CAIC, as well as the $AICC_m$, generally out-

WHITTAKER & FURLOW

Table 10: Non-Positive Definite Variance Component Matrix and Convergence Problems Encountered as a Function of Generating Condition

Condition	Simple		Complex	
	Replications Needed	% Usable Replications	Replications Needed	% Usable Replications
20 x 15; Parameter = .5; ICC = .1	1110	90.1	1733	57.7
20 x 15; Parameter = .5; ICC = .3	1000	100.0	1318	75.9
20 x 15; Parameter = .7; ICC = .1	1118	89.4	1794	55.7
20 x 15; Parameter = .7; ICC = .3	1002	99.8	1397	71.6
20 x 30; Parameter = .5; ICC = .1	1018	98.2	1184	84.5
20 x 30; Parameter = .5; ICC = .3	1000	100.0	1096	91.2
20 x 30; Parameter = .7; ICC = .1	1009	99.1	1185	84.4
20 x 30; Parameter = .7; ICC = .3	1000	100.0	1117	89.5
40 x 15; Parameter = .5; ICC = .1	1006	99.4	1072	93.3
40 x 15; Parameter = .5; ICC = .3	1000	100.0	1019	98.1
40 x 15; Parameter = .7; ICC = .1	1010	99.0	1096	91.2
40 x 15; Parameter = .7; ICC = .3	1000	100.0	1035	96.6
40 x 30; Parameter = .5; ICC = .1	1000	100.0	1012	98.8
40 x 30; Parameter = .5; ICC = .3	1000	100.0	1001	99.9
40 x 30; Parameter = .7; ICC = .1	1000	100.0	1013	98.7
40 x 30; Parameter = .7; ICC = .3	1000	100.0	1004	99.6

performed the remaining criteria examined. Still, the performance of these criteria was dependent upon the conditions examined in the current study. None of the criteria performed very well in the smallest total sample size with low ICC value conditions. Thus, in this situation, researchers may want to employ the BIC_m and

the $CAIC_m$ along with the AIC, regardless of model complexity. When total sample sizes are larger with higher ICC values, the BIC_N , $CAIC_N$, and the $AICC_m$ together may be used to select among a set of multilevel models. Researchers should be cautioned, however, that the $AICC_m$ performs less accurately when the competing

MODEL SELECTION CRITERIA WITH HLM

models are complex, unless the number of groups is large.

The models and conditions examined in the current study do not reflect all possible models and conditions found when analyzing real-world data. Hence, it is difficult to generalize the findings to every situation that may be encountered by applied researchers. Future research still needs to be conducted in order to more fully understand the characteristics of model selection criteria in the HLM arena. For example, the models examined in this study were limited to two levels; it would be interesting to examine how well these selection criteria perform with three-level models, particularly when calculating the criteria using N , m at level-2, and m at level-3.

Future research should also examine the sensitivity of the model selection criteria to non-normally distributed data as well as data that are missing at level-1, level-2, or both. Based on Gurka's (2006) finding that the model selection criteria worked well when models were misspecified by fixed effects using REML estimation, future research could also examine how well these criteria work using REML estimation under additional conditions.

In recent years, HLM has grown widely popular in its use. Indeed, our search in PsycInfo between January 2002 and March 2007 for articles in which HLM was used uncovered 220 articles. Our content analysis also indicated that model selection criteria were used in the model selection/comparison process in 45 of the 220 articles, with only 10 of those consisting of information criteria. Thus, most HLM research does not incorporate any type of model selection criteria. This could be a result of a lack of literature informing researchers as to the performance of these criteria and a lack of literature pointing to the necessity of these criteria when deciding between several competing models. In addition, while major software packages like SAS and SPSS include a number of information criteria in their output, other packages that estimate multilevel models, such as HLM 6 (Raudenbush, Bryk & Congdon, 2007) and MLwiN (Rasbash, et al., 2000), do not provide any information criteria in their output. While the deviance statistic is provided in these software packages, applied researchers

may be less likely, or aware of, the different information criteria available. This may also possibly be contributing to the lack of utilization of these criteria in the applied literature. Therefore, the current study provides valuable information concerning the existing practices of applied researchers when comparing and selecting among hierarchical models as well as the performance of existing and alternative criteria when selecting among hierarchical models.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory*. Budapest: Akademiai Kiado.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-370.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: A practical information-theoretic approach*. NY: Springer.
- Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician*, 60(1), 19-26.
- Hannon, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41(2), 190-195.
- Hedeker, D., & Gibbons, R. D. (1999). MIXREG (Version 1.2). [Computer Software]. Chicago, IL: Hedeker & Gibbons.
- Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, 72(2), 297-307.
- Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.
- McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection*. River Edge, NJ: World Scientific Publishing Co.
- Muthén, L. K., & Muthén, B. O. (2007). Mplus (Version 4.2). [Computer Software]. Los Angeles, CA: Muthén & Muthén.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN (Version 2.02). [Computer Software]. Bristol, UK: University of Bristol, Centre for Multilevel Modeling.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S., Bryk, A., & Congdon, R. (2007). HLM (Version 6.04) [Computer Software]. Lincolnwood, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(3), 199-213.

SAS Institute Inc. (2007). SAS (Version 9.2) [Computer Software]. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Shi, P., & Tsai, C. L. (2002). Regression model selection - a residual likelihood approach. *Journal of the Royal Statistical Society, Series B*, 64(2), 237-252.

SPSS Inc. (2007). SPSS (Version 16.0) [Computer Software]. Chicago, IL: SPSS Inc.

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

On The Expected Values of Distribution of the Sample Range of Order Statistics from the Geometric Distribution

Sinan Calik
Firat University, Turkey

Cemil Colak
Inonu University, Turkey

Ayşe Turan
Firat University, Turkey

The expected values of the distribution of the sample range of order statistics from the geometric distribution are presented. For n up to 10, algebraic expressions for the expected values are obtained. Using the algebraic expressions, expected values based on the p and n values can be easily computed.

Key words: Order statistics, expected value, moment, sample range, geometric distribution.

Introduction

Let X_1, X_2, \dots, X_n be a random sample of size n from a discrete distribution with a probability mass function (*pmf*) $f(x)$ ($x = 0, 1, 2, \dots$) and a cumulative distribution function $F(x)$. Let $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ be the order statistics obtained from the above random sample by arranging the observations in increasing order of magnitude. When spacing is denoted as $W_{i,j:n} = X_{j:n} - X_{i:n}$, and $i = 1$ and $j = n$, that is, in the case of the sample range W_n , then $W_n = X_{n:n} - X_{1:n}$. Denote the expected values of distribution of the sample range $E(W_n)$ by $\mu_{W_n}^{(k)}$ ($n \geq 2$). For convenience, denote $\mu_{W_n}^{(1)}$ simply by μ_{W_n} .

Order statistics from the geometric distribution have been studied by many authors, for example, see Abdel-Aty (1954) and Morgolin and Winokur (1967). In particular,

characterizations of the geometric distribution using order statistics have received great attention; for example, see Uppuliri (1964), Ferguson (1965, 1967), Crawford (1966), Srivastava (1974), Galambos (1975), El-Newehi and Govindarajulu (1979), and Govindarajulu (1980). Expressions for the first two single moments of order statistics have been obtained by Morgolin and Winokur (1967).

The calculation of the exact sampling distribution of ranges from a discrete population was obtained by Burr (1955). The distribution of the sample range from a discrete order statistics were given by Arnold, et al. (1992). Additional details on discrete order statistics can be found in the works of Khatri (1962), David (1981), Nagaraja (1992), and Balakrishnan and Rao (1998). In this study, for n up to 10, algebraic expressions for the expected values of the distribution of the sample range of order statistics from the geometric distribution are obtained.

Methodology

Marginal Distribution of Order Statistics

If $F_{r:n}(x)$ ($r = 1, 2, \dots, n$) denotes the cumulative distribution function (*cdf*) of $X_{r:n}$, then the following results:

Cemil Colak is Assistant Professor in the Department of Biostatistics. Email: cemilcolak@yahoo.com. Sinan Calik is Assistant Professor of Statistics. Email: scalik@firat.edu.tr. Ayşe Turan is research assistant in the Department of Statistics. Email: ayseturan23@hotmail.com.

$$\begin{aligned}
 F_{r:n}(x) &= P\{X_{r:n} \leq x\} \\
 &= P\{\text{at least } r \text{ of } X_1, X_2, \dots, X_n \text{ are at most } x\} \\
 &= \sum_{i=r}^n P\{\text{exactly } i \text{ of } X_1, X_2, \dots, X_n \text{ are at most } x\} \\
 &= \sum_{i=r}^n \binom{n}{i} [F(x)]^i [1-F(x)]^{n-i} \\
 &= \int_0^{F(x)} \frac{n!}{(r-1)!(n-r)!} t^{r-1} (1-t)^{n-r} dt
 \end{aligned} \tag{1}$$

for $-\infty < x < \infty$.

For a discrete population, the probability mass function (*pmf*) of $X_{r:n}$ may be obtained from (1) by differencing as

$$\begin{aligned}
 f_{r:n}(x) &= F_{r:n}(x) - F_{r:n}(x-1) \\
 &= \frac{n!}{(r-1)!(n-r)!} \int_{F(x-1)}^{F(x)} t^{r-1} (1-t)^{n-r} dt
 \end{aligned}$$

(Arnold, et al., 1992; Balakrishnan, 1986).

Order Statistics from the Geometric Distribution

To explore the properties of the geometric distribution order statistics, begin by stating that X is a Geometric (p) random variable. Note that its *pmf* is given by $f(x) = pq^{x-1}$, and its *cdf* is $F(x) = 1 - q^x$, for $x = 1, 2, \dots$. Consequently the *cdf* of the r th order statistic is given by

$$F_{r:n}(x) = \sum_{i=r}^n \binom{n}{i} (1-q^x)^i (q^x)^{n-i}, \quad x = 1, 2, \dots$$

Joint Distribution of Order Statistics

The joint distribution of order statistics can be similarly derived. For example, the joint cumulative distribution function of $X_{i:n}$ and $X_{j:n}$ ($1 \leq i \leq j \leq n$) can be shown to be

$$F_{i,j:n}(x_i, x_j) = F_{j:n}(x_j), \quad \text{for } x_i \geq x_j.$$

For $x_i < x_j$,

$$\begin{aligned}
 F_{i,j:n}(x_i, x_j) &= \\
 &= \sum_{s=j}^n \sum_{r=i}^s \frac{n!}{r!(s-r)!(n-s)!} \{F(x_i)\}^r \\
 &\quad \{F(x_j) - F(x_i)\}^{s-r} \{1-F(x_j)\}^{n-s}
 \end{aligned} \tag{2}$$

This expression holds for any arbitrary population whether continuous or discrete.

For discrete populations, the joint probability mass function of $X_{i:n}$ and $X_{j:n}$ ($1 \leq i \leq j \leq n$) may be obtained from (2) by differencing as:

$$\begin{aligned}
 f_{i,j:n}(x_i, x_j) &= P(X_{i:n} = x_i, X_{j:n} = x_j) \\
 &= F_{i,j:n}(x_i, x_j) - F_{i,j:n}(x_i - 1, x_j) - \\
 &\quad F_{i,j:n}(x_i, x_j - 1) + F_{i,j:n}(x_i - 1, x_j - 1)
 \end{aligned}$$

Theorem 1. For $1 \leq i_1 \leq i_2 \leq \dots \leq i_k \leq n$, the joint *pmf* of $X_{i_1:n}, X_{i_2:n}, \dots, X_{i_k:n}$ is given by

$$\begin{aligned}
 &f_{i_1, i_2, \dots, i_k:n}(x_{i_1:n}, x_{i_2:n}, \dots, x_{i_k:n}) \\
 &= [C(i_1, i_2, \dots, i_k : n) \times \\
 &\quad \int_D \left\{ \prod_{r=1}^k (u_{i_r} - u_{i_{r-1}})^{i_r - i_{r-1} - 1} \right\} (1 - u_k)^{n - i_k} du_{i_1} \dots du_{i_k}],
 \end{aligned}$$

where $i_0 = 0, u_0 = 0$,

$$\begin{aligned}
 C(i_1, i_2, \dots, i_k : n) &= \\
 &= \frac{n!}{\left\{ (n - i_k)! \prod_{r=1}^k (i_r - i_{r-1} - 1)! \right\}^2}
 \end{aligned}$$

and D is k -dimensional space given by

$$D = \left\{ \begin{array}{l} (u_{i_1}, \dots, u_{i_k}) : u_{i_1} \leq u_{i_2} \leq \dots \leq u_{i_k}, \\ F(x_{r-1}) \leq u_r \leq F(x_r), \\ r = i_1, i_2, \dots, i_k \end{array} \right\}$$

(Nagaraja, 1986; Arnold, et al., 1992; Balakrishnan & Rao, 1988). Khatri (1962) presented this result for $k \leq 3$, but only proved it for $k \leq 2$ for the case of no ties.

Distribution of the Sample Range

Starting with the pmf of the spacing $W_{i,j:n} = X_{j:n} - X_{i:n}$, and using Theorem 1, results in

$$P(W_{i,j:n} = w) = \sum_{x \in D} \int_{F(x-1)}^{F(x)} \int_{F(x+w-1)}^{F(x+w)} \int_{u_i < u_j} u_i^{i-1} (u_j - u_i)^{j-i-1} (1 - u_j)^{n-j} du_j du_i \tag{3}$$

Substantial simplification of the expression in (3) is possible when $i = 1$ and $j = n$, that is, in the case of the sample range W_n , this results in:

$$P(W_n = w) = C(1, n : n) \sum_{x \in D} \int_{F(x-1)}^{F(x)} \int_{F(x+w-1)}^{F(x+w)} (u_n - u_1)^{n-2} du_n du_1$$

Thus, the pmf of W_n is given by

$$\begin{aligned} P(W_n = 0) &= n(n-1) \sum_{x \in D} \int_{F(x-1)}^{F(x)} \int_{F(x-1)}^{F(x)} (u_n - u_1)^{n-2} du_n du_1 \\ &= \sum_{x \in D} \{F(x) - F(x-1)\}^n \\ &= \sum_{x \in D} \{f(x)\}^n \end{aligned} \tag{4}$$

and, for $w > 0$, from Arnold, et al., 1992,

$$\begin{aligned} P(W_n = w) &= n(n-1) \sum_{x \in D} \int_{F(x-1)}^{F(x)} \int_{F(x+w-1)}^{F(x+w)} (u_n - u_1)^{n-2} du_n du_1 \\ &= \sum_{x \in D} \left\{ \begin{array}{l} [F(x+w) - F(x-1)]^n \\ - [F(x+w) - F(x)]^n \\ - [F(x+w-1) - F(x-1)]^n \\ + [F(x+w-1) - F(x)]^n \end{array} \right\} \end{aligned} \tag{5}$$

Expressions (4) and (5) can also be obtained without using the integral expression from Theorem 1, and a multinomial argument can also be used to obtain an alternative expression for the pmf of W_n .

Expected Values of the Sample Range

The m^{th} moments of W_n can be written as

$$\mu_{w_n}^{(m)} = E(W_n^m) = \sum_{w=0}^{\infty} w^m P(W_n = w), \tag{6}$$

where $P(W_n = w)$ is as given in (5).

When X is a geometric (p) random variable, as in the case of the expected values of the sample range, (6) yields

$$\begin{aligned} \mu_{w_n}^{(1)} &= E(W_n) \\ &= \sum_{w=0}^{\infty} w P(W_n = w) \\ &= \sum_{w=1}^{\infty} w P(W_n = w), \end{aligned} \tag{7}$$

where $P(W_n = w)$ is as given in (7).

Distribution of the Sample Range from the Geometric Distribution

The distribution of higher order statistics is not as simple for the geometric distribution. For the sample range W_n , from (2),

$$\begin{aligned}
 P(W_n = 0) &= \sum_{x=1}^{\infty} (pq^{x-1})^n \\
 &= p^n \sum_{x=1}^{\infty} (q^n)^{x-1} \\
 &= \frac{p^n}{1-q^n}
 \end{aligned}$$

and from (3) the following is obtained:

$$\begin{aligned}
 P(W_n = w) &= \sum_{x=1}^{\infty} \left\{ \begin{aligned} &(q^{x-1}(1-q^{w+1}))^n + (q^x(1-q^{w-1}))^n \\ &- (q^x(1-q^w))^n - (q^{x-1}(1-q^w))^n \end{aligned} \right\} \\
 &= \frac{1}{1-q^n} \left\{ \begin{aligned} &(1-q^{w+1})^n - (1-q^w)^n - \\ &q^n \left[(1-q^w)^n - (1-q^{w-1})^n \right] \end{aligned} \right\}, \tag{7}
 \end{aligned}$$

for $w > 0$.

In particular,

$$\begin{aligned}
 P(W_2 = w) &= \frac{1}{1-q^2} \left\{ \begin{aligned} &(1-q^{w+1})^2 - (1-q^w)^2 - \\ &q^2 \left[(1-q^w)^2 - (1-q^{w-1})^2 \right] \end{aligned} \right\}
 \end{aligned}$$

for $w > 0$, thus

$$\begin{aligned}
 P(W_2 = w) &= \frac{q^w(2-4q+2q^2)}{1-q^2}
 \end{aligned}$$

for $w > 0$.

Using the above pmf, the moments of W_n can be determined. For example, when $n=2$, using the pmf in (7), the following results:

$$\begin{aligned}
 E(W_2) &= \sum_{w=1}^{\infty} w \frac{q^w(2-4q+2q^2)}{1-q^2} \\
 &= \frac{2q}{p(1+q)} \\
 &= \frac{2q}{1-q^2}
 \end{aligned}$$

For n up to 10, algebraic expressions for the expected values of the distribution of the sample range of order statistics from the geometric distribution are obtained; these are shown in Table 1.

Conclusion

Algebraic expressions are presented for n up to 10 for the expected values of distribution of the sample range of order statistics from the geometric distribution. Using the obtained algebraic expressions, these expected values can be computed. As it is shown in Table 1, different values can be obtained for q and n . For example, for $q=0.50$, using the value $n=2$ in Table 1, $\mu_{W_2} \approx 0,011765$ is obtained. Further studies may focus on a software program for estimating the expected values found in this study.

References

Abdel-Aty, S. H. (1954). Ordered variables in discontinuous distributions. *Statistica Neerlandica*, 8, 61-82.

Arnold, B. C., Balakrishnan, N. & Nagaraja, H. N. (1992). *A first course in order statistics*. NY: John Wiley & Sons.

Balakrishnan, N. & Rao, C. R. (1998). *Handbook of statistics 16-order statistics: Theory and methods*. NY: Elsevier.

Balakrishnan, N. (1986). Order Statistics from discrete distribution. *Communications in Statistics: Theory and Methods*, 15(3), 657-675.

SAMPLE RANGE ORDER STATISTICS FROM THE GEOMETRIC DISTRIBUTION

Table 1: The expected values of distribution of the sample range of order statistics from the geometric distribution

n	μ_{w_n}
2	$\frac{2q}{1-q^2}$
3	$\frac{3q}{1-q^2}$
4	$\frac{4q^5 + 2q^4 + 10q^3 + 2q^2 + 4q}{(1-q^4)(1+q+q^2)}$
5	$\frac{5q^5 + 15q^3 + 5q}{(1-q^4)(1+q+q^2)}$
6	$\frac{6q^{11} + 3q^{10} + 29q^9 - 4q^8 + 34q^7 + 13q^6 + 34q^5 - 4q^4 + 29q^3 - 3q^2 + 6q}{(1-q^4)(1+q+q^4)(1+q+q^2+q^3+q^4)}$
7	$\frac{7q^{11} - 7q^{10} + 42q^9 - 21q^8 + 49q^7 + 7q^6 + 49q^5 - 21q^4 + 42q^3 - 7q^2 + 7q}{(1-q^4)(1+q+q^4)(1+q+q^2+q^3+q^4)}$
8	$\frac{8q^{21} - 49q^{20} + 56q^{19} - 6q^{18} + 92q^{17} + 68q^{16} + 208q^{15} + 94q^{14} + 246q^{13} + 162q^{12} + 306q^{11} + 162q^{10} + 246q^9 + 94q^8 + 208q^7 + 68q^6 + 92q^5 + 6q^4 + 56q^3 - 4q^2 + 8q}{(1-q^8)(1+q^2+q^4)(1+q+q^2+q^3+q^4)(1+q+q^2+q^3+q^4+q^5+q^6)}$
9	$\frac{9q^{21} - 9q^{20} + 75q^{19} - 21q^{18} + 120q^{17} + 45q^{16} + 270q^{15} + 42q^{14} + 285q^{13} + 126q^{12} + 399q^{11} + 126q^{10} + 285q^9 + 42q^8 + 270q^7 + 45q^6 + 120q^5 - 21q^4 + 75q^3 - 9q^2 + 9q}{(1-q^8)(1+q+q^2+q^3+q^4+q^5+q^6)}$
10	$\frac{[10q^{31} - 25q^{30} + 125q^{29} - 180q^{23} + 337q^{27} - 233q^{26} + 536q^{25} - 337q^{24} + 724q^{23} - 212q^{22} + 991q^{21} - 290q^{20} + 1153q^{19} - 260q^{18} + 1381q^{17} - 311q^{16} + 1381q^{15} - 260q^{14} + 1153q^{13} - 290q^{12} + 99q^{11} - 212q^{10} + 724q^9 - 337q^8 + 536q^7 - 233q^6 + 337q^5 - 180q^4 + 125q^3 - 25q^2 + 99]}{[(1-q^4)(1+q^2+q^4) (1+q^3+q^6) (1+q+q^2+q^3+q^4)]}$

Burr, I. W. (1955). Calculation of exact sampling distribution of ranges from a discrete population. *Annals of Mathematical Statistics*, 26, 530-532. *Correction*, 38, 280.

Crawford, B. G. (1966). Characterization of geometric and exponential distributions. *Annals of Mathematical Statistics*, 37, 1790-95.

David, H. A. (1981). *Order statistics* (2nd Ed.), NY: John Wiley & Sons.

El-Newehi, E. & Gavindarajulu, Z. (1979). Characterization of geometric and exponential distribution and discrete ifr (dfr) distributions using order statistics. *Journal of Statistical Planning and Inference*, 3, 85-90.

Ferguson, T. S. (1965.) A characterization of the geometric distribution. *American Mathematical Monthly*, 72, 256-260.

Ferguson, T. S. (1967). On characterizing distributions by properties of order statistics. *Sankhya Series A*, 29, 265-278.

Galambos, J. (1975). Characterizations of probability distributions by properties of order statistics. In: G. P. Patil, S. Kotz and G. K. Ord, Eds., *Statistical distributions in scientific work, Characterization and Applications*, Reidel Publishing Company, Dordrecht- Holland, 2, 289-101.

Gavindarajulu, Z. (1980). Characterization of the Geometric Distribution using properties of order statistics *Journal of Statistical Planning and Inference*, 4, 237-47.

Khatri, C. G. (1962). Distributions of order statistics for discrete case. *Annals of Institute of Statistical Mathematics*, 14, 167-171.

Margolin, B. H. & Winokur, H. S., Jr. (1967). exact moments of the order statistics of the geometric distribution and their relation to inverse sampling and reliability of redundant systems. *Journal of the American Statistical Association*, 62, 915-925.

Nagaraja, H. N. (1986). Structure of discrete order statistics. *Journal of Statistical Planning and Inference*, 13, 165-177.

Nagaraja, H. N. (1992). Order statistics from discrete distribution (with discussion). *Statistics*, 23, 189-216.

Srivastava, R. C. (1974). Two characterizations of the geometric distribution. *Journal of the American Statistical Association*, 69, 267-269.

Uppuluri, V. R. R. (1964). A characterization of the geometric distribution. *Annals of Mathematical Statistics*. 4, 1841.

Approximations to Power When Comparing Two Small Independent Proportions

Michael Vorburger Breda Munoz
JMP Division, SAS Institute

Researchers often face the problem of accurately calculating power for tests of differences between two independent proportions. Four commonly used and accepted approximations are the arc sine, the Chi-squared, and the continuity-corrected versions of each. Comparisons of these are discussed for various sample sizes, ultimately focusing on small proportions.

Key words: Fisher's exact test, power calculation, power approximation, arc sine approximation, Chi-squared approximation, small proportions.

Introduction

The conditional probability of rejecting the null hypothesis, in an accept-reject test of hypothesis, given that the alternative hypothesis is true, is called the power of the test. Determining the power of a test is referred to as power calculation. For the purposes of this discussion, the alternative hypothesis is $P_1 > P_2$, where P_1 and P_2 are the larger and smaller proportions being compared, respectively. Many researchers use these hypothesis tests to determine the minimum detectable differences between two proportions, given desired power level $(1-\beta)$, sample size (n) , and significance level (α) . The method for calculating the exact power of these tests requires an extremely time-consuming, iterative process using 2×2 contingency tables. A common approach to circumventing this arduous process is to use an approximation of the power. Researchers often indiscriminately apply some of these formulas without questioning the reliability of the results obtained.

Two standard approximations used to calculate the power of a test of difference

between two independent proportions are the arc sine approximation, provided by Cochran and Cox (1957),

$$Z_\beta = Z_{1-\alpha} - \sqrt{2n} \left(\text{Sin}^{-1} \sqrt{P_1} - \text{Sin}^{-1} \sqrt{P_2} \right), \quad (1)$$

and the Chi-squared approximation, provided by Fleiss (1973),

$$Z_\beta = \frac{Z_{1-\alpha} \sqrt{(P_1 + P_2) \left(1 - \frac{P_1 + P_2}{2} \right)} - (P_1 - P_2) \sqrt{n}}{\sqrt{P_1(1-P_1) + P_2(1-P_2)}}. \quad (2)$$

A continuity-corrected version of the arc sine approximation was provided by Walters (1979),

$$Z_\beta = Z_{1-\alpha} - \sqrt{2n} \left(\begin{array}{l} \text{Sin}^{-1} \sqrt{\left(P_1 - \frac{1}{2n} \right)} \\ - \text{Sin}^{-1} \sqrt{\left(P_2 + \frac{1}{2n} \right)} \end{array} \right), \quad (3)$$

and a continuity-corrected version of the Chi-squared approximation has been provided by Fleiss, Tytun, and Ury (1980), as follows:

Michael Vorburger is a Systems Engineer at JMP Division, SAS Institute. Email: mike.vorburger@jmp.com. Breda Munoz is a Research Statistician at RTI International. Email: breda@rti.org.

$$Z_{1-\alpha} \sqrt{2 \left(\frac{P_1 + P_2}{2} \right) \left(1 - \frac{P_1 + P_2}{2} \right)}$$

$$Z_{\beta} = \frac{-\sqrt{n(P_1 - P_2)^2 - 2(P_1 - P_2)}}{\sqrt{P_1(1 - P_1) + P_2(1 - P_2)}}. \quad (4)$$

Each of these corrected approximations offers advantages and drawbacks, depending on the sample size and magnitude of the proportions. The corrected arc sine formula (Equation 3) is a simpler formula but requires the use of the arc sine function for $(P_1 - 1/2n)$, so P_1 must be greater than $1/2n$. Additionally, the corrected Chi-squared formula is invalid when $(P_1 - P_2)$ is less than $2/n$.

Ury (1981) and Dobson and Gebski (1986) showed that the corrected approximations (Equations 3 and 4) yield a substantial improvement in the accuracy of the uncorrected approximations, as compared with Fisher's exact test for a 2 x 2 contingency table, when the sample size is equal to 30, and the proportions are relatively large (i.e., P_1 of 0.6-0.9, P_2 of 0.1-0.8, with minimum difference of 0.1). To the best of our knowledge, the accuracy of results from these corrected approximations when testing differences between smaller proportions has not been previously evaluated.

Power calculations for detecting differences between smaller proportions, using Fisher's exact test, all the approximations and a sample size of 30, are presented and discussed. Also presented is a discussion of power results comparisons for detecting differences between relatively small proportions, where the larger proportion is between 0.01 and 0.05 and the smaller proportion ranges from 0.001 to 0.007, for a sample size of 300. Finally, power results are compared for detecting differences of relatively small proportions for sample sizes of 300, 750, and 1,500 using both corrected approximations, and the accuracy of these approximations is discussed.

Methodology

In a preliminary analysis, Fisher's exact test, corrected and uncorrected approximations were

used to calculate the power needed to detect the differences between smaller proportions (ranging between 0.001 and 0.15). It was found that all of these approximations overestimate power for small proportions when the sample size is small, but the corrected approximations can be very accurate when the sample size is 300 or greater.

The power needed to detect differences of relatively small proportions using all four approximations, as well as Fisher's exact test, were calculated and compared. Table 1 shows the power, as calculated using Fisher's exact method and the two uncorrected approximations (Equations 1 and 2), associated with detectable differences where the larger proportion ranges from 0.075 to 0.15, the smaller proportion ranges from 0.001 to 0.008, and sample size is 30.

Table 2 is a replication of Table 1, substituting the two corrected approximations (Equations 3 and 4) for the uncorrected approximations. Tables 3 and 4 compare the power levels, as calculated using Fisher's exact method and both the uncorrected and corrected versions of each approximation, associated with detectable differences where the larger proportion ranges from 0.02 to 0.03, the smaller proportion ranges from 0.001 to 0.007, and sample size is 300. Table 3 compares exact vs. arc sine (Equations 1 and 3), and Table 4 compares exact vs. Chi-squared (Equations 2 and 4). Tables 5, 6, and 7 compare the power, calculated using the same methods as in Table 2, associated with detectable differences where the larger proportion is between 0.01 and 0.05, and the smaller proportion ranges from 0.001 to 0.007, for sample sizes of 300, 750, and 1,500, respectively.

Results

All four approximations overestimate power, sometimes by as much as 1,000% when P_1 is less than 0.2, P_2 is less than 0.1, and $n = 30$ (see Tables 1 and 2). However, the corrected approximations can be very accurate in determining power when the proportions are small and the sample size approaches 300. Additionally, the corrected approximations are more accurate than the uncorrected versions

POWER APPROXIMATIONS COMPARING TWO INDEPENDENT PROPORTIONS

Table 1: Power of Fisher’s Exact Test, with Both Uncorrected Approximations
($n = 30, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.003	0.005	0.007	0.008
Exact Power	0.075	0.06	0.06	0.06	0.05	0.05
Corrected Arc Sine Approximation		0.60	0.53	0.48	0.44	0.42
Corrected Chi-Squared Approximation		0.44	0.42	0.40	0.37	0.36
Exact Power	0.100	0.17	0.16	0.15	0.14	0.14
Corrected Arc Sine Approximation		0.73	0.66	0.62	0.58	0.56
Corrected Chi-Squared Approximation		0.54	0.52	0.50	0.48	0.47
Exact Power	0.150	0.46	0.40	0.42	0.41	0.40
Corrected Arc Sine Approximation		0.88	0.84	0.81	0.78	0.77
Corrected Chi-Squared Approximation		0.71	0.70	0.68	0.67	0.66

Table 2: Power of Fisher’s Exact Test, with Both Corrected Approximations
($n = 30, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.003	0.005	0.007	0.008
Exact Power	0.075	0.06	0.06	0.06	0.05	0.05
Corrected Arc Sine Approximation		0.22	0.20	0.18	0.17	0.16
Corrected Chi-Squared Approximation		0.12	0.10	0.09	0.07	0.06
Exact Power	0.100	0.17	0.16	0.15	0.14	0.14
Corrected Arc Sine Approximation		0.34	0.32	0.30	0.28	0.27
Corrected Chi-Squared Approximation		0.25	0.24	0.22	0.21	0.20
Exact Power	0.150	0.46	0.40	0.42	0.41	0.40
Corrected Arc Sine Approximation		0.59	0.56	0.54	0.52	0.51
Corrected Chi-Squared Approximation		0.49	0.47	0.46	0.44	0.43

Table 3: Power of Fisher's Exact Test, with Arc Sine Approximations
($n = 300, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.002	0.003	0.005	0.007
Exact Power	0.020	0.62	0.53	0.46	0.34	0.24
Corrected Arc Sine Approximation		0.66	0.58	0.50	0.37	0.27
Uncorrected Arc Sine Approximation		0.86	0.77	0.69	0.54	0.41
Exact Power	0.025	0.79	0.72	0.65	0.51	0.40
Corrected Arc Sine Approximation		0.80	0.73	0.67	0.54	0.43
Uncorrected Arc Sine Approximation		0.93	0.87	0.82	0.70	0.58
Exact Power	0.030	0.89	0.84	0.78	0.67	0.57
Corrected Arc Sine Approximation		0.89	0.84	0.79	0.69	0.58
Uncorrected Arc Sine Approximation		0.97	0.94	0.90	0.81	0.71

Table 4: Power of Fisher's Exact Test, with Chi-Squared Approximations
($n = 300, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.002	0.003	0.005	0.007
Exact Power	0.020	0.62	0.53	0.46	0.34	0.24
Corrected Chi-Squared Approximation		0.58	0.51	0.45	0.34	0.25
Uncorrected Chi-Squared Approximation		0.74	0.68	0.62	0.50	0.40
Exact Power	0.025	0.79	0.72	0.65	0.51	0.40
Corrected Chi-Squared Approximation		0.71	0.66	0.61	0.50	0.40
Uncorrected Chi-Squared Approximation		0.83	0.79	0.74	0.64	0.54
Exact Power	0.030	0.89	0.84	0.78	0.67	0.57
Corrected Chi-Squared Approximation		0.81	0.77	0.73	0.64	0.55
Uncorrected Chi-Squared Approximation		0.89	0.86	0.83	0.76	0.67

POWER APPROXIMATIONS COMPARING TWO INDEPENDENT PROPORTIONS

Table 5: Power of Fisher's Exact Test, with Both Corrected Approximations
($n = 300, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.002	0.003	0.005	0.007
Exact Power	0.020	0.62	0.53	0.46	0.34	0.24
Corrected Arc Sine Approximation		0.66	0.58	0.50	0.37	0.27
Corrected Chi-Squared Approximation		0.58	0.51	0.45	0.34	0.25
Exact Power	0.025	0.79	0.72	0.65	0.51	0.40
Corrected Arc Sine Approximation		0.80	0.73	0.67	0.54	0.43
Corrected Chi-Squared Approximation		0.71	0.66	0.61	0.50	0.40
Exact Power	0.030	0.89	0.84	0.78	0.67	0.57
Corrected Arc Sine Approximation		0.89	0.84	0.79	0.69	0.58
Corrected Chi-Squared Approximation		0.81	0.77	0.73	0.64	0.55
Exact Power	0.050	0.99	0.99	0.98	0.96	0.93
Corrected Arc Sine Approximation		0.99	0.99	0.98	0.96	0.93
Corrected Chi-Squared Approximation		0.97	0.99	0.95	0.93	0.90

Table 6: Power of Fisher's Exact Test, with Both Corrected Approximations
 ($n = 750, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.002	0.003	0.005	0.007
Exact Power	0.010	0.68	0.51	0.38	0.19	0.09
Corrected Arc Sine Approximation		0.70	0.54	0.40	0.21	0.10
Corrected Chi-Squared Approximation		0.63	0.50	0.38	0.19	0.08
Exact Power	0.015	0.92	0.84	0.74	0.53	0.34
Corrected Arc Sine Approximation		0.92	0.84	0.74	0.53	0.35
Corrected Chi-Squared Approximation		0.86	0.79	0.70	0.51	0.33
Exact Power	0.020	0.98	0.96	0.92	0.80	0.64
Corrected Arc Sine Approximation		0.99	0.96	0.92	0.80	0.64
Corrected Chi-Squared Approximation		0.96	0.93	0.88	0.77	0.62
Exact Power	0.025	0.99	0.99	0.98	0.93	0.85
Corrected Arc Sine Approximation		0.99	0.99	0.98	0.93	0.85
Corrected Chi-Squared Approximation		0.99	0.98	0.96	0.91	0.82

POWER APPROXIMATIONS COMPARING TWO INDEPENDENT PROPORTIONS

Table 7: Power of Fisher’s Exact Test, with Both Corrected Approximations
($n = 1,500, \alpha = 0.05$)

Approximation	Larger Proportion (P_1)	Smaller Proportion (P_2)				
		0.001	0.002	0.003	0.005	0.007
Exact Power	0.010	0.96	0.87	0.72	0.40	0.17
Corrected Arc Sine Approximation		0.96	0.86	0.72	0.40	0.17
Corrected Chi-Squared Approximation		0.92	0.83	0.69	0.39	0.16
Exact Power	0.015	0.99	0.99	0.96	0.84	0.62
Corrected Arc Sine Approximation		0.99	0.99	0.97	0.84	0.62
Corrected Chi-Squared Approximation		0.99	0.98	0.95	0.82	0.61
Exact Power	0.020	0.99	0.99	0.99	0.98	0.91
Corrected Arc Sine Approximation		0.99	0.99	0.99	0.98	0.91
Corrected Chi-Squared Approximation		0.99	0.99	0.99	0.97	0.90
Exact Power	0.025	0.99	0.99	0.99	0.99	0.98
Corrected Arc Sine Approximation		0.99	0.99	0.99	0.99	0.99
Corrected Chi-Squared Approximation		0.99	0.99	0.99	0.99	0.98

when the proportions are small and $n = 300$ (see Tables 3 and 4).

When $n = 300$ (see Table 5), the corrected Chi-squared approximation (Equation 4) is more accurate for smaller proportions, whereas the corrected arc sine approximation (Equation 3) overestimates the exact power. As the proportions and differences become larger, the corrected arc sine approximation (Equation 3) becomes more accurate, although still slightly overestimating the exact power.

As n reaches 750 (see Table 6), the accuracy of both corrected approximations for calculating the power of tests of differences between relatively small proportions increases.

Again, with smaller proportions the corrected Chi-squared approximation (Equation 4) provides a more accurate and conservative calculation of power. However, once P_1 reaches 0.015, the corrected arc sine approximation (Equation 3) provides power calculations identical (to 2 decimal points) to Fisher’s exact test, whereas the corrected Chi-squared approximation (Equation 4) still slightly underestimates the power.

Furthermore, as n reaches 1,500 (see Table 7), the corrected arc sine approximation (Equation 3) is more accurate regardless of the magnitude of the proportions considered, and it

no longer overestimates the power for smaller proportions. Thus, these analysis results suggest that the corrected arc sine approximation (Equation 3) should be used exclusively to determine the power of tests of differences between two proportions once n reaches 1,500.

Conclusion

Analysis of results suggest that the continuity-corrected approximations provided by Walters (1979) and Fleiss, et al. (1980) result in more accurate power levels than the uncorrected versions previously provided by Cochran and Cox (1957), and Fleiss (1973), for determining the power of tests of differences between small proportions when sample size is at least 300. The uncorrected approximations greatly overestimate the power of these tests. Specifically, when $n = 300$ or 750 the corrected Chi-squared approximation (Equation 4) is more accurate for smaller proportions, whereas the corrected arc sine approximation (Equation 3) becomes more accurate as the size of the proportions increases. When $n = 1,500$ the corrected arc sine approximation (Equation 3) is more accurate for all proportions presented above.

Acknowledgement

Thanks to Dr. James Chromy for his valuable comments and suggestions.

References

- Cochran, W. G., & Cox, G. M. (1957). *Experimental design* (2nd Ed.). NY: Wiley.
- Dobson, A. J., & Gebski, V. J. (1986). Sample sizes for comparing two independent proportions using the continuity-corrected arc sine transformation. *The Statistician*, 35, 51-53.
- Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. NY: Wiley.
- Fleiss, J. L., Tytun, A., & Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, 36, 343-346.
- Ury, H. K. (1981). Continuity-corrected approximations to sample size or power when comparing two proportions: Chi-squared or arc sine? *The Statistician*, 30, 199-203.
- Walters, D. E. (1979). In defense of the arc sine approximation. *The Statistician*, 28, 219-222.

Improved Confidence Intervals for the Difference between Two Proportions

James F. Reed III
Christiana Care Hospital System, Newark, Delaware

Wald-z asymptotic methods, with and without a continuity correction, have less than nominal coverage probability characteristics but continue to be used. Newcombe's hybrid method and the Agresti-Caffo methods have coverage probabilities that are near nominal for either equal or unequal samples. Newcombe's hybrid and Agresti-Caffo methods demonstrate superior coverage properties.

Key words: Wald-z asymptotic, Newcombe's hybrid, Agresti-Caffo.

Introduction

In reporting the results of medical studies the problem of comparing two binomial success probabilities p_1 and p_2 , $p_1 > 0$ and $p_2 > 0$ is often encountered. Implicit in this comparison are the independent observations $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$. The most common comparison is the hypothesis $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$. Accompanying the hypothesis test is the construction of a confidence interval for the difference between p_1 and p_2 . Nearly all introductory statistics textbooks include a method for computing this confidence interval and issue a warning - usually in a footnote - when not to use the common method: this commonly described method is the Wald-z method. Occasionally, a continuity corrected version is given (Wald-c).

The problems associated with the confidence interval for the difference between two independent proportions are similar to the confidence interval of a single proportion. Despite these properties, the Wald-z and Wald-c methods continue to dominate. We review the coverage probability functions of the Wald methods and a set of alternative methods for computing a confidence interval for the difference between two independent proportions.

James F. Reed III, PhD, is a Senior Biostatistician. Email him at: JaReed@ChristianaCare.org.

Methodology

The Wald-z and Wald-c confidence interval lower upper bounds for the difference between two independent proportions are defined as (See Appendix A for a typical data structure):

Wald-z:

$$\begin{aligned} \text{LB} &= (p_1 - p_2) - z_{\alpha/2} \sqrt{ac/m^3 + bd/n^3} \\ \text{UB} &= (p_1 - p_2) + z_{\alpha/2} \sqrt{ac/m^3 + bd/n^3} \end{aligned}$$

Wald-c:

$$\begin{aligned} \text{LB} &= (p_1 - p_2) - [z_{\alpha/2} \sqrt{\{ac/m^3 + bd/n^3\} + (1/m + 1/n)/2}] \\ \text{UB} &= (p_1 - p_2) + [z_{\alpha/2} \sqrt{\{ac/m^3 + bd/n^3\} + (1/m + 1/n)/2}] \end{aligned}$$

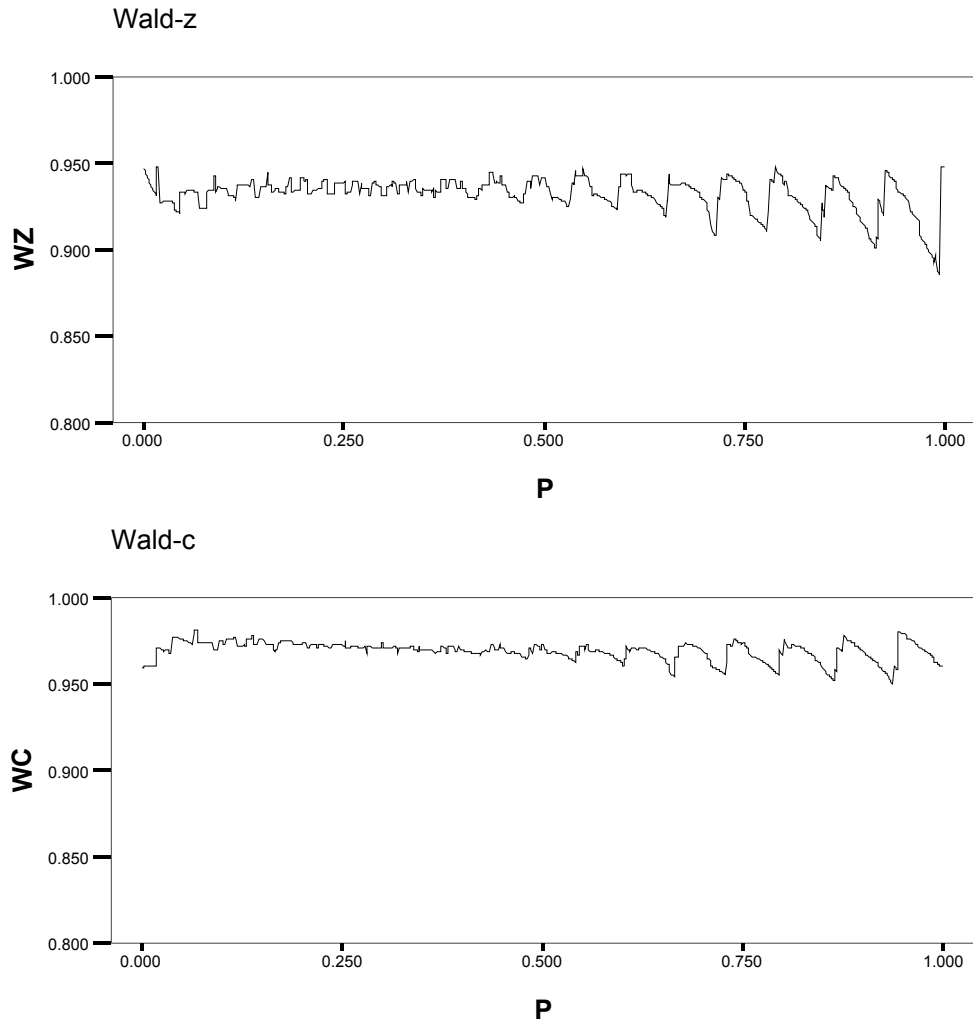
The primary criteria for evaluating a confidence interval method is the coverage probability function. This coverage probability for the difference between two independent proportions, $C(\pi_1, \pi_2 | n_1, n_2, \alpha)$, is found by fixing n_1, n_2, π_1 , and π_2 , then computing the confidence interval for each $x_i = 0, \dots, n_i$ for $i = 1, 2$. The coverage probability is then defined by:

$$\begin{aligned} C(\pi_1, \pi_2 | n_1, n_2, \alpha) &= \\ &= \sum \Pr(X_1 = x_1 | n_1, \pi_1) \Pr(X_2 = x_2 | n_2, \pi_2) \\ &= \delta(\pi_1, \pi_2 | x_1, x_2, n_1, n_2, \alpha). \end{aligned}$$

If $(\pi_1 - \pi_2) \in [\text{LB}(x_1, x_2, n_1, n_2, \alpha), \text{UB}(x_1, x_2, n_1, n_2, \alpha)]$, $\delta(\pi_1, \pi_2 | x_1, x_2, n_1, n_2, \alpha) = 1$, and 0 otherwise.

Figure 1 shows the 95% confidence interval coverage probability function for the Wald-z and Wald-c methods as a function of π_1 , $\pi_1 \in [0, 1]$ for $n_1 = n_2 = 20$ and $p_2 = 0.3$. The sawtooth appearance of the coverage functions

Figure 1: Coverage probabilities for nominal 95% Wald-z and Wald-c as a function of p_1 when $p_2=0.3$ with $n_1=n_2=20$



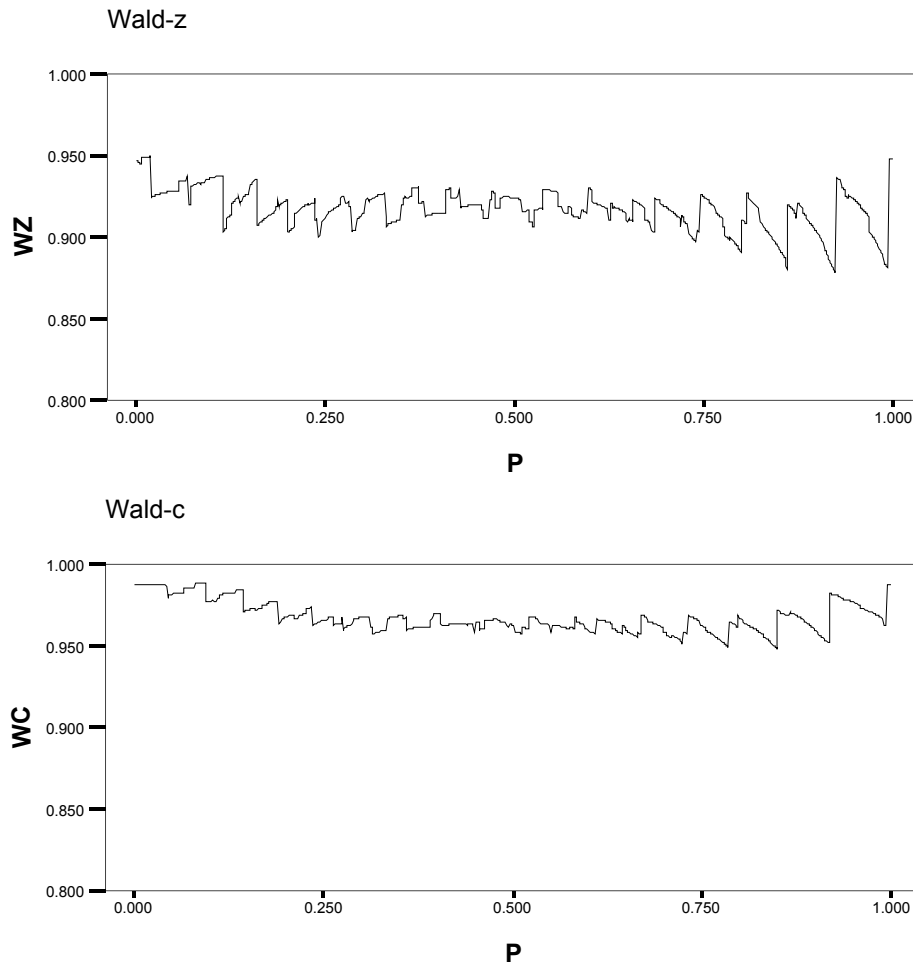
is due to the discontinuities for values of p_1 corresponding to any lower or upper limits in the set of confidence intervals. Like its one sample cousin, the Wald-z coverage probability curve is subnominal and less than 0.95 overall. The Wald-c coverage probability always exceeds 0.95 overall with interval widths larger than Wald-z.

Figure 2 shows the 95% confidence interval coverage probability function for the Wald-z and Wald-c methods as a function of π_1 , $\pi_1 \in [0,1]$ for $n_1 = 20$, $n_2 = 10$ and $p_2 = 0.3$. The Wald-z coverage probability curve is subnominal for differences in proportions near 0 and 1 and less than 0.95 overall.

Beal evaluated several asymptotic methods for computing a confidence interval between the differences of two independent proportions. All involved identifying the interval within which $(\theta - \theta')^2 \leq z^2 V(\psi, \theta')$, where $\theta' = p_1 - p_2$, and $V(\psi, \theta') = \psi(1 - \psi) = \pi_1(1 - \pi_1)/m + \pi_2(1 - \pi_2)/n$ (Beal, 1987). Beal examined two methods, labeled the Haldane (H) and Jeffreys-Perks (JP) methods. The JP method provides non-degenerate confidence intervals for all values of p_1 and p_2 unlike Wald-z or Wald-c. H and JP generally performed better than the Wald-z and Wald-c and of the two, JP was preferred (Beal, 1987; Radhakrishna, et. al., 1992).

CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

Figure 2: Coverage probabilities for nominal 95% Wald-z and Wald-c as a function of p_1 when $p_2=0.3$ with $n_1=20, n_2=10$



The Haldane and Jeffreys-Perks lower and upper limits are defined by:

H

$$LB=\theta^* - w,$$

and

$$UB=\theta^* + w,$$

where

$$\theta^*=(\theta'+z^2v(1-2\psi'))/(1+z^2u),$$

$$w=[z/(1+z^2u)]\sqrt{[u\{4\psi'(1-\psi')-\theta'^2\}+2v(1-2\psi')\theta'+4z^2u^2(1-\psi')\psi'+z^2v^2(1-2\psi')^2]}$$

$$\psi'=(a/m+b/n)/2,$$

$$u=(1/m+1/n)/4,$$

and

$$v=(1/m-1/n)/4.$$

JP

$$LB=\theta^* - w,$$

and

$$UB=\theta^* + w,$$

where ψ' from the Haldane method is:

$$\psi'=[(a+0.5)/(m+1)+(b+0.5)/(n+1)]/2.$$

Newcombe (1998) compared eleven methods for estimating the difference between independent proportion. Similar to the single proportion, the virtues of Wald-z and Wald-c

methods are in their simplicity, but overshoot and inappropriate intervals are still common. The Haldane and Jeffreys-Perks methods attempt to overcome the overshoot and inappropriate intervals while maintaining closed-form tractability. Newcombe concluded that both H and JP were improvements over the Wald-z and Wald-c methods, but both were still inadequate. Newcombe recommended a hybrid method based on Wilson's score method for a single proportion without continuity correction (NS). The LB and UB for the NS method are:

NS

$$LB=(p_1-p_2)-\delta,$$

where

$$\begin{aligned} \delta &= \sqrt{\{(a/m-1_1)^2+(u_2-b/n)^2\}} \\ &= z_{\alpha/2} \sqrt{\{l_1(1-l_1)/m+u_2(1-u_2)/n\}}. \end{aligned}$$

$$UB = (p_1 - p_2) + \varepsilon,$$

where

$$\begin{aligned} \varepsilon &= \sqrt{\{(u_1-a/m)^2+(b/n-1_2)^2\}} \\ &= z_{\alpha/2} \sqrt{\{u_1(1-u_1)/m+l_2(1-l_2)/n\}}, \end{aligned}$$

and l_1, l_2, u_1, u_2 are the lower and upper bounds for the two proportions p_1 and p_2 using Wilson's score method.

Agresti & Coull's (1998) adjustment to the Wald method for a single proportion adds $t/2$ successes and $t/2$ failures. Agresti & Caffo (2000) later suggested that by adding two successes and two failures (total) to the two-sample method would improve the simple Wald

method. This is an adjustment that adds a pseudo observation of each type to each sample. For instance, for sample i , $p_i = (r_i+1)/(n_i+2)$.

Results

Figure 3 shows the 95% confidence interval coverage probability function for the Newcombe NS, Haldane, Jeffreys-Perks, and Agresti-Caffo methods as a function of $\pi_1, \pi_1 \in [0,1]$ for $n_1 = n_2 = 20$ and $p_2 = 0.3$. The NS and Agresti-Caffo methods demonstrate coverage probabilities that are near nominal over $\pi_1 \in [0, 1]$.

Figure 4 shows the 95% confidence interval coverage probability function for the Newcombe NS, Haldane, Jeffreys-Perks, and Agresti-Caffo methods as a function of $\pi_1, \pi_1 \in [0,1]$ for $n_1 = 20, n_2 = 10$ and $p_2 = 0.3$. In the unequal sample size situation, Newcombe NS and Agresti-Caffo coverage probability functions are near nominal over $\pi_1 \in [0, 1]$.

Conclusion

In the case of differences between two independent proportions the Wald-z confidence interval behaves poorly with coverage probabilities below nominal values. Considering the coverage probability criterion, two alternative methods demonstrate superior coverage properties and both are easily programmable. Based on these results, the recommendation is to use either the NS or the Agresti-Caffo methods.

References

Agresti, A., & Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52, 119-126.

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, 54, 280-288.

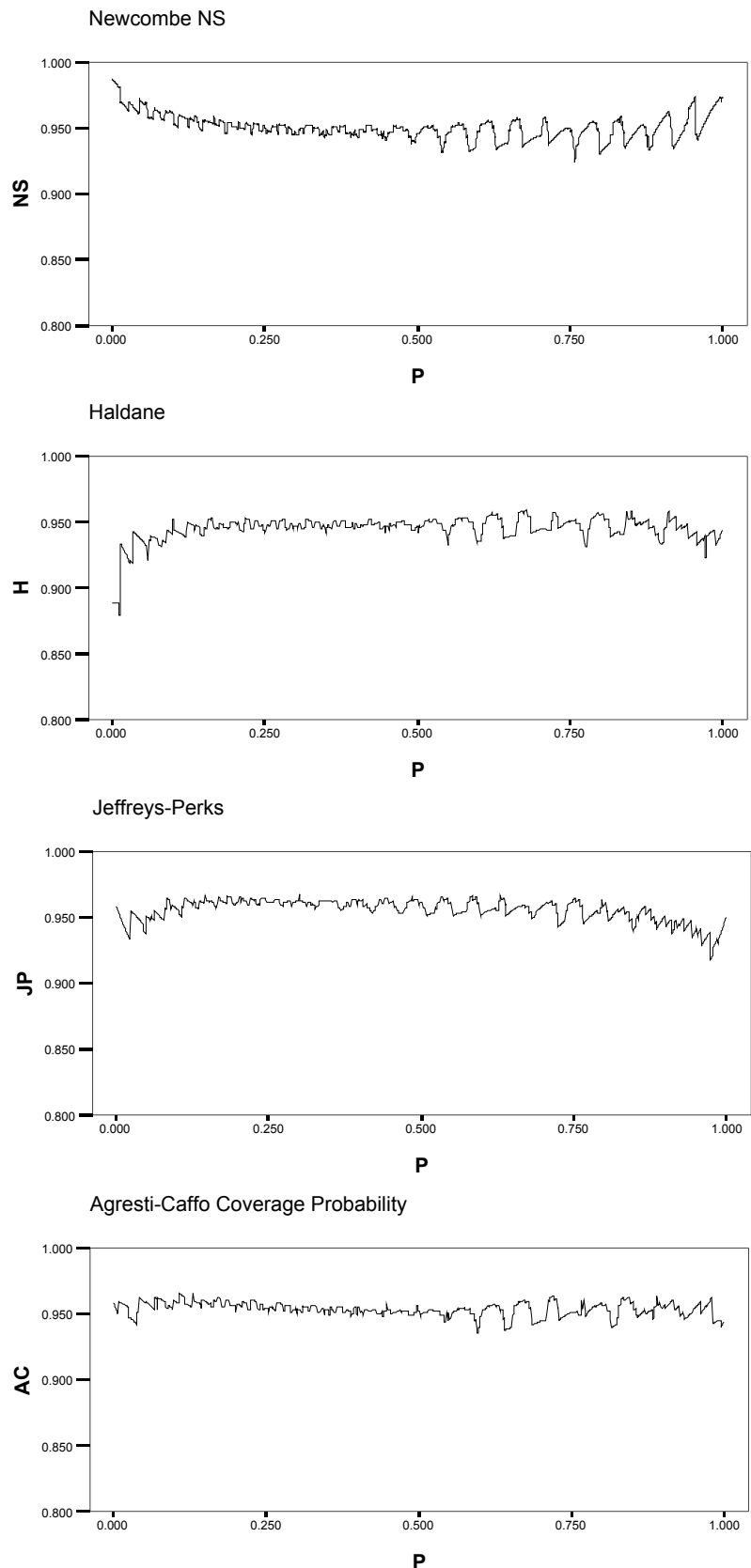
Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, 43, 941-950.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med*, 17, 873-890.

Radhakrishna, S., Murthy, B. N., Nair, N. G. K., Jayabal, P., & Jayasri, R. (1991). Confidence intervals in medical research. *Indian J Med Res [B]*, 96, 199-205.

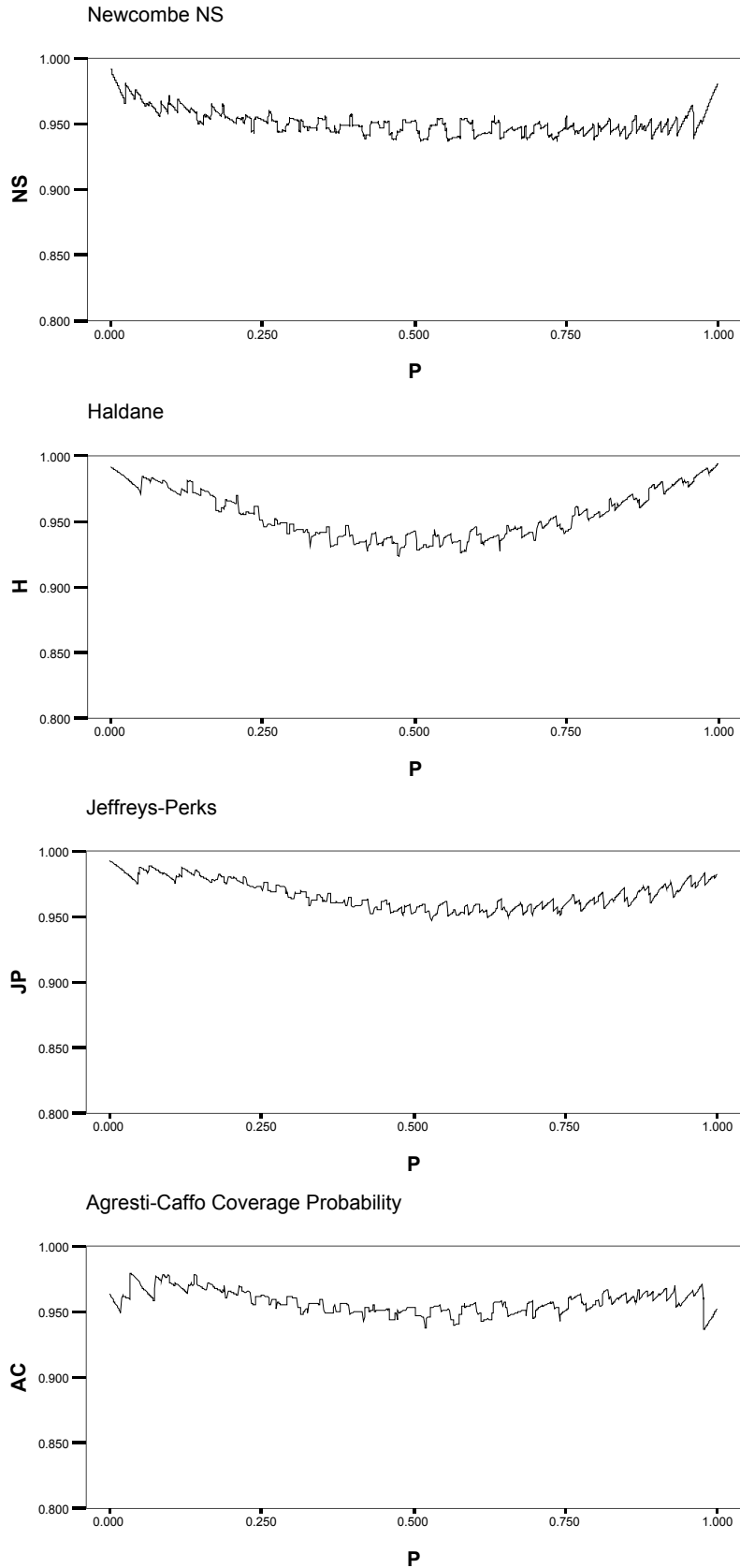
CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

Figure 3: Coverage probabilities for nominal 95% Newcombe NS, Haldane, Jeffreys-Perks, and Agresti-Caffo as a function of p_1 when $p_2=0.3$ with $n_1=n_2=20$



REED

Figure 4: Coverage probabilities for nominal 95% Newcombe NS, Haldane, Jeffreys-Perks, and Agresti-Caffo as a function of p_1 when $p_2=0.3$ with $n_1=20$, $n_2=10$



CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

Appendix A: Methods for calculation of confidence intervals for the difference between independent proportions

	Sample 1	Sample 2	
+	a	b	$p_1 = a/m$
-	c	d	$p_2 = b/n$
Total	m	m	$\theta = \pi_1 - \pi_2$ $\theta' = p_1 - p_2$

Method	Formula
Wald-z	LB= $(p_1 - p_2) - z_{\alpha/2} \sqrt{(ac/m^3 + bd/n^3)}$ UB= $(p_1 - p_2) + z_{\alpha/2} \sqrt{(ac/m^3 + bd/n^3)}$
Wald-c	LB= $(p_1 - p_2) - [z_{\alpha/2} \sqrt{\{ac/m^3 + bd/n^3\} + (1/m + 1/n)/2}]$ UB= $(p_1 - p_2) + [z_{\alpha/2} \sqrt{\{ac/m^3 + bd/n^3\} + (1/m + 1/n)/2}]$
Haldane-H	LB= $\theta^* - w$ UB= $\theta^* + w$, where $\theta^* = (\theta' + z^2 v(1 - 2\psi')) / (1 + z^2 u)$, $w = [z / (1 + z^2 u)] \sqrt{[u \{4\psi'(1 - \psi') - \theta'^2\} + 2v(1 - 2\psi')\theta' + 4z^2 u^2 (1 - \psi')\psi' + z^2 v^2 (1 - 2\psi')^2]}$ $\psi' = (a/m + b/n) / 2$, $u = (1/m + 1/n) / 4$, and $v = (1/m - 1/n) / 4$
Jeffreys-Perks-JP	LB= $\theta^* - w$ UB= $\theta^* + w$, where ψ' (from Haldane method) is: $\psi' = [(a + 0.5) / (m + 1) + (b + 0.5) / (n + 1)] / 2$
Newcombe-NS	LB= $(p_1 - p_2) - \delta$, where $\delta = \sqrt{\{(a/m - l_1)^2 + (u_2 - b/n)^2\}} = z_{\alpha/2} \sqrt{\{l_1(1 - l_1) / m + u_2(1 - u_2) / n\}}$ UB= $(p_1 - p_2) + \varepsilon$, where $\varepsilon = \sqrt{\{(u_1 - a/m)^2 + (b/n - l_2)^2\}} = z_{\alpha/2} \sqrt{\{u_1(1 - u_1) / m + l_2(1 - l_2) / n\}}$ l_1, l_2, u_1, u_2 are the LB and UB for p_1 and p_2 using Wilson's score method
Agresti & Caffo	LB = $(p_1 - p_2) - z_{\alpha/2} \sqrt{(ac/m^3 + bd/n^3)}$ UB = $(p_1 - p_2) + z_{\alpha/2} \sqrt{(ac/m^3 + bd/n^3)}$

Robustness to Non-Independence and Power of the I Test for Trend in Construct Validity

John Cuzzocrea Shlomo Sawilowsky
Wayne State University

The Multitrait-Multimethod Matrix is used to evaluate construct validity; Sawilowsky (2002) created the I test to analyze the matrix. This article examined the robustness and power of the Sawilowsky I test. Ad hoc critical values were determined to improve the statistical power of the technique for analyzing the Multitrait-Multimethod Matrix.

Key words: Multitrait-Multimethod matrix, convergent validity, discriminant validity, I test, robustness, power.

Introduction

“A construct is a fiction that is used to explain reality” (Sawilowsky, lecture notes). Nearly half a century ago, Campbell and Fiske (1959) developed the Multitrait-Multimethod Matrix as a means of analyzing convergent and divergent validity, the two integral parts of construct of validity. Analysis of the matrix is hinged on the concept that the greater the degree of convergent and discriminant validity; the greater the evidence of construct validity. The matrix is the classical approach to construct validation and has received considerable attention. According to Sternberg (1992), it had received over 2,000

citations over the years, making it the most cited paper published by *Psychological Bulletin*. Yet, the matrix remains troubled by the same issues that plagued it when it was initially conceived. According to Sawilowsky (2002), the “interpretation of the matrix is subjective ... (and) not amenable to straightforward interpretation” (p.78).

Campbell and Fiske (1959) recognized that further study was required and that “various statistical treatments for Multitrait-Multimethod matrices might be developed...However, the development of such statistical methods is beyond the scope of this paper” (p.103). The development of the Multitrait-Multimethod Matrix was viewed as a necessary first step in determining construct validity, from which it was believed that further research would resolve these issues over time. The recognized limitations of their study, as presented in their original article, turned to exasperation as little progress had been made in evaluating the matrix. Fiske and Campbell (1992) expressed their frustration by stating that scholarly journals and researchers alike continue to accept articles that provide no greater evidence of convergent and discriminant validity than from the time their original article was first published, and that there was still no general consensus of how to statistically evaluate convergent and discriminant validity.

The matrix is subdivided into various components that contribute to the analysis which

John Cuzzocrea is a Lecturer in Educational Evaluation and Research. His research interests include nonparametric statistics, Monte Carlo simulations, and evaluation and measurement. Email: jcuzzocrea@hotmail.com. Shlomo Sawilowsky is Assistant Dean of the Division of Administrative and Organizational Studies, and the Division of Theoretical and Behavioral Foundations, and Professor of Educational Evaluation and Research. His research interests include nonparametric statistics, Monte Carlo Simulations, and classical measurement theory. Email: shlomo@wayne.edu.

ROBUSTNESS AND POWER OF THE I TEST IN CONSTRUCT VALIDITY

include the: a) reliability diagonal, b) validity diagonal, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block. Campbell and Fiske (1992) provided a guideline for interpreting the matrix and determining the degree of convergent and discriminant validity. Figure 1 provides an illustration of the various components of the Multitrait-Multimethod Matrix.

To evaluate convergent validity, the values found in the validity diagonal “should be significantly different from zero and sufficiently large to encourage further examination of validity” (Campbell & Fiske, 1959, p.82). Conversely, the process in determining discriminant validity is more involved. To begin, the values in the validity diagonal should be higher than the values found in the corresponding heterotrait-monomethod block. Second, the values in the heterotrait-monomethod block should be higher than the values found in the heterotrait-heteromethod block. In applying the rationale outlined by Campbell and Fiske (1959), there should be an ascending trend from the heterotrait-heteromethod values to the reliability diagonal.

Various statistics have been employed as a means of analyzing the matrix (Hubert & Baker, 1978, Stanley, 1961, Jöreskog, 1971). However, these approaches are not without their own set of difficulties ranging from the complexity of the procedures to restrictive assumptions that are difficult to satisfy (Schmitt & Stults, 1986; Widaman, 1985). As a result, Sawilowsky (2002) created a quick, distribution-free test that does not suffer the same pitfalls of its predecessors. It was called the I statistic because it focuses on the number of inversions found within the matrix. The I statistic is relatively simple to compute, it incorporates the entire matrix, and it does not have the restrictive assumptions that have hampered previous efforts.

The I statistic is a combination of the Jonckheere’s distribution-free k-sample test against ordered alternatives (Jonckheere, 1954) and Mann’s test for randomness in a single sample (Neave & Worthington, 1988). According to Sawilowsky (2002), “The I statistic combines the counting function of the Mann’s test with the logic of Jonckheere’s

statistic.” (p.85). Whereas Jonckheere’s test uses all of the values within the matrix, which increases the power of the test, but also increases the probability of violating the independence assumption; the I statistic is limited to three values at each level of the matrix: a) minimum coefficient, b) median coefficient, and c) maximum coefficient. As a result, a minimum, median, and maximum value is derived from the each of the following components of the Multitrait-Multimethod Matrix: a) reliability diagonals, b) validity diagonals, c) heterotrait-monomethod block, and d) heterotrait-heteromethod block.

The hypothesis tested by the I statistic is the upward trend of values, from the heterotrait-heteromethod values to the reliability diagonal, as evidence of construct validity. This approach incorporates the criteria outlined by Campbell and Fiske (1959), in that the values in the heterotrait-heteromethod block should be lower than the values found in the heterotrait-monomethod block, which in turn should be lower than those found in the validity diagonals, and so forth. Therefore, construct validity is supported through fewer inversions. A nominal number of inversions are easily regarded as evidence of construct validity; however, the decision becomes more difficult and subjective as the number of inversions increase.

The internal correlation structure of the I test makes it susceptible to the independence assumption and although the risk of violating this assumption is minimized by using a limited number of the values in the matrix (i.e. minimum, median, and maximum coefficients with a three-point I statistic), the risk of violating this assumption increases as the number of values used in the test increases (i.e. four-point I statistic). However, the question becomes whether a violation of independence will impact adversely impact the Type I error rate.

Statement of the Problem

As a result, a modified version of the Sawilowsky I test is proposed to incorporate more data points. The three-point I statistic is comprised of four groups, representing the

Figure 1: An Example of a Multitrait-Multimethod Matrix (Campbell & Fiske, 1959, p.82)

	Method One			Method Two			Method Three		
	A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Method One									
A ₁	(.89)								
B ₁	.51	(.89)							
C ₁	.38	.37	(.76)						
Method Two									
A ₂	[.57]	.22	.09	(.93)					
B ₂	.22	[.57]	.10	.68	(.91)				
C ₂	.11	.11	[.46]	.59	.58	(.81)			
Method Three									
A ₃	[.56]	.22	.11	[.67]	.42	.33	(.94)		
B ₃	.23	[.58]	.12	.43	[.66]	.34	.67	(.92)	
C ₃	.11	.11	[.45]	.34	.32	[.58]	.58	.60	(.85)

Note. A = assertive; B = cheerful; C = serious. Values in parentheses represent the reliability diagonal. Values in the squared brackets represent the validity diagonal. Boldface type represents the heterotrait-monomethod values and regular type represents the heterotrait-heteromethod values.

different facets of the Multitrait-Multimethod Matrix, with three values in each (i.e., minimum coefficient, median coefficient, and maximum coefficient). A modified four-point version of the I statistic will encompass four data points at each level of the matrix (minimum coefficient, lower quartile, upper quartile, and maximum coefficient). Both versions of the I statistic will be examined to determine the impact upon each when independence has been violated. The study will also examine the power properties of both the three-point and four-point versions of the test to determine if an increasing number of data points will (comparing the three-point version to the four-point version) will lead to greater power.

Although Campbell and Fiske (1959) provided a heuristic approach for evaluating construct validity, a statistical approach that incorporates these guidelines is necessary in order to eliminate the subjectivity involved in this process. Fiske and Campbell (1992) argued that “editors and readers are accepting matrices showing limited convergence or discrimination, or both, perhaps because these are so typical, so common in the published literature” (p. 393).

Sawilowsky (2002) showed that the I statistic provided comparable results to those achieved by Campbell and Fiske (1959) using a quick test that eliminates the subjectivity that has plagued this process in the past.

Methodology

The study involved a Monte Carlo simulation whereby data were obtained through repeated sampling from the uniform distribution, as opposed to collecting data from a group of test subjects. The uniform distribution was selected because the data collected from this distribution would be similar in nature to the correlation coefficients that are found within the Multitrait-Multimethod matrix. A program was written in Intel Visual Fortran (Version 10) to compute the three-point and four-point versions of the I test. Specifically, the programs were written with the intent of examining the robustness of each test with regard to the internal correlation structure and the power properties of each version of the test. The design layouts used in the analysis were modeled on the matrices provided in Campbell and Fiske (1959). As a result, both the

ROBUSTNESS AND POWER OF THE I TEST IN CONSTRUCT VALIDITY

three-point and four-point versions of the I test were computed using a 2x3, 3x2, 3x3, 2x4, and 3x5 matrix.

The number of values obtained was dependent upon the design layout modeled. As an example, with a 2x3 matrix, the total number of values obtained from the random number generator would be 21. These values were then placed into one of four groups corresponding to the different levels of the Multitrait-Multimethod matrix. Therefore, in a 2x3 matrix, there are 6 heterotrait-heteromethod values, 6 heterotrait-monomethod values, 3 validity diagonal values, and 6 reliability diagonal values. The three-point version of the I test required three data points at each level: a) minimum, b) median, and c) maximum values. The four-point version of the I test required four data points at each level: a) minimum, b) lower quartile, c) upper quartile, and d) maximum values. These data points were obtained by sorting the data placed within each level to determine the minimum and maximum values and then computing the median for the three-point I test and the lower and upper quartiles for the four-point version of the I test.

In analyzing the robustness of I test, separate subroutines were programmed to calculate both the three-point and the four-point versions of the test. A counter was written into the program to check for the number of significant results at the 0.05 alpha level. This process was repeated for 1,000,000 repetitions and the number of times that the null hypothesis was rejected was then divided by 1,000,000; thereby providing the Type I error rate. This process was in turn repeated for the 0.01 alpha level.

These results were compared to those obtained by computing the I test using random, as opposed to sorted values. Specifically, a program was written to compute both the three-point and four-point versions of the I test, whereby values were placed within each level at random. Therefore, there is no internal correlation structure within each level. As a result, the program to be used to calculate the three-point I test using random data, only obtained 12 random values from the uniform distribution, as opposed to 21 (assuming a 2x3 matrix). The first three values were placed in the

heterotrait-heteromethod level; the next three values were placed in the heterotrait-monomethod level, and so forth. The four-point I test program using random data obtained 16 random values from the uniform distribution, as opposed to 21. The first four values were placed in the heterotrait-heteromethod level; the next four values were placed in the heterotrait-monomethod level, and so forth. As a result, the values were not sorted and the minimum, median, and maximum values were not calculated for the three-point I test, nor the minimum, lower quartile, upper quartile, and maximum values for the four-point I test. This process was in turn repeated for the 0.01 alpha level.

Despite the fact that the values are not ascending within each level of the randomized version of the I test, the number of comparisons remained constant for both the randomized and sorted versions of the I test. As a result, there were still 54 comparisons made for the three-point version and 96 comparisons made for the four-point version. There were no comparisons made within each level in determining the number of inversions. By maintaining the same number of comparisons, the critical values remained the same and thus a comparison could be made for the random and sorted versions of both the three-point and four-point I tests regarding the Type I error rate.

The next phase of the study examined the power properties of both the three-point and four-point versions of the I test. First, focus was placed on the Type I error rate, whereby significance was based solely on the number of inversions, without regard for the types of values comprised within each of the levels. In an applied setting, an analysis of the Multitrait-Multimethod matrix may be found to be significant; however, the results would be valid only if the reliability diagonal values were greater than or equal to 0.8. As a result, in determining the power properties of the I test, the reliability diagonal values were kept above a predetermined standard. Specifically, a series of programs were written for both the three-point and four-point versions of the I test that would ensure that the reliability diagonal values used in the analysis are greater than or equal to 0.7, 0.8, and 0.9 respectively. For each program, the

number of significant results were divided by the total number of repetitions to determine the power of the test. This process was completed for both the 0.05 and 0.01 alpha levels.

The number of repetitions used in this phase of the analysis was 2,000. Fewer repetitions were used because of the time involved in processing 1,000,000 repetitions when the values are required to be above a predetermined standard. As a result, if the random number generator returns values that are below this predetermined standard, then the program will be prompted to loop back to the beginning to find a new random set of values from the distribution. As an example, if the reliability diagonal values are required to be greater than or equal to 0.9, then the program will be required to cycle through numerous times before it will return values that conform to this requirement.

The results were compared to those obtained by computing the I test using random, as opposed to sorted values. Once again, a program was written to compute both the three-point and four-point versions of the I test, whereby values were placed within each level at random. As a result, there was no internal correlation structure within each level. The program was set to 2000 repetitions and the number of significant results was divided by the number of repetitions to determine the power of the test. This process was completed for both the 0.05 and 0.01 alpha levels.

In order to establish a baseline for comparison, the relative efficiency was calculated to quantify and thereby allow for a comparison between the power of the four-point I test and the three-point I test. The relative efficiency was calculated by dividing the three-point randomized values by the three-point sorted values. As well, the four-point randomized values were divided by the four-point sorted values. The next step was to divide the quotient from the four-point calculation by the quotient from the three-point calculation. This provided the relative efficiency of the four-point I test versus the three-point I test and this calculation was repeated for the 0.7, 0.8, and 0.9 thresholds for each of the experimental design layouts at both the 0.05 and 0.01 alpha levels.

The critical values used for the analysis of the three point I statistic were obtained from Sawilowsky (2002). It was found that the critical values for the three-point I statistic at the 0.05 and 0.01 alpha levels were 14 and 10, respectively. In contrast, the critical values for the four-point I statistic were obtained from Jonckheere (1954). Critical values for the 0.05 and 0.01 alpha levels were obtained by counting the number of inversions starting from the bottom of the table (refer to his Table 3, p.145). This is due to the fact that the Jonckheere test works in reverse order to the Sawilowsky I statistic. It was found that the critical values for the four-point I statistic at the 0.05 and 0.01 alpha levels were 29 and 23 respectively.

Results

Type I Error

It was predicted by Sawilowsky (2002), that the Type I error rate would increase with an increasing number of data points (i.e. the three-point versus the four-point versions of the test). Although it was predicted that the Type I error rate would be adversely affected, the severity in violating this assumption remained unknown. As a result, the Type I error rate for both the three-point and four-point versions of the I test were examined at both the 0.05 and 0.01 alpha levels. The three-point and four-point sorted versions of the I test were compared to the three-point and four-point randomized versions of the I test for various experimental design layouts (i.e. 2x3, 2x4, 3x2, 3x3, and 3x5 matrices).

In Table 1, it is shown that the randomized versions of both the three-point and four-point versions of the test performed as expected, with a Type I error rate that was close to 0.05; specifically, 0.042514 for the three-point randomized version and 0.042045 for the four-point randomized version. In examining the three-point and four-point sorted versions of the I test, it was found that the Type I error rate did increase with an increasing number of data points. Using the 2x3 matrix as an example, the Type I error rate for the three-point sorted version of the I test was 0.002193 and the Type I error rate for the four-point sorted version of the I test was 0.007527. This result was consistent

ROBUSTNESS AND POWER OF THE I TEST IN CONSTRUCT VALIDITY

across each of the experimental design layouts tested.

Table 2 examined the robustness of both the three-point and four-point versions of the I test at the 0.01 alpha level. Once again, it was found that the randomized versions of the test performed as expected, with a Type I error rate that was close to 0.01 (i.e. 0.009254 for the three-point randomized version and 0.009789 for the four-point randomized version). As well, it was found that the Type I error rate increased with an increasing number of data points. Using the 2x3 experimental design layout, it was found that the Type I error rate for the three-point sorted version of the I test was 0.000106 and the Type I error rate for the four-point sorted version of the I test was 0.000842. Once again, the result was consistent across each of the experimental design layouts tested.

Power Results

The second phase of the research examined the power of the I test by maintaining a predetermined threshold for the reliability diagonal values used in the analysis. The I test was computed with minimum reliability diagonal values set at 0.7, 0.8, and 0.9. It was expected that the power of both the three-point and four-point versions of the test would increase as the predetermined threshold for the reliability diagonal values increased, because it was logical to assume that there would be fewer inversions. As a result, focus was instead placed upon the examination of the three-point versus the four-point I test in terms of power.

Tables 3, 4, and 5 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.05 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts. However, it must be noted that these power equations are in closed form; therefore, a lack of resolution only indicates a limitation of resources. These values would compute given the proper time and resources to complete the analysis.

Tables 3, 4, and 5 each display an increased efficiency of the four-point over the three-point versions of the I test. In Table 3, the relative efficiency of the four-point test is nearly double (1.88) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 4, the relative efficiency is more than four times greater (4.16) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 5 as well with a value that is double that of the three-point version with a minimum reliability diagonal value of 0.7. The gains in relative efficiency do tend to decrease as the minimum reliability diagonal values increase. Despite this fact, the four-point I test was proven to be a more powerful test because it draws on a greater number of data points.

Tables 6, 7, and 8 illustrate the comparative power of both the three-point and four-point versions of the I test at the 0.01 alpha level, using various experimental design layouts (i.e. 2x3, 2x4, and 3x2 matrices respectively). Once again, programs were written to compute the three-point and four-point versions of the I test using a 3x3 and 3x5 matrix; however, due to limitations in the processing speed of the computer used, the programs did not resolve values for these design layouts.

The trend regarding the increased efficiency of the four-point I test versus the three-point I test is again displayed in Tables 6, 7, and 8. In Table 6, the relative efficiency of the four-point test is more than three times greater (3.02) in comparison to the three-point test with a minimum reliability diagonal value of 0.7. In Table 7, the relative efficiency is nearly seventeen times greater (16.96) with a minimum reliability diagonal value of 0.7. A higher relative efficiency was displayed in Table 8 as well with a relative efficiency nearly three and half times greater with a minimum reliability diagonal value. Once again, the difference in relative efficiency did decrease as the minimum reliability diagonal values increased; however, the fact remained that the four-point I test is more powerful than its three-point counterpart.

The I test is to be extremely conservative. As a result, although the critical values used in the analysis were mathematically correct based on elementary combinatorial

Table 1: Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.042514	0.002193	0.042045	0.007527
3x2	0.042514	0.001807	0.042045	0.006161
2x4	0.042514	0.000039	0.042045	0.000285
3x3	0.042514	0.000001	0.042045	0.000036
3x5	0.042514	0.000000	0.042045	0.000000

Note: Values obtained using 1,000,000 repetitions

Table 2: Type I Error Rate for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-point Sorted Values
2x3	0.009254	0.000106	0.009789	0.000842
3x2	0.009254	0.000081	0.009789	0.000585
2x4	0.009254	0.000001	0.009789	0.000006
3x3	0.009254	0.000000	0.009789	0.000000
3x5	0.009254	0.000000	0.009789	0.000000

Note: Values obtained using 1,000,000 repetitions

Table 3: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1430	0.3920	0.3460	1.88
≥ 0.8	0.3980	0.2305	0.5305	0.5020	1.63
≥ 0.9	0.5405	0.3600	0.6510	0.6830	1.57

Note: Values obtained using 2,000 repetitions

ROBUSTNESS AND POWER OF THE I TEST IN CONSTRUCT VALIDITY

Table 4: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.0390	0.3920	0.2090	4.16
≥ 0.8	0.3980	0.0625	0.5305	0.3365	4.03
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 5: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.05 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.3040	0.1315	0.3920	0.3490	2.06
≥ 0.8	0.3980	0.2165	0.5305	0.5120	1.77
≥ 0.9	0.5405	Did not resolve	0.6510	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 6: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x3 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0205	0.1525	0.0949	3.02
≥ 0.8	0.1410	0.0435	0.2435	0.1755	2.34
≥ 0.9	0.2280	0.0839	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 7: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 2x4 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0005	0.1525	0.0130	16.96
≥ 0.8	0.1410	0.0025	0.2435	Did not resolve	n/a
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

Table 8: Comparative Power Between the Three-point and Four-point Versions of the I Test Using a 3x2 Matrix Design Layout at the 0.01 Alpha Level

Reliability Diagonal Values	Three-point Randomized Values	Three-Point Sorted Values	Four-point Randomized Values	Four-Point Sorted Values	Relative Efficiency
≥ 0.7	0.0995	0.0160	0.1525	0.0845	3.45
≥ 0.8	0.1410	0.0299	0.2435	0.1535	2.97
≥ 0.9	0.2280	Did not resolve	0.3395	Did not resolve	n/a

Note: Values obtained using 2,000 repetitions. n/a = not applicable

analysis (i.e. 14 and 10 for the three-point I test at the 0.05 and 0.01 alpha levels respectively, and 29 and 23 for the four-point I test at the 0.05 and 0.01 alpha levels respectively), the lack of independence within each level of the I test results in a depressed false positive rate.

Ad hoc critical values were tested to determine the critical values that should be used in an applied setting to optimize the power of the test. They were obtained for both the three-point and four-point versions of the I test at both the 0.05 and 0.01 alpha for the following experimental design layouts: a) 2x3 matrix, b) 2x4 matrix, c) 3x2 matrix, d) 3x3 matrix, and e) 3x5 matrix.

The ad hoc critical values for both the three-point and four-point versions of the I at the 0.05 alpha level are presented in Table 9. It was found that the ad hoc values were quite different from those taken from the cumulative

distribution function. As an example, the optimal critical value for a 2x3 matrix at the 0.05 alpha level is 19 for the three-point I test and 35 for the four-point I test. These values are different from those taken from the suggested values of 14 and 29 respectively. The difference is greater as the matrix becomes larger. In analyzing a 3x5 matrix, it was found that the optimal critical values were 22 for the three-point I test and 41 for the four-point I test.

These findings were consistent with ad hoc critical values tested at the 0.01 alpha level. The ad hoc critical values for both the three-point and four-point versions of the I at the 0.01 alpha level are presented in Table 10. Once again, these values were quite different from those taken from the suggested values of 10 for the three-point I test and 23 for the four-point I test. Using a 2x3 matrix as an example, the optimal critical value at the 0.01 alpha level is

ROBUSTNESS AND POWER OF THE I TEST IN CONSTRUCT VALIDITY

Table 9: Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.05 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	19	0.0418	35	0.0491
3x2	21	0.0426	38	0.0389
2x4	19	0.0389	35	0.0445
3x3	21	0.0285	39	0.0387
3x5	22	0.0343	41	0.0405

Note: Values obtained using 1,000,000 repetitions

Table 10: Ad Hoc Critical Values for both the Three-Point and Four-Point I Test at the 0.01 Alpha Level

Matrix	Ad Hoc Critical Values	Three-Point Sorted Values	Ad Hoc Critical Values	Four-point Sorted Values
2x3	16	0.0080	29	0.0075
3x2	19	0.0083	35	0.0094
2x4	16	0.0069	30	0.0088
3x3	19	0.0037	36	0.0069
3x5	20	0.0033	39	0.0089

Note: Values obtained using 1,000,000 repetitions

16 for the three-point I test and 29 for the four-point I test. Once again, these differences grew larger as the matrix grew more complex.

Conclusion

According to Sawilowsky (2002), the problem with using the Jonckheere test in analyzing the Multitrait-Multimethod Matrix is the use of all of the values in the matrix increases the risk of violating the assumption of independence, and would thereby lead to inflation in the Type I error rate. By using only three data points within each level, the three-point I test was conceived as an alternative test of trend that would limit the severity of violating this assumption.

The four-point I test was found to have a higher Type I error rate that more closely matched nominal alpha. Nevertheless, the test remains quite conservative, with concomitant

depressed power that should be achievable for the stated nominal alpha level. Nevertheless, the I test is still a better alternative to evaluating the Multitrait-Multimethod Matrix than an using the guidelines established by Campbell and Fiske (1959) and alternatives such as confirmatory factor analysis which has restrictive underlying assumptions. Further developments on this approach to the analysis of construct validity is warranted, with goal of increasing its statistical power.

References

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Fiske, D. W. & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112(3), 393-395.

Hubert, L. J. & Baker, F. B. (1978). Analyzing the multitrait-multimethod matrix. *Multivariate Behavioral Research*, 13, 163-179.

Jonckheere, A. R. (1954). A distribution-free k -sample test against ordered alternatives. *Biometrika*, 41, 133-143.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36(2), 109-133.

Sawilowsky, S. S. (Personal communications, 1987).

Sawilowsky, S. S. (2002). A quick distribution-free test for trend that contributes evidence of construct validity. *Measurement and Evaluation in Counseling and Development*, 35, 78-88.

Sawilowsky, S. S. (Ed.) (2007). *Real data analysis*. Charlotte, N.C.: Information Age Publishing.

Schmitt, N. & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10(1), 1-22.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 26(2), 205-219.

Sternberg, R. J. (1992). Psychological Bulletin's top 10 "hit parade". *Psychological Bulletin*, 112(3), 387-388.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9(1), 1-26.

Which Is the Best Parametric Statistical Method For Analyzing Delphi Data?

Hiral A. Shah
St. Cloud State University

Sema. A. Kalaian
Eastern Michigan University

This study compares the three parametric statistical methods: coefficient of variation, Pearson correlation coefficient, and F-test to obtain reliability in a Delphi study that involved more than 100 participants. The results of this study indicated that coefficient of variation was the best procedure to obtain reliability in such a study.

Key words: Survey Research, Reliability, Delphi Technique, Decision-making.

Introduction

The Delphi Technique is a method for systematic solicitation and collation of judgments on a particular topic through a set of carefully designed sequential questionnaires interspersed with summarized information and feedback of opinions derived from earlier responses (Delbecq, Van de Ven, & Gustafson, 1975). The Delphi technique can be considered as an important tool to bring the knowledge and intuition of a group of qualified individuals to bear upon the future possibilities in a given field. Therefore, the technique can be used at a micro-level to arrive at a qualitative forecast which may vary from past trends in an organization.

The Delphi process consists of a series of rounds of questionnaires. The first round is characterized by exploration of the subject under discussion, wherein each individual contributes with information that he/she believes is pertinent. The second round involves the process of reaching an understanding of how the group views the issue (i.e., what group members mean by relative terms such as importance, desirability

or feasibility). The Delphi rounds of questionnaires should continue until a predetermined level of consensus is reached or no new information is gained (Ludwig, 1997; Linstone & Turoff, 1975; Delbec, Van de Ven & Gustafson, 1975). In most instances it is found that three iterations are sufficient, and not enough new information is gained to warrant the cost of more iterations (Ludwig, 1997).

Parametric statistical methods such as the coefficient of variation (CV) and the F-test have been used in Delphi studies with sample size below 50. The CV is a statistical measure of the deviation of a variable from its mean. The F-test is performed to determine the ratio of squares of two variances or, in other words, to test if the standard deviations of two populations are equal.

English and Kernan (1976, cited in Yang 2003) used the coefficient of variation (CV) to determine the stopping rule. If the magnitude of CV for an item was found to be too large (e.g., greater than 0.8), the corresponding statement was needed to be modified and required an additional round(s) of questionnaire administration.

Yang (2003) suggested using the F-test to compare two variances. The F-value is determined by the ratio of the variances of item scores among panelists between the two successive rounds. If no significant difference is found in the F-test, the questionnaire item will be dropped from further rounds. Questionnaire items where significant between-round difference in variances is found are retained in a

Hiral A. Shah is an Assistant Professor in Engineering Management. Email: hashah@stcloudstate.edu. Sema. A. Kalaian is a Professor in Technology Studies. Email: skalaian@emich.edu.

subsequent round. Yang (2003) described this method as being suggested by Jolson and Rossow (1971) with the problem being that assumptions made for the F-test may be seriously violated when using data collected from the Delphi rounds.

This study compared the three parametric statistical methods: coefficient of variation (CV), Pearson correlation coefficient, and F-test to obtain consensus and reliability in a Delphi study using data from Shah (2004) and Shah and Kalaian (2006) to find out the best method that fits the study that involves a large number of participants.

The data used in this study was obtained from research conducted by Shah (2004). The purpose of this study was to gather data using Delphi technique to discover and describe what experts in the field consider important to know in the discipline of engineering management, and use that information to update the curriculum for Eastern Michigan University's Engineering Management masters program. The Delphi panel in the study consisted of 194 panelists. These panelists were asked to rate the competency areas on a 5-point Likert-type scale and provide qualitative comments through mailed questionnaires. The following criterion of importance was assigned to the responses provided on the questionnaire, along with an example of how to respond: 5 = of very high importance, 4 = of high importance, 3 = of medium importance, 2 = of low importance, 1 = of very low importance.

This study consisted of three rounds of questionnaires. The sample comprised of individuals who belonged to any of these four categories: (a) Professor/instructor of Engineering Management, (b) Industry Professional, (c) Author of published text/papers/articles related to the breadth of Engineering Management discipline, and (d) Certified Engineering Manager/Certified Enterprise Integrator. Moreover, the competency areas were also grouped into four categories, namely: (a) Technical, (b) R&D/Design, (c) System/Organization/Project Management, and (d) Human Issues.

Methodology

The Round 1 Delphi questionnaire was sent to 707 subjects. Based on the information obtained from Round 1, Delphi panel members were selected and the Round 2 Delphi questionnaire was developed. In the second round, an analysis of the group's modal response and percentage concurrence for each degree of importance from the first round was provided to the Delphi panel for reference. Specific comments to a particular competency provided by the Round 1 subjects were reported in the Round 2 Delphi questionnaire. A space for comments was provided after every competency area for the respondents to respond to the comments made by other panel members from Round 1 or to give their own comment. Additional competency areas suggested by the Round 1 respondents were added to the existing list of competencies.

Panel members were asked to consider respondent comments and the percents of concurrence obtained from Round 1, rate each competency area on a five-point Likert-type scale, and explain their choice if it was two or more categories away from the Round 1 respondent's modal rating. An example of how and where to record their responses and comments was also provided. Additional comments made by Round 1 respondents were reported in Round 2 questionnaire for their reference. Space for additional or general comments was provided at the end of questionnaire.

The Round 3 Delphi questionnaire was developed using Round 2 results and was administered in the same manner as Round 2. Based on the category in which the Delphi panel members categorized themselves, a six-digit (rCodexxx) alphanumeric code was assigned to each of them. The first digit - r - represented the Delphi round (2 or 3) to which they responded; the code represented the category to which they belonged to in the form of letters A-for authors, C-for Certified Engineering Manager/Certified Enterprise Integrator, I-Industry professionals, and P-Professors teaching Engineering Management; and xxx represented the panel member's assigned number. Round 3 was also sent to the individuals who participated in Round 1 but did not participate in Round 2.

BEST PARAMETRIC METHOD FOR ANALYZING DELPHI DATA

Because the codes could not be assigned to the panelists who did not participate in Round 2, an additional sheet was sent to these panelists asking them to participate in the final Round 3 and also to checkmark the category to which they belonged.

The Round 3 Delphi questionnaire that was sent to the panel members who participated in Round 2 had individual codes. Moreover, additional questions were asked on the front page of the questionnaire asking the Delphi panel members to: Rate the overall importance of the results of this to the discipline of Engineering Management as a guide for others for curriculum development, rate the overall quality of this study, rate their own level of expertise in the field of Engineering Management, and additional space was provided to comment on the importance/quality of this study and suggestions for possible improvements. Table 1 shows the participation in the study and the response rates at the end of each round of Delphi study.

Table 1: Response Rates from Three Rounds of Delphi Study

Delphi Round	Number Sent	Number Received	Response Rate (%)
1	707	194	27.4%
2	194	148	76.3%
3	194	136	70.1%

Data Analysis using Parametric Statistical Methods

The data was entered for each of the rounds using SPSS software. Due to missing values for one or more competency areas in several cases, those cases were excluded from the study. Thus, the sample size for this study was 52. The mean and standard deviation corresponding to each of the competency areas in Rounds 1, 2 and 3 were calculated using SPSS and Microsoft Excel software. Because coding for each panel member was applied from Round 2, the data obtained from Rounds 2 and 3 of the Delphi study could be corresponded case-wise. Hence, for this study, Rounds 2 and 3 will be considered for analysis purposes.

Coefficient of Variation

The Coefficient of Variation (CV), which is the ratio of standard deviation (σ) of a competency area to its corresponding mean (μ) among the panelists, was calculated using the formula:

$$CV = \sigma / \mu. \quad (1)$$

The CV was obtained for Rounds 2 and 3, and in order to determine if additional rounds were required, the absolute difference was calculated by subtracting the CV obtained from Round 3 from that obtained from Round 2. A small CV value was an indication that the data scatter or variation compared to the mean was small. A large CV value compared to the mean was an indication that the amount of variation was large.

As shown in Table 2, the absolute value of the difference in CV between Rounds 2 and 3 was less than 0.2, which can be considered to be a minor difference according to Dajani (1979, cited in Yang, 2003). Though negative values of difference was obtained for competency areas such as: Information systems, Linear programming, Materials engineering, Metrology-Measurement Science, Six sigma black belt certification and others, the absolute difference was still less than 0.2. Hence, it can be assumed that stability was reached for each of the competency areas and no further rounds of Delphi were required.

F-test to compare Two Variances

The F value for each competency area was obtained by calculating the ratio of the variances (σ^2) of item scores among panelists between Rounds 2 and 3. Hence,

$$F\text{-Ratio} = \frac{\sigma^2 \text{Round } 3}{\sigma^2 \text{Round } 2} \quad (2)$$

It is important to note that the degrees of freedom have not been taken into consideration in the F-test as they are already a part of variances. When no significant difference in the F-test is obtained, the questionnaire item will be dropped from further rounds.

The F-ratio of 1 implies that the variance of Round 3 is equal to the variance of

Table 2: Results of the Three Parametric Procedures from Round 2 to Round 3

Statistic	Absolute difference in $CV = CV(R2) - CV(R3)$	F-ratio = $Var(R3/R2)$	Pearson's r
Mean	0.025	0.789	0.397
Median	0.025	0.746	0.416
Minimum Value	0.070	0.000	-0.240
Maximum Value	0.130	2.070	0.730
% Reliability Obtained	100%	79%	83%
Skewness Value Using Z scores	0.080	0.093	-0.429

Note: R3=Round 3, R2=Round 2

Round 2. Hence, a F-ratio less than or equal to 1 is desirable. The results from the F-test suggested that 79% of the competency areas had F-ratios less than or equal to 1 (see Table 2), indicating that stability was established in Round 3.

Pearson's Product-Moment Correlation

Correlation is a technique used to determine the relationship between two quantitative, continuous variables. A correlation is often called a bivariate correlation to designate a simple correlation between two variables, as opposed to relationships between more than two variables (George & Mallery, 2005). A correlation, also known as Pearson's Product-Moment Coefficient of Correlation, or the Pearson r , is one such measure of the strength of the association between two variables. George and Mallery (2005) stated, "although the Pearson r is predicted on the assumption that the two variables involved are approximately normally distributed, the formula often performs well even when assumptions of normality are violated or when one of the variables is discrete" (p. 124). A correlation value of +1.00 indicates a perfect, positive correlation, whereas, a correlation of zero

indicates no relationship between the two variables. A negative correlation indicates a relation in which one variable tends to increase as the other variable tends to decrease. The closer a correlation coefficient is to zero, the weaker the relationship between the two variables.

The correlation value, r , was obtained for each competency area using SPSS software. If the correlation coefficient for a particular competency area varied significantly from zero and was very high, it indicated that the ratings of panel members on the competency area were stable and less fluctuating.

The Pearson's correlation coefficient obtained, indicated that there was a negative relationship for competency areas: Management of technology, Communications, Customer issues, and People and teamwork. Values of these coefficients were closer to zero, indicating a weaker tendency of increase in value of one competency with the decrease of value in the subsequent round. Thus, the panel members who responded lower in Round 2 for these competency areas, tended to respond higher in Round 3. The relationship was found to be weak and hence it was an indication that stability was obtained in Round 3. The results from Pearson's correlation indicated that 83% of the

BEST PARAMETRIC METHOD FOR ANALYZING DELPHI DATA

competency areas had correlation values, which were either greater than or equal to zero (Table 2). Thus, it can be implied that there was a good correlation between the competency areas in Round 2 and Round 3.

Results

As the results of all the three parametric procedures used to obtain reliability in the Delphi study indicated similar results, it was important to determine the best procedure among the three. Hence, further analysis was performed on the results of the three parametric procedures: CV, F-ratio, and Pearson's r . Because the values of the three procedures were on a different scale, transformation of the values to similar scales for all the three procedures was completed using z scores (a measure of the distance in standard deviations of a sample from the mean). The z transformation is calculated as $(X - \mu)/\sigma$; where X is the observation, μ is the mean and σ is the standard deviation of the observations. A positive z score indicates that an

observation is greater than the mean whereas a negative z score indicates that an observation is below the mean.

A box plot comparing the z scores of the three parametric procedures for the 76 competency areas contained three outliers: case numbers 32, 38 and 69. As the outliers tended to skew the normal distribution, these cases were deleted and a box plot was derived. Figure 1 shows the box plots comparing the three parametric procedures without outliers and Figure 2 shows the histogram obtained from the data.

Because skewness is a measure of symmetry of the distribution, a positive value shows the distribution is positively skewed and a negative value shows that the data is negatively skewed. A comparative look at the values of skewness for all the three parametric procedures as shown in Figure 2 and Table 2 was the procedure to determine the best parametric procedure. Coefficient of variation had a smaller positive value of skewness (0.080) compared to Pearson r (-0.429), and F-ratio (0.093).

Figure 1: Box Plot Comparing the Z-Scores of the Three Parametric Procedures: Coefficient of Variation (CV), F-test, and Pearson's r

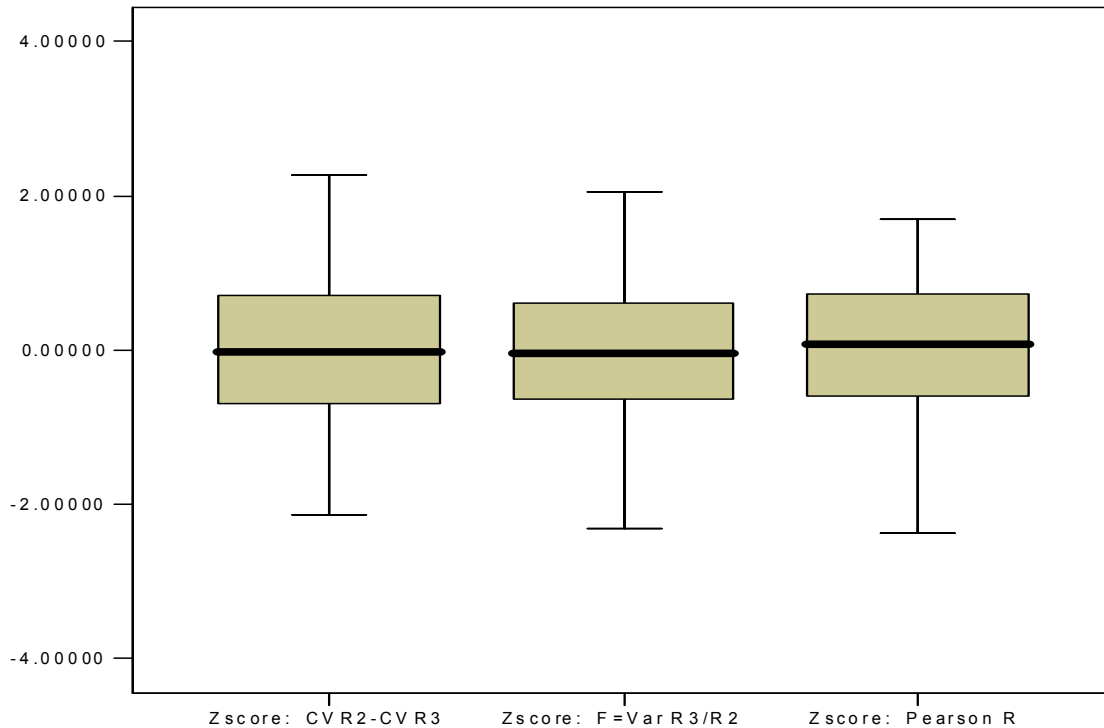
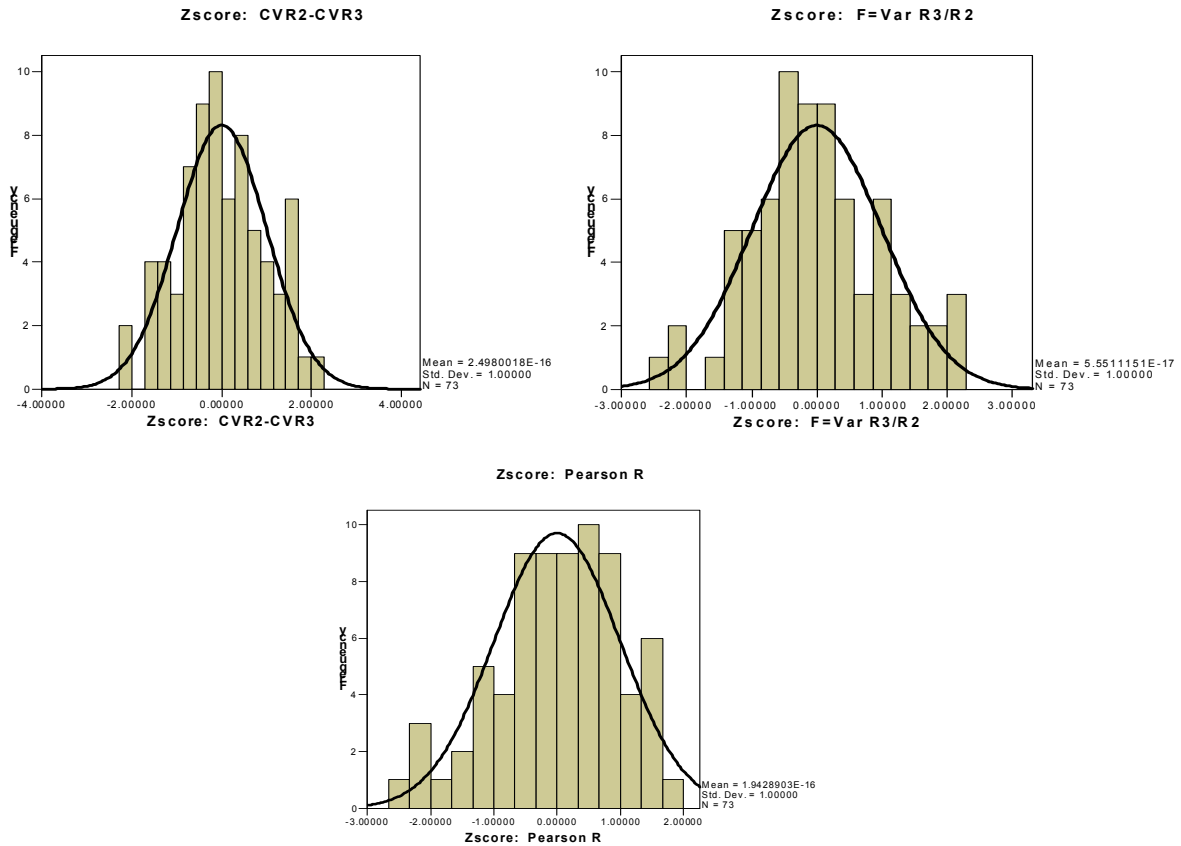


Figure 2: Histogram Obtained From the Z-Scores of the Three Parametric Procedures: Coefficient of Variation (CV), F-test, and Pearson's r



Conclusion

In summary, the results of the three parametric procedures indicated: 100% of the competency areas in Round 3 obtained stability and hence reliability was achieved by the coefficient of variation method; 79% of the competency areas had F-ratios less than or equal to 1, which indicated that stability has been established in Round 3; and 83% of the competency areas had Pearson r correlation values either greater than or equal to zero, depicting a good correlation between Round 2 and Round 3. As all the three parametric procedures were a good indication of obtaining reliability in a Delphi study a z scores were calculated and box plot was graphed.

The values of the skewness obtained from the descriptive values of the box plots, it was found that the coefficient of variation (CV) had a smaller positive value of skewness (0.080)

compared to Pearson's r (-0.429), and F-ratio (0.093). From these values, it could be concluded that the coefficient of variation was the best procedure to obtain reliability in a Delphi study that included more than 100 participants. The second best procedure to obtain reliability in a Delphi study is F-ratio and the third one is Pearson's r. As the literature related to Delphi procedure describes, it can be further confirmed that stability is obtained at the third round of Delphi and hence, three rounds of questionnaire are enough in a Delphi study.

References

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Glenview, IL: Scott Foresman and Company.

BEST PARAMETRIC METHOD FOR ANALYZING DELPHI DATA

Dunham, R. B. (1996). *The Delphi technique*. Retrieved September 20, 2004 from http://www.slais.ubc.ca/resources/research_methods/group.htm#delphi

George, D. & Mallery, P. (2005). *SPSS for Windows step by step: A simple guide and reference, 12.0 update*. Boston, MA: Pearson Education.

Jones, C. M. (1994). *The component skills of workplace literacy and the utilization of computer assisted instruction to achieve it*. Unpublished doctoral dissertation, Kent State University.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Boston, MA: Addison-Wesley.

Ludwig, B. (1997). Predicting the future: Have you considered using the Delphi methodology? *Journal of Extension*, 35(5). <http://www.joe.org/joe/1997october/tt2.html>.

Riggs, W. E. (1983). *The Delphi technique: An experimental evaluation. Technological Forecasting and Social Change*, (23), 89-94.

Rowe, S. E. (2001). *Development of a test blueprint for the National Association of Industrial Technology certification exam*. Doctoral dissertation, Iowa State University, Ames, IA.

Shah, H. A. (2004). *A Delphi study to develop engineering management curriculum at Eastern Michigan University*. Unpublished Master's Thesis, Eastern Michigan University, Ypsilanti, MI.

Tillman, T. (1989). *A Delphi study to identify fundamental competency areas for Certification Testing of Manufacturing Technologists and entry-level manufacturing engineers*. Unpublished Doctoral thesis, Purdue University.

Yang, Y. N. (2003). *Convergence on the guidelines for designing a web-based art-teacher education curriculum: A Delphi study*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

EXPONENTIAL DISTRIBUTION ESTIMATION OF THE RELIABILITY FUNCTION

The First Stage Estimation

The two-stage shrinkage estimation procedure for $R(t) = \exp(-t/\beta)$ is as follows:

- 1) Select a sample of size n_1 on T .
Let T_{1i} , $i = 1, 2, \dots, n_1$ be the first sample.
Let $T_1 = \sum T_{1i}$.
Then, $\bar{T}_1 = T_1 / n_1$ is the mean of the first stage sample.

- 2) Test the prior knowledge about $\beta = \beta_0$, i.e. test $H_0 : \beta = \beta_0$ versus $H_a : \beta \neq \beta_0$ at level α .

- 3) The rejection region is given by $T_1 \leq a_1$ or $T_1 \geq a_2$, where a_1 and a_2 are given by:

$$\Gamma_1(n_1, a_2 / \beta_0) - \Gamma_1(n_1, a_1 / \beta_0) = 1 - \alpha, \text{ and}$$

$$\Gamma_1(n_1 + 1, a_2 / \beta_0) - \Gamma_1(n_1 + 1, a_1 / \beta_0) = 1 - \alpha$$

where $\Gamma_1(\cdot)$ denotes the incomplete gamma function.

[See Bain (1991) for details.]

- 4) If $H_0 : \beta = \beta_0$ is not rejected, then the shrinkage estimator of reliability is:

$$\hat{R}(t) = k \exp(-t/\bar{T}_1) + (1-k) \exp(-t/\beta_0) \quad (3)$$

where the shrinkage factor k , $0 < k < 1$, is given by

$$k = |\bar{T}_1 - \beta_0| / [(a_2 - a_1) / n_1]. \quad (4)$$

The Second Stage Estimation

If H_0 is rejected then select the second sample of size

$$n_2, T_{2i}, i = 1, 2, \dots, n_2.$$

Let \bar{T}_2 be the sample mean of the second sample.

Calculate:

$$\bar{T} = \frac{n_1 \bar{T}_1 + n_2 \bar{T}_2}{n_1 + n_2}, \quad (5)$$

and define the estimator of the reliability as

$$\hat{R}(t) = \exp(-t/\bar{T}).$$

Thus, the two-stage shrinkage estimator of the reliability function denoted by is given by $\hat{R}_s(t)$:

$$\hat{R}_s(t) = k \exp(-t/\bar{T}_1) + (1-k) \exp(-t/\beta_0)$$

if H_0 is not rejected,

and

$$\hat{R}_s(t) = \exp(-t/\bar{T})$$

if H_0 is rejected. (6)

Bootstrapping the Shrinkage Factor k and Related Two-stage Estimator of Reliability

The shrinkage estimators and the choice of shrinkage factor have been studied for over the last five decades for various applications. In what follows, the use of bootstrap technique for selecting a shrinkage factor k in the above estimator (6) is investigated.

First, note that the efficiency of (6) is a function of k defined above in (4). Further, for given α , the factor k is a function of \bar{T}_1 , the mean of the first-stage sample, and hence is a random variable. Therefore, the bootstrapping for the first-stage sample T_{1i} , $i = 1, 2, \dots, n_1$ and the corresponding k is considered as follows.

Generating a Set of k 's Using Bootstrap Method

First, proceed as in the steps (1)-(4) described above in the methods section. If H_0 is not rejected, then the bootstrap method is used as follows for the observed data on T .

1. Generate a bootstrap sample T_{1i}^* , $i = 1, 2, \dots, n_1$, from the first stage sample T_{1i} , $i = 1, 2, \dots, n_1$. (The * denotes the bootstrapping sample operation). Let \bar{T}_1^* denote the bootstrap mean and let k^* denote the corresponding shrinkage factor. Thus, $k^* = |\bar{T}_1^* - \beta_0| / [(a_2 - a_1) / n_1]$ with the property $0 < k^* < 1$.

2. Repeat the bootstrap procedure and calculate k^* until a set of predetermined B values of k^* (where $0 < k^* < 1$) is generated.
3. Several ways of using this sequence of k^* values are available for defining the shrinkage factor. Here, the mean of B values of k^* is selected. Let \bar{k}^* denote this mean. Now, the two-stage bootstrap shrinkage estimator of the reliability function, denoted by $\hat{R}_b(t)$, is defined as,

$$\begin{aligned} \hat{R}_b(t) &= \bar{k}^* \exp(-t / \bar{T}_1) + (1 - \bar{k}^*) \exp(-t / \beta_0) \\ &\text{if } H_0 \text{ is not rejected, and} \\ \hat{R}_b(t) &= \exp(-t / \bar{T}) \\ &\text{if } H_0 \text{ is rejected.} \end{aligned} \tag{7}$$

Since, the derivations for the mean and the mean squared errors for $\hat{R}_s(t)$ and $\hat{R}_b(t)$ are not straightforward the values of $(\hat{R}_s(t), \hat{R}_b(t))$ were simulated for the comparison of bias and the MSE's of these estimators.

Results

Fifteen thousand repetitions were carried out for different combinations of the parameter $\beta = 1$ and specified values of $(R(t), \beta_0, \alpha, n_1, n_2)$. For each repetition, B = 100 bootstrap samples were selected from the first stage sample. The simulation results are shown in Table 1.

Conclusion

The simulation results show that the estimator of the reliability function based on the mean of the values of the shrinkage factor obtained using the bootstrap procedure is more efficient as compared to the one without such bootstrapping. The same conclusion for the other values of the parameters and the sample sizes n_1, n_2 is applicable. For brevity, the other simulations results are not included here.

This article demonstrates the use of a bootstrap method for generating a set of

shrinkage factors. Using this, a final shrinkage factor can be defined based on these bootstrapped shrinkage factors as appropriate for a given problem. In the above discussion the mean of the set of shrinkage factor values is used, however, other possible selections are the median, the maximum or the minimum of the set of k^* 's. In fact, the bootstrapping of the shrinkage factor can be used in many other shrinkage estimator settings where such factor is a function of a sample statistic.

References

- Adke, S. R., Waikar, V. B., & Schuurmann, F. J. (1987). A two-stage shrinkage estimator for the mean of an exponential distribution. *Commun. Statist.- Theory and Methods*, 16, 1821-1834.
- Bain, L. J., & Engelhardt, M. (1991). *Statistical analysis of reliability and life-testing models- theory and methods*, (2nd Ed.). NY: Marcel Dekker, Inc.
- Thompson. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63(321), 113-122.
- Tse, S., & Tso, G. (1996). Shrinkage estimation of reliability for exponentially distributed life times. *Commun. Statist.- Simulations*, 25, 415-430.
- Waikar, V. B., Schuurmann, F. J., & Raghunathan, T. E. (1984). On a two-stage shrinkage estimator of the mean of a normal distribution. *Commun. Statist.- Theory and Methods*, 13, 1901-1913.

EXPONENTIAL DISTRIBUTION ESTIMATION OF THE RELIABILITY FUNCTION

Table 1: The Bias and Mean Squared Error for Estimators $\hat{R}_s(t)$ and $\hat{R}_b(t)$
 ($\beta=1.00, \beta_0=1.00, \alpha=0.05$)

$R(t) = 0.9, n_1 = 10, n_2 = 10$			
	$\hat{R}_s(t)$	$\hat{R}_b(t)$	$\hat{R}_s(t) / \hat{R}_b(t)$
Bias	-0.00229	-0.00213	-
MSE	0.00019	0.00016	1.19
$R(t) = 0.8, n_1 = 10, n_2 = 10$			
	$\hat{R}_s(t)$	$\hat{R}_b(t)$	$\hat{R}_s(t) / \hat{R}_b(t)$
Bias	-0.00387	-0.00374	-
MSE	0.00065	0.00055	1.20
$R(t) = 0.9, n_1 = 10, n_2 = 15$			
	$\hat{R}_s(t)$	$\hat{R}_b(t)$	$\hat{R}_s(t) / \hat{R}_b(t)$
Bias	-0.00189	-0.00182	-
MSE	0.00015	0.00012	1.20

A New Approximate Bayesian Approach for Decision Making About the Variance of a Gaussian Distribution Versus the Classical Approach

Vincent A. R. Camara
University of South Florida

Rules of decision-making about the variance of a Gaussian distribution are obtained and compared. Considering the square error loss function, an approximate Bayesian decision rule for the variance of a normal population is derived. Using normal data and SAS software, the obtained approximate Bayesian test results were compared to their counterparts obtained with the well-known classical decision rule. It is shown that the proposed approximate Bayesian decision rule relies only on observations. The classical decision rule, which uses the Chi-square statistic, does not always yield the best results: the proposed approach often performs better.

Key words: Hypothesis testing, loss function, Type II error, statistical analysis.

Introduction

Life testing in reliability has received a substantial amount of interest from theorists as well as reliability engineers. Their concern was a product of the increased complexity and sophistication in electronic and structural systems, which came into existence very rapidly during this time. In the early 1950's, Epstein and Sobel began to explore the field of parametric life testing. Under the assumption of an exponential time-to-failure distribution, they produced a series of papers (1953, 1954, 1955) which were to influence future work in reliability and life parameter testing.

Shortly thereafter other failure distributions more complex than the exponential were used as failure models. For example, Kao (1956) brought attention to the Weibull probability distribution, while Birnbaum and Saunders (1958) suggested the gamma distribution. In this study, the normal probability

distribution - which has been and is still widely used in industry and in academia - is considered. The normal distribution is defined as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad (1)$$
$$-\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

A test of hypothesis consists in testing a given theory or belief about a population parameter based on some sample information. Once the underlying model is found to be normal or approximately normal, the classical approach considers the following decision rule for a level of significance of alpha and a sample of size n (Mario F. Triola, 2007):

Two-Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 \neq c$$

Non-rejection region:

$$(\chi_{n-1,1-\alpha/2}^2, \chi_{n-1,\alpha/2}^2)$$

Rejection region:

$$(-\infty, \chi_{n-1,1-\alpha/2}^2] \cup [\chi_{n-1,\alpha/2}^2, \infty)$$

Vincent A. R. Camara earned a Ph.D. in Mathematics/Statistics. His research interests include the theory and applications of Bayesian and empirical Bayes analyses with emphasis on the computational aspect of modeling. E-mail: gvcamara@ij.net.

Right Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 \succ c$$

Non-rejection region:

$$(-\infty, \chi^2_{n-1, \alpha})$$

Rejection region:

$$[\chi^2_{n-1, \alpha}, \infty)$$

Left Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 \prec c$$

Non-rejection region:

$$(\chi^2_{n-1, \alpha}, \infty)$$

Rejection region:

$$(-\infty, \chi^2_{n-1, \alpha}]$$

The Chi-square test statistic that is used to conduct the above tests will be denoted by Chi, with:

$$Chi = \frac{(n-1)s^2}{\sigma^2}.$$

Methodology

Although no specific analytical procedure exists that allows identification of the appropriate loss function to be used in Bayesian analysis, the most commonly used is the square error loss function. One of the reasons for selecting this loss function is due to its analytical tractability in Bayesian analysis. The square error loss function places a small weight on estimates near the true value and proportionately more weight on extreme deviation from the true value of the parameter. The square error loss is defined

$$L_{SE}(\hat{\theta}, \theta) = \left(\hat{\theta} - \theta \right)^2 \quad (2)$$

The use of the square error loss function along with a suitable approximation of the Pareto prior leads to the following approximate Bayesian confidence bounds for the normal population variance (Camara, 2003):

$$L_{\sigma^2(SE)} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n-2-2\ln(\alpha/2)}$$

$$U_{\sigma^2(SE)} = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n-2-2\ln(1-\alpha/2)}$$

or

$$L_{\sigma^2(SE)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-2-2\ln(\alpha/2)}$$

$$U_{\sigma^2(SE)} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-2-2\ln(1-\alpha/2)} \quad (3)$$

To obtain the approximate Bayesian decision rule for the variance of a normal population, the close relationship that exists between confidence intervals and hypothesis testing is used. Considering the above mentioned approximate Bayesian confidence intervals along with the test statistic Chi, the following approximate Bayesian decision rule is derived:

Two-Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 \neq c$$

Non-rejection region:

$$(n-2-2\ln(1-\alpha/2), n-2-2\ln(\alpha/2))$$

Rejection region:

$$(-\infty, n-2-2\ln(1-\alpha/2)] \cup [n-2-2\ln(\alpha/2), \infty)$$

Right Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 > c$$

Non-rejection region:

$$(-\infty, n - 2 - 2 \ln(\alpha))$$

Rejection region:

$$[n - 2 - 2 \ln(\alpha), \infty)$$

Left Tailed Test

Hypotheses:

$$H_0 : \sigma^2 = c$$

$$H_a : \sigma^2 < c$$

Non-rejection region:

$$(n - 2 - 2 \ln(\alpha), \infty)$$

Rejection region:

$$(-\infty, n - 2 - 2 \ln(\alpha)]$$

To compare the classical and approximate Bayesian decision rules and evaluate their performances, the absolute difference, AD, between the parameter and the claim is used and is defined by:

$$AD = |Parameter - Claim|$$

From the calculated results of the absolute difference between the parameter and the claim, the following are able to be concluded:

- For a reasonably large value of AD, the test that will perform better than its counterpart will be the one that will reject the null hypothesis.
- For a reasonably small value of AD, the test that will perform better than its counterpart will be the one that will fail to reject the null hypothesis.
- A test and its counterpart will perform equally well, if both reject the null hypothesis for a reasonably large value of AD or both fail to reject the null

hypothesis for a reasonably small value of AD.

- A test and its counterpart will perform poorly if, for a reasonably large value of AD, both fail to reject the null hypothesis, or both reject the null hypothesis for a reasonably small value of AD.

Results

In order to compare the proposed approximate Bayesian decision rule with the classical approach, samples obtained from normally distributed populations (e.g., 1, 2, 3, .4, 7) as well as approximately normal populations (e.g., 5, 6) are considered. SAS software was used to obtain the normal population parameters corresponding to each sample data set.

The observed value, which is the value of the test statistic Chi under the assumption that the null hypothesis is true, will be denoted by Chio. If this observed value, Chio, falls into the rejection region, the null hypothesis will be rejected at a level of significance selected beforehand. If the observed value falls into the non-rejection region, the null hypothesis will not be rejected at the selected level of significance

Data Set #1:

24, 28, 22, 25, 24, 22, 29, 26, 25, 28, 19, 29 (Mann, 1998, p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 25.083, \sigma = 3.1176).$$

The population and sample variances are: $\sigma^2 = 9.71943$, and $s^2 = 9.719696$. For the following test of hypothesis,

$$H_0 : \sigma^2 = c,$$

$$H_a : \sigma^2 \neq c,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 1. Table 1 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the first data set.

BAYESIAN APPROXIMATION FOR THE VARIANCE OF A GAUSSIAN DISTRIBUTION

Table1: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	5.578 – 17.275	10.211 – 14.605
90	4.575 – 19.675	10.101 – 15.991
95	3.8159 – 21.92	10.051 – 17.378
99	2.603 – 26.757	10.010 – 20.597

Test of Hypothesis #1:

$$H_0 : \sigma^2 = 9.71943 ,$$

$$H_a : \sigma^2 \neq 9.71943 ,$$

AD = 0.

The observed value is Chio = 11.0003. Therefore, both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2. These are good decisions since the normal population variance under study is equal to 9.71943

Test of Hypothesis #2:

$$H_0 : \sigma^2 = 8 ,$$

$$H_a : \sigma^2 \neq 8 ,$$

AD = 1.71943.

The observed value is Chio = 13.364582. Therefore, both the classical and our proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #3:

$$H_0 : \sigma^2 = 4 ,$$

$$H_a : \sigma^2 \neq 4 ,$$

AD=5.71943.

Considering the observed value Chio = 26.729164, the classical approach fails to reject the null hypothesis at a level of significance equal to 0,01, while the approximate Bayesian approach rejects the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #4:

$$H_0 : \sigma^2 = 20 ,$$

$$H_a : \sigma^2 \neq 20 ,$$

AD = 10.28057.

In this case, considering the observed value Chio = 5.345832, the classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1, while the approximate Bayesian approach reject the null hypothesis at any level of significance smaller or equal to 0.2

Test of Hypothesis #5:

$$H_0 : \sigma^2 \geq 23 ,$$

$$H_a : \sigma^2 < 23 ,$$

AD greater or equal to 13.28057.

Considering the observed value Chio = 4.64855, the classical approach fails to reject the null hypothesis at a level of significance smaller or equal to equal to 0.05. The approximate Bayesian approach rejects the null hypothesis at any level of significance smaller or equal to 0.2.

Data Set #2:

13, 11, 9, 12, 8, 10, 5, 10, 9, 12, 13 (Mann, 1998 p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 10.182, \sigma = 2.4008) .$$

The population and sample variances are $\sigma^2 = 5.76384$, and $s^2 = 5.763636$. For the following two tailed test of hypothesis:

$$H_0 : \sigma^2 = c ,$$

$$H_a : \sigma^2 \neq c ,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 2. Table 2 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the second data set.

Table 2: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	4.865 – 15.987	9.211 – 13.605
90	3.94 – 18.307	9.102 – 14.991
95	3.247 – 20.483	9.051 – 16.378
99	2.156 – 25.188	9.010 – 19.597

Test of Hypothesis #6:

$$H_0 : \sigma^2 = 5.76384,$$

$$H_a : \sigma^2 \neq 5.76384,$$

AD = 0.

The observed value is Chio = 9.999645. Considering Table 2, it is observed that both, the classical and the approximate Bayesian approaches, fail to reject the null hypothesis at any levels of significance smaller or equal to 0.2.

Test of Hypothesis #7:

$$H_0 : \sigma^2 = 4.5,$$

$$H_a : \sigma^2 \neq 4.5,$$

AD = 1.26384.

The observed value is Chio = 12.80808. Therefore both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #8:

$$H_0 : \sigma^2 = 10,$$

$$H_a : \sigma^2 \neq 10,$$

AD = 4.23616.

In this case, Chio = 5.763636. Contrary to the classical approach, the proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2.

Test of Hypothesis #9:

$$H_0 : \sigma^2 = 15,$$

$$H_a : \sigma^2 \neq 15,$$

AD = 9.23616.

In this case, Chio = 3.8424. The proposed approach rejects the null hypothesis at any level of significance smaller than or equal to 0.2, while the classical approach fails to reject the same null hypothesis only at significance levels smaller or equal to 0.05.

Test of Hypothesis #10:

$$H_0 : \sigma^2 \geq 14,$$

$$H_a : \sigma^2 < 14,$$

AD is greater or equal to 8.23616.

Here the Chio = 4.11688. The proposed approach rejects the null hypothesis at levels of significance smaller or equal to 0.2. The classical approach fails to reject the null hypothesis at a level of significance of 0.05.

Data Set #3:

16, 14, 11, 19, 14, 17, 13, 16, 17, 18, 19, 12 (Mann, 1998 p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 15.5, \sigma = 2.6799).$$

The population and sample variances are $\sigma^2 = 7.18186$, and $s^2 = 7.181818$. For the following test of hypothesis:

$$H_0 : \sigma^2 = c,$$

$$H_a : \sigma^2 \neq c,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 3. Table 3 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the third data set.

Test of Hypothesis #11:

$$H_0 : \sigma^2 = 7.18186,$$

$$H_a : \sigma^2 \neq 7.18186,$$

AD = 0, Chio=10.999935.

BAYESIAN APPROXIMATION FOR THE VARIANCE OF A GAUSSIAN DISTRIBUTION

Both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Table 3: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	5.578 – 17.275	10.211 – 14.605
90	4.575 – 19.675	10.103 – 15.991
95	3.8159 – 21.92	10.051 – 17.378
99	2.603 – 26.757	10.010 – 20.597

Test of Hypothesis #12:

$$H_0 : \sigma^2 = 6,$$

$$H_a : \sigma^2 \neq 6,$$

AD = 1.18186.

The observed value is $\text{Chio} = 13.166666$. Therefore both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #13:

$$H_0 : \sigma^2 = 14,$$

$$H_a : \sigma^2 \neq 14,$$

AD = 6.81814, $\text{Chio} = 5.64285$.

Contrary the classical approach, the proposed approximate Bayesian approach rejects the null hypothesis at levels of significance respectively small or equal to 0.2.

Test of Hypothesis #14:

$$H_0 : \sigma^2 = 18,$$

$$H_a : \sigma^2 \neq 18,$$

AD=10.81814, $\text{Chio}=4.388888$.

The proposed approximate Bayesian approach rejects the null hypothesis at any significance level smaller or equal to 0.2. The classical approach fails to reject the null hypothesis at levels of significance respectively smaller or equal to 0.05.

Test of Hypothesis #15:

$$H_0 : \sigma^2 \geq 17,$$

$$H_a : \sigma^2 < 17,$$

AD is greater or equal to 9.81814.

The observed value $\text{Chio} = 4.647058$. Based on Table 3, the proposed decision rule rejects the null hypothesis at any level of significance smaller or equal to 0.1. The classical approach fails to reject the null hypothesis at levels of significance smaller or equal 0.05.

Data Set #4:

27, 31, 25, 33, 21, 35, 30, 26, 25, 31, 33, 30, 28 (Mann, 1998 p. 504).

Normal population distribution obtained with SAS:

$$N(\mu = 28.846, \sigma = 3.9549)$$

The population and sample variances are $\sigma^2 = 15.64123$, and $s^2 = 15.641025$. For the following test of hypothesis:

$$H_0 : \sigma^2 = c,$$

$$H_a : \sigma^2 \neq c,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 4. Table 4 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the fourth data set.

Table 4: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	6.304 – 18.549	11.211 – 15.605
90	5.226 – 21.026	11.103 – 16.991
95	4.404 - 23.337	11.051 – 18.378
99	3.074 – 28.300	11.010 – 21.597

Test of Hypothesis #16:

$$H_0 : \sigma^2 = 15.64123,$$

$$H_a : \sigma^2 \neq 15.64123,$$

AD = 0, Chio=11.999842.

Both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #17:

$$H_0 : \sigma^2 = 16.5,$$

$$H_a : \sigma^2 \neq 16.5,$$

AD = 0.85877.

The observed value is Chio = 11.3752909. Therefore both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #18:

$$H_0 : \sigma^2 = 30,$$

$$H_a : \sigma^2 \neq 30,$$

AD = 14.35877, Chio = 6.2564

The classical approach fails to reject the null hypothesis at a level of significance smaller or equal to 0.1. The proposed decision rule rejects the null hypothesis for any level of significance smaller or equal to 0.2.

Test of Hypothesis #19:

$$H_0 : \sigma^2 = 8,$$

$$H_a : \sigma^2 \neq 8,$$

AD = 7.64123, Chio = 23.461536.

The proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2. The classical approach fails to reject the null hypothesis at a level of significance of 0.01.

Test of Hypothesis #20:

$$H_0 : \sigma^2 \geq 25,$$

$$H_a : \sigma^2 < 25,$$

AD=9.35877, Chio=7.50779.

Based on Table 4, the classical approach fails to reject the null hypothesis at any significance level smaller or equal to 0.1. The

proposed approximate Bayesian decision rule rejects the null hypothesis for any level of significance smaller or equal to 0.1.

Data Set #5:

52, 33, 42, 44, 41, 50, 44, 51, 45, 38, 37, 40, 44, 50, 43 (McClave & Sincich, 1997 p. 301).

Normal population distribution obtained with SAS:

$$N(\mu = 43.6, \sigma = 5.4746)$$

The population and sample variances are $\sigma^2 = 29.97124$, and $s^2 = 29.971428$. For the following test of hypothesis:

$$H_a : \sigma^2 \neq c,$$

$$H_0 : \sigma^2 = c,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 5. Table 5 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the fifth data set.

Table 5: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	7.790–21.064	13.211–17.605
90	6.571–23.685	13.103–18.991
95	5.629–26.119	13.051–20.378
99	4.075–31.319	13.010–23.597

Test of Hypothesis #21:

$$H_0 : \sigma^2 = 29.97124,$$

$$H_a : \sigma^2 \neq 29.97124,$$

AD = 0, Chio = 14.000882.

Both, the classical and the proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

BAYESIAN APPROXIMATION FOR THE VARIANCE OF A GAUSSIAN DISTRIBUTION

Test of Hypothesis #22:

$$H_0 : \sigma^2 = 31.5 ,$$

$$H_a : \sigma^2 \neq 31.5 ,$$

AD = 1.52876.

The observed value is Chio = 13.32063467. Therefore both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2

Test of Hypothesis #23:

$$H_0 : \sigma^2 = 60 ,$$

$$H_a : \sigma^2 \neq 60 ,$$

AD = 30.02876, Chio = 6.99333.

The proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2. The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1.

Test of Hypothesis #24:

$$H_0 : \sigma^2 = 17 ,$$

$$H_a : \sigma^2 \neq 17 ,$$

AD = 12.97124, Chio = 24.682352.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.05, while the proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2.

Test of Hypothesis #25:

$$H_0 : \sigma^2 = 18 ,$$

$$H_a : \sigma^2 \neq 18 ,$$

AD = 11.97124, Chio = 23.31111.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1, while the proposed approximate Bayesian approach only fails to reject the null hypothesis at levels of significance smaller or equal to 0.01.

Data Set #6:

52, 43, 47, 56, 62, 53, 61, 50, 56, 52, 53, 60, 50, 48, 60, 55 (McClave & Sincich, 1997 p. 301).

Normal population distribution obtained with SAS:

$$N(\mu = 53.625, \sigma = 5.4145)$$

The population and sample variances are $\sigma^2 = 29.31681$, and $s^2 = 29.316666$. For the following test of hypothesis:

$$H_0 : \sigma^2 = c ,$$

$$H_a : \sigma^2 \neq c ,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 6. Table 6 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the sixth data set.

Table 6: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	8.547–22.307	14.211–18.605
90	7.261–24.996	14.103–19.991
95	6.262–27.488	14.051–21.378
99	4.601–32.801	14.010–24.597

Test of Hypothesis #26:

$$H_0 : \sigma^2 = 29.31681 ,$$

$$H_a : \sigma^2 \neq 29.31681 ,$$

AD = 0, Chio = 14.99992.

Both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #27:

$$H_0 : \sigma^2 = 26 ,$$

$$H_a : \sigma^2 \neq 26 ,$$

AD = 3.31681.

The observed value is Chio = 16.91346115. Therefore both, the classical and

proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #28:

$$H_0 : \sigma^2 = 60,$$

$$H_a : \sigma^2 \neq 60,$$

AD = 30.68319, Chio=7.329166.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1. The proposed approximate Bayesian approach rejects the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #29:

$$H_0 : \sigma^2 = 17,$$

$$H_a : \sigma^2 \neq 17,$$

AD=12.31681, Chio=25.867646.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.05. On the other hand, the proposed approximate Bayesian approach rejects the null hypothesis at any level of significance smaller equal to 0.2.

Test of Hypothesis #30:

$$H_0 : \sigma^2 \geq 50,$$

$$H_a : \sigma^2 < 50,$$

AD is greater or equal to 20.68319.

Using Table 6 it can be inferred that the classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1, while the proposed approximate Bayesian approach o reject the null hypothesis at levels of significance smaller or equal to 0.1.

Data Set #7:

The following observations have been obtained from the collection of SAS data sets: 50, 65, 100, 45, 111, 32, 45, 28, 60, 66, 114, 134, 150, 120, 77, 108, 112, 113, 80, 77, 69, 91, 116, 122, 37, 51, 53, 131, 49, 69, 66, 46, 131, 103, 84, 78.

Normal population distribution obtained with SAS:

$$N(\mu = 82.861, \sigma = 33.226)$$

The population and sample variances are $\sigma^2 = 1103.96716$, and $s^2 = 1103.951587$.

For the following test of hypothesis:

$$H_0 : \sigma^2 = c,$$

$$H_a : \sigma^2 \neq c,$$

the classical and approximate Bayesian non-rejection regions are presented in Table 7. Table 7 was used to conduct the following five tests of hypothesis about the normal population variance corresponding to the seventh data set.

Table 7: Classical and Approximate Bayesian Non-Rejection Regions

C. L. %	Non-Rejection Regions	
	Classical Method	Approximate Bayesian Approach
80	24.825–46.031	34.211–38.605
90	22.501–49.765	34.103–39.991
95	20.612–53.160	34.051–41.378
99	17.247–60.219	34.010–44.597

Test of Hypothesis #31:

$$\sigma^2 = 1103.96716,$$

$$\sigma^2 \neq 1103.96716,$$

Chio = 34.9995.

Both, the classical and the proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

Test of Hypothesis #32:

$$H_0 : \sigma^2 = 1110,$$

$$H_a : \sigma^2 \neq 1110,$$

The observed value is Chio = 4.809284. Therefore both, the classical and proposed approximate Bayesian approaches, fail to reject the null hypothesis at any level of significance smaller or equal to 0.2.

BAYESIAN APPROXIMATION FOR THE VARIANCE OF A GAUSSIAN DISTRIBUTION

Test of Hypothesis #33:

$$H_0 : \sigma^2 = 1800 ,$$

$$H_a : \sigma^2 \neq 1800 ,$$

AD = 0, Chio = 21.46572.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.05, The proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2

Test of Hypothesis #34:

$$H_0 : \sigma^2 = 800 ,$$

$$H_a : \sigma^2 \neq 800 ,$$

AD = 1000, Chio = 48.297879.

The classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.1. The proposed approximate Bayesian approach rejects the null hypothesis at levels of significance smaller or equal to 0.2.

Test of Hypothesis #35:

$$H_0 : \sigma^2 \leq 800 ,$$

$$H_a : \sigma^2 > 800 ,$$

AD is greater or equal to 1000, Chio = 48.297879.

Using Table 7 it is inferred that the classical approach fails to reject the null hypothesis at any level of significance smaller or equal to 0.05. On the other hand the proposed approximate Bayesian approach o reject the null hypothesis at levels of significance smaller or equal to 0.1.

Conclusion

All randomly selected thirty-five tests of hypothesis show that the proposed approximate Bayesian decision rule performs well: The approximate Bayesian approach yields a non-rejection region that is strictly included in its classical counterpart.

In the present study, a new approximate Bayesian decision rule for the variance of a normal population has been derived with the use of the square error loss function. Based on the

above numerical results we can conclude the following:

1. The classical decision rule for the variance of a normal population does not always yield the best results. In fact, contrary to our proposed Bayesian decision rule, the classical approach fails, at times , to reject claims that are far from being good estimates of the population variance
2. The classical decision rule does not always yield a smaller Type II error than the approximate Bayesian decision rule. In fact the numerical simulation shows that the Bayesian approach performs better when it comes to rejecting a wrong null hypothesis.
3. Contrary to the classical rejection and non-rejection regions that are defined with the use the Chi-square table, their approximate Bayesian counterparts rely only on the observations
4. The approximate Bayesian decision rule can be easily applied to any normal or approximately normal data, irrespective of the size of the sample that is used for the study.
5. With the approximate Bayesian decision rule, tests of hypothesis about a normal population variance are easily conducted at any level of significance.

Bayesian analysis contributes to reinforcing well-known statistical theories such as the Decision Theory.

Acknowledgment

This research was sponsored by the Research Center for Bayesian Applications, Inc.

References

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

- Birnbaum, Z. W., & Saunders S. C. (1958). A statistical model for life-length of material, *Journal of the American Statistical Association*, 53, 151-160.
- Camara, V. A. R., & Tsokos, C. P. (1999). Bayesian, reliability modeling with a new function. *STATISTICA*, 61(4), 619-630.
- Camara, V. A. R., & Tsokos, C. P. (1999). the effect of loss functions on empirical bayes reliability analysis. *Journal of Engineering Problems*, 373-378.
- Camara, V. A. R. (2003). Approximate Bayesian confidence intervals for the variance of a Gaussian distribution. *Journal of Modern Applied Statistical Methods*, 2(2), 350-358.
- Drake, A. W. (1966). Bayesian statistics for the reliability engineer. *Annual Symposium on Reliability*, Proc. 1966, 315-320.
- Epstein, B. & Sobel, M. (1953). Life testing. *Journal of the American Statistical Association*, 48, 486-502.
- Epstein, B. & Sobel, M. (1954). Some theorems relevant to life testing from an exponential distribution. *Annals of Mathematical Statistics*, 25, 373-381.
- Epstein, B., & Sobel, M. (1955). Sequential life tests in the exponential case. *Annals of Mathematical Statistics*, 26, 82-93.
- Kao, J. H. K. (1956). A new life quality measurer for electron tubes. *TRE Transactions on Reliability and Quality Control*, 1(4), 389-407.
- McClave, J. T., & Sincich, T. A. (1997). *A first course in Statistics*, (6th Ed.). San Francisco, CA: Dellen Publishing Co.
- Prem, S. M. (1998). *Introductory Statistics*, (3rd Ed.). NY: Wiley.
- Triola, M. F. (2007). *Elementary statistics*, (10th Ed.). Boston: Addison Wesley.
- Winkler, R. L. (1972), *Introduction to Bayesian inference and decision making*. NY: Probabilistic Publishing.

Bias in Stabilized Sieve Sampling

Liming Guan John P. Wendell
University of Hawaii at Manoa

The stabilized sieve sample selection method (SSM) is considered to be a probability proportional to size (PPS) sampling method with an unbiased estimator (Horgan 1997, 1998). This article demonstrates that SSM does not select items with PPS and that the point estimator is biased.

Key words: Sampling with probability proportional to size; Hansen-Hurwitz estimator; Horvitz-Thompson estimator.

Introduction

Consider a situation where it is desired to make an inference about an unknown population parameter, Y , such that

$$Y = \sum_{I=1}^N y_I \quad (1)$$

where N is the population size, $I = (1, 2, \dots, N)$, and y_I is unknown but can be determined exactly by applying some procedure. An unbiased estimate of Y can be obtained when sampling with replacement using the Hansen-Hurwitz estimator (Brewer & Hanif 1983, p. 5)

$$\hat{Y}_{HH} = \frac{1}{n} \sum_i^n \frac{y_i}{p_i} \quad (2)$$

where n is the sample size, y_i is the value y_I that is determined for the i th item in the sample, p_i is the probability of inclusion as the i th item in the sample of the population item I and p_i is the value of p_I for the i th item selected for the sample. Note that under sampling with replacement an individual population item, I , can be included in the sample more than once.

Liming Guan is an Associate Professor of Accounting in the Shidler College of Business. Email: lguan@hawaii.edu. John P. Wendell is a Professor of Accounting in the Shidler College of Business. Email: john.wendell@hawaii.edu.

When sampling without replacement, an unbiased estimate of Y can be obtained using the Horvitz-Thompson estimator (Brewer & Hanif 1983, p. 6)

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (3)$$

where π_i is the probability of inclusion in the sample of the population item I and y_i is the value of y_I for the i th item in the sample.

Equations (2) and (3) are general and allow for an unbiased estimate of Y regardless of how p_i or π_i are determined. For sampling with equal probabilities $p_i = 1/N$ and $\pi_i = n/N$. Sampling with unequal probabilities is often a good choice and may be the only possible method given the sampling frame. Examples of sampling with unequal probabilities are stratified sampling and cluster sampling. Another method for sampling with unequal probabilities is probability proportional to size (PPS) sampling. The size variable can be any variable x for which every x_i satisfies

$$0 < x_i < \frac{X}{n} \text{ where } X = \sum_{I=1}^N x_I. \quad (4)$$

The right side of the inequality is a requirement only when sampling without replacement. If these conditions are met then a PPS sample can be drawn by setting

$$p_i = \frac{x_i}{X} \quad (5)$$

when sampling with replacement and

$$\pi_I = n \frac{x_I}{X} \tag{6}$$

when sampling without replacement.

PPS sampling methods are generally applicable to any population where it is desired to estimate Y using either (2) or (3) and there is a size variable available conforming to (4). This article examines the properties of two such methods, the sieve method and the stabilized sieve method (SSM).

Sieve Sampling

The sieve method is a PPS sampling without replacement method that was developed by Rietveld (1978, 1979a,b). The presentation of the method given here is based on Horgan (1998). A population item is selected for inclusion in the sample if it satisfies the inequality

$$r_I \leq x_I \tag{7}$$

where r_I is a random variable uniformly distributed on the interval $(0, X/n)$ and each r_I is independently generated. It is important to note that the realized sample size, n_r , is a random variable that will not always be the same as n . Equation (3) with the sum over n_r and π_I defined as in (6) will yield an unbiased estimate of Y . The properties of the sieve method and the SSM will be illustrated by sampling from a hypothetical population with $N = 5$ and $n = 2$ used by Wright (1991) to demonstrate that systematic PPS samples lose their PPS property when augmented by systematically sampling the remaining population. The details of this population are given in Table 1.

Table 1: Test Population. $N = 5, n = 2, X = 20$.

I	x_I	π_I	$1 - \pi_I$	p_I
1	2	0.2	0.8	0.10
2	3	0.3	0.7	0.15
3	4	0.4	0.6	0.20
4	5	0.5	0.5	0.25
5	6	0.6	0.4	0.30

Table 2: Probabilities of sample outcomes for the test population in Table 1 for sieve sampling. Column j is an identification variable for each of the 32 outcomes. The second column indicates which population items were included in a particular sample outcome and p is the probability of that outcome.

j	I_s	p
1	Null	0.0672
2	1	0.0168
3	2	0.0288
4	3	0.0448
5	4	0.0672
6	5	0.1008
7	1,2	0.0072
8	1,3	0.0112
9	1,4	0.0168
10	1,5	0.0252
11	2,3	0.0192
12	2,4	0.0288
13	2,5	0.0432
14	3,4	0.0448
15	3,5	0.0672
16	4,5	0.1008
17	1,2,3	0.0048
18	1,2,4	0.0072
19	1,2,5	0.0108
20	1,3,4	0.0112
21	1,3,5	0.0168
22	1,4,5	0.0252
23	2,3,4	0.0192
24	2,3,5	0.0288
25	2,4,5	0.0432
26	3,4,5	0.0672
27	1,2,3,4	0.0048
28	1,2,3,5	0.0072
29	1,2,4,5	0.0108
30	1,3,4,5	0.0168
31	2,3,4,5	0.0288
32	1,2,3,4,5	0.0072

BIAS IN STABILIZED SIEVE SAMPLING

These sample outcome probabilities are calculated as

$$p_j = \prod_{I \in s_j} \pi_I \prod_{I \notin s_j} 1 - \pi_I \quad (8)$$

where s_j is the j th sample outcome in Table 2. For example, the probability of getting sample outcome 11, item 2 and 3, is $0.3 \times 0.4 \times 0.8 \times 0.5 \times 0.4 = 0.0192$. That the sieve method is indeed PPS for this population can be checked by summing the probabilities for each sample outcome containing a particular population item, I , and verifying that it is equal to the value for π_I in Table 1.

Table 3 shows the probability of achieving a particular n_r . These probabilities can be calculated from Table 2 by summing all probabilities for outcomes of a given size.

Table 3: Probabilities of a realized sample size for the test population in Table 1.

n_r	p
0	0.0672
1	0.2584
2	0.3644
3	0.2344
4	0.0684
5	0.0072

Table 4 shows the probabilities of inclusion in n_r for each combination of population item and realized sample size for the test population in Table 1. These conditional probabilities are not proportional to x_I .

Table 4: Conditional probabilities of inclusion.

	n_r					
I	0	1	2	3	4	5
1	0	0.0650	0.1658	0.3242	0.5789	1
2	0	0.1115	0.2700	0.4863	0.7544	1
3	0	0.1734	0.3908	0.6314	0.8421	1
4	0	0.2601	0.5247	0.7389	0.8947	1
5	0	0.3901	0.6487	0.8191	0.9298	1

The stabilized sieve method (SSM) (Horgan 1997, 1998) is a modification of the sieve method that ensures that the final sample size is always equal to n . This section details how the method selects items for a sample and then considers the properties of point estimators of Y for samples selected using the SSM.

The SSM is selected in two stages. First an initial sample, S_1 , is selected using (7). In the second stage the sampling process is conditioned upon the number of items in S_1 (Horgan 1998)

$$S_2 = \begin{cases} S_1 + A(n - n_r) & \text{if } n_r < n \\ S_1 & \text{if } n_r = n \\ S_1 - R(n_r - n) & \text{if } n_r > n \end{cases} \quad (9)$$

where S_2 is the final sample and $A(m)$ and $R(m)$ are defined as follows: $A(m)$ selects m items one at a time (with replacement) by taking a simple random sample of one from the entire population (including items in S_1) and this item is selected for inclusion in the sample if

$$r \leq x_I \quad (10)$$

where r is a uniformly distributed random number in the interval $(0, \max(x_I)]$. The process is repeated, generating a new value for r each time, until m items are selected. All items selected using $A(m)$ satisfy (5). $R(m)$ selects m items to remove from S_1 by taking a simple random sample of size m from S_1 .

Table 5 gives the probabilities for each sample outcome and n_r for the population in Table 1 sampled using SSM. Because of the complexity of (9) some explanation of how individual cells in this table were calculated may be useful. The simplest case is when $n_r = 2$ where the values are taken directly from Table 2. Outcome 9 (2,5) with $n_r = 4$ will be used to illustrate the cases when $n_r > 2$. First, the probabilities of all the outcomes where $n_r = 4$ and both item 2 and 5 are present (outcomes 28, 29, and 31) are summed and then divided by the number of combinations of two items that can be drawn from a population of four items. This gives $(0.0072 + 0.0108 + 0.0288)/6 = 0.0078$. When $n_r < 2$, there may be more than one path to a sample outcome. For example, outcome 7 (2,3)

with $n_r = 1$ can occur when the initial sieve sample contains only item 2 or only item 3. Outcome 3 in Table 2 gives the probability of S_l containing only item 2 and Table 1 gives the value of p_l for selecting item 3 in the second stage. The probabilities that S_l contains only item 3 and that item 2 is selected in the second stage can be determined in the same manner. Thus, the probability for outcome 7 when $n_r = 1$ is $0.0288 \times 0.20 + 0.0448 \times 0.15 = 0.01248$.

Table 5 shows probabilities for sample outcomes for the test population in Table 1 for stabilized sieve sampling. Column j is an identification variable for each of the 15 outcomes. The second column indicates which population items were included in a particular sample outcome, n_r is the realized sample size in stage one and the cells contain the joint probability of the sample outcome and n_r . Table 6 provides p_l for the population in Table 1 when sampling with SSM and demonstrates that these probabilities are not PPS. These probabilities are derived from Table 5 by summing the probabilities for each sample outcome that contains a particular I divided by n , which is 2 in this case. For outcomes where I is included twice it is counted twice.

Horgan (1998, equations 8, 17, and 19) provides an estimator for Y that is conditional upon n_r :

$$\hat{Y}_s = wX \sum_{i=1}^n \frac{y_i}{x_i} \text{ where } w = \begin{cases} \frac{1}{2n - n_r} & \text{if } n_r < n \\ \frac{1}{n} & \text{if } n_r = n \\ \frac{n_r}{n^2} & \text{if } n_r > n \end{cases} \quad (11)$$

Although (11) conditions on n_r , it does not take into consideration that the probabilities of inclusion in the sample given n_r are not proportional to x_l (see Table 4). Consequently, \hat{Y}_s is a biased estimator. The expected value of \hat{Y}_s for the test population can be calculated by determining \hat{Y}_s for every cell in Table 5, multiplying the result by the probability in the cell and then taking the sum of those products. The result is $1.0917 y_1 + 1.0877 y_2 + 1.0824 y_3 + 1.0749 y_4 + 1.0637 y_5$.

Because the SSM method is a sampling with replacement method based on the sieve method, and the sieve method is a PPS method without replacement it seems reasonable to use \hat{Y}_{HH} with p_l calculated according to (5). This will also give a biased estimate of Y , the expected value of which is $0.9491 y_1 + 0.9693 y_2 + 1.0029 y_3 + 1.0111 y_4 + 0.9791 y_5$ for the Table 1 population.

It is possible to construct an unbiased estimator of Y when using the SSM with the population in Table 1. This is done by first setting each p_l to the corresponding value in Table 6 and then calculating \hat{Y}_{HH} accordingly. Unfortunately, the use of this estimator is limited to very small populations because it requires an enumeration of all 2^N possible sample outcomes for the stage 1 sieve sample.

Conclusion

The stabilized sieve method does not sample with PPS and that both \hat{Y}_s and \hat{Y}_{HH} with p_l calculated according to (5) are biased estimators of Y . Further, the calculation of the unbiased estimator is prohibitively expensive to compute for any but the smallest populations. Nonetheless, the SSM performed well in the simulations in Horgan (1997 and 1998) in comparison to the sieve method and the probability proportionate to size with replacement method (PPR).

All three of these methods have drawbacks, either the possibility of items showing up more than once in the sample (SSM, PPR) or variable sample size (sieve), or bias (SSM). Systematic PPS sampling methods utilizing a random sort of the population before application have none of these drawbacks because they select fixed size samples without replacement with probabilities that are exactly proportional to x_l (see Brewer & Hanif 1983, procedures 2 and 3). These selection methods are easily applied with modern computers if both I and x_l are available in a computer accessible file. Consequently, with these sampling frames the systematic procedures should be preferred over either the sieve, SSM, or PPR methods. However, not all sampling frames make the entire population x_l conveniently accessible by computer and the sieve, SSM, and PPR methods

BIAS IN STABILIZED SIEVE SAMPLING

may have some practical advantages with these sampling frames that offset their disadvantages. With such challenging sampling frames, the SSM method should not be ruled out simply because of the difficulty in achieving a completely unbiased estimate of Y , particularly if the population characteristics and sample sizes are similar to those used for the simulations in Horgan (1997 and 1998).

References

Brewer, K. R. W., & Hanif, M. (1983). *Sampling with unequal probabilities*. NY: Springer-Verlag.

Horgan, J. M. (1997). Stabilizing the sieve sample using PPS. *Auditing: A Journal of Practice and Theory*, 16, 40-51.

Horgan, J. M. (1998). Stabilized sieve sampling: A point-estimator analysis. *Journal of Business and Economic Statistics*, 16, 42-51.

Rietveld, C. (1978). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (I). *Compact: Computer en Accountant*, 15, 2-11.

Rietveld, C. (1979a). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (II) en (III). *Compact: Computer en Accountant*, 16, 2-13.

Rietveld, C. (1979b). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepaktijk (IV). *Compact: Computer en Accountant*, 17, 9-18.

Wright, D. W. (1991). Augmenting a sample selected with probability proportional to size. *Auditing: A Journal of Practice and Theory*, 10, 145-158.

Table 5: Probabilities of sample outcomes for the test population in Table 1 for stabilized sieve sampling.

j	I_s	n_r					
		0	1	2	3	4	5
1	1,1	0.00067	0.00168	0	0	0	0
2	1,2	0.00202	0.00540	0.00720	0.00760	0.00380	0.00072
3	1,3	0.00269	0.00784	0.01120	0.01093	0.00480	0.00072
4	1,4	0.00336	0.01092	0.01680	0.01453	0.00540	0.00072
5	1,5	0.00403	0.01512	0.02520	0.01760	0.00580	0.00072
6	2,2	0.00151	0.00432	0	0	0	0
7	2,3	0.00403	0.01248	0.01920	0.01760	0.00680	0.00072
8	2,4	0.00504	0.01728	0.02880	0.02320	0.00740	0.00072
9	2,5	0.00605	0.02376	0.04320	0.02760	0.00780	0.00072
10	3,3	0.00269	0.00896	0	0	0	0
11	3,4	0.00672	0.02464	0.04480	0.03253	0.00840	0.00072
12	3,5	0.00806	0.03360	0.06720	0.03760	0.00880	0.00072
13	4,4	0.00420	0.01680	0	0	0	0
14	4,5	0.01008	0.04536	0.10080	0.04520	0.00940	0.00072
15	5,5	0.00605	0.03024	0	0	0	0

Table 6. Probabilities of inclusion in a sample draw for each item in the test population in Table 1 compared to the probability under PPS.

I	p_I	
	actual	PPS
1	0.09491	0.10
2	0.14540	0.15
3	0.19805	0.20
4	0.25277	0.25
5	0.30886	0.30

Some Estimators for the Population Mean Using Auxiliary Information Under Ranked Set Sampling

Walid A. Abu-Dayyeh
Sultan Qaboos University

M .S. Ahmed
Sultan Qaboos University

R. A. Ahmed
Yarmouk University

Hassen A. Muttalak
King Fahd University of Petroleum & Minerals

Auxiliary information is used along with ranking information to derive several classes of estimators to estimate the population mean of a variable of interest based on RSS (ranked set sample). The properties of these newly suggested estimators were examined. Comparisons between special cases of these estimators and other known estimators are made using a real data set. Some of the new estimators are superior to the old ones in terms of bias and mean square error.

Keywords: Auxiliary variables, efficiency, ranking, ranked set sample.

Introduction

Many authors have discussed the use of supplementary information of auxiliary variables in survey sampling to improve the existing estimators (for example, Cochran, 1977). The ratio estimator is among the most commonly adopted to estimate: (1) population means, or (2) the total of some variable of interest from a finite population with the help of an auxiliary variable when the correlation coefficient between the two variables is positive. When the correlation coefficient between the two variables is negative, the product estimator is used. These estimators are more efficient, i.e. have smaller

variances than the usual estimators of the population mean based on the sample mean of a simple random sample (SRS).

Ranked set sampling (RSS) can be used when the measurement of sample units drawn from a population of interest is very laborious or costly, but several elements can be easily arranged (ranked) in the order of magnitude. Takahasi and Wakimoto (1968) established the theory of RSS. They showed that the mean of the RSS is an unbiased estimator for the population mean and is more efficient than the mean of SRS. Dell and Clutter (1972) studied the effect of ranking error on the efficiency of RSS. The RSS has many statistical applications in biology and environmental studies (Barabesi & El-Sharaawi, 2001), for example, McIntyre (1952) first suggested using RSS to estimate the yield of pasture. In addition, RSS has been investigated by many researchers (Stokes, 1977; Stokes & Sager, 1988; Lam, et al., 1994, 1980; Mode, et al., 1999; Al-Saleh & Al-Shrafat, 2001; Al-Saleh & Zheng, 2000; Al-Saleh & Al-Omary, 2000), for more details about RSS, see Kaur, et al., 1995.

The RSS method can be summarized as follows: Select m random samples of size m units each and rank the units within each sample with respect to the variable of interest by a visual inspection or some other simple method.

Walid Abu-Dayyeh is an associate Professor in the Department of Mathematics and Statistics at Sultan Qaboos University/Sultanate of Oman. Email: abudayyehw@yahoo.com. M .S. Ahmed is an Associate Professor in the Department of Mathematics and Statistics at Sultan Qaboos University, Sultanate of Oman. Email: msahmed@squ.edu.om. R. A. Ahmed is a Lecturer at Almosul University/Iraq. Hassen A. Muttalak is a Professor in the Department of Mathematics and Statistics at King Fahd University of Petroleum & Minerals, Saudi Arabia. Email: hamstat@kfupm.edu.sa.

ESTIMATORS FOR THE MEAN UNDER RANKED SET SAMPLING

Next, select for actual measurement the i^{th} smallest unit from the i^{th} sample for $i=1, 2, \dots, m$. In this way, a total of m measured units are obtained, one from each sample. The cycle may be repeated r times to get a sample of size $n = rm$. These $n = rm$ units form the RSS data. Note that in RSS, rm^2 elements are identified, but only rm of them are quantified. Thus, comparing this sample with a simple random sample (SRS) of size rm is reasonable.

Some Notions and Preliminaries

Let Y denote the variable of interest whose population mean and variance are μ_y and σ_y^2 respectively. Estimate μ_y using the information provided by one or two auxiliary variables X_1 and X_2 based on SRS and RSS will be considered. Let μ_{x_i} and $\sigma_{x_i}^2$ be the population mean and variance for X_i , $i=1, 2$. Let Y_j, X_{1j} and X_{2j} denote the values of the variables Y, X_1 and X_2 respectively, on the j^{th} unit of the population. The population means μ_{x_1} and μ_{x_2} of the auxiliary variables are assumed to be known.

Let $Y_{(i)j}, X_{1(i)j}$ and $X_{2(i)j}$ represent the i^{th} order statistics of a sample of size m in the j^{th} cycle of the variables Y, X_1 and X_2 respectively based on a RSS of size $n = rm$ drawn from the population. The sample mean for each variable using RSS data are defined as follows:

$$\bar{Y}_{(n)} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m Y_{(i)j},$$

$$\bar{X}_{1(n)} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m X_{1(i)j},$$

and

$$\bar{X}_{2(n)} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m X_{2(i)j}.$$

Consider the following notations:

$$T_{y(i)} = (\mu_{y(i)} - \mu_y), T_{x_{1(i)}} = (\mu_{x_{1(i)}} - \mu_{x_1}),$$

$$T_{x_{2(i)}} = (\mu_{x_{2(i)}} - \mu_{x_2}),$$

$$T_{yx_{1(i)}} = (\mu_{y(i)} - \mu_y)(\mu_{x_{1(i)}} - \mu_{x_1}),$$

$$T_{yx_{2(i)}} = (\mu_{y(i)} - \mu_y)(\mu_{x_{2(i)}} - \mu_{x_2}),$$

$$T_{x_{1x_{2(i)}}} = (\mu_{x_{1(i)}} - \mu_{x_1})(\mu_{x_{2(i)}} - \mu_{x_2}),$$

$$\sigma_{yx_{1(i)}} = E(Y_{(i)} - \mu_{y(i)})(X_{1(i)} - \mu_{x_{1(i)}}),$$

$$\sigma_{yx_{2(i)}} = E(Y_{(i)} - \mu_{y(i)})(X_{2(i)} - \mu_{x_{2(i)}}),$$

$$\sigma_{x_{1x_{2(i)}}} = E(X_{1(i)} - \mu_{x_{1(i)}})(X_{2(i)} - \mu_{x_{2(i)}}).$$

then:

$$\sum_{i=1}^n T_{y(i)} = 0, \quad \sum_{i=1}^n T_{x_{1(i)}} = 0, \quad \sum_{i=1}^n T_{x_{2(i)}} = 0,$$

$$\sum_{i=1}^n \sigma_{y(i)}^2 = n\sigma_y^2 - \sum_{i=1}^n T_{y(i)}^2,$$

$$\sum_{i=1}^n \sigma_{x_{1(i)}}^2 = n\sigma_{x_1}^2 - \sum_{i=1}^n T_{x_{1(i)}}^2,$$

$$\sum_{i=1}^n \sigma_{yx_{1(i)}} = n\sigma_{yx_1} - \sum_{i=1}^n T_{yx_{1(i)}},$$

$$\sum_{i=1}^n \sigma_{yx_{2(i)}} = n\sigma_{yx_2} - \sum_{i=1}^n T_{yx_{2(i)}}$$

$$\sum_{i=1}^n \sigma_{x_{1x_{2(i)}}} = n\sigma_{x_1x_2} - \sum_{i=1}^n T_{x_{1x_{2(i)}}}$$

and

$$\sum_{i=1}^n \sigma_{x_{2(i)}}^2 = n\sigma_{x_2}^2 - \sum_{i=1}^n T_{x_{2(i)}}^2.$$

The following classes of estimators of the mean of the variable Y based on RSS are:

$$\tilde{Y}_{a_1, a_2} = \bar{Y}_{(n)} \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} \left(\frac{\bar{X}_{2(n)}}{\mu_{x_2}} \right)^{a_2} \quad (2.1)$$

and

$$\tilde{Y}_{w_1, w_2} = \bar{Y}_{(n)} \left[w_1 \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} + w_2 \left(\frac{\bar{X}_{2(n)}}{\mu_{x_2}} \right)^{a_2} \right] \quad (2.2)$$

where a_1, a_2, w_1, w_2 are constants and $w_1 + w_2 = 1$.

Estimators Based on RSS and One or Two Auxiliary Variables

It is not possible to rank two or more dimensional data, therefore, ranking one of the variables and taking the corresponding values of other variables is an option. Assuming that the variable can be ranked perfectly - there are no errors in ranking the units, there will be errors in ranking the other variables.

Ranking on Study Variable Y

Assume that the ranking on variable Y is perfect while the ranking on variables X_1 and X_2 will have errors; the estimators (2.1) and (2.2) are respectively given by:

$$\tilde{Y}_a = \bar{Y}_{(n)} \left(\frac{\bar{X}_{1[n]}}{\mu_{x_1}} \right)^{a_1} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2}, \quad (3.1)$$

and

$$\tilde{Y}_w = w_1 \bar{Y}_{(n)} \left(\frac{\bar{X}_{1[n]}}{\mu_{x_1}} \right)^{a_1} + w_2 \bar{Y}_{(n)} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2}. \quad (3.2)$$

where

$$\bar{X}_{1[n]} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m X_{1[i]j}$$

and

$$\bar{X}_{2[n]} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m X_{2[i]j}$$

are the sample means of the RSS for X_1 and X_2 respectively and $X_{1[i]j}$ and $X_{2[i]j}$ are the i^{th} judgment order statistic of the i^{th} sample of

the j^{th} cycle, of the variables X_1 and X_2 respectively.

Let

$$e_0 = \frac{\bar{Y}_{(n)} - \mu_y}{\mu_y}, \quad e_1 = \frac{\bar{X}_{1[n]} - \mu_{x_1}}{\mu_{x_1}}$$

and

$$e_2 = \frac{\bar{X}_{2[n]} - \mu_{x_2}}{\mu_{x_2}}.$$

Obtain the bias and the MSE of the estimators \tilde{Y}_a and \tilde{Y}_w respectively up to the order of n^{-1} as follows:

$$\begin{aligned} B(\tilde{Y}_a) &= \frac{a_1}{rm^2 \mu_{x_1}} (m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_{1[i]}}) + \frac{a_2}{rm^2 \mu_{x_2}} (m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_{2[i]}}) \\ &+ \frac{\mu_y a_1 (a_1 - 1)}{2rm^2 \mu_{x_1}^2} (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_{1[i]}}^2) + \\ &\frac{\mu_y a_2 (a_2 - 1)}{2rm^2 \mu_{x_2}^2} (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_{2[i]}}^2) \\ &+ \frac{\mu_y a_1 a_2}{rm^2 \mu_{x_1} \mu_{x_2}} (m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_{2[i]}}) \end{aligned} \quad (3.3)$$

The MSE of \tilde{Y}_a when ranking on variable Y is:

$$\begin{aligned} MSE(\tilde{Y}_a) &= \frac{m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2}{rm^2} + \frac{\mu_y^2 a_1^2 (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_{1[i]}}^2)}{rm^2 \mu_{x_1}^2} + \\ &\frac{\mu_y^2 a_2^2 (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_{2[i]}}^2)}{rm^2 \mu_{x_2}^2} + \frac{2\mu_y^2 a_1 (m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_{1[i]}})}{rm^2 \mu_y \mu_{x_1}} \\ &+ \frac{2\mu_y^2 a_2 (m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_{2[i]}})}{rm^2 \mu_y \mu_{x_2}} + \frac{2\mu_y^2 a_1 a_2 (m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_{2[i]}})}{rm^2 \mu_{x_1} \mu_{x_2}} \end{aligned} \quad (3.4)$$

ESTIMATORS FOR THE MEAN UNDER RANKED SET SAMPLING

up to the order of n^{-1} . The optimum values of a_1 and a_2 , which minimize the MSE of \tilde{Y}_a , are obtained by the derivation of (3.4) with respect to a_1 and a_2 respectively

$$a_1^* = \frac{\mu_{x_1} (m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]}) (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2) - (m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]}) (m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})}{\mu_y (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2) (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2) - (m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})} \quad (3.5)$$

$$a_2^* = \frac{\mu_{x_2} (m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]}) (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2) - (m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]}) (m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})}{\mu_y (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2) (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2) - (m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})} \quad (3.6)$$

The minimum MSE up to terms of n^{-1} for the class \tilde{Y}_{a^*} is:

$$MSE_{\min}(\tilde{Y}_{a^*}) = \frac{1}{rm^2} [(m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]})^2 (m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2) + (m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]})^2 (m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2) - (m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2) - \frac{2(m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2)(m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2)(m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})}{(m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2)(m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2) - (m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]})^2}] \quad (3.7)$$

If a_1 and a_2 take the values in (3.5) and (3.6) respectively, the bias of \tilde{Y}_a from (3.3) is given by:

$$B_{\min}(\tilde{Y}_{a^*}) = \frac{g_1}{rm^2 \mu_y^2 \{ [m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2] [m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2] - [m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]}]^2 \}} \quad (3.8)$$

where g_1 is given in the Appendix.

The bias and the MSE of the estimators of (3.2) are given by:

$$B(\tilde{Y}_w) = \mu_y \left\{ \frac{a_1 w_1}{rm^2 \mu_y \mu_{x_1}} [m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]}] + \frac{a_2 w_2}{rm^2 \mu_y \mu_{x_2}} [m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]}] + \frac{w_1 a_1 (a_1 - 1)}{2rm^2 \mu_{x_2}^2} [m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2] + \frac{w_2 a_2 (a_2 - 1)}{2rm^2 \mu_{x_1}^2} [m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2] \right\}, \quad (3.9)$$

up to the order of n^{-1} . The MSE of the estimator \tilde{Y}_w if ranking on variable Y is:

$$\begin{aligned}
 MSE(\tilde{Y}_w) = & \mu_y^2 \left\{ \frac{m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2}{rm^2\mu_y^2} + \frac{w_1^2 a_1^2 [m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2]}{rm^2\mu_{x_1}^2} + \right. \\
 & \frac{w_2^2 a_2^2 [m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2]}{rm^2\mu_{x_2}^2} + \frac{2w_1 a_1 [m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]}]}{rm^2\mu_y\mu_{x_1}} + \\
 & \left. \frac{2w_2 a_2 [m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]}]}{rm^2\mu_y\mu_{x_2}} + \frac{2w_1 w_2 a_1 a_2 [m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]}]}{rm^2\mu_{x_1}\mu_{x_2}} \right\} \quad (3.10)
 \end{aligned}$$

if a_1 and a_2 are both known and take the values in (3.5) and (3.6) respectively, up to order n^{-1} .

The optimum values of w_1 and w_2 , which minimize the MSE of \tilde{Y}_w , obtained by the derivation of equation (3.10) with respect to w_1 under the restriction $w_1 + w_2 = 1$, are given by:

$$\begin{aligned}
 w_1^* = & \frac{a_2^2 \left(m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right) - a_1 \left(m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right) + a_2 \left(m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right) - a_1 a_2 \left(m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right)}{\left(a_1^2 m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right) + a_2^2 \left(m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right) - 2a_1 a_2 \left(m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right)}
 \end{aligned}$$

and the MSE of

$$\tilde{Y}_{w^*} = w_1^* \bar{Y}_{(n)} \left(\frac{\bar{X}_{1[n]}}{\mu_{x_1}} \right)^{a_1} + w_2^* \bar{Y}_{(n)} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2} \quad (3.12)$$

is the minimum MSE up to terms of n^{-1} . As for the class \tilde{Y}_w :

$$\begin{aligned}
 MSE_{\min}(\tilde{Y}_{w^*}) = & \frac{1}{rm^2} \left\{ \left[m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right] + w_1^{*2} \left(a_1^2 \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \right. \right. \\
 & \left. \left. + a_2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] - 2a_1 a_2 \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] \right) \right. \\
 & - 2w_1^* \left(a_2^2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] - a_1 \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] \right. \\
 & \left. \left. + a_2 \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] - a_1 a_2 \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] \right) \right. \\
 & \left. \left. + a_2^2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] + 2a_2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \right\} \quad (3.13)
 \end{aligned}$$

If w takes the value in (3.11), then bias of \tilde{Y}_w from (3.9) is given by:

$$\begin{aligned}
 B(\tilde{Y}_{w^*}) = & \mu_y \left\{ w_1^* \left(a_1 \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] - a_2 \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right. \right. \\
 & \left. \left. + \frac{a_1(a_1-1)}{2} \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] - \frac{a_2(a_2-1)}{2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \right) \right. \\
 & \left. \left. + \frac{a_2(a_2-1)}{2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] + a_2 \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right\} \quad (3.14)
 \end{aligned}$$

Ranking on One Auxiliary Variable

If the ranking of X_1 is perfect, then the two estimators (2.1) and (2.2) are given by:

$$\tilde{Y}_a = \bar{Y}_{[n]} \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2} \quad (3.15)$$

and

$$\tilde{Y}_w = w_1 \bar{Y}_{[n]} \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} + w_2 \bar{Y}_{(n)} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2} \quad (3.16)$$

The formulas for the bias and *MSE* of estimators (3.15) and (3.16) respectively will be the same as in 3.1 except for the current estimators replace [] by () in X_1 , and () by [] in Y .

Similarly if the ranking on X_2 is perfect, then the estimators:

$$\tilde{Y}_a = \bar{Y}_{[n]} \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2} \quad (3.17)$$

$$\tilde{Y}_w = w_1 \bar{Y}_{[n]} \left(\frac{\bar{X}_{1(n)}}{\mu_{x_1}} \right)^{a_1} + w_2 \bar{Y}_{(n)} \left(\frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right)^{a_2} \quad (3.18)$$

result.

The formulas for the bias and the *MSE* of estimators (3.17) and (3.18) respectively, will be the same as in 3.1 except for the case of replacing [] by () in X_2 , and () by [] in Y ($a_1 = 0$ or $a_2 = 0$ in (2.1) and $w_1 = 0$ or $w_2 = 0$ correspond to the case of one auxiliary variable).

Comparisons of Estimators

Consider the following known estimators. The RSS sample mean of the data:

$$\bar{Y}_{(n)} = \frac{1}{n} \sum_{j=1}^r \sum_{i=1}^m Y_{(i)j}$$

is an unbiased estimator for the population mean and its variance is given by:

$$Var(\bar{Y}_{(n)}) = \frac{1}{rm^2} [m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2]$$

(Takahasi & Wakimoto, 1968).

The Ratio estimator using RSS data is defined as:

$$\bar{Y}_R = \bar{Y}_{(n)} \frac{\mu_x}{\bar{X}_{[n]}}$$

This estimator is a special case of the estimator in equation (1) where $a_1 = -1$ and $a_2 = 0$. The bias and the *MSE* of this estimator are respectively given by:

$$B(\bar{Y}_R) = \frac{\mu_y}{rm^2 \mu_x} \left[\left(m\sigma_{yx} - \sum_{i=1}^m T_{yx[i]} \right) - \frac{1}{\mu_x} \left(m\sigma_x^2 - \sum_{i=1}^m T_{x[i]}^2 \right) \right]$$

$$MSE(\bar{Y}_R) =$$

$$\left[\frac{\mu_y^2}{rm^2} \left(\frac{1}{\mu_y^2} \left(m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right) + \frac{1}{\mu_x^2} \left(m\sigma_x^2 - \sum_{i=1}^m T_{x[i]}^2 \right) - \frac{2}{\mu_y \mu_x} \left(m\sigma_{yx} - \sum_{i=1}^m T_{yx[i]} \right) \right]$$

(Samawi & Muttlak, 1996).

The product estimator using RSS data is defined as:

$$\bar{Y}_P = \bar{Y}_{(n)} \frac{\bar{X}_{[n]}}{\mu_x}$$

This estimator is a special case for the estimator in equation (1) where $a_1 = 1$ and $a_2 = 0$. The new estimator is called the product estimator and its bias and *MSE* respectively are given by

$$B(\bar{Y}_P) = \frac{\mu_y}{rm^2} [m\sigma_{yx} - \sum_{i=1}^m T_{yx[i]}]$$

$$MSE(\bar{Y}_P) = \frac{\mu_y^2}{rm^2} \left\{ \frac{1}{\mu_y^2} \left[m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right] + \frac{1}{\mu_x^2} \left[m\sigma_x^2 - \sum_{i=1}^m T_{x[i]}^2 \right] + \frac{2}{\mu_y \mu_x} \left[m\sigma_{yx} - \sum_{i=1}^m T_{yx[i]} \right] \right\}$$

If $a_1 = a_2 = -1$ is set in the estimator of equation (2) the following new estimator results:

$$\hat{Y}_a = \bar{Y}_{(n)} \frac{\mu_{x_1}}{\bar{X}_{1[n]}} \frac{\mu_{x_2}}{\bar{X}_{2[n]}}$$

The bias and the *MSE* are respectively given by

$$B(\hat{Y}_a) = \frac{\mu_y}{rm^2} \left\{ \frac{1}{\mu_{x_1}^2} \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] + \frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] + \frac{1}{\mu_{x_1} \mu_{x_2}} \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] - \frac{1}{\mu_y \mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] - \frac{1}{\mu_y \mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right\}$$

and

$$MSE(\hat{Y}_a) = \frac{\mu_y^2}{rm^2} \left\{ \frac{1}{\mu_y^2} \left[m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right] + \frac{1}{\mu_{x_1}^2} \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] + \frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] - \frac{2}{\mu_y \mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] - \frac{2}{\mu_y \mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] + \frac{2}{\mu_{x_1} \mu_{x_2}} \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] \right\}$$

Setting $a_1 = a_2 = 1$ in the estimator of equation (2) results in a new estimator defined as:

$$\check{Y}_a = \bar{Y}_{(n)} \frac{\bar{X}_{1[n]}}{\mu_{x_1}} \frac{\bar{X}_{2[n]}}{\mu_{x_2}}$$

The bias and the *MSE* are respectfully given by

$$B(\check{Y}_a) = \frac{\mu_y}{rm^2} \left\{ \frac{1}{\mu_{x_1} \mu_{x_2}} \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] + \frac{1}{\mu_y \mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] + \frac{1}{\mu_y \mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right\}$$

and

$$MSE(\check{Y}_a) = \frac{\mu_y^2}{rm^2} \left\{ \frac{1}{\mu_y^2} \left[m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right] + \frac{1}{\mu_{x_1}^2} \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] + \frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] + \frac{2}{\mu_y \mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] + \frac{2}{\mu_y \mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] + \frac{2}{\mu_{x_1} \mu_{x_2}} \left[m\sigma_{x_1 x_2} - \sum_{i=1}^m T_{x_1 x_2[i]} \right] \right\}$$

If $a_1 = a_2 = 1$ in the estimator of equation 3 is set, a new estimator called the Multivariate ratio estimator using RSS can be defined as

$$\check{Y}_w = \bar{Y}_{(n)} \left\{ w_1 \frac{\bar{X}_{1[n]}}{\mu_{x_1}} + w_2 \frac{\bar{X}_{2[n]}}{\mu_{x_2}} \right\}$$

The bias and the *MSE* are respectively given by

$$B(\check{Y}_w) = \mu_y \left\{ \frac{w_1}{\mu_y \mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] + \frac{w_2}{\mu_y \mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right\}$$

and

$$\begin{aligned}
 MSE(\tilde{Y}_w) &= \frac{\mu_y^2}{rm^2} \left\{ \frac{1}{\mu_y^2} \left[m\sigma_y^2 - \sum_{i=1}^m T_{y(i)}^2 \right] \right. \\
 &+ w_1^2 \left(\frac{1}{\mu_{x_1}^2} \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1(i)}^2 \right] + \frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2(i)}^2 \right] \right. \\
 &- \frac{2}{\mu_{x_1}\mu_{x_2}} \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2(i)} \right] \left. \right) - 2w_1 \left(\frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2(i)}^2 \right] \right. \\
 &- \frac{1}{\mu_y\mu_{x_1}} \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1(i)} \right] + \frac{1}{\mu_y\mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2(i)} \right] \\
 &- \left. \left. \frac{1}{\mu_{x_1}\mu_{x_2}} \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2(i)} \right] \right) \right\} \\
 &+ \frac{1}{\mu_{x_2}^2} \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2(i)}^2 \right] + \frac{2}{\mu_y\mu_{x_2}} \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2(i)} \right] \left. \right\}.
 \end{aligned}$$

The comparison between the estimators proposed is illustrated by using a real data set. The data for the illustration was taken from Ahmed (1995); the population consists of 332 villages. Consider the variables, Y , X_1 and X_2 where Y is number of cultivators, X_1 is the area of the village and X_2 is the number of household in the village.

The following steps summarize the simulation procedure to find the bias and MSE of an estimator for the population mean using perfect ranking on the variable of interest Y .

Step 1:

Simulate rm^2 observations from the 332 real data values with replacement and perform the RSS procedure with $m=5$ and $r=16$ to get sample of size $n=rm=80$.

Step 2:

Use the data in Step 1 to calculate

$$\hat{Y}_{(n)} = \frac{1}{mr} \sum_{j=1}^r \sum_{i=1}^m \hat{Y}_{(i:m)j} = \frac{1}{80} \sum_{j=1}^{16} \sum_{i=1}^5 \hat{Y}_{(i:m)j}$$

where $\hat{Y}_{(i:m)j}$ is the i^{th} smallest in the sample of size $m=5$ in the j^{th} cycle.

Step 3:

Repeat steps 1 and 2 (30,000) times, using these 30,000 values to obtain

$$\bar{\hat{Y}}_{(n)} = \frac{1}{30000} \sum_{i=1}^{30000} \hat{Y}_{(n)i}$$

Step 4:

Find the approximate bias and MSE for $\hat{Y}_{(n)}$. The bias is obtained by

$$B(\hat{Y}_{(n)}) = \frac{1}{30000} \sum_{i=1}^{30000} \hat{Y}_{(n)i} - \mu_y,$$

and the MSE of $\hat{Y}_{(n)}$ is obtained as

$$MSE(\hat{Y}_{(n)}) = \frac{1}{30000} \sum_{i=1}^{30000} (\hat{Y}_{(n)i} - \bar{\hat{Y}}_{(n)})^2.$$

The above simulation was preformed for all other estimators suggested ranking on one of the variables Y , X_1 or X_2 . Calculate the efficiency of these estimators with respect to the $MSE(\hat{Y}_{(n)}) = Var(\bar{\hat{Y}}_{(n)})$ estimator using

$$e(\hat{Y}) = \frac{MSE(\hat{Y}_{(n)})}{MSE(\hat{Y})},$$

where \hat{Y} represents any of the estimators given.

In Tables 1-3, MSE , bias, and efficiency have been calculated for each of the suggested estimators. In Table 1, ranking on the variable Y is shown (i.e., the ranking of variable Y will be perfect while the ranking of the other variables will be with errors in ranking). Tables 2 and 3 show the ranking on the variables X_1 and X_2 respectively.

Considering the results of Tables 1-3 it is observed that \tilde{Y}_{a^*} dominates all other estimators and achieved the highest efficiency. Its efficiency is more than 22 times higher than the

Table 1: The Bias, MSE and the Efficiency for all Estimators Based on Ranking of the Variable Y

Estimator	Auxiliary variable	MSE	Efficiency	Bias
$\bar{Y}_{(n)}$	None	8374.579	1	0
\tilde{Y}^*	x_1	820.252	10.2041	0.843
\tilde{Y}^*	x_2	909.625	9.17431	0.752
\tilde{Y}_a^*	x_1, x_2	379.579	22.2222	0.253
\tilde{Y}_w^*	x_1, x_2	582.065	14.4928	-0.521
\bar{Y}_R	x_1	4403.464	1.8939	0.422
\bar{Y}_R	x_2	2217.261	3.7594	0.998
\bar{Y}_P	x_1	33092.8	0.25163	12.522
\bar{Y}_P	x_2	30979.3	0.2688	7.121
\hat{Y}_a	x_1, x_2	8479.4	0.98231	14.153
\ddot{Y}_a	x_1, x_2	69893.8	0.119147	15.332
\check{Y}_w	x_1, x_2	598.243	14.0845	7.151

ESTIMATORS FOR THE MEAN UNDER RANKED SET SAMPLING

Table 2: The Bias, MSE and the Efficiency for all Estimators Based on Ranking of the Variable X_1

Estimator	Auxiliary variable	MSE	Efficiency	Bias
$\bar{Y}_{[n]}$	None	8311.06	1	0
\tilde{Y}^*	x_1	899.012	9.2592	0.899
\tilde{Y}^*	x_2	1250.112	6.6666	0.822
\tilde{Y}_{a^*}	x_1, x_2	378.865	22.2222	0.299
\tilde{Y}_{w^*}	x_1, x_2	581.231	14.4928	-0.675
\bar{Y}_R	x_1	4403.511	1.8903	0.533
\bar{Y}_R	x_2	2213.96	3.7594	1.228
\bar{Y}_P	x_1	33077.6	0.2513	14.532
\bar{Y}_P	x_2	30895.8	0.26903	9.217
\hat{Y}_a	x_1, x_2	8711.43	0.98231	15.533
\ddot{Y}_a	x_1, x_2	69790.4	0.11909	17.222
\check{Y}_w	x_1, x_2	612.103	13.6986	10.511

Table 3: The Bias, MSE and the Efficiency for all Estimators Based on Ranking of the Variable X_1

Estimator	Auxiliary Variable	MSE	Efficiency	Bias
$\bar{Y}_{[n]}$	None	8311.06	1	0
\tilde{Y}^*	x_1	899.012	9.2592	0.899
\tilde{Y}^*	x_2	1250.112	6.6666	0.822
\tilde{Y}_{a^*}	x_1, x_2	378.865	22.2222	0.299
\tilde{Y}_{w^*}	x_1, x_2	581.231	14.4928	-0.675
\bar{Y}_R	x_1	4403.511	1.8903	0.533
\bar{Y}_R	x_2	2213.96	3.7594	1.228
\bar{Y}_P	x_1	33077.6	0.2513	14.532
\bar{Y}_P	x_2	30895.8	0.26903	9.217
\hat{Y}_a	x_1, x_2	8711.43	0.98231	15.533
\ddot{Y}_a	x_1, x_2	69790.4	0.11909	17.222
\check{Y}_w	x_1, x_2	612.103	13.6986	10.511

RSS estimator. Some other estimators achieved higher efficiency than $\hat{Y}_{(n)}$, these estimators are: \tilde{Y}_{w^*} , \check{Y}_w , \tilde{Y}^* , and \bar{Y}_R .

The estimators achieved about the same efficiency no matter which variable was ranked on. This provides greater flexibility in choosing the variable to rank on, since some of the variables are more difficult to rank than others.

References

Ahmed, M. S. (1995) Some Estimation Procedure Using Multivariate Auxiliary Information in Sample Surveys (Ph. D. Thesis, Department of Statistics & Operations Research, Aligarh Muslim University, India).

Al-Saleh, M. Fraiwan & Al-Sharf, K. (2001). Estimation of average milk yield using ranked set sampling. *Environmetrics*, 12, 395-399.

Al-Saleh, M. Fraiwan & Al-Omari, Amer. (2002). Multi-Stage RSS. *Journal of Statistical Planning and Inference*, 102, 273-286.

Al-Saleh, M. Fraiwan and Zheng, Gang. (2002). Estimation of Bivariate Characteristics Using Ranked Set Sampling. *The Australian & New Zealand Journal of Statistics*, 44, 221-232.

Barabesi, L. and El-Sharaawi A. (2001). The efficiency of ranked set sampling for parameter estimation. *Statistics and Probability Letters*, 53, 189-199.

Cochran, W. G. (1977) *Sampling techniques*, 3rd edition (John Wiley, NY).

ESTIMATORS FOR THE MEAN UNDER RANKED SET SAMPLING

Dell, D. R. and Clutter, J. L. (1972) Ranked set sampling theory with order statistics background, *Biometrics*, 28, 545-55.

Kaur, A., Patil, G. P., Sinha, B. K. and Taillie, C. (1995) Ranked set sampling: an annotated bibliography, *Environmental and Ecological Statistics*, 2, 25-54.

Lam, K., Sinha, B.K. and Wu, Z.(1994). Estimation of parameters in two-parameter Exponential distribution using ranked set sampling, *Annals of the Institute of Statistical Mathematics*, 46(4), 723-736.

McIntyre, G. A. (1952) A method of unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Research*, 3, 385-390.

Mode, N., Conquest, L. & Marker, D. (1999). Ranked set sampling for ecological research: Accounting for the total cost of sampling. *Environmetrics*, 10, 179-194.

Samawi, H. M. & Muttalak, H. A. (1996) Estimation of ratio using rank set sampling, *Biometrical Journal*, 38, 753-764.

Stokes, S. L. & Sager, T.(1988). Characterization of ranked set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83, 374-381.

Stokes, S. L. (1977): Ranked set sampling with concomitant variables, *Communications in statistics*, A6, 1207-1211.

Takahasi K. & Wakimoto K. (1968) On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics*, 21, 249-55.

Appendix

$$\begin{aligned}
 \mathbf{g}_{1=} & \left\{ \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right] \left(2 \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \right. \right. \\
 & - 2 \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right]^2 - \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \\
 & \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right] + \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right]^2 \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right] + \\
 & \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right]^2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] + \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \\
 & \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right]^2 - \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right] + \\
 & \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right]^2 \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right] + \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right]^2 \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \\
 & \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] + \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1x_2} - \sum_{i=1}^m T_{x_1x_2[i]} \right]^2 \left. - \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \right. \\
 & \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left(\left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right]^2 + \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right]^2 \right. \\
 & + \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \left[m\sigma_{yx_1} - \sum_{i=1}^m T_{yx_1[i]} \right] + \left[m\sigma_{x_2}^2 - \sum_{i=1}^m T_{x_2[i]}^2 \right] \left[m\sigma_{x_1}^2 - \sum_{i=1}^m T_{x_1[i]}^2 \right] \\
 & \left. \left. \left[m\sigma_{yx_2} - \sum_{i=1}^m T_{yx_2[i]} \right] \right) \right\}
 \end{aligned}$$

On the BLUE of the Population Mean for Location and Scale Parameters of Distributions Based on Moving Extreme Ranked Set Sampling

Walid Abu-Dayyeh
Sultan Qaboos University
Muscat, Oman

Lana Al-Rousan
Yarmouk University
Jordan

The best linear unbiased estimator (BLUE) for the population mean under moving extreme ranked set sampling (MERSS) is derived for general location and scale parameters of distributions which generalizes Al-Odat and Al-Saleh (2001). It is compared with the sample mean of simple random sampling (SRS). The efficient sample size under the MERSS for which the BLUE estimator dominates the usual sample mean under SRS for estimating the population mean is also computed for several distributions.

Key words: Best linear unbiased estimator; location parameter; scale parameter; moving extreme ranked set sampling, simple random sampling.

Introduction

Ranked set sampling (RSS) as introduced by McIntyre (1952) is useful for cases when the variable of interest can be more easily ranked than quantified. The aim of RSS is to increase the efficiency of the sample mean as an estimator for the population mean μ . Takahasi and Wakimoto (1968) established a very important statistical foundation for the theory of RSS. They showed that the mean of the RSS is an unbiased estimator for the population mean and has smaller variance than the mean of SRS. Dell and Clutter (1972) studied the effect of ranking error on the procedure. The RSS has many statistical applications in biological and environmental studies and reliability theory (e.g. Dell & Clutter, 1972; Stokes, 1977, 1980; Mode et al., 1999; Barabesi & El-Sharaawi, 2001; Al-Saleh & Zheng, 2002; & Al-Saleh & Al-Omary, 2002). Sinha, et al., (1996) explored the concept

of RSS when the population is partially known using the parameters of normal and exponential distributions. They found that the use of knowledge of the distribution along with RSS provides improvement in estimation over SRS, as well as over nonparametric RSS. Li and Chuiv (1997) discussed the issue of the efficiency of RSS compared to SRS in many parametric estimation problems. They found an improvement in estimation of many common parameters of interest with smaller numbers of measurements compared to SRS.

RSS has been investigated extensively (see for example, Stokes, 1977; Stokes & Sager, 1988; Lam, et al., 1994; Barabesi & El-Sharaawi, 2001). Al-Saleh and Al-Kadiri (2000) introduced Double RSS to increase the efficiency of RSS estimates without increasing the set size m and Al-Saleh and Al-Omary (2002) generalized it to multistage RSS. Samawi, et al., (1996) used extreme ranked set sample (ERSS), which is easier to use than the usual RSS procedure, when the set size is large to estimate the population mean in the case of symmetric distributions. Al-Odat and Al-Saleh (2001) introduced the concept of varied set size RSS, which is coined here as Moving Extreme Ranked Set Sampling (MERSS). They investigated this modification non-parametrically and found that the

Walid Abu-Dayyeh is an associate Professor in the Department of Mathematics and Statistics at Sultan Qaboos University/Sultanate of Oman. Email: abudayyehw@yahoo.com. Lana Al-Rousan is a statistician in the Department of Statistics/Jordan. Email: lanaal211@yahoo.com.

procedure can be more efficient and applicable than the simple random sampling technique (SRS). The MERSS procedure is as follows:

1. Select m random samples of size 1, 2, 3, ..., m respectively.
2. Identify the maximum of each set by eye or by some other relatively inexpensive method without actually measuring the characteristic of interest.
3. Measure accurately the selected judgment identified maximum.
4. Repeat steps 1, 2, 3, but for the minimum.
5. Repeat the above steps r times until the desired sample size, $n = 2rm$ is obtained.

Clearly, the procedure of MERSS is easier to use than the usual RSS procedure.

Methodology

The BLUE of the Mean for Distributions with a Location Parameter

Let $\{X_{i1}^1, X_{i2}^1, \dots, X_{ii}^1\}$ and

$\{X_{i1}^2, X_{i2}^2, \dots, X_{ii}^2\}$ be simple random samples each of size i , for $i = 1, 2, \dots, m$ from a population with distribution function F and a probability density function f . Let μ and σ^2 be the mean and variance of the population respectively. If

$$y_{i1} = \text{Min}\{X_{i1}^1, X_{i2}^1, \dots, X_{ii}^1\},$$

and

$$y_{i2} = \text{Max}\{X_{i1}^2, X_{i2}^2, \dots, X_{ii}^2\},$$

$$i = 1, 2, \dots, m,$$

then

$$\{y_{11}, y_{21}, \dots, y_{m1}, y_{12}, y_{22}, \dots, y_{m2}\}$$

is a MERSS of size $2m$.

The BLUE for μ for a population can be derived with a pdf of the form:

$$f(x - \theta), \quad -\infty < \theta < \infty, \quad (2.1)$$

where f is a pdf.

Result 1

Let y_1, y_2, \dots, y_{2m} be $2m$ independent ordered statistics of simple random samples each of size less than m from an underlying distribution with a pdf as in (2.1). Then the BLUE of the population μ is then given by:

$$\hat{\mu}_{Blue} = \sum_{i=1}^{2m} \frac{1}{2\sigma_i^2 d} \{k - bt + bw c_i - tc_i\} y_i \quad (2.2)$$

where

$$k = \sum_{i=1}^{2m} \frac{c_i^2}{\sigma_i^2}, \quad t = \sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2}, \quad w = \sum_{i=1}^{2m} \frac{1}{\sigma_i^2},$$

$$d = \sum_{i=1}^{2m} \frac{1}{\sigma_i^2} \sum_{i=1}^{2m} \frac{c_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} \right)^2,$$

and C_i and σ_i^2 are the mean and the variance of Z_i respectively, where $Z_i = y_i - \theta$ and $\mu = E_{\theta} X = \theta + b$. (Note that y_1, y_2, \dots, y_{2m} are not necessarily identically distributed.)

Proof

Starting with a class of unbiased linear estimators of μ of the form

$$\hat{\mu} = \sum_{i=1}^{2m} a_i y_i, \quad (2.3)$$

implies that

$$E(\hat{\mu}) = \theta \sum_{i=1}^{2m} a_i + \sum_{i=1}^{2m} a_i c_i = \mu = \theta + b,$$

which, in turn, implies that

$$\sum_{i=1}^{2m} a_i = 1$$

and

$$\sum_{i=1}^{2m} a_i c_i = b. \quad (2.4)$$

Applying the method of the Lagrange multiplier to minimize

$$Var(\hat{\mu}) = \sum_{i=1}^{2m} a_i^2 \sigma_i^2,$$

subject to (2.4), results in:

$$a_i^* = \frac{\lambda_1 + \lambda_2 c_i}{2\sigma_i}, \quad \sum_{i=1}^{2m} a_i^* = 1, \quad \sum_{i=1}^{2m} a_i^* c_i = b,$$

$$\lambda_1 = \frac{\left\{ -\sum_{i=1}^{2m} \frac{c_i^2}{\sigma_i^2} + b \sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} \right\}}{\left(\sum_{i=1}^{2m} \frac{1}{\sigma_i^2} \sum_{i=1}^{2m} \frac{c_i^2}{\sigma_i^2} - \left\{ \sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} \right\}^2 \right)},$$

$$\lambda_2 = \frac{\left\{ -b \sum_{i=1}^{2m} \frac{1}{\sigma_i^2} + \sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} \right\}}{\left(\sum_{i=1}^{2m} \frac{1}{\sigma_i^2} \sum_{i=1}^{2m} \frac{c_i^2}{\sigma_i^2} - \left\{ \sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} \right\}^2 \right)},$$

where λ_1 and λ_2 are the Lagrange multipliers and

$$a_i^* = \frac{1}{2\sigma_i^2 d} \{k - bt + c_i bw - c_i t\}. \quad (2.5)$$

Then

$$\hat{\mu}^* = \sum_{i=1}^{2m} \frac{1}{2\sigma_i^2 d} \{k - bt + c_i bw - c_i t\} y_i, \quad (2.6)$$

is the BLUE of μ with variance

$$Var(\hat{\mu}^*) = \sum_{i=1}^{2m} \left(\frac{1}{2d\sigma_i^2} \{k - bt + c_i bw - c_i t\} \right)^2 \sigma_i^2. \quad (2.7)$$

If $y_i = y_{i1}$, for $i = 1, 2, \dots, m$ and $y_i = y_{i2}$, for $i = m + 1, m + 2, \dots, 2m$.

If $E(y_i) = E(y_{i1}) = c_{i1} + \theta$, $i = 1, 2, \dots, m$

$E(y_i) = E(y_{i2}) = c_{i2} + \theta$, $i = m + 1, m + 2, \dots, 2m$

$Var(y_i) = Var(y_{i1}) = \sigma_{i1}^2$, $i = 1, 2, \dots, m$

and

$Var(y_i) = Var(y_{i2}) = \sigma_{i2}^2$, $i = m + 1, m + 2, \dots, 2m$,

where $c_{i1} = E(u_i)$, $Var(u_i) = \sigma_{i1}^2$, and u_i is the minimum of a SRS of size i , and $c_{i2} = E(w_i)$, $Var(w_i) = \sigma_{i2}^2$ and w_i is the maximum of a SRS of size i , under $\theta = 0$. It then follows that:

$$\begin{aligned} \hat{\mu}_{MEBLUE} &= \sum_{i=1}^m \frac{1}{2\sigma_{i1}^2 d} \{k - bt + c_{i1} bw - c_{i1} t\} y_{i1} \\ &+ \sum_{i=m+1}^{2m} \frac{1}{2\sigma_{i2}^2 d} \{k - bt + c_{i2} bw - c_{i2} t\} y_{i2} \end{aligned} \quad (2.8)$$

and

$$\begin{aligned} Var(\hat{\mu}_{MEBLUE}) &= \sum_{i=1}^m \frac{1}{4d^2 \sigma_{i1}^2} (k - bt + c_{i1} bw - c_{i1} t)^2 \\ &+ \sum_{i=m+1}^{2m} \frac{1}{4d^2 \sigma_{i2}^2} (k - bt + c_{i2} bw - c_{i2} t)^2 \end{aligned} \quad (2.9)$$

Al-Odat and Al-Saleh (2001) introduced MERSS and studied the linear estimators of the form: $\sum_{i=1}^m a_i (y_{i1} + y_{i2})$. They derived the

BLUE among such linear combinations for the population mean. The BLUE derived by Al-Odat and Al-Saleh (2001) is not the BLUE estimator based on $(y_{11}, y_{21}, \dots, y_{m1}, y_{12}, y_{22}, \dots, y_{m2})$,

but the BLUE based on (k_1, k_2, \dots, k_m) where

$k_i = y_{i1} + y_{i2}$ for $i=1,2,\dots, m$. If the underlying distribution is symmetric about its mean μ , then (2.9) coincides with the results obtained by Al-Odat and Al-Saleh (2001).

The BLUE estimator based on MERSS, obtained with the sample mean based on SRS in case of uniform $U(\theta, \theta + 1)$ and $Exp(\theta, 1)$ distributions are compared. The first is symmetric about its mean $\theta + \frac{1}{2}$ and the second

is skewed to the right with mean $\theta + 1$. Both families are location parameter families of distributions, so the BLUE's are the same as given in (2.8), with $b = \frac{1}{2}$ for $U(\theta, \theta + 1)$ and

$b=1$ for $Exp(\theta, 1)$. Balakrishnan and Cohen (1990) computed the variances of the estimators in this case and in the following cases.

The estimators compared are both unbiased for μ . Therefore, they will be compared through their variances. The efficiency between two estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ is defined as:

$$eff(\hat{\mu}_2, \hat{\mu}_1) = Var(\hat{\mu}_1) [Var(\hat{\mu}_2)]^{-1}$$

The larger the efficiency, the better the estimator $\hat{\mu}_2$ will be. The efficiency of $\hat{\mu}_{MEBlue}$ with respect to the sample mean under SRS was computed for both distributions for $m = 2, \dots, 10$. The results are summarized in Tables 1 and 2. From these tables, it may be concluded that the variance of the BLUE decreases as m increases and $eff(\hat{\mu}_{MEBlue}, \bar{X}_{2m}) \geq 1$ for both distributions. Also, the efficiency is more than 2 for $m \geq 4$ in the uniform case and for $m \geq 9$ in the exponential case.

Efficiency of $\hat{\mu}_{MEBlue}$ with respect to \bar{X}_{2m}

Table 1		Table 2	
U($\theta, \theta+1$)		Exp($\theta, 1$)	
m	$eff(\hat{\mu}_{MEBlue}, \bar{X}_{2m})$	m	$eff(\hat{\mu}_{MEBlue}, \bar{X}_{2m})$
2	1.333	2	1.167
3	1.765	3	1.333
4	2.200	4	1.483
5	2.863	5	1.639
6	3.150	6	1.647
7	3.683	7	1.8397
8	4.288	8	1.996
9	4.932	9	2.087
10	5.620	10	2.177

The BLUE of the mean for distributions with a scale parameter

Let $\{y_{11}, y_{21}, \dots, y_{m1}, y_{12}, y_{22}, \dots, y_{m2}\}$ be a MERSS from a population with a pdf of the form:

$$\frac{1}{\theta} f\left(\frac{x}{\theta}\right), \theta > 0 \tag{3.1}$$

where f is a pdf. Then as shown previously, if

$$y_{i1} = \theta \min\left\{\frac{X_{i1}^1}{\theta}, \frac{X_{i2}^1}{\theta}, \dots, \frac{X_{ii}^1}{\theta}\right\},$$

then

$$E(y_{i1}) = \theta \text{Min}\left\{\frac{X_{i1}^1}{\theta}, \frac{X_{i2}^1}{\theta}, \dots, \frac{X_{ii}^1}{\theta}\right\} = C_{i1} \theta$$

where $C_{i1} = E(U_i)$ and U_i is the first order statistic of a SRS of size i from the pdf in (3.1), under $\theta = 1$. Similarly, $E(y_{i2}) = C_{i2} \theta$, for $C_{i2} = E(W_i)$ where W_i is the maximum order statistic of a SRS of size i from the pdf in (3.1),

BLUE LOCATION AND SCALE PARAMETERS OF DISTRIBUTIONS BASED ON MERSS

under $\theta = 1$. Also, $\text{Var} (y_{i1}) = \theta^2 \sigma_{i1}^2$ and $\text{Var} (y_{i2}) = \theta^2 \sigma_{i2}^2$ where σ_{i1}^2 and σ_{i2}^2 are the variances of u_i and w_i respectively, for $i = 1, 2, \dots, m$. (The BLUE of the mean of the population with pdf (3.1) proof is similar to that of Result (1) and therefore is omitted.)

Result 2

Let y_1, y_2, \dots, y_{2m} be $2m$ independent order statistics each of size less than m from an underlying distribution with a pdf as in (3.1). Then the BLUE of the population μ is given by:

$$\hat{\mu}_{Blue} = \frac{\sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2} y_i}{\sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2}} \quad (3.2)$$

with variance

$$\text{Var}(\hat{\mu}_{Blue}) = \frac{\theta^2}{\sum_{i=1}^{2m} \frac{c_i}{\sigma_i^2}} \quad (3.3)$$

where $\mu = b \theta$ and $b = E_{\theta=1} X$.

The BLUE of μ using MERSS is given by:

$$\hat{\mu}_{MEBlue} = \frac{\sum_{i=1}^m \frac{c_{i1}}{\sigma_{i1}^2} y_{i1} + \sum_{i=m+1}^{2m} \frac{c_{i2}}{\sigma_{i2}^2} y_{i2}}{\sum_{i=1}^m \frac{c_{i1}}{\sigma_{i1}^2} + \sum_{i=m+1}^{2m} \frac{c_{i2}}{\sigma_{i2}^2}} \quad (3.4)$$

and

$$\text{Var}(\hat{\mu}_{MEBlue}) = \frac{\theta^2}{\sum_{i=1}^m \frac{c_{i1}}{\sigma_{i1}^2} + \sum_{i=m+1}^{2m} \frac{c_{i2}}{\sigma_{i2}^2}} = \frac{\theta^2}{\left\{ \sum_{i=1}^m \frac{c_{i1}}{\sigma_{i1}^2} + \sum_{i=m+1}^{2m} \frac{c_{i2}}{\sigma_{i2}^2} \right\}} \quad (3.5)$$

Comparing the BLUE estimator based on MERSS with the sample mean based on SRS in

case of uniform $\text{Exp}(\theta)$ and $U(0, \theta)$ distributions. The first is skewed to the right with mean θ and the second is symmetric about its mean $\frac{\theta}{2}$. So, the BLUE's are the same as given in (3.2). The estimators are unbiased and therefore are compared using their variances for $m = 2 \dots 10$. The results are summarized in Tables (3) and (4). Similar conclusions to those presented for Tables (1) and (2) can be given.

Efficiency of $\hat{\mu}_{MEBlue}$ with respect to \bar{X}_{2m}

Table 3		Table 4	
Exp(θ)		U(0, θ)	
m	eff($\hat{\mu}_{MEBlue}, \bar{X}_{2m}$)	m	eff($\hat{\mu}_{MEBlue}, \bar{X}_{2m}$)
2	1.200	2	1.331
3	1.380	3	1.815
4	1.540	4	2.264
5	1.690	5	2.955
6	1.820	6	3.574
7	1.950	7	4.593
8	2.070	8	5.713
9	2.190	9	6.935
10	2.300	10	8.261

Saving by using MERSS to estimate the population mean

Measuring the units of a sample costs money, time, and effort. The previous tables show that the BLUE for estimating the population mean μ under MERSS is more efficient (less variance) than the sample mean of SRS, which is usually used for estimating μ . Therefore, $\hat{\mu}_{MEBlue}$ will be as good as \bar{X}_{2m} by using a smaller number of observations which will result in saving time, money and effort. Table (5) shows the smallest $2m$ such that the variance of the BLUE under MERSS using $2m$ observations is smaller than the variance of the sample mean of SRS using a specified sample size in case of the normal, logistic, uniform, and exponential distributions. The first two

distributions are location parameter families of distributions while the other two are scale parameter families.

Table (5), shows how the BLUE, under MERSS for estimating the population mean, requires a smaller number of observations than \bar{X}_{2m} based on SRS. This indicates a reduction in the sample size required for estimating the mean. As m increases then the savings will be greater for all the cases studied. According to Table (5), the savings in sample sizes range from 0% to 70%. For example, $\hat{\mu}_{MEBlue}$ based on 12 observations is better than \bar{X}_{2m} based on 40 observations in the case of $U(\theta, \theta + 1)$ for estimating the mean, resulting in saving 70% of the sample size from using the MERSS compared to SRS.

Conclusion

If ordering the data can be done more easily than quantifying it, then the BLUE under MERSS can be used instead of the mean of SRS for estimating the population mean because the BLUE under MERSS provides better results than the mean of SRS with fewer numbers of observations.

References

Al-Odat, M. T., & Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, 10(2), 137-146.

Al-Saleh, M. F., & AL-Hadramy, S. (2003a). Estimation of the mean of the normal distribution using moving extreme ranked set sampling. *Environmetrics*, 14(7), 651-664.

Al-Saleh, M. F., and AL-Hadramy, S. (2003b). Estimation of the mean of the exponential distribution using moving extreme ranked set sampling. *Statistical Papers*, 44, 367-382.

Al-Saleh, M. F., & Al-Kadiri, M. (2000). Double ranked set sampling. *Statistics and Probability Letters*, 48, 205-212.

Al-Saleh, M. F., & Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inferences*, 102, 273-286.

Al-Saleh, M. F., & Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *The Australian and New Zealand Journal of Statistics*, 44(2), 221-232.

Arnold, B. C, Balakrishnan, N., & Nagaraja, H. N. (1992). A first course in order statistics. *New York: John Wiley & Sons. Inc.*

Balakrishnan, N., & Cohen, A. (1990). Order statistics and inference, estimation method. *New York: Academic Press, Inc.*

Barabesi, L., & El-Sharaawi, A. (2001). The efficiency of ranked set sampling for parameter estimation. *Statistics and Probability Letters*, 53, 189-199.

Dell, T. R., & Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28, 545-555.

Fei, H., Sinha, B. K., & Wu, Z. (1994). Estimation of parameters in two-parameter Weibull and extreme-value distributions using ranked set sampling. *Journal of Statistical Research*, 28, 149-161.

Lam, K., Sinha, B. K., & Wu, Z. (1994). Estimation of parameters in two-parameter exponential distribution using ranked set sample. *Annals of the Institute of Statistical Mathematics*, 46, 723-736.

Li, D., & Chuiv, N. (1997). On the efficiency of ranked set sampling strategies in parametric estimation. *Calcutta Statistical Association Bulletin*, 47, 185-186.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385-390.

Mode, N., Conquest, L. & Marker, D. (1999). Ranked set sampling for ecological research: Accounting for the total cost of sampling. *Environmetrics*, 10, 179-194.

Muttalak, H.A. (1997). Median ranked set sampling. *Journal of Applied Statistical Sciences*, 6(4), 245-255.

Patil, G., Sinha, A., & Taillie, C. (1999). Ranked set sampling: Bibliography. *Environmental and Ecological Statistics*, 6, 91-98.

BLUE LOCATION AND SCALE PARAMETERS OF DISTRIBUTIONS BASED ON MERSS

Samawi, H., Ahmed, M., & Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrics*, 30, 577-586.

Sinha, B. K., Sinha, B. K., & Purkayastha, S. (1996). On some aspects of ranked set sampling for estimation of normal and exponential parameters. *Statistics and Decisions*, 14, 223-240.

Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics*, 36, 35-42.

Stokes, S.L.(1977). Ranked set sampling with concomitant variables. *Communications in Statistics- Theory and Methods A6*, 1207-1211.

Stokes, S. L. (1976). *An investigation of the consequences of ranked set sampling*. Ph.D. Thesis, Department of Statistics, University of North Carolina, Chapel Hill, NC.

Stokes, S. L., & Sager, T. W. (1988). Characterization of ranked set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, 83, 374-381.

Takahasi, K. (1970). Practical note on estimation of population means based on samples stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 22, 421-428.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.

Zheng, G., & Al-Saleh, M. F. (2002). Modified Maximum Likelihood Estimator based on ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 54, 641-658.

Table 5: Efficiency of the Smallest Number of Observations for MERSS Compared to the SRS of Size 2m

SRS 2m	MERSS					
	$N(\theta,1)$	$L(\theta,1)$	$Exp(\theta,1)$	$Exp(\theta)$	$U(\theta, \theta+1)$	$U(0, \theta)$
2	2	2	2	2	2	2
4	4	4	4	4	4	4
6	6	6	6	6	6	6
8	6	6	8	6	6	6
10	8	8	8	8	6	6
12	8	10	10	8	8	8
14	10	10	10	10	8	8
16	10	12	10	10	8	8
18	12	12	12	12	10	10
20	12	14	12	12	10	10
22	14	14	14	14	10	10
24	14	16	14	14	10	10
26	14	16	16	14	10	10
28	16	18	16	16	10	10
30	16	20	16	16	12	12
32	16	20	18	16	12	12
34	18	21	18	18	12	12
36	18	21	18	18	12	12
38	19	22	20	18	14	12
40	19	22	20	20	14	12

Bayesian Inference on the Variance of Normal Distribution Using Moving Extremes Ranked Set Sampling

Said Ali Al-Hadhrami

College of Applied Sciences, Nizwa, Oman

Amer Ibrahim Al-Omari

Al al-Bayt University, Mafrqa, Jordan

Bayesian inference of the variance of the normal distribution is considered using moving extremes ranked set sampling (MERSS) and is compared with the simple random sampling (SRS) method. Generalized maximum likelihood estimators (GMLE), confidence intervals (CI), and different testing hypotheses are considered using simple hypothesis versus simple hypothesis, simple hypothesis versus composite alternative, and composite hypothesis versus composite alternative based on MERSS and compared with SRS. It is shown that modified inferences using MERSS are more efficient than their counterparts based on SRS.

Key words: Moving extremes ranked set sampling (MERSS), confidence interval, test hypothesis, Bayesian approach.

Introduction

Ranked set sampling (RSS) for estimating a population mean was suggested by McIntyre (1952) as a cost efficient alternative to simple random sampling (SRS) if the units of a sample can be easily ranked according to the variable of interest rather than actual measurements. The RSS involves randomly selecting m^2 units from the population and randomly allocating them into m sets, each of size m . The m units of each sample are ranked visually (or by any inexpensive method) with respect to the variable of interest. From the first set of m units, the smallest unit is measured. From the second set of m units, the second smallest unit is measured, the process continues until the largest unit is measured from the m^{th} set of m units. Repeating

the process r times results in a set of size mr from initial m^2r units.

Takahasi and Wakimoto (1968) provided the mathematical theory for RSS. Muttlak (1996) proposed pair ranked set sampling instead of RSS, and Samawi, et al. (1996) suggested using extreme ranked set sampling to estimate the population mean. Muttlak (1997) also suggested using median ranked set sampling. Al-Saleh and Al-Kadird (2000) considered double ranked set sampling (DRSS). Al-Saleh and Al-Omari (2002) generalized DRSS to multistage RSS. Muttlak (2003) proposed quartile ranked set sampling. Weighted modified RSS was put forward by Muttlak and Abu-Dayyeh (2004).

Al-Odat and Al-Saleh (2001) introduced the concept of varied set size RSS. They investigated this modification non-parametrically and found that the procedure can be more efficient than the simple random sampling technique. Al-Saleh and Al-Hadhrami (2003a) considered the work of Al-Odat and Al-Saleh (2001) and investigated parametrically the mean of exponential distribution; they coined their method of moving extremes ranked set sampling (MERSS). Investigation of the mean of the normal distribution under MERSS was considered by Al-Saleh and Al-Hadhrami

Said Ali Al-Hadhrami is an Assistant Professor in the Department of Mathematics, College of Applied Sciences. Email him at: abur1972@yahoo.co.uk. Amer Ibrahim Al-Omari is an Assistant Professor in the Department of Mathematics. Email him at: alomari_amer@yahoo.com.

(2003b). They showed that the suggested estimators of the population mean are unbiased and more efficient than those based on SRS. Abu-Dayyeh and Al-Sawi (2007) studied the scale parameter of exponential distribution based on MERSS. (For more about RSS see Chen, et al., 2004; Al-Saleh & Al Ananbeh, 2007; Al-Omari & Jaber, 2008; Al-Nasser, 2007; Tseng & Wu, 2007; and Balakrishnan & Li, 2008.)

Methodology

The MERSS General Process

The MERSS can be described as follows:

- Step 1: Select m random samples sized 1, 2, 3, ..., m , respectively.
- Step 2: Identify the maximum of each set by eye or by some other inexpensive method, without actually measuring the characteristic of interest.
- Step 3: Accurately measure the selected judgment identified maxima.
- Step 4: Repeat Steps 1, 2, 3 but for the minimum.
- Step 5: Repeat the above steps r times until the desired sample size, $n = 2rm$ is obtained. The sample of these units is called moving extremes ranked set sample (MERSS).

For one cycle, let

$$\left\{ \begin{matrix} X_{m:m}, X_{m-1:m-1}, X_{m-2:m-2}, \dots, \\ X_{1:1}, Y_{1:m}, Y_{1:m-1}, Y_{1:m-2}, \dots, Y_{1:1} \end{matrix} \right\}$$

be a MERSS from a normal distribution mean μ and variance σ^2 . If judgment ranking is perfect, then for $i = 1, 2, \dots, m$, $X_{i:i}$ has the same density as the i^{th} order statistic of a SRS of size i from $f(x; \theta)$, i.e., $X_{i:i}$ has the density:

$$f_{i:i}(x) = if(x; \theta)[F(x; \theta)]^{i-1}. \quad (2.1)$$

In addition, $Y_{1:i}$ has the same density as the first order statistic of a SRS of size i from $f(y; \theta)$, i.e., $Y_{1:i}$ has the density

$$f_{1:i}(y) = if(y; \theta)[1 - F(y; \theta)]^{i-1}, \quad (2.2)$$

and the likelihood function of θ is given by

$$L(\theta) = \prod_{i=1}^m if(x_{i:i}, \theta)[F(x_{i:i}, \theta)]^{i-1} if(y_{1:i}, \theta)[1 - F(y_{1:i}, \theta)]^{i-1}. \quad (2.3)$$

Assuming the random variable X is normally distributed with mean μ and variance σ^2 , then the probability density function (pdf) of X is given by

$$\begin{aligned} f_X(x, \theta) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right), \end{aligned} \quad -\infty < x < \infty, \quad (2.4)$$

and the cumulative distribution function is

$$F_X(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(u-\theta)^2}{2\sigma^2}} du = \Phi\left(\frac{x-\theta}{\sigma}\right), \quad (2.5)$$

where ϕ and Φ are the density and cumulative distribution of the standard normal distribution, respectively.

Generalized Maximum Likelihood Estimator (GMLE)

In the case of estimating the population variance, the information number is proportional to $1/\sigma^2$ (see Al-Hadhrami, et al., 2009), allowing the Jeffery prior for σ to be written as

$\pi(\sigma)\alpha 1/\sigma$. The posterior distribution for σ is then given by

$$h(\sigma | x, y) \propto \frac{1}{\sigma} \prod_{i=1}^m i^2 \left(\frac{1}{\sigma}\right) \phi(z_i) \left(\frac{1}{\sigma}\right) \phi(w_i) [\Phi(z_i)]^{i-1} [1-\Phi(w_i)]^{i-1}. \quad (3.1)$$

The log of both sides of (3.1) is

$$L^*(\sigma) = C - \log \sigma + \sum_{i=1}^m \left[\log \left(\frac{1}{\sigma}\right) \phi(z_i) + \log \left(\frac{1}{\sigma}\right) \phi(w_i) \right] + \sum_{i=1}^m (i-1) \left[\log \Phi(z_i) + \log(1-\Phi(w_i)) \right], \quad (3.2)$$

where C is a constant. The first derivative of (3.2) is given by

$$\frac{\partial L^*}{\partial \sigma} = -\frac{1}{\sigma} + \left[\frac{1}{\sigma} \sum_{i=1}^m z_i^2 - \frac{m}{\sigma} \right] + \left[\frac{1}{\sigma} \sum_{i=1}^m w_i^2 - \frac{m}{\sigma} \right] + \sum_{i=1}^m (i-1) \left[\frac{\left(\frac{w_i}{\sigma}\right) \phi(w_i)}{1-\Phi(w_i)} - \frac{\left(\frac{z_i}{\sigma}\right) \phi(z_i)}{\Phi(z_i)} \right].$$

Let $\frac{\partial L^*}{\partial \sigma} = 0$, then the likelihood equation is defined as

$$-\frac{2m+1}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^m [z_i^2 + w_i^2] + \sum_{i=1}^m (i-1) \left[\frac{w_i \phi(w_i)}{1-\Phi(w_i)} - \frac{z_i \phi(z_i)}{\Phi(z_i)} \right] = 0,$$

which may be written as

$$1 - \frac{1}{2m+1} \sum_{i=1}^m (z_i^2 + w_i^2) - \frac{1}{2m+1} \sum_{i=1}^m (i-1) \left[\frac{w_i \phi(w_i)}{1-\Phi(w_i)} - \frac{z_i \phi(z_i)}{\Phi(z_i)} \right] = 0. \quad (3.3)$$

If the second derivative of the likelihood with respect to σ is negative at the solution of $\frac{\partial L^*}{\partial \sigma} = 0$, then this solution is the GMLE of σ .

The second derivative of the log likelihood with respect to σ is

$$\frac{\partial^2 L^*}{\partial \sigma^2} = T_1 + T_2 + T_3, \quad (3.4)$$

where

$$T_1 = \frac{2m+1}{\sigma^2} - \frac{3}{\sigma^2} \sum_{i=1}^m (z_i^2 + w_i^2) + \frac{3}{\sigma^2} \sum_{i=1}^m (i-1) \left[\frac{z_i \phi(z_i)}{\Phi(z_i)} - \frac{\phi(w_i)}{1-\Phi(w_i)} \right] = \frac{2m+1}{\sigma^2} \left[1 - \frac{3}{2m+1} \sum_{i=1}^m (z_i^2 + w_i^2) + \frac{3}{2m+1} \sum_{i=1}^m (i-1) \left(\frac{z_i \phi(z_i)}{\Phi(z_i)} - \frac{\phi(w_i)}{1-\Phi(w_i)} \right) \right].$$

The value of T_1 at the solution of Equation (3.3) is

$$T_1 = -2 \left(\frac{2m+1}{\sigma^2} \right),$$

and

$$T_2 = \sum_{i=1}^m (i-1) \left[\left(-\frac{z_i^2}{\sigma^2} \right) \frac{\phi(z_i)}{\Phi(z_i)} \left(z_i + \frac{\phi(z_i)}{\Phi(z_i)} \right) \right],$$

$$T_3 = \sum_{i=1}^m (i-1) \left[\frac{w_i^2}{\sigma^2} \left(\frac{\phi(w_i)}{1-\Phi(w_i)} \right) \left(w_i - \frac{\phi(w_i)}{1-\Phi(w_i)} \right) \right]$$

which are both negative. Therefore,

BAYESIAN INFERENCE USING RANKED SET SAMPLING

$\frac{\partial^2 L^*}{\partial \sigma^2} < 0$. Thus, the GMLE of σ is the solution of Equation (3.3) and the GMLE of variance is the square of this solution. Note that the GMLE of σ using SRS, when μ is known is given by:

$$\hat{\sigma}_{SRS} = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n+1}}. \quad (3.5)$$

As shown in Table (1), the GMLE using MERSS is more efficient than its counterparts based on SRS, and the efficiency increases as the sample size increases.

Table 1: The Efficiency of $eff = MSE_{SRS} / MSE_{MERSS}$, $\sigma \sim \pi(\sigma) \propto 1/\sigma$, and $X \sim N(0,1)$

m	MSE_{SRS}	M_{MERSS}	eff
3	0.2721	0.2405	1.1313
5	0.1740	0.1230	1.4150
7	0.1304	0.0769	1.6951
11	0.0820	0.0375	2.1848
14	0.0661	0.0251	2.6261

Confidence Interval

From the sampling distribution of the variance, Table 2 shows the interval width (IW), lower bound (LB), upper bound (UB), and the approximated two-sided 95% confidence intervals (CI) for the variance of the normal distribution $N(3,1)$ using both MERSS and SRS methods. Table 3 shows the approximated two-sided 95% confidence intervals (CI) for the variance of $N(4,4)$ based on MERSS and SRS.

Based on Tables 2 and 3, it may be noted that the intervals using MERSS are shorter than that those based on SRS. Also, the width of the intervals becomes shorter as the set size increases. The width also depends on the population variance; the smaller the variance, the smaller the width.

Table 2: 95% Confidence Intervals for σ^2 of the Normal Distribution, $N(3,1)$, Using MERSS and SRS

CI for σ^2 using SRS			
m	IW	LB	UB
3	2.3216	0.1676	2.4892
7	1.5719	0.3810	1.9530
10	1.2898	0.4701	1.7599
15	1.0330	0.5456	1.5787
CI for σ^2 using MERSS			
m	IW	LB	UB
3	1.8525	0.3952	2.2477
7	1.1120	0.4939	1.6059
10	0.8358	0.6188	1.4546
15	0.5628	0.7352	1.2981

Table 3: 90% Confidence Intervals for the Variance of the Normal Distribution, $N(4,4)$, Using MERSS and SRS with MLE

CI for σ^2 using SRS			
m	IW	LB	UB
3	8.0225	0.9010	8.9235
7	4.4965	2.0174	6.5139
10	3.9709	2.2040	6.1749
15	3.2805	2.4799	5.7605
CI for σ^2 using MERSS			
m	IW	LB	UB
3	6.3062	1.9864	8.2926
7	3.2390	2.5284	5.7675
10	2.3814	2.8574	5.2388
15	1.8909	3.1147	5.0056

Testing Hypothesis

Once a confidence interval about the parameter is obtained, a test hypothesis about this parameter can be constructed. For a two-sided hypothesis the two-sided confidence interval may be used and the upper or lower

bound confidence interval is for one-sided hypotheses with same significance level.

Consider the test hypothesis about the variance σ^2 of the normal distribution with known mean based on Bayesian paradigm when the sample is drawn using MERSS. The decision is based on the Bayes factor which is of the form

$$B = \frac{p_0 / p_1}{\pi_0 / \pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0}, \quad (4.1)$$

where

$\pi_0 = p(\theta \in \Theta_0)$: Prior probability for $\theta \in \Theta_0$.

$\pi_1 = p(\theta \in \Theta_1)$: Prior probability for $\theta \in \Theta_1$.

$P_0 = P(\theta \in \Theta_0 | x)$: Posterior probability for $\theta \in \Theta_0$.

$P_1 = P(\theta \in \Theta_1 | x)$: Posterior probability for $\theta \in \Theta_1$.

π_0 / π_1 : Prior odds on H_0 versus H_1 .

p_0 / p_1 : Posterior odds on H_0 versus H_1 .

Two Simple Hypotheses

Consider testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 = \sigma_1^2$, where σ^2 is the variance of a normal distribution with known mean, μ . The Bayes factor in this case is $B = p(x, y | \sigma_0^2) / p(x, y | \sigma_1^2)$ which can be written for a sample from a normal distribution using MERSS as

$$B = \frac{\prod_{i=1}^m i^2 \left(\frac{f(x_i; \mu, \sigma_0^2) F^{i-1}(x_i; \mu, \sigma_0^2)}{f(y_i; \mu, \sigma_0^2) [1 - F(y_i; \mu, \sigma_0^2)]^{i-1}} \right)}{\prod_{i=1}^m i^2 \left(\frac{f(x_i; \mu, \sigma_1^2) F^{i-1}(x_i; \mu, \sigma_1^2)}{f(y_i; \mu, \sigma_1^2) [1 - F(y_i; \mu, \sigma_1^2)]^{i-1}} \right)} \quad (4.2)$$

To test the null hypothesis, 1,000 numerical comparisons were made between MERSS and SRS. Results for tests of rejection of the true null hypothesis are summarized in Tables 4 and 5 for two normal distributions $N(4,1)$ and $N(-6,3)$, respectively, using SRS and MERSS methods.

Tables 4 and 5 show that the error in rejecting the null hypothesis using MERSS is less than the error when using SRS; the error in rejecting the true hypothesis also becomes smaller as the sample size increases. In addition, the error becomes smaller as the alternative moves farther from the value assumed for the null hypothesis.

Simple Null Hypothesis versus Composite Hypothesis

Next a simple hypothesis was tested against a composite hypothesis about the variance of normal distribution using MERSS. That is $H_0 : \sigma^2 = \sigma_0^2$ was tested against $H_1 : \sigma^2 \neq \sigma_0^2$ when the population mean was known. The following Bayes factor was used

$$B = \frac{p(x, y | \sigma_0^2)}{\int_{\sigma^2 \neq \sigma_0^2} p(x, y | \sigma^2) \pi(\sigma^2) d\sigma^2}, \quad (4.3)$$

where

$$\begin{aligned} p(x, y | \sigma_0^2) &= \prod_{i=1}^m i^2 \left(\frac{f(x_i; \sigma_0^2) F^{i-1}(x_i; \sigma_0^2)}{f(y_i; \sigma_0^2) [1 - F(y_i; \sigma_0^2)]^{i-1}} \right) \\ &= \sum_{k_1=0}^0 \sum_{k_2=1}^1 \dots \sum_{k_m=0}^{m-1} \left\{ \prod_{i=1}^m [a(i, k_i) G(x, y, \sigma_0^2, i, k_i)] \right\}, \\ G(x, y, \sigma_0^2, i, k_i) &= F^{i-1}(x_i; \sigma_0^2) F^{k_i}(y_i; \sigma_0^2) f(x_i; \sigma_0^2) f(y_i; \sigma_0^2), \end{aligned}$$

and

$$a(i, k_i) = i^2 (-1)^{k_i} \binom{i-1}{k_i}, \quad k_i = 0, 1, 2, \dots, i-1.$$

BAYESIAN INFERENCE USING RANKED SET SAMPLING

Table 4: Comparison Between MERSS and SRS When a Simple Hypothesis about the Variance of the Normal Distribution, $N(4,1)$ was Tested 1,000 Times

Number of Rejections of the Null Hypothesis						
H_1	$m = 3$		$m = 6$		$m = 15$	
	MERSS	SRS	MERSS	SRS	MERSS	SRS
1.2	287	302	252	284	111	218
1.4	223	237	123	163	14	85
1.6	151	154	62	92	0	40
1.8	88	131	28	75	0	7
2	67	79	16	40	0	6
2.2	43	69	6	18	0	2
2.4	37	47	5	15	0	0

Table 5: Comparison Between MERSS and SRS When a Simple Hypothesis about the Variance of the Normal Distribution, $N(-6,3)$ was Tested 1,000 Times

Number of Rejections of the Null Hypothesis						
H_1	$m = 6$		$m = 10$		$m = 15$	
	MERSS	SRS	MERSS	SRS	MERSS	SRS
3.2	378	396	363	380	327	346
3.4	317	321	243	307	167	290
3.6	245	271	172	231	107	228
3.8	213	251	122	186	54	159
4	173	205	83	172	23	109
4.2	119	141	59	129	8	79
4.4	99	127	32	95	2	52
4.6	82	120	22	71	1	32
4.8	53	91	6	63	0	31

Also,

$$\int_{\sigma^2 \neq \sigma_0^2} p(x, y | \sigma^2) \pi(\sigma^2) d\sigma^2 = \left(\begin{array}{c} \sum_{k_1=0}^0 \sum_{k_2}^1 \dots \sum_{k_m=0}^{m-1} \prod_{i=1}^m a(i, k_i) \\ \int_{\sigma^2 \neq \sigma_0^2} \prod_{i=1}^m G(x, y, \sigma^2, i, k_i) \pi(\sigma^2) d\sigma^2 \end{array} \right),$$

therefore, the Bayes factor can be written as:

$$B = \frac{\sum_{k_1=0}^0 \sum_{k_2}^1 \dots \sum_{k_m=0}^{m-1} \left\{ \prod_{i=1}^m \left[a(i, k_i) G(x, y, \sigma_0^2, i, k_i) \right] \right\}}{\left(\begin{array}{c} \sum_{k_1=0}^0 \sum_{k_2}^1 \dots \sum_{k_m=0}^{m-1} \prod_{i=1}^m a(i, k_i) \\ \int_{\sigma^2 \neq \sigma_0^2} \prod_{i=1}^m G(x, y, \sigma^2, i, k_i) \pi(\sigma^2) d\sigma^2 \end{array} \right)} \quad (4.4)$$

Using Monte Carlo methods, an approximation for the denominator of the Bayes factor is given by

$$\int_{\sigma^2 \neq \sigma_0^2} p(x, y | \sigma^2) \pi(\sigma^2) d\sigma^2 = \frac{1}{r} \sum_{k_1=0}^0 \sum_{k_2=0}^1 \dots \sum_{k_m=0}^{m-1} \sum_{i=1}^r \left[\prod_{i=1}^m a(i, k_i) \right] \prod_{i=1}^m G(x, y, \sigma_i^2, i, k_i) \quad (4.5)$$

If the underlying distribution is $N(2, 1)$, assuming that $\pi_0 = \pi_1 = 0.5$, the test is executed 1,000 times using computer simulation using SRS and MERSS for $m = 5, 10, 15$; results are presented in Table 6 based on the constant prior.

Table 6: Numerical Comparison Between MERSS and SRS when Testing Hypothesis about the Variance of the Normal Distribution

Method	Number of rejections the null hypothesis while it is true		
	$m = 5$	$m = 10$	$m = 15$
MERSS	300	168	119
SRSS	384	278	206

From Table 6, it is observed that the error in testing the hypothesis using MERSS is less than the error when using SRS, also the error becomes smaller as sample size increases.

Composite Null Hypothesis versus Composite Alternative Hypothesis

If the null and alternative hypotheses are composite, the Bayes factor

$$B = \frac{\int_{\sigma^2 \in \Theta_0} p(x | \sigma_0^2) \pi_0(\sigma^2) d\sigma^2}{\int_{\sigma^2 \in \Theta_1} p(x | \sigma_1^2) \pi_1(\sigma^2) d\sigma^2}, \quad (4.6)$$

may be used, where

$$\int_{\sigma^2 \in \Theta_0} p(x, y | \sigma_0^2) \pi_0(\sigma^2) d\sigma^2 = \left(\sum_{k_1=0}^0 \sum_{k_2=0}^1 \dots \sum_{k_m=0}^{m-1} \prod_{i=1}^m a(i, k_i) \int_{\sigma^2 \in \Theta_0} \prod_{i=1}^m G(x, y, \sigma_0^2, i, k_i) \pi_0(\sigma^2) d\sigma^2 \right)$$

and

$$\int_{\sigma^2 \in \Theta_1} p(x, y | \sigma_1^2) \pi_1(\sigma^2) d\sigma^2 = \left(\sum_{k_1=0}^0 \sum_{k_2=0}^1 \dots \sum_{k_m=0}^{m-1} \prod_{i=1}^m a(i, k_i) \int_{\sigma^2 \in \Theta_1} \prod_{i=1}^m G(x, y, \sigma_1^2, i, k_i) \pi_1(\sigma^2) d\sigma^2 \right)$$

with

$$G(x, y, \sigma^2, i, k_i) = F^{i-1}(x_i; \sigma^2) F^{k_i}(y_i; \sigma^2) f(x_i; \sigma^2) f(y_i; \sigma^2)$$

where

$$a(i, k_i) = i^2 (-1)^{k_i} \binom{i-1}{k_i}, \quad k_i = 0, 1, 2, \dots, i-1.$$

Suppose that the hypothesis to be tested is a one-sided hypothesis $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$. For example, let $\pi_0 = \pi_1 = 0.5$, $m = 5, 10, 15$, $H_0 : \sigma^2 \leq 9$ versus $H_1 : \sigma^2 > 9$, and assume that the hypothesis is tested 1,000 times. Table 7 shows the simulation comparison between MERSS and SRS based on Bayes factors.

Table 7 indicates that the error in rejecting the null hypothesis using MERSS is less than using SRS based on the same sample size. Also, the error decreases as the sample size increases. Furthermore, because $H_0 : \sigma^2 \leq 9$, the error decreases as the true value moves farther from 9.

BAYESIAN INFERENCE USING RANKED SET SAMPLING

Table 7: Results for Testing $H_0 : \sigma^2 \leq 9$ versus $H_1 : \sigma^2 > 9$ Using SRS and MERSS

The Population Variance	Number of Rejections of the Null Hypothesis					
	$m = 5$		$m = 10$		$m = 15$	
	MERSS	SRS	MERSS	SRS	MERSS	SRS
$\sigma^2 = 7.84$	379	418	253	339	150	289
$\sigma^2 = 6.76$	246	317	112	185	41	177
$\sigma^2 = 5.76$	159	194	31	107	4	56
$\sigma^2 = 4.84$	16	65	5	43	0	18
$\sigma^2 = 4$	19	72	0	13	0	0

Conclusion

Bayesian inferences regarding the population variance of the normal distribution were considered based on the MERSS method. Results indicate that the confidence intervals based on MERSS are shorter than those from SRS. These intervals will be shorter as the set size and the width increases, and they depend on the population variance. For the hypothesis testing considered in this study, it was shown that the error in rejecting the null hypothesis using MERSS is less than the error observed when using SRS.

References

Abu-Dayyeh, W., & Abu-Sawi, E. (2009). Modified inference about the mean of exponential distribution using moving extreme ranked set sampling. *Statistical Papers*, 50, 249-259.

Al-Hadhrani, S. A., Al Omari, A. E., & Al-Saleh, M. F. (2009). Estimation of standard deviation of normal distribution using moving extreme ranked set sampling. *PWASET*, 37.

Al-Nasser, A. (2007). L ranked set sampling: A generalization procedure for robust visual sampling. *Communication in Statistics: Simulation and Computation*, 36, 33-43.

Al-Odat, M. T., & Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, 10, 137-146.

Al-Omari, A. I., & Jaber, K. (2008). Percentile double ranked set sampling. *Journal of Mathematics and Statistics*, 4, 60-64.

Al-Saleh, M. F., & Al-Ananbeh, A. M. (2007). Estimation of the means of the bivariate normal using moving extreme ranked set sampling with concomitant variable. *Statistical Papers*, 48(2), 179-195.

Al-Saleh, M. F., & Al-Hadhrani, S. A. (2003a). Estimation of the mean of exponential distribution using moving extremes ranked set sampling. *Statistical Papers*, 44, 367-382.

Al-Saleh, M. F., & Al-Hadhrani, S. A. (2003b). Parametric estimation for the location parameter for symmetric distributions using moving extremes ranked set sampling with application to trees data. *Environmetrics*, 14(7), 651-664.

Al-Saleh, M. F., & Al-Kadiri, M. (2000). Double ranked set sampling. *Statistics and Probability Letters*, 48, 205-212.

Al-Saleh, M. F., & Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inferences*, 102, 31-44.

Balakrishnan, N., & Li, T. (2008). Ordered ranked set samples and applications to inference. *Journal of Statistical Planning and Inference*, 138, 3512-3524.

Chen, Z., Bai, Z. D., & Sinha, B. K. (2004). Ranked set sampling: Theory and Application. NY: Springer-Verlag.

McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets. *Australian Journal Agricultural Research*, 3, 385-390.

Muttlak, H. A. (2003). Investigating the use of quartile ranked set sampling for estimating the population mean. *Applied Mathematics and Computation*, 146, 437-443.

Muttlak, H. A. (1997). Median ranked set sampling. *Journal of Applied Statistical Sciences*, 6, 245-255.

Muttlak, H. A. (1996). Pair ranked set sampling. *The Biometrical Journal*, 38, 879-885.

Muttlak, H. A., & Abu-Dayyeh, W. (2004). Weighted modified ranked set sampling. *Applied Mathematics and Computation*, 151, 645-657.

Samawi, H., Abu-Dayyeh, W., & Ahmed, S. (1996). Extreme ranked set sampling. *The Biometrical Journal*, 30, 577-586.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals Institute of Statistics and Mathematics*, 20, 1-31.

Tseng, Y., & Wu, S. (2007). Ranked set sample based tests for normal and exponential means. *Communications in Statistics: Simulation and Computation*, 36, 761-782.

Estimating Task Duration in PERT using the Weibull Probability Distribution

Edward L. McCombs
The Charles Machine Works
Perry, Oklahoma

Matthew E. Elam
Texas A&M University
Commerce

David B. Pratt
Oklahoma State University

The Weibull probability distribution can be used as an alternative model for task time estimates in the PERT estimating methodology. It has the same advantages as the traditional beta distribution for this application. It has additional benefits, however, that make it a preferred option.

Key words: PERT; Weibull Probability Distribution; Beta Probability Distribution; Pearson Skew Plot.

Introduction

Malcolm, Roseboom, Clark, and Fazar (1959) published the project time estimating methodology that they developed for Project PERT (Program Evaluation Research Task) under the Polaris Ballistic Missile Program. The development of their methodology was motivated by the fact that there was little or no historical data available upon which to base estimates of task durations. In subsequent years, this methodology has been applied in wide variety of fields. However, various authors have identified five significant issues with PERT (e.g., Cottrell, 1999; Premachandra, 2001; Pleguezuelo et al., 2003):

1. Accurately estimating the optimistic, most likely and pessimistic durations of an activity is, in general, difficult.
2. The calculated mean and variance of the specific activity durations are estimates of the actual mean and variance.

3. The beta distribution is assumed to provide an adequate model for activity durations.
4. PERT focuses on the critical path when computing project completion time probabilities.
5. The methodology requires that multiple time estimates be developed. These estimates can be costly.

Focus on items two and three in the above list. Specifically, consider the Weibull distribution as an alternative to the traditionally used beta distribution. It is shown, among other advantages, that the Weibull distribution does not require approximations for the mean and variance, as does the beta distribution.

Beta Probability Distribution

The beta probability distribution has traditionally been used as the distribution of choice in PERT analyses based on the following advantages (Fente, Schexnayder, & Knutson, 2000; Lu & AbouRizk, 2000):

1. It is continuous.
2. It has finite endpoints.
3. It has a defined mode between its endpoints.
4. It is capable of describing both skewed and symmetric activity time distributions.

For the current discussion, consider stated advantage two. The second advantage makes sense from a practical point of view in that every activity must have a maximum completion time. The difficulty with this stated advantage,

Edward L. McCombs is the Director of Operations Support. Email him at: emccombs@ditchwitch.com. Matthew E. Elam is an Associate Professor of Industrial Engineering and is an ASQ Certified Quality Engineer. Email him at: Matthew_Elam@tamu-commerce.edu. David B. Pratt is an Associate Professor in the School of Industrial Engineering and Management. Email him at: david.pratt@okstate.edu.

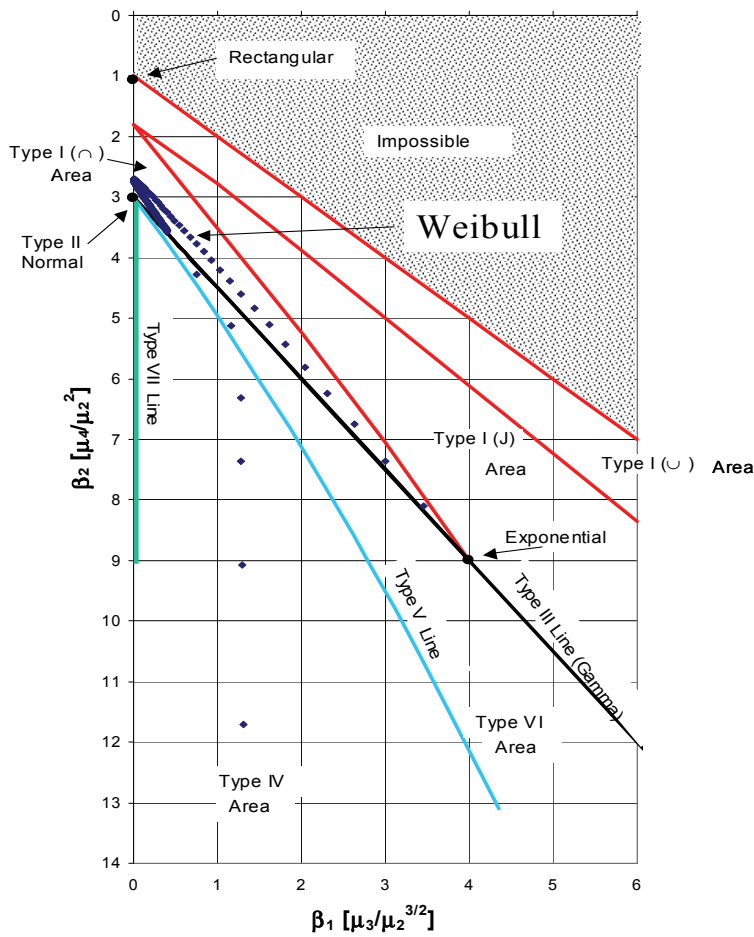
however, is determining the value of this maximum. For example, the truck travel time study described by Fente, et al. (2000). The maximum travel time is computed as two times the mode. This assumption is supported by the reasoning that management would notice the slow moving truck and take actions necessary to reduce its travel time. Undoubtedly, this type of assumption is necessary when a decision maker is constrained to using the beta probability distribution. However, it may be more reasonable to consider a distribution that can accommodate a longer tail probability than is allowed by the beta distribution.

Weibull Probability Distribution

The Weibull probability distribution can accommodate this longer right tail probability. Additionally, the Weibull distribution has advantages one, three, and four as listed above for the beta probability distribution.

Figure 1 shows a Pearson skew plot (Pearson, 1920; Pearson & Tukey, 1965) with the Weibull probability distribution plotted. The Type I areas shown in Figure 1 can be represented by the beta probability distribution. Figure 1 shows that the Weibull distribution can approximate distributions ranging from close to the normal to the exponential, can accommodate

Figure 1: Pearson's Skew Curves Plot Showing the Weibull Distribution



ESTIMATING TASK DURATION USING THE WEIBULL PROBABILITY DISTRIBUTION

distributional skewness (β_1), and can approximate activity duration models with fatter distributional tails (β_2) than can be accommodated by the beta probability distribution. Note that the Weibull probability distribution divides the triangular Type I (\cap shaped) region. It is expected that the Weibull probability distribution can satisfactorily describe those Type I (\cap shaped) models that are coincident with the beta models located in this region. Further research may also show that the Weibull probability distribution can serve as a proxy for the entire Type I (\cup shaped) region. An additional advantage of the Weibull distribution is that it should also satisfactorily model some Type III, IV, V, and VI distributions. This would be useful considering the review by Maio, et al. (2000), which shows that the beta probability distribution is not the best model for all construction operations.

Equations (1) – (5) show the Weibull probability density function, reliability, mode, variance, and mean formulas, respectively, from Ebeling (1997):

$$f(x) = (\beta/\theta)(x/\theta)^{\beta-1} \exp[-(x/\theta)^\beta] \quad (1)$$

$$R(x) = \exp[-(x/\theta)^\beta] \quad (2)$$

$$M = \text{mode} = \theta(1-1/\beta)^{1/\beta} \text{ for } \beta > 1 \quad (3)$$

$$\sigma^2 = \theta^2 \left\{ \Gamma(1+2/\beta) - [\Gamma(1+1/\beta)]^2 \right\} \quad (4)$$

$$\mu = \text{mean} = x_0 + \theta\Gamma(1+1/\beta) \quad (5)$$

where β is the shape parameter, θ is the scale parameter, Γ is the gamma function, and x_0 shifts the mean on the x-axis.

Methodology

Let x_a be the lower expert judgment percentile estimate, x_b be the upper expert judgment percentile estimate, and M be the most likely expert judgment estimate. Equation (2) can, therefore, be rewritten to solve for x_a and x_b ,

with the results as Equations (6) and (7), respectively:

$$x_a = \theta \left[\ln(1/R(x_a)) \right]^{1/\beta} \quad (6)$$

$$x_b = \theta \left[\ln(1/R(x_b)) \right]^{1/\beta} \quad (7)$$

The traditional form of the Weibull distribution has defined left and right bounds of zero and infinity, respectively. As a result, only two of the three expert opinion estimates are required to calculate the distributional parameters. If x_a and x_b , as well as their respective percentiles, are known, then Equations (6) and (7) can be used to calculate the shape parameter β in Equation (8):

$$\begin{aligned} \frac{x_a}{x_b} &= \frac{\theta \left[\ln(1/R(x_a)) \right]^{1/\beta}}{\theta \left[\ln(1/R(x_b)) \right]^{1/\beta}} \\ \Rightarrow \frac{x_a}{x_b} &= \left(\frac{\ln(R(x_a))}{\ln(R(x_b))} \right)^{1/\beta} \\ \Rightarrow \beta &= \frac{\ln \left[\ln(R(x_a)) / \ln(R(x_b)) \right]}{\ln(x_a/x_b)} \end{aligned} \quad (8)$$

Substituting the calculated value of β into Equation (6) or (7) allows the scale parameter to be calculated.

If x_a or x_b and also M are known, then Equation (6) or (7) and also Equation (3) can be used to calculate the shape parameter β as in Equation (9):

$$\begin{aligned} \frac{M}{x_b} &= \frac{\theta[1-1/\beta]^{1/\beta}}{\theta \left[\ln(1/R(x_b)) \right]^{1/\beta}} \\ \Rightarrow \frac{M}{x_b} &= \left(\frac{1/\beta - 1}{\ln(R(x_b))} \right)^{1/\beta} \\ \Rightarrow \left(\frac{x_b}{M} \right)^\beta &= \left(\frac{1}{\beta} - 1 \right) = \ln(R(x_b)) \end{aligned} \quad (9)$$

Finally, the scale parameter θ can be calculated using Equation (6) as in Equation (10):

$$\theta = M / (1 - 1/\beta)^{1/\beta} \quad (10)$$

An additional advantage to using the Weibull distribution exists. Specifically, a user is allowed to use whichever percentiles he/she feels are the most appropriate. Moreover, not only is a user now able to use percentiles other than the 5 and 95 percentiles with equal accuracy, the percentiles need not be symmetric; i.e., the 5 and 90 percentiles could be used.

Consider the situation in which there is a zero probability of an event occurring before a certain threshold time. For the Weibull distribution, a threshold value, x_0 , can be included as in Equations (11) – (14) from Ebeling (1997):

$$f(x - x_0) = (\beta/\theta) \left((x - x_0)/\theta \right)^{\beta-1} \times \exp \left[- \left((x - x_0)/\theta \right)^\beta \right] \quad (11)$$

$$R(x - x_0) = \exp \left[- \left((x - x_0)/\theta \right)^\beta \right] \quad (12)$$

$$M = \text{mode} = x_0 + \theta (1 - 1/\beta)^{1/\beta} \text{ for } \beta > 1 \quad (13)$$

$$x = \theta \left[\ln(1/R(x)) \right]^{1/\beta} + x_0 \quad (14)$$

The equation for the variance remains unchanged. The addition of a threshold value does not change the basic shape of the distribution, only its location on the x-axis. Because the left boundary is no longer known and there is an additional parameter, additional information needs to be incorporated.

The calculation of the ratio

$$\frac{(M - x_A)}{(x_B - x_A)}$$

in terms of its respective components from Equations (13) and (14) is shown in Equation (15):

$$\frac{M - x_A}{x_B - x_A} = \frac{x_0 + \theta (1 - 1/\beta)^{1/\beta} - \left\{ x_0 + \theta \left[\ln(1/R(x_a)) \right]^{1/\beta} \right\}}{x_0 + \theta \left[\ln(1/R(x_b)) \right]^{1/\beta} - \left\{ x_0 + \theta \left[\ln(1/R(x_a)) \right]^{1/\beta} \right\}} \quad (15)$$

The threshold value cancels, as do the scale parameters, with the result in Equation (16):

$$\frac{M - x_A}{x_B - x_A} = \frac{(1 - 1/\beta)^{1/\beta} - \left[\ln(1/R(x_a)) \right]^{1/\beta}}{\left[\ln(1/R(x_b)) \right]^{1/\beta} - \left[\ln(1/R(x_a)) \right]^{1/\beta}} \quad (16)$$

The shape parameter β can be computed using a solver program (e.g., Microsoft Excel's Solver® function). Because the threshold value is unknown, the equation for the mode cannot be used to calculate the scale parameter θ as in Equation (10). However, the variance constant K can be calculated and used to calculate the variance.

Using the calculated shape parameter β and a scale parameter θ equal to 1.0, the temporary variance is calculated as in Equation (17):

$$\sigma_{temp}^2 = \theta_{temp}^2 \left\{ \Gamma(1 + 2/\beta) - \left[\Gamma(1 + 1/\beta) \right]^2 \right\} \quad (17)$$

where $\theta_{temp} = 1$. Next, the temporary x-axis values for the required lower and upper percentiles are calculated using Equations (6) and (7). The variance constant K can now be calculated as shown in Equation (18):

ESTIMATING TASK DURATION USING THE WEIBULL PROBABILITY DISTRIBUTION

$$K = (x_{b \text{ temp}} - x_{a \text{ temp}}) / \sigma_{\text{temp}} \quad (18)$$

The variance based on the actual data can now be calculated using Equation (19):

$$\sigma^2 = ((x_b - x_a) / K)^2 \quad (19)$$

With the variance known, the actual scale parameter θ can be calculated as shown in Equation (20):

$$\begin{aligned} \sigma^2 &= \theta^2 \left\{ \Gamma(1+2/\beta) - [\Gamma(1+1/\beta)]^2 \right\} \\ \Rightarrow \theta &= \sqrt{\sigma^2 / \left\{ \Gamma(1+2/\beta) - [\Gamma(1+1/\beta)]^2 \right\}} \end{aligned} \quad (20)$$

With β and θ known, the threshold value can be calculated using the most likely value M as in Equation (21):

$$\begin{aligned} M = \text{mode} &= x_0 + \theta(1-1/\beta)^{1/\beta} \\ \Rightarrow x_0 &= M - \theta(1-1/\beta)^{1/\beta} \end{aligned} \quad (21)$$

All of the parameters for the required Weibull distribution in Equation (11) can now be calculated.

Example

As an example of these parameter calculations, consider the truck travel example as shown in Fente, et al. (2000). The travel distance is 3.7 – 3.9 km. The traditional PERT information is as follows. The minimum possible travel time is based on the physical characteristics of the project site and the truck manufacturer's specifications and is equal to 7.67 minutes. The most likely travel time is 9.21 minutes. The maximum travel time is 18.42 minutes. Additionally, it is given that the 75th percentile estimate is 11.05 minutes. Fente, et al. (2000) report that a beta probability distribution with parameters $\alpha = 1.898$ and $\beta = 6.372$ is a reasonable model for the truck travel time distribution.

To use the methodology presented in this paper for the offset Weibull probability distribution, only two percentile estimates and

the mode estimate are required. Because the 75th percentile estimate is explicitly stated, it is an obvious choice for one of the required estimates. The required second boundary estimate requires an assumption with regard to the percentile that it represents. The lower boundary of 7.67 minutes was selected because it is a finite boundary. Specifically, the lower boundary is assumed to represent the 0.01 percentile. As a result, the following parameters were calculated in Table 1. Figure 2 shows the resulting Weibull distribution plotted with the resulting beta distribution as derived in Fente, et al. (2000). The two curves converge together as the value of the lower percentile converges to zero.

Because the proposed Weibull model and the resulting beta model, as presented by Fente, et al. (2000), are both estimates of the unknown underlying distribution, it is not useful to compare the fits via a goodness-of-fit test. However, visually it seems that either model could satisfactorily model the underlying distribution. So why consider the Weibull model over the beta model? First, the Weibull model required only three estimates, while the beta model required four. Second, the Weibull model can easily be developed in a Microsoft Excel® spreadsheet. Finally, when compared to the traditional PERT methodology, the Weibull model does not require an estimate of the variance -- this value is calculated exactly (as is the mean value). Moreover, with regard to this last point, the only errors associated with the Weibull model relate to the accuracy of the original estimates and whether the Weibull model can satisfactorily describe the underlying distribution.

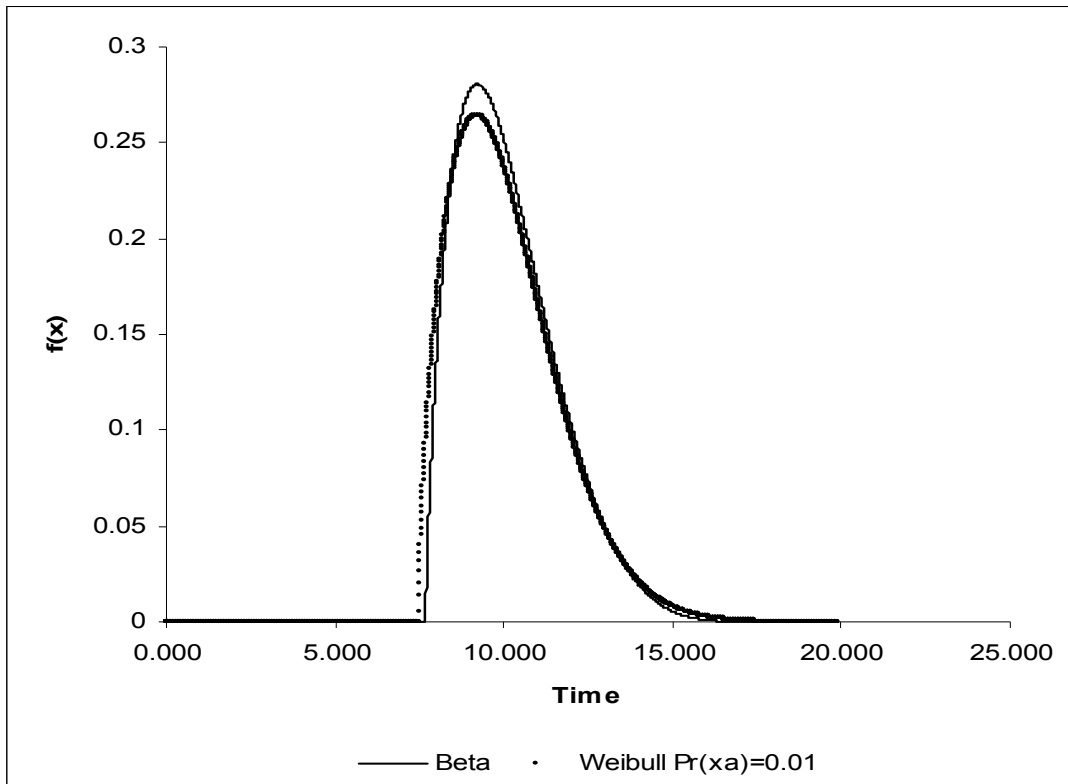
Conclusion

If an activity's duration time starts at $t=0$, and one can estimate at least two of three estimates (x_a , x_b , and M) of an unknown distribution, then one can estimate the unknown distribution with a Weibull probability distribution. This approach could be beneficial in situations where two of the three estimates (lower percentile, most likely, upper percentile) can be assumed to be known with greater certainty than the third estimate. If all three estimates are assumed known with equal certainty and/or an activity's

Table 1: Results of Fitting the Presented Weibull Model to the Data in Fente, et al. (2000)

Parameter	Eq.	Substituted Values	Result
β	(16)	$M = 9.21, x_a = 7.67$ $x_b = 11.05, R(x_a) = 1-0.01$ $R(x_b) = 1-0.75$	1.6900
$x_{a \text{ temp}}$	(6)	$\theta = 1, R(x_a) = 1-0.01, \beta = 1.6900$	0.0657
$x_{b \text{ temp}}$	(7)	$\theta = 1, R(x_b) = 1-0.75, \beta = 1.6900$	1.2132
σ^2_{temp}	(17)	$\theta = 1, \beta = 1.6900$	0.2953
K	(18)	$x_{b \text{ temp}} = 1.2132, x_{a \text{ temp}} = 0.0657$ $\sigma^2_{\text{temp}} = 0.2953$	2.1118
σ^2	(19)	$x_b = 11.05, x_a = 7.67, K = 2.1118$	2.5618
θ	(20)	$\sigma^2 = 2.5618, \beta = 1.6900$	2.9456
x_0	(21)	$M = 9.21, \theta = 2.9456, \beta = 1.6900$	7.4764
μ	(5)	$x_0 = 7.4764, \theta = 2.9456, \beta = 1.6900$	10.1056

Figure 2: Plot of the Truck Travel Time from Fente, et al. (2000) and the Weibull Model



duration time does not begin at $t = 0$, then it is advisable to use the shifted Weibull distribution.

The objective of this article was to provide an alternative approach to the traditional Project PERT methodology using the Weibull probability distribution. It was shown that by using the Weibull probability distribution it is not necessary to estimate a future activity's mean or variance. These values are calculated exactly and have only the uncertainty inherent in the original subjective estimates and the uncertainty as to whether the Weibull probability distribution accurately models the underlying distribution of future activity times. The ease of use and the reduction in uncertainty with the proposed Weibull model will benefit both practitioners and researchers.

The beta distribution unarguably is more robust within the Pearson Type I (\cup shape) region than the Weibull distribution. However, as Lau, Lau, and Zhang (1996) have pointed out, there is a practical application for distributional models that are more robust to the third and fourth moments. The Weibull distribution satisfies this need. The true test with regard to the applicability of the Weibull distribution will be its ability to accurately model a broad range of actual problems.

References

Cottrell, W. D. (1999). Simplified program evaluation and review technique (PERT). *Journal of Construction Engineering and Management*, 125(1), 16-22.

Ebeling, C. (1997). *An introduction to reliability and maintainability engineering*. NY: McGraw-Hill.

Fente, J., Schexnayder, C., & Knutson, K. (2000). Defining a probability distribution function for construction simulation. *Journal of Construction Engineering and Management*, 123(3), 234-241.

Lau, A. H., Lau, H., & Zhang, Y. (1996). A simple and logical alternative for making PERT time estimates. *IIE Transactions*, 18, 183-192.

Lu, M. & AbouRizk, S. M. (2000). Simplified CPM/PERT simulation model. *Journal of Construction Engineering and Management*, 126(3), 219-226.

Maio, C., Schexnayder, C., Knutson, K., & Weber, S. (2000). Probability distribution functions for construction simulation. *Journal of Construction Engineering and Management*, 126(4), 285-292.

Malcolm, D. G., Roseboom, J. R., Clark, C. E., & Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations Research*, 7, 646-669.

Pearson, K. (1920). On the probable errors of frequency constants. *Biometrika*, 13(1), 113-132.

Pearson, E. S. & Tukey, J. W. (1965). Approximate means and standard deviations based on distance between percentage points of frequency curves. *Biometrika*, 52(3/4), 533-546.

Pleguezuelo, R. H., Perez, J. G., & Rambaud, S. C. (2003). A note on the reasonableness of PERT hypotheses. *Operations Research Letters*, 31, 60-62.

Premachandra, I. M. (2001). An approximation of activity duration distribution in PERT. *Computers and Operations Research*, 28, 443-452.

Industrialization in Animal Agriculture: A Kalman Filter Analysis

Oya S. Erdogdu Levent Ozbek
Ankara University, Turkey

Studies discussing the effects of technological developments on (animal) agricultural production argue that the effective usage of chemicals and genetic engineering increase control over production processes, which in turn decreases seasonality (one significant factor defining agricultural production) significantly and brings standardization to production. Studies on broilery also show that production is not limited by nature determined seasons. Supply side changes accompanied by changes in demand have led to more healthier, standardized products. Using tools of economics and statistics, this study documents this transformation in animal agricultural production of beef, pork and milk. Results indicate decreasing seasonality, thus the industrialization of animal agriculture.

Key words: Animal agriculture, seasonality, Kalman Filter.

Introduction

Agricultural production today is far different than it was 50 years ago. The social conditions and living standards in the 21st century has led consumer preferences to support more standardized, health concerned, and user friendly agricultural products. This change from the demand side opened the door to big corporations who are capable of producing different, standardized products to satisfy demand. As opposed to small family producers, these big corporations easily deal with economies of scope, economies of scale, market power and risk management problems, by using technology intensive, manufacturing-type production techniques. These demand and supply side changes have replaced small family production with large corporations and have led to the industrialization of agricultural production. This process is called industrialization due to the intensive usage of high technology which

increases control over nature and nurture, and standardization which increases size and quality of production.

Although it is important to analyze the demand and supply side factors that have caused significant changes in the sector, this article only attempts to document the decreasing seasonality in pork, beef and milk production that is the result of increased control achieved by using intensive high technology production techniques.

Control over nature and nurture

Allen and Lueck (2000) argued that nature is “the main feature that distinguishes farm organization from ‘industrial’ organization” (p. 14). Due to its very core of existence, agricultural production is defined and restricted by the forces of nature. Nature determines the properties, types, sequence, and timing of the stages of production, creating a certain amount of stability and predictability in the process. Nature determines the time to plant, harvest, breed, and furrow, and so creates a type of certainty in production. For example, in Iowa, USA, April-June is the time to sow, whereas September-November is the time to harvest, and spring has traditionally been the time to furrow for pigs. These are subject to weather conditions and so, contrary to standardization in manufacturing process, it can be different for

Oya S. Erdogdu is an Associate Professor at Ankara University, and is a faculty member in Political Sciences, Department of Economics, Email: oerdogdu@politics.ankara.edu.tr. Levent Ozbek, is in the Department of Statistics, Email: ozbek@science.ankara.edu.tr.

INDUSTRIALIZATION IN ANIMAL AGRICULTURE

different parts of the world and for different products.

Nature not only governs certainty but also uncertainty in agricultural production. The random forces of nature – unexpected changes in weather conditions, blizzards, and storms – create unpredictable and unpreventable shocks to the system.

The forces of nature and the concept of seasonality it creates, is significant to understand in the agricultural production process. For a producer of an agricultural product, a *season* is the specific period of the year during which a given activity takes place. Hence, shaped by the forces of nature, seasonality determines the stages, timing and time length of a specific process. As can be expected, this creates cycles in the production over a given period of time. As opposed to analyzing the properties or its effects of (decreasing) seasonality on production or managerial decisions, this article documents the decreasing seasonality in agricultural production over the last 50 years.

Mobility of livestock during growing stages allow it to be reared in controlled environments. Though seasonality is an issue for all types of agricultural production, compared to crop production, mobility of livestock allows a producer to exercise greater control over nature by using high-tech factory style production techniques. This article focuses on the effect of increased control over nature and nurture on animal production, specifically, beef, pork and milk.

Technological advancements are the primary factor in decreasing seasonality; they have facilitated human control on biological processes and the production environment by the effective use of veterinary medicine and by the use of genetically improved products. Thus, intensive use of technology has increased control over the production environment and biological development processes and allows producers to implement modern manufacturing principles to create less risky, more elastic production environments to produce more consistent, feed efficient, special nutrition enriched products. In other words, with the ability to control nature, producers have gained higher flexibility to respond to changes in consumer demand and have had an increased ability to set and sustain a

certain quality level and have given the ability to reduce risks concerning food safety and contamination.

In general terms, the ability to control nature, and thus the genetic input, allows a producer to change the order in the system through mixture or separation. The method of mixture/separation can be used at the farm level, which leads to herd heterogeneity, or at the processing level, which leads to heterogeneous raw produce. The profit maximizing producer performs a cost/benefit analysis to decide on separating (at cost) or working with the mixed types they purchased to satisfy the strong demand for consistent, preparation-friendly products.

On the cost side, the use of genetic engineering is subject to patent costs and costs associated with information and uncertainty. Patent costs being a large asset, are specific costs to achieve a genetic improvement of a given species. But more importantly, the biological improvement creates information costs due to uncertainty about the composition of the mixture or the uncertainty about the reaction of each type to stimulation. Moreover, these uncertainties create inefficiency in volume production, low quality and inconsistency in raw production, leading to unsatisfactory completion of the transformation process. However, besides these negative significant impacts on commercial gains, extensive use of controlled genetic inputs is expected to decrease costs and improve commercial gains.

Given incentives, variations in inputs lead to variations in the performance of the product brought to market at the same time (intra-temporal inconsistency) and at different times. Therefore, inconsistency in production due to variations in input, like nutrition and environment, is decreased by greater control of the production environment.

Confined production systems with increased control over the production environment such as improvements in nutrition, housing, handling equipment, and management have encouraged higher and more uniform supply. Factory-style corporate livestock farming, using veterinary medicines, healthier diets and indoor environmentally controlled sheds has satisfied the needs and improved the

health and production conditions of the animals. The result is a healthier, uniform, larger supply (Hurt, 1994).

Thus, the ability to control nature and nurture leads to structural changes in animal production and decreasing seasonality with more uniform and standard products. The remainder of this article aims to document this transformation using different analytical and statistical tools.

Data analysis

The data on the monthly production of pork, beef, and milk were obtained from the United States Department of Agriculture (USDA) website. Monthly milk production data was obtained for the period 1930-2000 (except 1960-1963), and monthly beef and hog production data are for the period 1944-1999 (except 1982).

The data series are monthly calculations from the first to the last day of the month. Monthly data was first normalized to 30 days per month to decrease noise in the system, in order to detect decreasing seasonality in production, the Herfindahl-Hirshman Index (HHI) was calculated, model stability/structural change tests were conducted and lastly the Kalman filter analysis was performed.

Figure 1 shows the normalized monthly production shares, calculated for 12-year averages for each month for different time periods. The shares getting closer to each other indicate increasing smoothness, which is clearly observed in the production of pork and milk. However, for beef production the variability continues; this may be due to the definition of the beef data group. Data on beef production includes data on all kinds of meat production, such as cattle and sheep. Since every production has its own timing of structural transformation, it is difficult to capture structural change from that data group, which is also expected to be a very slow process.

Figure 1 shows that the most dramatic change has occurred in milk production. The significant importance of summertime production in the 1930's is replaced with rather constant shares in 2000, indicating relatively stable production.

Methodology

In order to verify the industrialization process of animal agricultural production statistically, the Herfindahl-Hirshman (HHI) index was calculated and, to analyze the structural change in the system, Chow, CUSUMSQ and ARCH LM statistics were calculated.

Herfindahl-Hirshman Index (HHI)

HHI, is a market structure analysis tool that measures the degree of concentration in an industry. It has an advantage over other concentration measures since it works with all firms in the market and takes into account the relative distributional shares of the market held by all firms.

Based on the Jensen Inequality, the HHI is calculated using the sum of squares of the market shares of all firms. The HHI index is

$$HHI = 10,000 \sum_{i=1}^K w_i^2, i = 1, \dots, K,$$

where, w_i is the market share of the firm i .

In this study HHI was used to measure the degree of spread of production over 12 months for beef, pork, and milk production. HHI was calculated for each year by summing up the square of each month's share in total production; the 12-year averages of that sum were also calculated. Thus, for the time period 1945-1956 the HHI index was calculated as:

$$HHI = \frac{1}{12} \sum_{j=1945}^{1956} \sum_{i=1}^{12} s_{i,j}^2,$$

where s_{ij}^2 is the i^{th} monthly production share in the j^{th} year: the calculation is slightly different from its original form. Since decimals were not a concern, the summation result was not multiplied by 10,000, but it was preferred to take the averages to minimize the noise in the system.

Table 1 summarizes the calculation of the HHI for beef and pork averaged over the time periods: 1945-1956, 1958-1969, 1970-1981, 1983-1994, and 1988-1999. The HHI for milk production was averaged over the time

INDUSTRIALIZATION IN ANIMAL AGRICULTURE

Figure 1: Monthly U.S. Production Averages for Beef, Pork and Milk

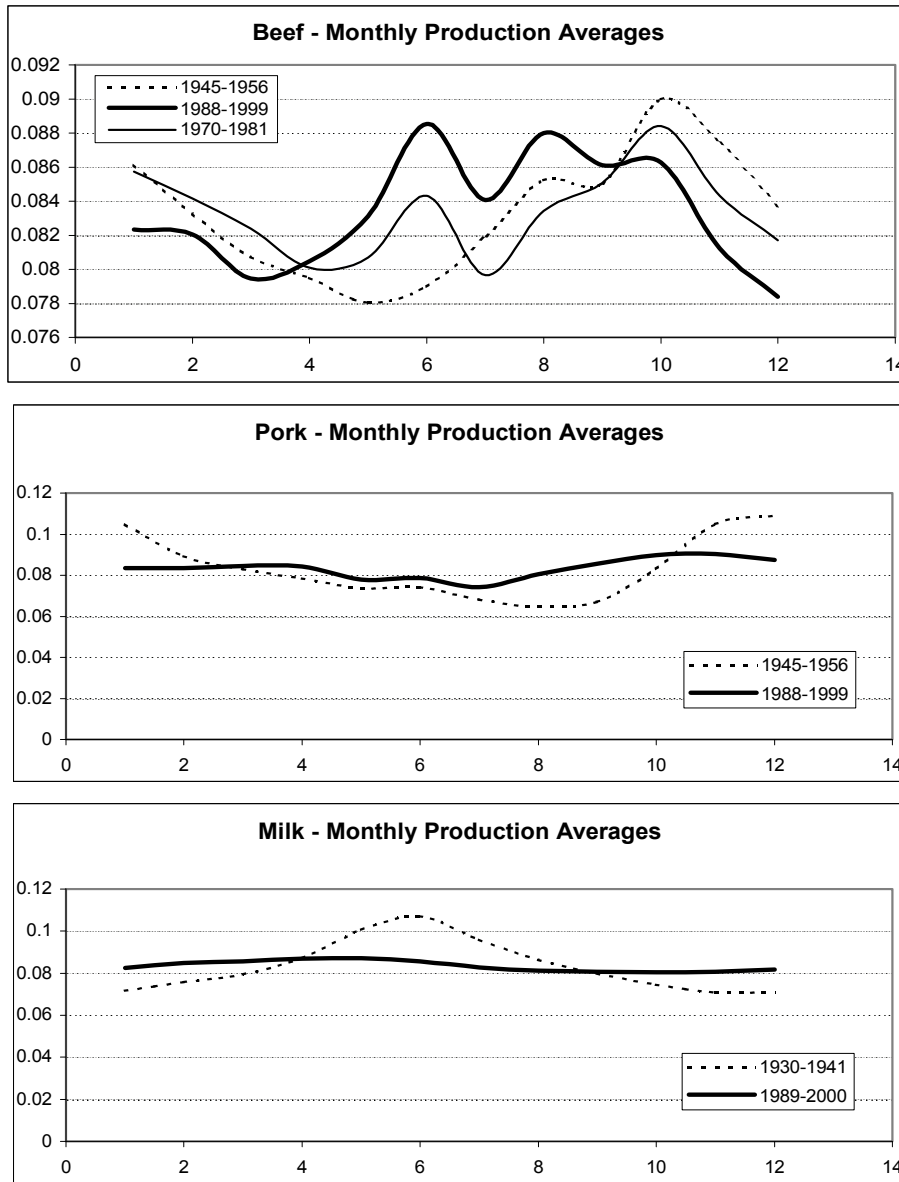


Table 1: HHI Index Values

HHI Beef		HHI Pork		HHI Milk	
1945-1956	0.084035	1945-1956	0.086944	1930-1941	0.085029
1958-1969	0.083530	1958-1969	0.084153	1941-1952	0.085543
1970-1981	0.083536	1970-1981	0.083995	1948-1959	0.084893
1983-1994	0.083542	1983-1994	0.083977	1963-1974	0.083811
1988-1999	0.083534	1988-1999	0.082635	1971-1982	0.083591
				1981-1992	0.083453
				1989-2000	0.083416

periods 1930-1941, 1941-1952, 1948-1959, 1963-1974, 1971-1982, 1981-1992, and 1989-2000.

If each month of each year had equal shares of production, 1/12, the index takes the value: $HHI = \frac{1}{12} \sum_{j=1945}^{1956} \sum_{i=1}^{12} (0.08333)^2 = 0.0833$.

At the other extreme, if production was composed in only one month at each year – $s_{ij} = 1$ and $s_{kj} = 0$ for all $k \neq i$, the index takes the value: $HHI = \frac{1}{12} \sum_{j=1945}^{1956} 1 = 1$.

As shown in Table 1, all indexes decrease over the time periods and move towards the value of 0.083. This indicates a change in the production process such that the production is spreading over the whole year equally.

Model stability tests

The decline in seasonality implies an underlying structural change in the production process and changing parameter values, which can be detected using the Chow, CUSUMSQ and ARCH LM statistics. The OLS regression analysis implicitly assumes that the coefficients do not change over time, however, Chow, CUSUMSQ and ARCH LM tests can detect the existence of time dependency in the model, if any are present.

To test for structural change in our problem, the monthly production shares were regressed on a constant term and the monthly dummy variables:

$$y_t = \beta_0 + \sum_{i=1}^{11} \beta_i M_i + \varepsilon_t (*)$$

In order to prevent the dummy trap, 11 dummies were used instead of 12. The dummy for the month with less production share is excluded from the regression. Thus, for beef production the dummy for November was excluded, for pork production the dummy for October was excluded, and for milk production the dummy for March was excluded.

The monthly production shares getting closer to each other is a satisfactory indicator of

decreasing seasonality. Therefore, it was expected that a structural change had occurred and the coefficients of the model have changed over time.

To document these changes, structural change statistics including Chow, CUSUMSQ and ARCHLM were calculated. To calculate the Chow test statistics, the time of structural change must be defined. However, the graphical analysis indicates a very slow change; no specific shock is given, thus the statistics for different time periods were calculated. For beef and pork production the statistics are calculated to determine if the coefficients of the regressions are different for the periods 1944-1961, 1962-1998, 1944-1974, and 1975-1998. For milk production Chow statistics are calculated for the periods 1930-1961 and 1962-2000. These results are summarized in Table 2.

Each Chow statistic for pork and milk production was greater than the critical value 1.75 at the 5% significance level. Therefore, the null hypothesis of same coefficients was rejected, and it was concluded that the coefficients obtained on regression for the given two time periods were significantly different from each other. That is, a structural change has occurred in pork and milk production in the last 50 years.

As for beef production, similar to the case in Figure 1, the Chow test results are the image of the definition of the beef data group. The test statistics for beef production indicate a structural change between 1944-1981 and 1983-1999. The same result was achieved when the sample is divided into three different time periods, but a more detailed analysis indicated that no structural change has occurred. The Chow calculation did not result in rejecting the null of no structural change for the time periods, 1944-1961 and 1962-1981, and similarly for the periods 1983-1992 and 1993-1999. This reflects a significant, but slow, transformation in beef production.

The Chow test statistics search for structural changes in the specified markets for specified periods of time. In this study the CUSUMSQ statistics were also calculated without restricting the cut off time periods in the data when searching for the existence of stability. In addition, the CUSUMSQ test has a

lower power than the Chow test; results are shown in Figure 2.

The CUSUMSQ statistics for beef and pork production move outside the confidence bounds until the 1980's, indicating a structural change in the production process. However, the statistic moves inside the confidence bounds in the 1990's. This same confusing result was observed in milk production. Although the lack of milk production data may provide the explanation regarding the generality of the null hypothesis, the CUSUMSQ statistics are not very helpful in determining a structural change. This is surprising given that previous results indicated a very slow transformation process, which may be ongoing even now.

Besides searching for structural changes in the model using the Chow and CUSUMSQ statistics, ARCH LM statistics were also calculated to test whether the coefficients of the model were time varying. The results shown in Table 3 reject the null hypothesis of constant variance and thus certifies that beef, pork and milk production coefficients are time varying.

Based on these analyses the models for beef, pork and milk production were estimated again under the assumption that parameters were time varying: the Kalman filter was used for that purpose.

Kalman filter analysis

Because the model stability/structural change test results indicated that the parameters of the equation (*) are not constant due to the ongoing industrialization process of animal agricultural production, the equation is modified to allow for parameters varying over time.

$$y_t = \beta_{0t} + \sum_{i=1}^{11} \beta_{it} M_i + \varepsilon_t \quad (2)$$

The Kalman filter estimation results from equation (**) reported in Figures 3, 4 and 5 show convergent monthly shares and thus decreasing seasonality in production. The beef production estimation results are not as clear in defining structural change, but pork and milk production estimation results show that monthly production shares are getting closer to each other. Figures show that the constant term converges to 0.1 and the dummy variable coefficient values converge to zero. As in Figure

1, the most significant change is observed in milk production. The increase in summer production and relatively low winter production is replaced by production spreading equally across all year. This change occurring in the late 1990's indicates the effect of greater control over nature and nurture in animal agricultural production.

Discussion

This study focused on decreasing seasonality to document the structural change in animal agricultural production. To satisfy consumers' preferences for healthier, user-friendly products, high technology is used intensively in production, thus increasing control over nature and nurture. The demand and supply side factors leading to decreasing seasonality have caused a significant transformation in the sector, creating factory style large manufacturing firms instead of small family farms. That process is named the *industrialization* of animal agricultural production.

In this study analytical (HHI) and statistical (Chow, CUSUMSQ and ARCH LM) tools were used with Kalman Filter methodology to document the industrialization process of animal agricultural production. However, many questions remain that must be answered by economists.

First, it is important to document how effective existent policies have been on the structural changes in animal agriculture. To document the impact of these policies on innovation, the implementation of scientific knowledge, and the role of policies to encourage/discourage vertical integration is crucial to decide on the direction of future actions.

Second, it is important to analyze the impacts of this new production structure on technological developments, bio-security, national and international market structure, prices, and the environment.

It is argued that the use of technological developments in animal agriculture have created uniformity in production. Is this a two-way road? Does uniformity encourage or discourage technological developments and innovative attempts? If so, what would the effect on market

Table 2: Chow Test Results

<u>Beef Production</u>	
<u>Hypothesis</u>	<u>Chow Statistics</u>
$H_0 : \beta_{1944-1981} = \beta_{1983-1999}$	Chow Test: 4.19
$H_0 : \beta_{1944-1961} = \beta_{1962-1981}$	Chow Test: 1.19
$H_0 : \beta_{1983-1992} = \beta_{1993-1999}$	Chow Test: 0.53
$H_0 : \beta_{1944-1981} = \beta_{1983-1991} = \beta_{1992-1999}$	Chow Test: 46.65
$H_0 = \beta_{1944-1971} = \beta_{1972-1981} = \beta_{1983-1999}$	Chow Test: 5.23
<u>Pork Production</u>	
<u>Hypothesis</u>	<u>Chow Statistics</u>
$H_0 : \beta_{1944-1981} = \beta_{1983-1999}$	Chow Test: 5.85
$H_0 : \beta_{1944-1961} = \beta_{1962-1981}$	Chow Test: 27.55
$H_0 : \beta_{1944-1981} = \beta_{1983-1991} = \beta_{1992-1999}$	Chow Test: 4452.44
$H_0 = \beta_{1944-1971} = \beta_{1972-1981} = \beta_{1983-1999}$	Chow Test: 622.41
<u>Milk Production</u>	
<u>Hypothesis</u>	<u>Chow Statistics</u>
$H_0 : \beta_{1930-1959} = \beta_{1963-2000}$	Chow Test: 228.45
$H_0 : \beta_{1930-1945} = \beta_{1946-1959}$	Chow Test: 5.37
$H_0 : \beta_{1963-1982} = \beta_{1983-2000}$	Chow Test: 43.92
$H_0 : \beta_{1930-1959} = \beta_{1963-1981} = \beta_{1982-2000}$	Chow Test: 315.65
$H_0 : \beta_{1930-1945} = \beta_{1946-1959} = \beta_{1963-2000}$	Chow Test: 259.04

INDUSTRIALIZATION IN ANIMAL AGRICULTURE

Figure 2: CUSUMSQ Results

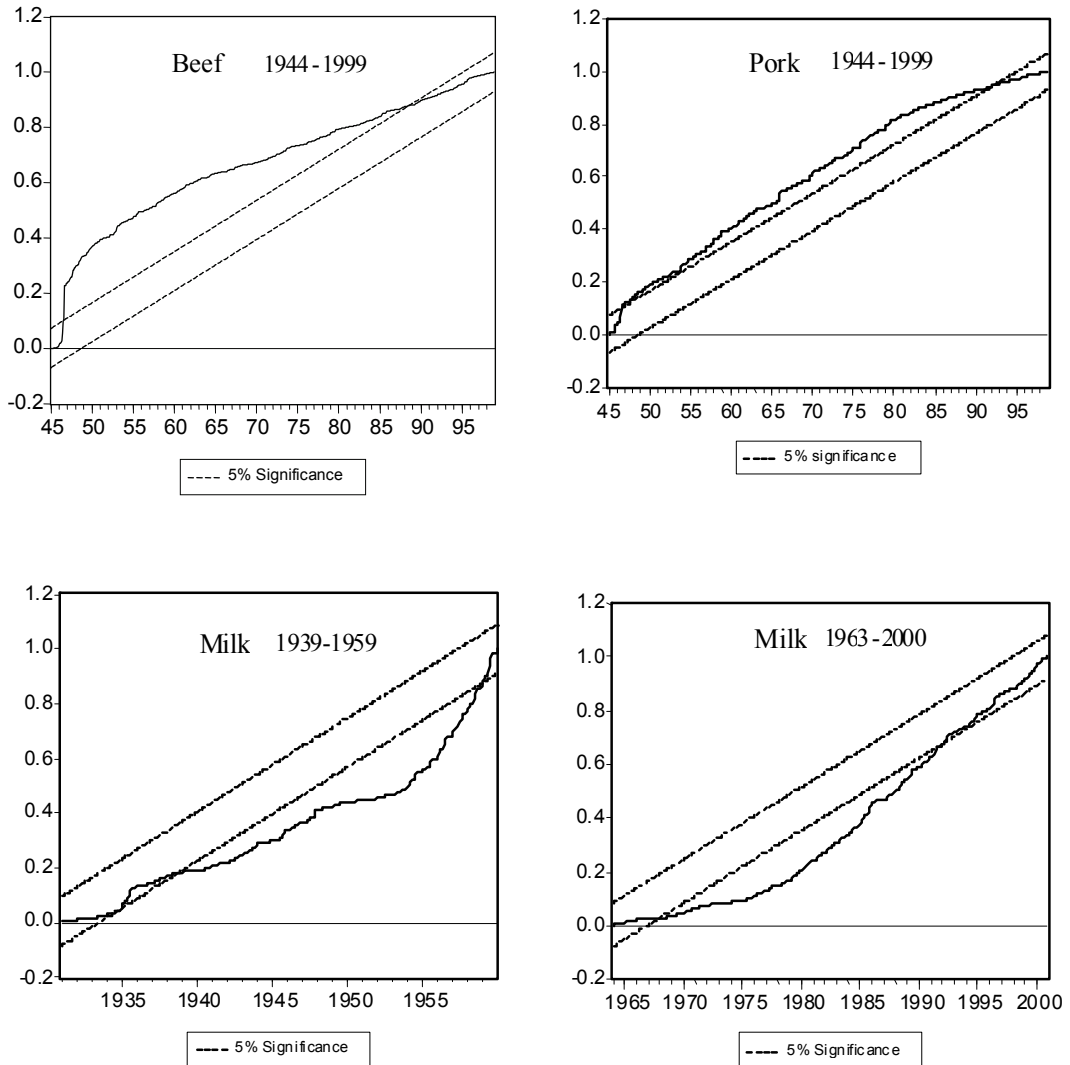


Table 3: Arch LM Test Results

	Milk 1930-1959	Milk 1963-2000	Beef 1944-1999	Pork 1944-1999
ARCH LM	93.68 (0.00)	70.23 (0.00)	24.49 (0.00)	15.16 (0.00)

Figure 4: Milk Production Estimation Results

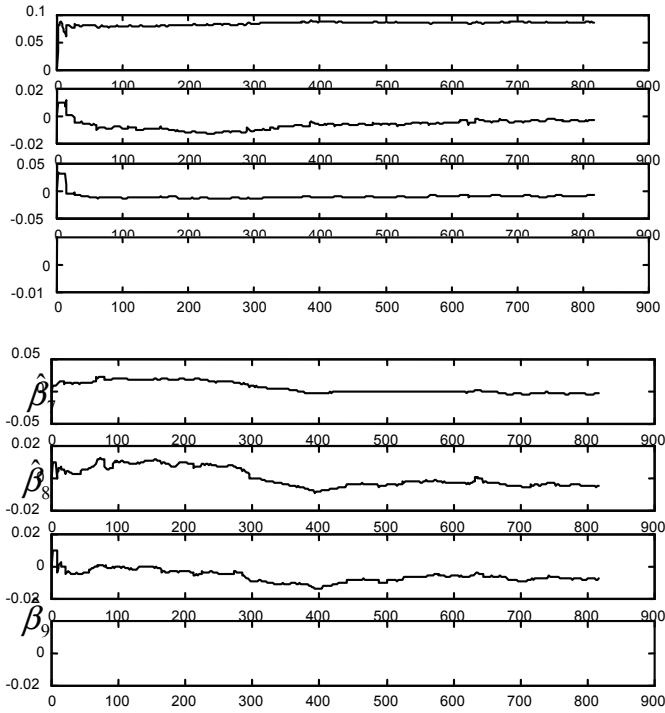


Figure 5: Pork Production Estimation Results

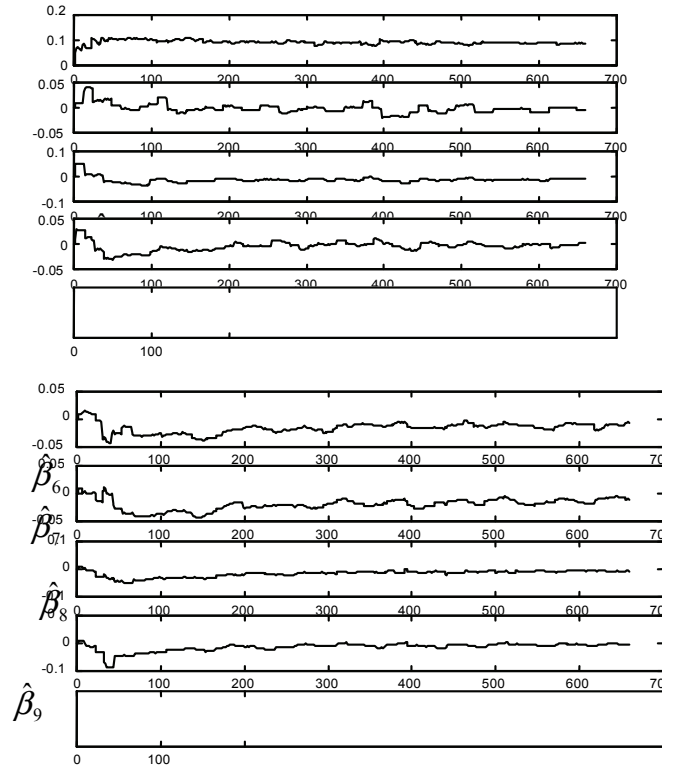


Figure 6: Beef Production Estimation Results

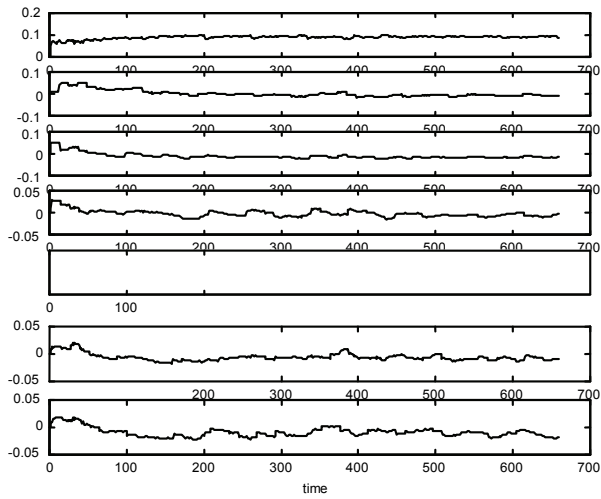
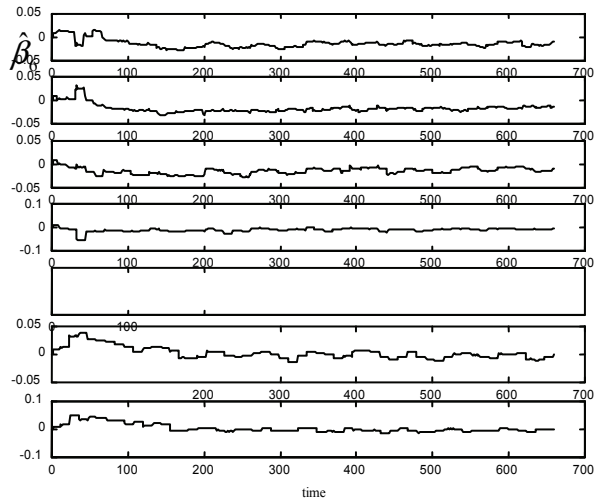


Figure 6: continued



INDUSTRIALIZATION IN ANIMAL AGRICULTURE

structure, quality, quantity, prices, and the role of government? How and how much regulation should be there? As the Dioxin case in Belgium and Starlink case in Iowa pointed out, there exist important bio-security issues regarding the usage of veterinary medicines and genetic improvement techniques in large corporations with high division of labor. What would be the regulations on the usage of veterinary medicines, genetic inputs, and patent rights? Do these regulations affect the pattern of seasonality in animal agriculture?

With globalization, the international effects of decreasing seasonality in domestic markets have also become an important issue.

The effects of seasonality on the price, quantity, and quality in the international markets should be analyzed as well as the consequences of policies on the usage of biological improvement techniques and medicines.

Finally, similar to arguments regarding the use of genetics in human development processes, arguments on the effect of high control of nature and nurture on animal welfare exist. Animal rights activists question if it is fair to genetically and environmentally restrict the natural development process, as in the case of factory style animal production. All of these present areas for further research.

References

Allen, D. & Dean, L. (2000). Family farm, inc. *Choices*, 1, 13 –17.

Boehlje, M. (1999). Structural changes in the agricultural industries: How do we measure, analyze and understand them? *American Journal of Agricultural Economics*, 81(5), 1028-1041.

Brewer, C. (1999). The trend of concentration in the pork industry: A brief explanation of why it's happening. *Unpublished Notes*.

Brown, R., Durbin, J. & Evans, J. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 37, 149-172.

Bugos, G. (1992). Intellectual property protection in the American chicken-breeding industry. *Business History Review*, 66, 127-168.

Drabenstott, M. (1994). Industrialization: Steady, Current or Tidal Wave. *Choices*, 4, 4-8.

Greene, W. (1997). *Econometric Analysis*. Prentice Hall Int.

Hall R, & Lieberman, M. (2001). *Microeconomics: Principles and Applications*. Southwestern College Publishing.

Hennessy, D. & Roosen, J. (2001). A Model of Seasonal Production, with Application to EU Milk Policy. *Tartışma Metini*.

Hurt, C. (1994). Industrialization in the Pork Industry. *Choices*, 4, 9-13.

United States Department of Agriculture Economic Research Service: <http://www.ers.usda.gov>.

Applying Census Data for Small Area Estimation in Community and Social Service Planning

Michael Wolf-Branigin Hyon-Sook Suh Star Muir Emily S. Ihara
George Mason University

Small area estimation provides a tool for community analysis. A procedure for accessing, selecting, joining and analyzing US Census data is provided. Skills acquired while completing the procedure include accessing census data, downloading boundary files and displaying themes. Such skills are valuable tools for students to possess as they enter the workforce.

Key words: Small area estimation, US Census, social policy analysis, geographic information systems (GIS), systems evaluation.

Introduction

Social services planners and evaluators have developed their planning abilities through improved access to data, methods for analyzing data, and techniques for visualizing the results. This article presents a method for data application. Small area estimation involves using outcome data that joins to a set of predictor variables within small domains or geographic areas in order to generate estimates.

Although small area estimates and geographic information systems (GIS) have been used by human service workers for years (Wolf-Branigin, LeRoy & Miller, 2001), the study of human environment interactions often fail to consider individual-level information or cross-discipline data, resulting in a lack of explanatory

and predictive power (An, Linderman, Qi, Shortridge & Liu, 2005). Calls from the field advocate for governmental and non-governmental organizations to improve the collecting, linking, and sharing of microdata in order to improve decision-making (Weitzman, Silver & Brazill, 2006).

For more than a century, the United States Census has collected data based on census tracts in selected areas (Krieger, 2006). These tracts, which typically include approximately 4,600 individuals, provide the basis for conducting small area estimations. Census tracts were defined for the entire country for the first time in 2000. This allows for better access to data needed for social policy and planning and to for documenting need and making informed decisions about the allocation of resources in various communities (Krieger, 2006). In addition to an increase in the scientific use of census tract data, the American FactFinder function of the US Census has increased consumer access to this vast database. The improvement in coverage of census tracts as well as easier access to this data allows for more precise service planning.

Given these improvements, social workers have the potential to be at the forefront of policy decisions by including GIS and mapping skills in their toolboxes. These skills allow health and social service workers to strengthen the social survey tradition, identify

Michael Wolf-Branigin is an Associate Professor in the School of Social Work. Email: mwolfbra@gmu.edu. Hyon-Sook Suh is the Head of Government Documents and Maps, and the Geography Liaison Librarian in the DP&S, Library. Email: hsu1@gmu.edu. Star Muir is an Associate Professor and the Hiring and Scheduling Director in the Department of Communication. Email: smuir@gmu.edu. Emily S. Ihara is an Assistant Professor in the School of Social Work. Email: eihara@gmu.edu.

APPLYING CENSUS DATA FOR SMALL AREA ESTIMATION

community needs and resources, improve the delivery of services, and empower communities and disenfranchised groups (Hoefler, Hoefler & Tobias, 1994; Robinson & Wier, 1998; Wier & Robinson, 1998; Hillier, 2007).

Introducing students to GIS methodologies can further enhance their ability to visualize and solve complex issues (Watkins, 2001). Potential issues to explore cover a vast range, for example, child maltreatment patterns based on neighborhoods (Ernst, 2000; Friesler, Lery, Gruenewald & Chow, 2006), housing patterns of persons with disabilities (Wolf-Branigin, 2002), or adult addiction epidemiology (Grant, Martinez & White, 1998; Gerwe, 2000; Maxwell, 2000). Empirical research has increasingly shown the significance of context in social problems, thereby pointing to a need for better understanding determinants such as place and time or community and neighborhood effects on outcomes for various populations.

Procedure

The following six steps provide access to software and data for creating and displaying small area estimates. These steps are: (1) downloading census data, (2) using MS Excel for descriptive analysis, (3) creating a DBF file from an Excel file, (4) downloading GIS boundary files, (5) importing Census data to a GIS environment, and (6) displaying themes.

The required software to complete these steps include: ArcView (GIS) software, Excel, Access, Internet Explorer (not Netscape) and Winzip (or any unzipping software). This method, designed for a course assignment entitled, Community Analysis Using GIS Technology, takes about two hours to complete.

Step 1: Downloading Census Data

Extract census variables such as low-income families (poverty), older people with low income, or other demographic variables (e.g., language speaking, Hispanic population) from an online database, called American Factfinder, and download them in MS Excel format. It is helpful to first create and name a folder on the C: drive in which to store the data before starting this exercise.

Census information needed is by census tract level in a county in northern Virginia.

- Visit the American Factfinder Website (<http://www.factfinder.census.gov>).
- Click on Data Sets (Decennial Census) in the left hand side menu of the website.
- In the 2000 section, select Census 2000 Summary File 3 (SF3)-Sample Data and click on the Detailed Tables on the right hand side.
- In the Select Geography type section, scroll down to Census Tracts in the drop down menu. In the state section, use the drop down menu to select Virginia; in the county section, use the drop down menu and select Fairfax County.

In the Select One or More Geographic Areas, select all census tracts and click on Add, then click Next.

The screenshot shows the American Factfinder website interface. At the top, there are navigation tabs: 'list', 'name search', 'address search', 'map', and 'geo within geo'. Below these is a header with 'Show all geography types' and a link to 'Explain Census Geography'. The main content area has several selection steps:

- Select a geographic type: A dropdown menu showing '..... Census Tract'.
- Select a state: A dropdown menu showing 'Virginia'.
- Select a county: A dropdown menu showing 'Fairfax County'.
- Select one or more geographic areas and click 'Add': A list box containing 'All Census Tracts' and a list of census tracts from 4151 to 4157. A 'Map It' button is next to the list.

At the bottom of the form is an 'Add' button with a dropdown arrow.

- Select Census Variables that for analysis and mapping. For example, P87 for poverty status in 1999 by age number of 65 years and over below the poverty level, P77 for median family income, P1 for total population, etc., and click on Add.
- Click on Show Result to view a summary of the table selected.
- In the Print/Download option in the top menu, click Download. This will provide various format options for downloading data.

Detailed Tables

You are here: Main > All Data Sets > Data Sets with Detailed Tables > Geography > Tables > Results

Use the links above to change your results | Options | Print / Download | Related

Note: use download to retrieve all selected tables and geographies

Print
Download
Load Query
Save Query
Download

P1: TOTAL POPULATION [1] - Universe: Total population
Data Set: Census 2000 Summary File 1 (SF 1) 100-Percent Data

geographies 1-10 of 165 Next

NOTE: For information on confidentiality protection, nonsampling error, and definitions, see <http://factfinder.census.gov/home/en/datatops/exps11u.htm>.

	Census Tract 4151, Fairfax County, Virginia	Census Tract 4152, Fairfax County, Virginia	Census Tract 4153, Fairfax County, Virginia	Census Tract 4154, Fairfax County, Virginia	Census Tract 4155, Fairfax County, Virginia	Census Tract 4156, Fairfax County, Virginia	Census Tract 4157, Fairfax County, Virginia	Census Tract 4158, Fairfax County, Virginia	Census Tract 4159, Fairfax County, Virginia
Total	3,237	3,032	3,783	7,744	5,744	2,478	3,622	4,434	3,034

U.S. Census Bureau
Census 2000

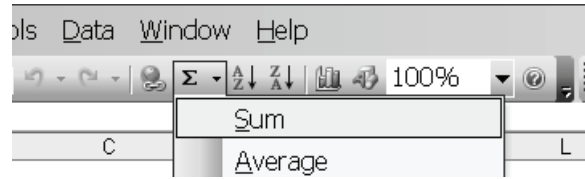
- Uncheck the Include Descriptive Data Element names box. Select MS Excel (.xls) and click OK. Next, save the zipped file to the folder created on the C: drive.
- Unzip the downloaded census file using Winzip software (or any other unzipping software) and save it in the created folder in C: as a file named census.

The census data that is downloaded gives two Excel sheets with information on the total population and median value with the names ending with geo and data1. For example, when the poverty by age data set is downloaded, it provides two excel sheets; one would be named dt_dec_2000_sf3_u_geo and the other would be named as dt_dec_2000_sf3_u_data1. The file data1 contains the population information, while the geo information contains the county geographic information such as census tracts, etc.

Step 2: Descriptive Analysis in Excel.

Open dt_dec_2000_sf3_u_data1 in Excel; this gives 18 columns starting with P087001. To make this exercise easier, change the name of these columns based on the description of the detailed table downloaded earlier (e.g., P087001 as total, etc.). In order to analyze the distribution of older people below the poverty level in Fairfax County, choose P087008 and P087008 and add these two columns for the total or sum to a new column (create a new column called, 65+) and add the last two columns of data (65+ and 75+) to this new column). To complete this, highlight the

columns to be summed (65+ and 76+) along with the newly labeled column. Click on the sigma sign (Σ) on the Excel toolbar as shown below while holding shift key.



This will add the two columns and give the total number of persons who are 65 years older in poverty status in each census tract.

	A	B	C	D	E	F	G
	GEO_ID	GEO_ID2	SUMLEVEL	GEO_NAME	P087008	P087009	sum
1							
2	140000551029415100	51029415100	140	Fairfax County, Virginia	0	10	10
3	140000551029415200	51029415200	140	Fairfax County, Virginia	12	9	21
4	140000551029415300	51029415300	140	Fairfax County, Virginia	21	19	40
5	140000551029415400	51029415400	140	Fairfax County, Virginia	3	0	3
6	140000551029415500	51029415500	140	Fairfax County, Virginia	26	25	51
7	140000551029415600	51029415600	140	Fairfax County, Virginia	7	29	36
8	140000551029415700	51029415700	140	Fairfax County, Virginia	0	0	0
9	140000551029415800	51029415800	140	Fairfax County, Virginia	0	0	0
10	140000551029415900	51029415900	140	Fairfax County, Virginia	0	7	7

If not already activated in Excel, activate the Analysis ToolPak in the Tools menu under Add-ins. Label one additional column to the right. Highlight the columns for analysis along with the new column and click on Tools and Data Analysis. Check Descriptive Statistics and then highlight the column to be analyzed and click on Summary Statistics. This will create a table with several summary statistics (e.g., mean, median, and mode).

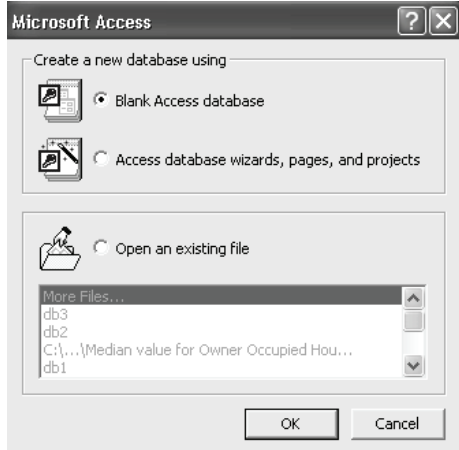
Step 3: Create a DBF file from an Excel file.

You must have MS Access to continue with this exercise. Before starting this step, check to see that the columns containing population information in the downloaded Excel sheets. This Excel file must be imported to the MS Access database in order to keep the Census tract number column, to relationally join the table to another table in GIS software such as ArcView, and to save the file to a dbf format.

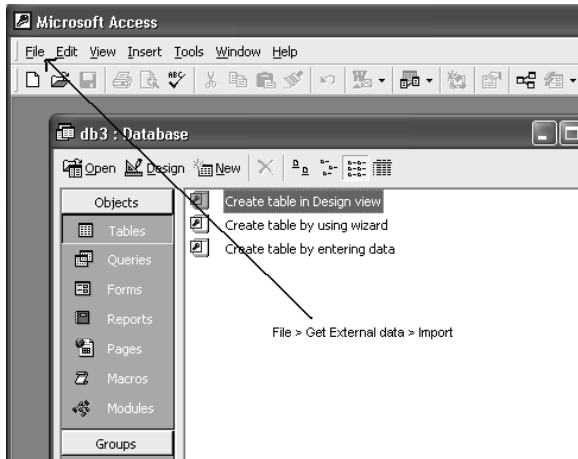
Importing Excel files into Access

- To import the Excel files into Access, open the Access database.

APPLYING CENSUS DATA FOR SMALL AREA ESTIMATION

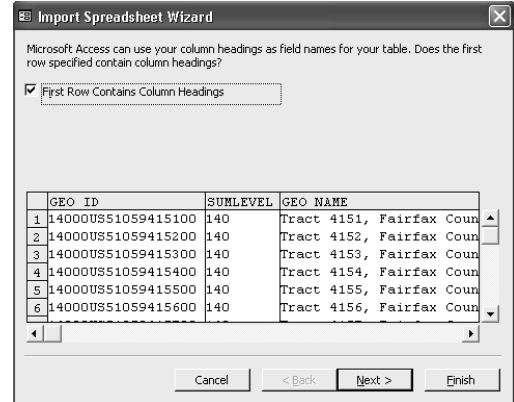


- Click on Blank Access database and on OK.
- Save db1 database in a temporary folder and click Create.
- Go to File>Get External Data>Import.



- Scroll to the folder where the Excel files are stored and open the Excel file containing the census tract data.

- Access opens a wizard to import Excel files. In the first screen make sure the First Row Contains Column Headings box is checked and then click Next.




- When asked where would you like to store data, check the In a New Table option, then click Next.
- In the current screen click Next.
- Click on No Primary Key and click Next.
- Name the table with the same name as the excel sheet containing the census data.
- The wizard will display that it has finished importing the excel sheet; click on OK.

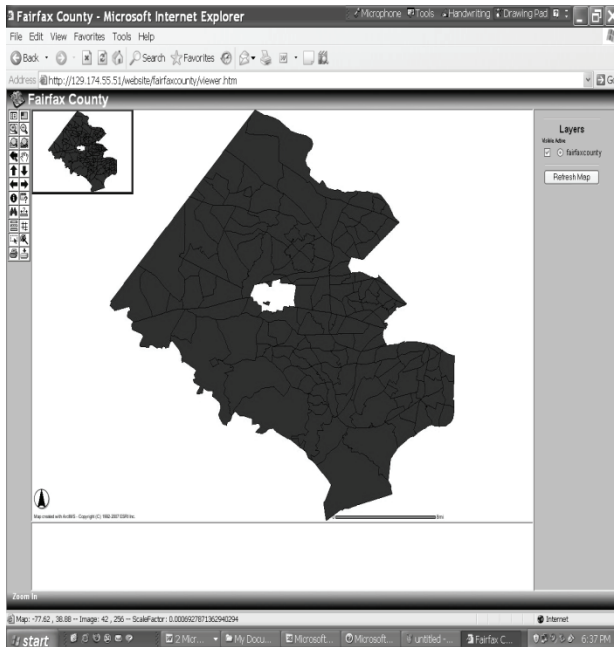
Delete any unnecessary columns and save the table in Access Database as a dbf format by going to File>Export and saving the table as census.dbf in the same folder created at the beginning of the exercise on the C: drive. Make sure the file is saved as a DBF IV format, which is ArcView compatible.

Exercise 2: GIS Mapping using Census Data

Step 1: Downloading Boundary Files for a County in Virginia

In order to map census data, a map boundary file which shows Census Tracts is needed.

- Go to <http://129.174.55.51/website> and choose the fairfaxcounty folder. Download this file using  sign from the menu on the left side of the screen. Unzip all three files (shp, shx and other) and save them all in a fairfaxCT folder on the C: drive.



Another boundary file may also be chosen, including Arlington, D.C. and other counties in Northern Virginia from this same web site. If downloading the boundary files from this site is problematic, go to the ESRI website at http://www.esri.com/data/download/census2000_tigerline. At this site, click on Freeview and Download on the left hand side of the page. Select a state, such as Virginia, and then select Census Tract 2000 from the Select by Layer category. Select a county, for example, Fairfax County, on which to conduct the analysis. Download the file (zipped). Unzip the three files (shp, shx, and other) and save them.

Step 2: Importing Census Data to GIS Environment.

In order to map census data (e.g. a distribution of elderly in the poverty level), the census data must be added to the boundary data file in GIS software such as ArcView. This involves importing the DBF file (Census.dbf) and joining it to the Fairfax County Boundary Data.

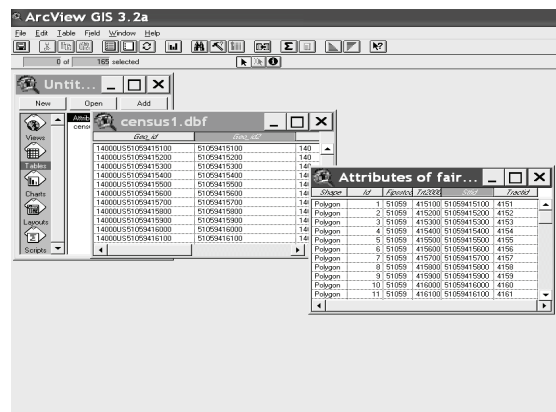
- To begin, open ArcView and choose the Select With a New View option.
- Say Yes to the question, would you like to add data now? and scroll to the folder where the Fairfax County shape files are stored. Open the census tracts folder (which is inside the fairfaxCT folder) and select the

tgr51059trt.shp (or 0.shp) file, which is the Fairfax county shape file. This creates a new view named View1 with the Fairfax County theme added to it (by using Theme-Property).


- Open the attribute table for the shape file by clicking on the Open Theme Table button.
- Next, go to the project window and click on the Tables icon on the left hand side: click the Add button to open a table.



- Browse and open the census.dbf file.
- To join the two tables first determine the common field (column) for the join. Looking at the two tables, the Tract column is common to both the tables. For the join, the attribute table for Fairfax County is the destination table and the dbf file is the source table.



APPLYING CENSUS DATA FOR SMALL AREA ESTIMATION

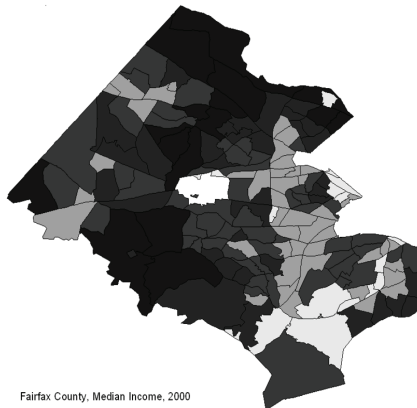
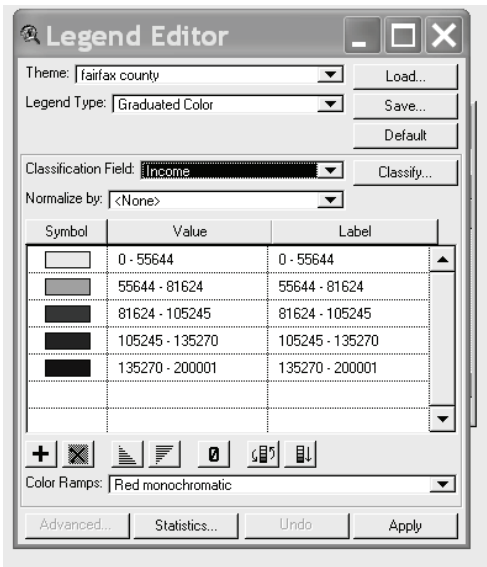
- Open the dbf table and click the Tract column name, then click on the Tract column name of the attribute table to highlight it. Make sure the attribute table is on top and click on the JOIN button . The dbf file table will automatically close after the join indicating that the join was successful.

- Save this project by going to File>Save as .apr extension and open the file in ArcView. This figure can be saved by going to File>Export and selecting the bitmap option. Open this .bmp extension file to MS Word by using the Insert menu (on the tool bar) and File Option (find your bmp file) in MS Word.

Step 3: Display Themes in ArcView.

To plot the calculated fields as maps, go to the project window and double click view1.

- Double click on this theme to view the legend editor. In the legend editor, select Graduated Color from the Legend Type field.
- Choose a suitable color ramp from the color ramps section and then click Apply.



Application and Discussion for Social Work

The range of education, health and family unit data readily available through the US Census, provides a valuable resource for persons working with a variety of populations. Local area data allows a user to focus on the identified needs within small geographic regions and to identify trends. Because the data can be used at the elementary and school district level, it further provides useful estimates for planners and evaluators dealing with a variety of complex issues such as resource allocation.

The procedure appears appropriate for advanced undergraduate and graduate levels in disciplines such as nursing, educational administration, social work and public health. Students and practitioners completing this procedure will acquire awareness and basic skills in downloading census data, using MS Excel for descriptive analysis, creating a DBF file from an Excel file, downloading GIS boundary files, importing Census data to a GIS environment, and displaying themes. Based on our experience, the typical student can complete the assignment in approximately two and one half hours assuming they have the access and have loaded the appropriate software.

References

- An, L., Linderman, M., Qi, J., Shortridge, A., & Liu, J. (2005). Exploring complexity in a human environment system: An agent-based spatial model for multidisciplinary and multiscale integration. *Annals of the Association of American Geographers*, 95(1), 54-79.
- Ernst, J. S. (2000). Mapping child maltreatment: Looking at neighborhoods in a suburban county. *Child Welfare*, 79(5). 555-572.

- Freishler, B., Lery, B., Gruenewald, P., & Chow, J. (2006). Methods and challenges of analyzing spatial data for social work problems: The case of examining child maltreatment geographically. *Social Work Research, 30*(4), 198-210.
- Gerwe, C. (2000). Chronic addiction relapse treatment: As study of the effectiveness of high-risk identification and prediction treatment model. *Journal of Substance Abuse Treatment, 19*(4), 415-444.
- Grant, D., Martinez, D., & White, B. (1998). Substance abuse among African-American children: A developmental framework for identifying intervention strategies. *Journal of Human Behavior in the Social Environment, 5*(2-3), 137-163.
- Hillier, A. (2007). Why social work needs mapping. *Journal of Social Work Education, 43*(2), 205-221.
- Hoefler, R.A., Hoefler, R.M., & Tobias, R.A. (1994). Geographic information systems and human services. *Journal of Community Practice, 1*(3), 113-128.
- Krieger, N. (2006). A century of census tracts: Health and the body politic. *Journal of Urban Health-Bulletin of the New York Academy of Medicine, 83*(3), 355-361.
- Maxwell, J. C. (2000). Methods for estimating the number of "hard-core" drug users. *Substance Use and Misuse, 35*(3), 399-420.
- Robertson, J.G., & Wier, K.R. (1998). Using geographic information systems to enhance community-based child welfare services. *Child Maltreatment, 3*(3), 224-234.
- Wier, K.R. & Robertson, J.G. (1998). Teaching geographic information systems for social work applications. *Journal of Social Work Education, 34*(1), 81-96.
- Weitzman, B., Silver, D., & Brazill, C. (2006). Efforts to improve public policy and programs through data practice: Experiences in 15 distressed American cities. *Public Administration Review, 66*(3), 386-399.
- Wilkins, R. L. (2001). Using geographic information system (GIS) technology to integrate research into the field practicum. *Journal of Technology in Human Services, 18*(1-2), 135-154.
- Wolf-Branigin, M., LeRoy, B., & Miller, J. (2001). Physical inclusion of people with developmental disabilities: An evaluation of the Macomb-Oakland Regional Center. *American Journal on Mental Retardation, 106*(4), 368-375.
- Wolf-Branigin, M. (2002). Applying spatial randomness to community inclusion. *Journal of Modern Applied Statistical Methods, 1*(1), 110-113.

Efficiency of Canonical Discriminant Function versus Mahalanobis Distance in Differentiating Groups: Screening Ovarian Cancer in a Multivariate System Analysis Using Enzyme Markers

Chinmoy K. Bose

Netaji Subhas Chandra Bose Cancer Research Institute, Kolkata, India

Due to its low prevalence, high mortality and uniquely hidden intrapelvic position, ovarian cancer remains a subject of intense interest to researchers. Statistical calculation and new technology both have major roles to play in the effort to screen this cancer at an early stage. Advanced statistics, such as multivariate analysis, remain at the root of screening endeavors. Multivariate analysis has the power to combine many tests and to produce better results in terms high specificity and positive predictive value. Multivariate analysis techniques include Mahalanobis distance (D^2), canonical stepwise discriminant function (Z) and Posterior Probability. These may have varied efficacy, but to date comparisons have not been conducted to determine which is best in the context of ovarian cancer screening.

Key words: Multivariate analysis, Mahalanobis distance (D^2), canonical stepwise discriminant function (Z), posterior probability, ovarian cancer screening, tumor marker.

Introduction

With an overall survival rate of 30%, ovarian cancer remains the fifth leading cause of cancer death. This disease, which is neither common nor rare (Bast, 2004), has remained enigmatic amongst gynecological cancers with agonizing prospects. Ovarian cancer is the second most common gynecologic malignancy, and little is known about the progression of its early changes (dysplasia).

Ovarian cancer has the highest mortality rate among gynecologic malignancies (70%) and its mortality rate has not lowered in the last 50 years. Only 25% of cases are diagnosed in an early stage and late case diagnosis survival is very poor. Though tests such as tumor markers and ultrasounds are available, no cost-effective

screening method with adequate sensitivity and specificity is available to detect early ovarian cancer.

Combining markers and tests results in higher sensitivity and specificity, thus, many scientists have used multivariate analysis in their experiments. In an ovarian cancer screening system, multivariate stepwise discriminant function analysis is described using different tumor markers, for example, CA125, TPA, IAP, CEA, and ferritin (Yabushita, et al., 1985; LaHousen, et al., 1987). Kobayashi and Terao (1992) combined CA 125, TPA, Ferritin, CEA, AFP and Sialyl Lewis Xi using Mahalanobis distance and were able to decrease both false positive and false negative cases. Bose and Mukherjea (1994) statistically combined several enzymatic tumor markers to increase specificity, positive predictive value (PPV) and to decrease false positive tests.

Other groups described combining multiple markers, but they either combined them in a statistically unacceptable way (Inoue, Fujita, Nakazawa, Ogawa & Tanizawa, 1992), in a simple Euclidian relationship, such as the risk of malignancy Index (RMI, Oram, et al., 1990; Jacobs, et al., 1990) or otherwise (Jacobs, et al.,

Chinmoy K Bose is a consultant gynecological oncologist presently working in Kolkata, India. He works on ovarian cancer and has special interest in statistical methods like multivariate analysis and non linear dynamics. He has many publications in his subject and publishing a book on endocrine feedback. Email him at: ckbose@hotmail.com.

1990). Jacobs, Oram & Bast (1992) used a multivariate system while also using apolipoprotein A1 (down-regulated in cancer); a truncated form of transthyretin (down-regulated) and a cleavage fragment of inter- α -trypsin inhibitor heavy chain H4 (up-regulated).

Zhang, Bast, et al. (2004) described the risk of ovarian cancer (ROC) algorithm. They combined the parameters Serial CA125 assay value, changes in CA125 levels over time and woman's age, and assay variability by a multivariate based software program, which they called the ROC algorithm. However, they did not describe the actual procedure they followed. They speculated sensitivity 86%, specificity 99.7%, and PPV up to 19% which is encouraging (Menon, et al., 2005). They are now conducting a massive trial on population screening in the UK, which will take another two years to complete.

Timmerman, et al. (2005) combined 12 useful independent prognostic variables in a logistic regression model and found a probability cut off value of 0.10 that gave a sensitivity of 93% and a specificity of 76%. Curling, et al. (1998) conducted a multivariate analysis of DNA ploidy, steroid hormone receptors and CA 125 as prognostic factors in ovarian carcinoma, and Kozak, et al. (2005) used multivariate analysis to greatly improve the detection of early stage ovarian tumors compared to cancer antigen CA125 alone with the help of differential expression of transthyretin (TTR), beta-hemoglobin (Hb), apolipoprotein AI (ApoAI) and transferrin (TF).

Multivariate procedures include Mahalanobis distance (D^2), canonical stepwise discriminant function (Z) and Posterior Probability, but no research has been conducted to determine whether they are equally effective in detection systems for screening ovarian cancer using multiple parameters.

Methodology

Serum levels of four enzyme makers (placental alkaline phosphatase, lactate dehydrogenase, 5' nucleotidase and Amylase) were measured using a commercially available kit, in 50 ovarian cancer patients and 31 patients with benign gynecological disease before initiation of any

treatment. These were compared with the levels in a control group of 30 healthy women using different multivariate parameters Mahalanobis distance (D^2), canonical stepwise discriminant function (Z) and Posterior Probability. The goal was to determine if any difference exists in the power of detection of disease state by these methods and if one is more or most efficient in detecting disease state.

Data for all enzyme levels in different groups were fed into a DIGITAL-VAX 8650 computer using a VMS operating system. BMDP 1990 version software program packages 3D and 7M were used to analyze the data. In BMDP 3D, mean, standard deviation, standard error of mean and pooled T test were used to show significant group differences separately for each enzyme. Sensitivity and specificity for each enzyme were determined at different cut off scores and a Receiver Operator Characteristic Curve (ROC) was prepared to compare the efficacy of individual enzyme.

In the same program, Hotelling's T^2 test, F, p for four enzymes taken together at a time (multivariate analysis) were obtained and were analyzed to observe significant differences between different groups. The F value was observed and, if it significantly exceeded unity, the two groups were assumed to be statistically significantly different.

If a random sample of size n yields the sample value $x_1, x_2, x_3, \dots, x_n$

$$\bar{X} = \sum_{i=1}^n \frac{x_i}{n},$$

and the sample estimate of variance is

$$s^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(n-1)},$$

then these are estimates of corresponding population parameters – the population mean μ and the population variance σ^2 .

In a similar way, multivariate population can be summarized by mean vectors and covariance matrices. These are defined as follows. If there are p variables $x_1, x_2, x_3, \dots, x_p$ and the values of these for the i^{th} individual in a

CANONICAL DISCRIMINANT FUNCTION VERSUS MAHALANOBIS DISTANCE

sample are $x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}$ respectively, then the sample mean of variable j is

$$\bar{X}_j = \sum_{i=1}^n \frac{X_{ij}}{n}$$

and the sample variance is

$$s_j^2 = \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)^2}{(n-1)}.$$

In addition the sample covariance between variable j and k is defined as

$$c_{jk} = \sum_{i=1}^n \frac{(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{(n-1)}.$$

The pooled estimate of variance from the two sample n_1 and n_2 is,

$$s^2 = \frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{(n_1 + n_2 - 2)},$$

the matrix of covariances (C_1 and C_2)

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{23} \\ \dots & \dots & \dots & \dots \\ c_{p1} & c_{p2} & \dots & c_{pp} \end{pmatrix},$$

the pooled estimate of covariance matrix is

$$C = \frac{[(n_1 - 1)C_1 + (n_2 - 1)C_2]}{(n_1 + n_2 - 2)},$$

and Hotelling's T^2 statistics is defined as

$$T^2 = \frac{n_1 n_2 (\bar{X}_1 - \bar{X}_2) C^{-1} (\bar{X}_1 - \bar{X}_2)}{(n_1 + n_2)}.$$

A significantly large value for these statistics is evidence that the mean vectors are different for the two sample populations. The significance or the lack of significance of T^2 is most simply determined by using the null hypothesis case of

equal population means for the transformed statistics.

The analysis of variance (also known as Snedecor's F or the Fisher-Snedecor F) test is based on the continuous F-distribution, which is a random variate arising as the ratio of two Chi-squared variates:

$$\frac{U_1/d_1}{U_2/d_2},$$

where U_1 and U_2 have Chi-square distributions with d_1 and d_2 degrees of freedom respectively, and U_1 and U_2 are independent. Thus,

$$F = \frac{(n_1 + n_2 - p - 1)T^2}{[(n_1 + n_2 - 2)p]}$$

Because T^2 is a quadratic form it is scalar, and can be written in as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^p (\bar{X}_{1i} - \bar{X}_{2i}) c_{ik} (\bar{X}_{1k} - \bar{X}_{2k}),$$

which is simpler to compute. Here \bar{x}_{ji} is the mean of the variable x_i in the j^{th} sample and c_{ik} is the element in the i^{th} row and the k^{th} column of the inverse matrix C^{-1} .

BMDP 7M was used for multivariate stepwise canonical discriminant function analysis. To separate the different groups of patients, following simple linear combination was used

$$Z = K + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4.$$

Z is the canonical discriminant function of variable enzymes namely $X_1 = \text{PLAP}$, $X_2 = \text{LDH}$, $X_3 = 5'N$, $X_4 = \text{Amylase}$, whereas $a_1, a_2, a_3,$ and a_4 are the coefficients of the above variable respectively and K is the constant. Coefficient and constant were determined by using BMDP.

Mahalanobis distance of individuals to group centers can be calculated by the following formula

$$D_i^2 = (x - \bar{X}_i)' C^{-1} (x - \bar{X}_i) \\ = \sum_{r=1}^p \sum_{s=1}^p (x_r - x_{ri})' c_{rs} (x_s - x_{si}),$$

while the posterior probability is

$$P(A : PT) = \frac{P(PT : A)P(A)}{P(PT : A)P(A) - P(PT : NotA)P(NotA)}$$

where p is probability, A is the abnormality, and PT is the positive test result. Thus, the expression on the left, P (A: PT) is equivalent to the probability P of abnormality A given the positive test result PT. This is described as posterior probability and is in Bayes' theorem, which relates the conditional and marginal probabilities of stochastic events A and B:

$$\Pr(A|B) = \frac{\Pr(A|B) \Pr(B|A)}{\Pr(B)} \propto L(A|B) \Pr(A)$$

where L(A|B) is the likelihood of A given fixed B. Each term in Bayes' theorem has a conventional name. Pr(A) is the prior probability or marginal probability of A. It is prior in the sense that it does not take into account any information about B. Pr(A|B) is the conditional probability of A given B; it is also called the posterior probability because it is derived from or depends upon the specified value of B. Pr(B|A) is the conditional probability of B given A. Pr(B) is the prior or marginal probability of B, and acts as a normalizing constant. The posterior probability is proportional to the prior probability times the likelihood.

Both D^2 and PP were determined through the same package. The ROC curve of individual marker enzymes showed LDH to be most sensitive and specific. Thus, LDH was compared with these three multivariate systems in Receiver Operator Characteristic (ROC) chart.

In healthy women discriminant function (Z), Mahalanobis distance (D^2) and posterior probability (PP) were determined for each case. Their mean and standard deviation were

compared with the values of corresponding multivariate parameters for each individual in both the benign gynecological disease (BGD) and ovarian cancer group. When plotted with chosen cut-off scores with their corresponding (1-specificity) in x axis and sensitivity in y axis the ROC curve will show comparative efficacy of Z over D^2 , PP and LDH in terms of largest area under the graph.

Results

Table 1 shows the serum concentration of four enzymes markers, placental alkaline phosphatase, lactate dehydrogenase, 5' nucleotidase and Amylase, with their mean \pm standard error of mean. Results of enzyme estimations of three groups, Healthy control, Benign gynecological disease (BGD) and Ovarian cancer are shown in three columns. Significant differences are also shown as p values. Although activities of all enzymes have been found to be significantly higher in ovarian cancer cases than in healthy women, positivity rates were not very high. The positivity rate measured for each enzyme in ovarian cancer, showed LDH to be the most sensitive (positivity 48%), whereas amylase showed least sensitivity at a positivity rate of 30%. However, these are much lower compared to the positivity rate (75%) of cancer antigen CA 125, the most sensitive tumor marker in ovarian cancer (Heinonen, Kallinoiemi & Koivula, 1987).

Test results in the ovarian cancer group were significantly different from the healthy control group, but showed no statistically significant difference with the benign gynecological disease (BGD) group.

Table 2 shows sensitivity and specificity of serum enzymes markers at different cut-off concentrations. Table 3 summarizes the cut-off scores of the markers that had the highest sensitivity and specificity. They were compared with CA125 at a suitable cut off level of 35 IU/L. LDH had the highest sensitivity and specificity at a cut off score of 157.88 IU/L but it still fell behind CA125.

With different sensitivity and specificity at cut-off values of those enzyme markers, a Receiver Operator Characteristic Curve (ROC) was prepared to compare the power and efficacy

CANONICAL DISCRIMINANT FUNCTION VERSUS MAHALANOBIS DISTANCE

of individual enzymes to differentiate between groups in a univariate system (Figure 1).

Four serum enzymes are used at a time in multivariate analysis by the BMDP package (program 3D), where a significant group difference was observed between healthy versus ovarian cancer and BGD versus ovarian cancer patients; in healthy versus BGD there was no statistically significant difference (Table 4). Z , D^2 and PP were obtained through program 7M for healthy versus ovarian cancer patients. Sensitivity and specificity of LDH, Z , D^2 and PP at different cut-off scores or action lines with their confidence interval (derived from the binomial distribution chart) for healthy women and ovarian cancer cases are shown in Table 5. Table 6 compares sensitivity and specificity of different multivariate parameters such as, LDH, Z , D^2 and PP.

The performance of the canonical discriminant function (Z) at various upper limits is illustrated in the ROC chart (Figure 2) and is observed to combine higher levels of sensitivity and specificity than those achieved by Mahalanobis distance (D^2), PP and LDH. Table 7 shows the positive and negative predictive values for malignancy of ovary for different levels of Z (cutoff scores 1.377, 2.907, 3.437 and 5.967). A cut-off score of 3.437 produced the best results. No statistically significant group difference was predicted between healthy and BGD, which was corroborated by the determination of Z value.

Conclusion

The population screening of ovarian cancer has remained elusive due to low disease prevalence and low positive predictive value of the tests. Some groups are trying to combine different test results in different software packages using algorithms based on multivariate systems of data processing, but many alternatives in this multivariate system exist which are not based on some type of mathematical calculation. As a result, finding more efficacious methods in terms of higher specificity and higher positive predictive value is a priority. This system is applicable in many areas in biology and medicine. This article presented an example of the use of multivariate analysis in ovarian cancer

screening to illustrate the comparative efficacy of stepwise discriminant function (Z) Mahalanobis distance (D^2) and posterior probability (PP).

It is expected that this example will be replicated in other experimental circumstances, but will need further verification and establishment of mathematical proof as to why it occurs. In the experiment presented, it was observed that the Multivariate stepwise discriminant function (Z) analysis of enzyme variables establishes an easy quantitative assessment method of the risk of malignancy in the ovary. A Z value with a cut-off score of 3.437 has a higher predictive value and relative risk than LDH, Mahalanobis distance (D^2) or posterior probability (PP). This system of combining four enzymes for improvement of ovary screening must be established in clinical practice through further research.

References

- Bast, R. C. Jr. (2004). Early detection of ovarian cancer: New technologies in pursuit of a disease that is neither common nor rare. *Trans American Clinical Climatological Association*, 115, 233-248.
- Bose, C. K., & Mukherjea, M. (1994). Enzymatic tumor markers in ovarian cancer: a multiparametric study. *Cancer Letters*, 77(1), 39-43
- Curling, M., Stenning, S., Hudson, C. N., & Watson, J. V. (1998). Multivariate analyses of DNA index, p62c-myc, and clinicopathological status of patients with ovarian cancer. *Journal of Clinical Pathology*, 51(6), 455-461.
- Heinonen, P. K., Kallinoiemi, O. P., & Koivula, T. (1987). Comparison of CA 125 and placental alkaline phosphatase as ovarian tumour markers. *Tumori*, 73(3), 301-302.
- Inoue, M., Fujita, M., Nakazawa, A., Ogawa, H., Tanizawa, O. (1992). Sialyl-Tn, sialyl-Lewis Xi, CA 19-9, CA 125, carcinoembryonic antigen, and tissue polypeptide antigen in differentiating ovarian cancer from benign tumors. *Obstetrics and Gynecology*, 79, 434-440.

Table 1: Serum Concentration of Enzymes Markers Mean \pm SEM

Enzymes	Healthy Control	Benign Gynecological Disease (BGD)	Ovarian Cancer
Placental alkaline phosphatase (IU/L)	0.81 \pm 0.09	1.76 \pm 0.47 (P< .0615)	4.47 \pm 0.89 (P< .0011)
Lactate dehydrogenase (IU/L)	157.88 \pm 8.61	155.65 \pm 7.88 (P< .8497)	255.44 \pm 16.19 (P< .0001)
5' nucleotidase (IU/L)	5.94 \pm 0.75	5.22 \pm 0.42 (P< .4098)	9.13 \pm 0.93 (P< .0191)
Amylase (IU/L)	79.1 \pm 3.83	77.8 \pm 3.38 (P< .6828)	121.54 \pm 8.23 (P< .0001)

Table 2: Sensitivity and Specificity of Serum Enzymes Markers at Different Cut-off Concentrations

Test and Action Line (cutoff score in IU/L)	Sensitivity	Specificity
LDH		
110.73	94	20
157.88	88	63
205.03	50	90
252.18	34	93.3
Amylase		
75.25	76	43.3
79.1	74	53.3
83.93	68	60
86.76	64	60
PLAP		
0.72	64	50
0.81	60	60
0.90	58	66.6
5'Nucleotidase		
5.19	56	56.6
5.94	54	63.6
6.69	42	66.6
7.44	39	66.6

CANONICAL DISCRIMINANT FUNCTION VERSUS MAHALANOBIS DISTANCE

Table 3: Cut-off Score Offering Highest Sensitivity and Specificity

Cutoff Scores (IU/L)	Enzyme Markers	Sensitivity	Specificity
0.90	Placental alkaline phosphatase (IU/L)	58	66.6
157.88	Lact dehydrogenase (IU/L)	88	63
83.93	Amylase (IU/L)	68	60
5.95	5' nucleotidase(IU/L)	54.8	64.3
35	CA125	72	75

Figure 1: Receiver Operator Characteristic Curve (ROC) to Compare the Power of Individual Enzymes

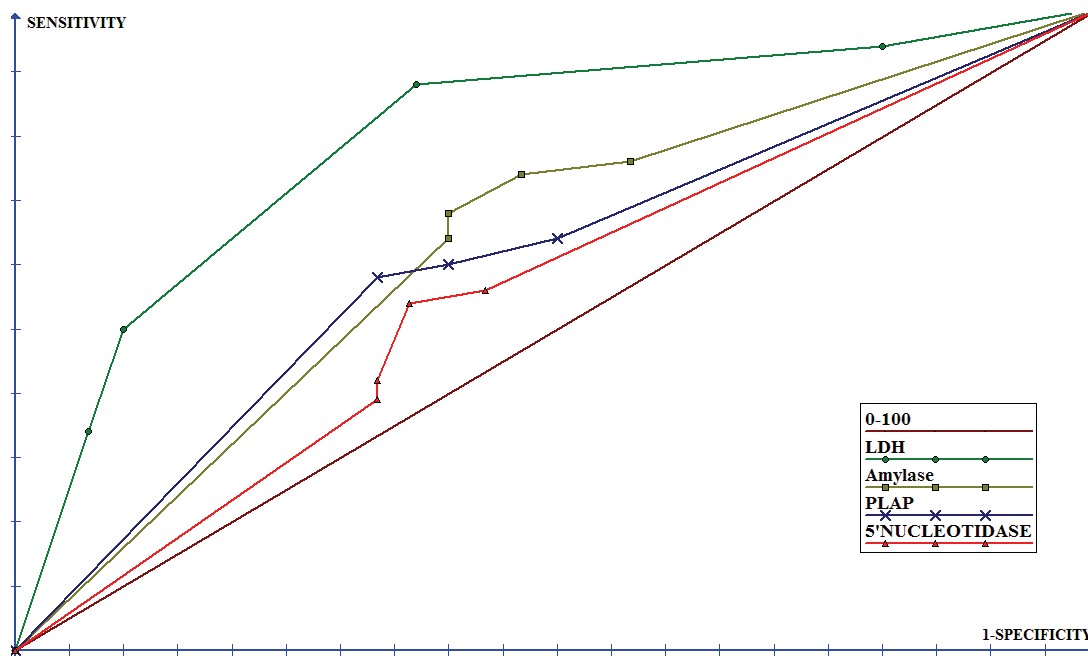


Table 4: Mahalanobis Distance (D^2) Hotelling's T^2 F & P Value to Predict Multivariate Based Statistical Significance of Different Between Groups

Group Difference	Mahalanobis Distance (D^2)	Hotelling's T^2	F	P
Healthy women vs. ovarian cancer	2.7061	50.7390	12.1969	0.0001
Healthy women vs. Benign gyn. disease	0.2992	4.5611	1.0823	0.3741
Ovarian cancer vs. Benign gyn. disease	2.4808	54.3602	13.0740	0.0001
Ovarian cancer vs. non-responder	0.0530	1.0721	0.2582	0.9028
Ovarian cancer vs. responder	1.3802	12.444	2.9528	0.0277

Table 5: Sensitivity and Specificity of Multivariate Based Statistical Parameters Compared With LDH at Different Cut-off Concentrations

Test and action line (cutoff score in IU/L)	Sensitivity		Specificity	
	%	(95% CI)	%	(95% CI)
LDH				
110.73	94	(82-99)	20	(82-99)
157.88	88	(82-99)	63	(82-99)
205.03	50	(82-99)	90	(82-99)
252.18	34	(82-99)	93.3	(82-99)
Z				
1.337	98	(88-100)	13	(82-99)
2.907	96	(86-99)	70	(82-99)
3.437	96	(86-99)	83	(82-99)
5.967	76	(61-87)	93.3	(82-99)
D^2				
0.09	100	(92-100)	0	(0-12)
0.93	86	(72-94)	73	(54-87)
1.77	82	(68-91)	86.6	(66-96)
2.61	74	(59-86)	93	(77-99)
PP				
0.525	80	(67-90)	90	(72-97)
0.726	92	(80-92)	76.6	(56-89)
0.887	96	(86-99)	20	(10-58)
1.048	100	(92-100)	0	(0-12)

Table 6: Sensitivity and Specificity of Different Multivariate Parameters Such as, LDH, Z, D^2 and PP

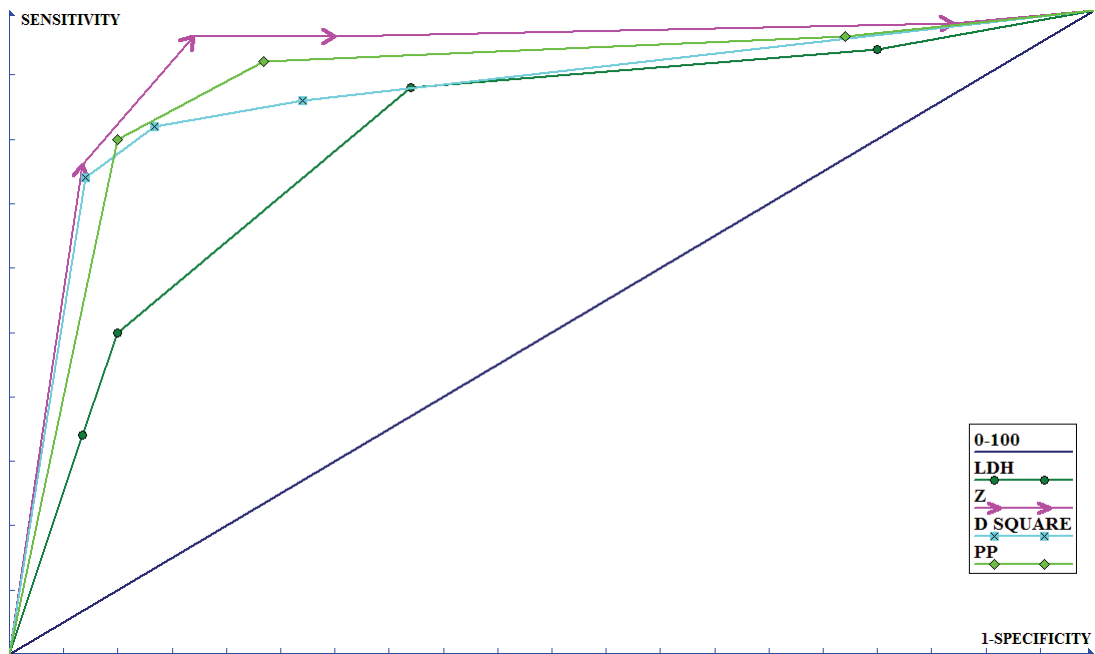
Statistics	Sensitivity	Specificity
Mahalanobis Distance $D^2 = 1.77$	82	86.3
Posterior Probability $P_p = 0.726$	92	76.6
Discriminant Function $Z = 3.437$	96	83

CANONICAL DISCRIMINANT FUNCTION VERSUS MAHALANOBIS DISTANCE

Table 7: Positive and Negative Predictive Value for Different Levels of Z

Z Score	Predictive Value (%)	
	Positive	Negative
1.337	91.3	3.7
2.907	91.3	9.6
3.437	84	92
5.967	70	95

Figure 2: Receiver Operator Characteristic Curve (ROC) to Compare the Power of Multivariate Based Statistical Parameters Compared with LDH



Jacobs, I. J., et al. (1990). A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *British Journal of Obstetrics and Gynaecology*, 97(10), 922-929.

Jacobs, I. J., Oram, D., & Bast, R. C. Jr. (1992). Strategies for improving the specificity of screening for ovarian cancer with tumor-associated antigens CA 125, CA 15-3, and TAG 72.3. *Obstetrics and Gynecology*, 80(3), Part 1, 396-399.

Kobayashi, H., & Terao, T. (1992). Field trial for the early detection of patients with ovarian cancer. *Rinsho Byori*, 40(2), 139-145.

Kozak, K. R., Su, F., Whitelegge, J. P., et al. (2005). Characterization of serum biomarkers for detection of early stage ovarian cancer. *Proteomics*, 5(17), 4589-4596.

LaHousen, M., Stettner, H., Prckel, J., Urdl, W., & Purstner, P. (1987). The predictive value of a combination of tumor-markers in monitoring patients with ovarian cancer. *Cancer*, 60(9), 2228-2232.

Menon, U., et al. (2005). Prospective study using the risk of ovarian cancer algorithm to screen for ovarian cancer. *Journal of Clinical Oncology*, 23(31), 7919-7926.

Oram, D. H., Jacobs, I. J., Brady, J. L., Prys-Davies, A. (1990). Early diagnosis of ovarian cancer. *Br J Hosp Med.*, 44(5), 320-324.

Timmerman, D., Testa, A. C., Bourne, T., et al (2005). Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology*, 23(34), 8794-8801.

Yabushita, H., et al. (1985). Diagnostic usefulness of stepwise discriminant analysis employing the values of CA 125, TPA, IAP, CEA and Ferritin in sera measured simultaneously for gynecological malignant neoplasm. *Nippon Sanka Fujinka Gakki Zasshi*, 37(9), 1883-1892.

Zhang, Z., Bast, R. C. Jr., et al. (2004). Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Research*, 64, 5882-5890.

A Socratic Dialogue

Vance Berger
National Institute of Health

Socrates has found some aspects of medical biostatistics a bit confusing, and wishes to discuss some of these issues with Simplicio, a prominent medical researcher. This Socratic dialogue will shed some light on the errant use of parametric analyses in clinical trials.

Key words: Exact test, parametric analysis, permutation test.

Introduction

Socrates: Good morning Simplicio, how are you today?

Simplicio: Doing well, thank you, and how are you Socrates?

Socrates: Not bad, thank you, but a bit confused by some of these newfangled ideas I am now seeing in the medical literature. Tell me, Simplicio, is it not the case that you also contribute to this medical literature? If so, then you must be somewhat of an expert, and certainly in a position to teach me some of the analyses so that I will no longer be confused.

Simplicio: Yes, Socrates, in fact I was part of a research team that recently published a clinical trial in a prestigious medical journal. Would you like a reprint?

Socrates: No thank you, I have already read it. And it contributed to my confusion.

Simplicio: How so, Socrates?

Socrates: In many ways, but let us focus, at least for now, on just one of these ways. You mention that you will compare the blood pressures between the treatment groups by using a t-test, is that right?

Simplicio: Yes, although I fear that, being a layman, you are not using sufficiently precise language. The primary endpoint in our cardiac trial was the diastolic blood pressure 12 weeks after treatment. It is this endpoint that we compared with a t-test.

Socrates: That is all very well, but my interest at the moment is in the t-test itself, and not in the specific details of the variable on which it was used. I thought that I had read somewhere that the t-test requires normality to be valid, is this not so? And I also read about permutation tests that do not require normality for their validity.

Simplicio: Technically, yes, but in practice the distributions are close enough to Gaussian that we can treat them as such. And we do not use permutation tests for a variety of reasons.

Socrates: Pray tell me these reasons, dear Simplicio.

Simplicio: For one thing, permutation tests use an overly restrictive null hypothesis, specifically that the entire distribution of outcomes is the same across treatment groups. In contrast, the t-test is testing only the equality of the means.

Vance W. Berger is a mathematical statistician with the National Cancer Institute. Email: Vance917@gmail.com.

Socrates: So the permutation test would be sensitive to changes in spread and/or shape, whereas the t-test would not?

Simplicio: Yes, I believe this to be true.

Socrates: But I also read that the t-test requires equal variances, or homogeneity, to be valid. Does this mean that without equal variances it is not valid, or might have a high probability of rejecting a true null hypothesis?

Simplicio: We compute the p-value under the assumption that the null hypothesis is true, so this would specify that the variances are equal.

Socrates: So the null hypothesis is that the means are the same and that the variances are the same, across the two treatment groups?

Simplicio: Quite so.

Socrates: Did you not tell me that the benefit of the t-test was the ability to test nothing more than the equality of the means?

Simplicio: I need to confer with my text book, but remember, that was only one reason. We also use the t-test because it is robust to violations of its assumptions.

Socrates: Robustness sounds nice. What does it actually mean? If the data are not normally distributed, and/or the variances are not equal, then the t-test p-value is the same as it would have been had the data been normally distributed and the variances equal?

Simplicio: Yes, I believe so.

Socrates: If the variances are unequal, then we can make them equal by increasing the smaller to match the larger, by decreasing the larger to match the smaller, by bringing them both in to the mean (or geometric mean or harmonic mean), or in any other of a myriad number of ways. The t-test p-value is the same as which one of these? Or are they all the same?

Simplicio: Yes, I would say that they will all be the same.

Socrates: Is it not the case that with larger variances the p-value will be larger, and with smaller variances the p-value will be smaller?

Simplicio: Yes, I am afraid so.

Socrates: So then would you agree that the t-test p-value cannot possibly agree with all possible values of the t-test p-value when the variances across groups are equal?

Simplicio: Yes, I am afraid so.

Socrates: Once again, what does this supposed robustness mean?

Simplicio: I was mistaken, but now I remember. Robustness means that even if the assumptions are violated, the t-test p-value will still be close to the exact one.

Socrates: Is there but one exact p-value to be close to?

Simplicio: There is only one way to conduct an exact permutation test when using the same randomization scheme as was used in the study and the t-test statistic.

Socrates: I will agree that this is a well-defined p-value, this exact t-test p-value. So your statement is beginning to take some form, but there is still ambiguity in the closeness concept. Can we say that the difference in p-values is bounded by some function of the extent to which the assumptions underlying the t-test are violated?

For example, if R is the ratio of variances across the two groups, and D is the difference between the t-test p-value and the exact t-test p-value, then can we say something to the effect that $|D| \leq \log(R)$? I should be quite interested in any theorem of this sort, especially if it accounts for and quantifies deviations from both normality and homoscedasticity.

Simplicio: I am not aware of any such theorems, but in practice the two p-values are usually close. That is, D is usually quite small.

A SOCRATIC DIALOGUE

Socrates: Do you have the values of D from prior studies to substantiate this assertion?

Simplicio: No.

Socrates: Do you even bother to compute the exact p -value?

Simplicio: We do if the assumptions are grossly violated.

Socrates: You mean if the assumptions are violated enough that D would be large?

Simplicio: Yes.

Socrates: Yet you never actually compute D ?

Simplicio: Correct.

Socrates: So you presume to know when D is large or small based on a cursory examination of the extent to which the assumptions are violated, then take the smallness of D in these cases as a known fact with which to justify continuing in this fashion? Is this not circular reasoning?

Simplicio: Perhaps so, but we use the exact test when we need to.

Socrates: You said you do this when the assumptions are violated enough that D would be expected to be large. Why not use the t -test even in these cases?

Simplicio: Socrates, you are not seriously suggesting that we use the t -test when its assumptions are known to be grossly violated? Especially after grilling me for using it when the assumptions are violated to a lesser degree?

Socrates: My good man, I am not suggesting anything. Recall that you are the clinical trials expert, and I am merely trying to learn from you. Right now I want to learn why you do not use the t -test when the assumptions are badly violated.

Simplicio: I am afraid that this is a trap, and you are asking me an obvious question just to see what I will say, but the reason is that we do not

want to use the t -test if its assumptions are badly violated because then it may give distorted results.

Socrates: When you say “distorted” you are referring implicitly to deviation from some gold standard, presumably the exact test?

Simplicio: Yes, that is correct.

Socrates: Is it the exact p -value, and not the t -test p -value, that is of interest? It was conceivable that the t -test itself was the quantity of interest, but now it appears that this is not the case, and that when you use the t -test, you do so only so that it can serve as an approximation to the exact p -value?

Simplicio: Quite right Socrates.

Socrates: I understand the need for approximations in some cases. For example, one could compute the number of defective items in a large batch by examining each one, but this would consume large amounts of resources, so a sample is taken and an estimate based on this sample is offered as an approximation so as to save time and money.

Simplicio: Yes, that is a good example.

Socrates: Similarly, when you want to compute the area under the curve of some function that is not written explicitly in closed form, you could graph the function on your computer screen, trace the region below it with a marker, get a glass cutter, cut out the glass from the screen to correspond to this area, then weigh the glass. But instead you rely on an approximation so as to save the computer screen, is that correct?

Simplicio: Yes, I suppose so.

Socrates: Do you see the common element in these two examples?

Simplicio: Yes, in both cases we needed to use an approximation.

Socrates: No Simplicio, we did not need to use an approximation, but we chose to do so in order to save resources.

Simplicio: Yes, that was what I meant.

Socrates: When you use the t-test as an approximation, what resources are you saving?

Simplicio: What do you mean?

Socrates: What great cost is involved in computing the exact test p-value? Clearly, you can compute it, since you just told me that you would compute it if the situation so warranted. I am trying now to get some sense of the cost-benefit ratio in doing so. Do you need to rent time on the university super computer to compute the exact p-value.

Simplicio: No, Socrates, computing has gotten to the point that I can compute the exact p-value instantaneously on my PC.

Socrates: Is the exact test patented, so that you need to pay royalties to use it?

Simplicio: No Socrates, that is not it either.

Socrates: Why don't you just tell me the reason?

Simplicio: There is no additional cost in computing the exact p-value.

Socrates: I see. But I am not sure that I like what I hear. You have no reason not to compute the exact p-value, yet choose not to do so even though your decision to use it or not to use it is based on how well an approximation approximates it. And you assess this closeness not by computing both quantities and simply comparing them but rather by using some vague notion of how well the assumptions of the approximation seem to hold, even though you readily admit that this has no implications for an upper bound on the difference between the two p-values.

Then you count the times that you ostensibly do not need to compute the exact p-value and offer this as further evidence of successes without the exact p-value, so more reason not to have to use it in the future. Tell

me, Simplicio, can you offer a valid reason for this approach instead of simply computing both p-values and assessing the difference in this way?

Simplicio: No, I am afraid that I cannot.

Socrates: Would you agree that it would be better to dispense with this nonsense about testing the assumptions underlying the t-test, or similarly checking that expected cell counts exceed five for the chi-square test, and instead just compute both p-values, and note how close or far they are to each other? After all, how much power would you expect these tests to have to detect deviations from normality (or some other distribution) when the sample sizes are chosen not for this purpose but rather to detect a treatment effect?

Simplicio: Yes, this would be better.

Socrates: Let us anticipate your doing this in the future. You will then have an exact p-value as the gold standard, and you will have an approximation to it, the t-test p-value. How will you use these two to render a decision as to the suitability of the t-test?

Simplicio: Socrates, as we already said, I would use the approximation only if it is close enough to the exact p-value.

Socrates: When you go to the market for groceries, and the cashier totals the price of your selected merchandise, do you pay this amount, or some other amount that is close enough to this amount? I mean, one could obtain the dollar amount for the items in question, then toss two dice, and add (in cents) the value showing on the first die and subtract the value showing on the second die. The deviation would be no more than six cents either way.

Simplicio: Of course, I pay the requested amount.

Socrates: If you had a wrist watch with the approximate time, but also were able to see a clock with the exact time (which I could not

A SOCRATIC DIALOGUE

see), then what would you do if I, with no watch, were to ask you the time?

Simplicio: I would imagine that I would tell you the time.

Socrates: But how would you obtain the time?

Simplicio: You just told me that there is a watch and a clock, so I can't imagine having too much difficulty in telling the time. You seem to be belittling my intelligence, Socrates, but I assure you that even I can tell time.

Socrates: I meant no offense, Simplicio, and rather meant to ask only which measure of time you would use.

Simplicio: Because the clock has the exact time, I would use that one when it were available, as you said it would be in this case. I would use my watch only when I could not see the clock, or some other clock with a more precise measure of the time.

Socrates: You would not check both the watch and the clock, and then decide to report the time on the watch if it were sufficiently close to the exact time on the clock?

Simplicio: No, Socrates, this seems to me rather silly. If I can just check the exact time and tell you that, then why would I also check an approximation to a quantity I can observe?

Socrates: If you can observe the exact p-value, then why would you go on to attempt to approximate it? How close must an approximation be before it is preferred to the very quantity it is attempting to approximate?

Simplicio: I hear your point.

Socrates: Is it not the case that decision analysts concern themselves with the value of perfect information? And do they not sometimes decide to exchange resources for additional information? It is unclear to me why someone would have perfect information, in the form of an exact value, and then choose to instead use imperfect information, in the form of an

approximation. Have you considered the ramifications of this loss of information?

Simplicio: It would not really matter too much if the two p-values are close, especially if they are both on the same side of alpha (0.05).

Socrates: If the t-test p-value is 0.03 and, for the same data, the exact p-value is 0.04, then there is no harm in using the t-test?

Simplicio: None that I can imagine.

Socrates: Would there be any harm in using the exact p-value in this case?

Simplicio: No, of course not!

Socrates: Hence, we have one analysis that is always right, and another that is right or wrong depending on the extent to which it agrees with the first one. Because it is often close, we use the approximate one, is that it?

Simplicio: At least when they are on the same side of alpha.

Socrates: And alpha is always 0.05?

Simplicio: Yes, this is an industry standard.

Socrates: My dear Simplicio, at my age I suffer many ailments, including arthritis. Now suppose that a new medication comes along that can offer relief for my symptoms. How certain would I need to be that this new treatment is effective before I decide to take it? Surely this question cannot be answered in a vacuum, but rather requires careful consideration of the frequency and severity of side effects, would you agree?

Simplicio: Most certainly.

Socrates: Is it conceivable that, after considering the side effect profile, I would come up with a personal alpha level of 0.035?

Simplicio: I cannot see why not.

Socrates: In such a case, I would take the medication if the primary efficacy p-value were 0.03, but not if it were 0.04. Use of the t-test could change what should be 0.04 to 0.03. In other words, I would be misled into taking a medication that, were I to know all the facts, I would not take. I would be denied the ability to render an informed decision.

Simplicio: I suppose so.

Socrates: Are you familiar with dense sets, Simplicio?

Simplicio: Are you calling me dense again Socrates?

Socrates: No Simplicio, dense sets are a formal construct in mathematics. For example, the rational numbers are a dense subset of the real numbers, because between any two real numbers, no matter how close together, one can find a rational number. Is it not also the case that the set of potential personal alpha levels is a dense subset of the set of potential p-values?

Simplicio: Yes, I suppose that it is.

Socrates: In that case, no matter how close the approximation is, somebody could have an alpha level that falls between the two p-values. In other words, the distortion in p-values created by the use of the approximation has consequences, not only abstractly, but also for real patients, the very patients who are relying on the researchers to provide unbiased information.

Simplicio: I never looked at it that way.

Socrates: Given the extent to which your research is funded by taxpayers, do you feel any obligation to deal with them honestly?

Simplicio: Yes, Socrates, thank you for bringing these issues to my attention. From now on I will use nothing but exact p-values.

EMERGING SCHOLARS
A Heteroscedastic, Rank-Based Approach for Analyzing
2 x 2 Independent Groups Designs

Laura Mills Robert A. Cribbie Wei-Ming Luh
York University National Cheng Kung
University

The ANOVA F is a widely used statistic in psychological research despite its shortcomings when the assumptions of normality and variance heterogeneity are violated. A Monte Carlo investigation compared Type I error and power rates of the ANOVA F , Alexander-Govern with trimmed means and Johnson transformation, Welch-James with trimmed means and Johnson Transformation, Welch with trimmed means, and Welch on ranked data using Johansen's interaction procedure. Results suggest that the ANOVA F is not appropriate when assumptions of normality and variance homogeneity are violated, and that the Welch/Johansen on ranks offers the best balance of empirical Type I error control and statistical power under these conditions.

Key words: Factorial ANOVA, Welch factorial test, non-normality, variance heterogeneity.

Introduction

The factorial independent groups design investigates the effects of two or more factors on an outcome variable and usually considers both the main and interactive effects. For example, Pegg et al. (2005) investigated therapeutic methods for military personnel who had experienced traumatic brain injury. The researchers were interested in how information offered (personal vs. general) and information preference (high vs. low preference for health care information) would influence therapeutic outcome. The design was a 2 x 2 independent groups factorial design and the results indicated that regardless of preference for information, information offered positively affected treatment

outcome. This type of design is common in psychological studies and the analysis of variance (ANOVA) F statistic is most often employed to analyze the results.

The ANOVA F test may not be appropriate when the data do not meet the validity assumptions that accompany the test (e.g., homogeneity of variance). These assumptions are discussed in most if not all texts but are largely ignored in applied research. This is especially problematic as previous studies have found that the assumptions of the ANOVA F are rarely met (e.g., Micceri, 1989; Wilcox, 1989). This article focuses on three objectives. The first is to discuss the assumptions associated with the ANOVA F statistic. The second is to examine recommended procedures for analysis of factorial designs when assumptions are violated. Finally, these previously recommended procedures will be compared to a new procedure to determine the method that provided the best balance between Type I error control and power. Ultimately, the goal is for applied researchers to regard alternatives to the ANOVA F test as necessary tools that need to be considered for implementation when assumptions are violated.

Laura Mills is a PhD candidate in the Department of Psychology at York University. Email: lmills@yorku.ca. Robert Cribbie is an Associate Professor in the Department of Psychology at York University. Email: cribbie@yorku.ca. Wei-Ming Luh is a Professor in the Institute of Education at the National Cheng Kung University.

Assumption Violation

The first assumption of the ANOVA F test is that the observations are independent of one another; this is ascertained during the design stage and established during sampling. The second assumption is that data from each population are normally distributed. When non-normality is a characteristic of the data in the cells, the deleterious effects on the ANOVA F can be quite serious. As a distribution becomes increasingly skewed, the mean of that distribution will be misrepresented because it will be pulled toward the tail and away from the middle of the data. Further, extreme scores in skewed distributions can elevate the variances of the distributions.

The third assumption of the ANOVA F is that the data are drawn from populations with equal variances. The standard error of the ANOVA F is based on a pooled variance term which weights the variances of the cells by their sample sizes. The cells with the larger sample sizes will contribute more information about variability to the computation of the standard error than the cells with the smaller sample sizes. For example, when sample sizes and variances are positively paired (larger sample sizes with larger variances and smaller sample sizes with smaller variances), empirical Type I error rates for the ANOVA F will be deflated and power will be compromised. When sample sizes and variances are negatively paired (larger sample sizes with smaller variances and smaller sample sizes with larger variances), empirical Type I error rates will be inflated.

Criteria for Robustness

The current study investigates how well different procedures perform, and thus a measure of how well warrants a brief discussion. The threshold for acceptable empirical Type I error rate adopted in the current study was $\pm .2 \alpha$, meaning a statistical procedure was considered robust if it maintained empirical Type I error rates between .04 and .06 when $\alpha = .05$. This was deemed a reasonable middle ground between Bradley's (1978) conservative ($\pm .1\alpha$) and liberal ($\pm .5\alpha$) criteria.

Robust Test Statistics

When assumptions are violated, the empirical Type I error rates of the ANOVA F vary in terms of robustness. The following summarizes conditions where the ANOVA F holds acceptable empirical Type I error rates and offers suggestions for alternatives when it does not. When data are normal in shape and have equal variance, the ANOVA F has accurate empirical Type I error rates and maximal power. In this situation, it merits the popularity it enjoys.

Non-normality

When distributions are non-normal but have equal variance, Hsuing and Olejnik (1996) found that the empirical Type I error rates for ANOVA F satisfied the threshold of $\pm .2 \alpha$. However, Wilcox (2003) argued that non-normality has deleterious effects on statistical power and that these effects are exacerbated by unequal sample size and heterogeneity (see, for example, Keselman, Wilcox, & Lix, 2003; Wilcox & Keselman, 2003). In these cases, the Welch on trimmed means (W_t) is recommended.

Variance Heterogeneity

The presence of unequal variances with normal distributions resulted in empirical Type I error rates for the ANOVA F that deviated considerably from the nominal level (Hsuing & Olejnik, 1996). Recommended alternatives for data that violate the assumption of variance homogeneity include the James, Welch, and Alexander-Govern (A-G) tests (Hsuing & Olejnik, 1996; Luh, 1999). Each of these procedures had acceptable empirical Type I error rates under heterogeneity.

Variance Heterogeneity and Non-normality

When non-normality was coupled with heterogeneous variances, the empirical Type I error rates for the ANOVA F become extremely unreliable (Hsuing & Olejnik, 1996). In this case, trimmed version of the James, Welch or A – G procedures have acceptable Type I error rates for several nonnormal distributions (Luh, 1999). Further, the use of a Johnson transformation improves the empirical Type I error rates of these procedures (Luh & Guo, 2001).

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

In general, the ANOVA F test is inappropriate when variance equality is compromised and especially so in combination with non-normality and unequal sample sizes. Researchers have the option of choosing from robust alternatives, but it remains unclear which choice is optimal. One method of simplifying the alternatives is to determine the procedures that maintain acceptable Type I error control, and then seek the procedure with the highest statistical power. Below is an overview of the reported power findings for procedures that maintained acceptable Type I error control.

Power Findings

When distributions were normal but variances heterogeneous, the James, Welch, and A-G tests reported by Luh (1999) all had similar power findings. When both normality and variance homogeneity were violated, trimmed versions of the the A-G, Welch and James tests had very similar power (Luh, 1999).

The primary goal of this study is to identify test statistics for 2 x 2 factorial designs that are best suited to psychological research whether the data meets the assumptions of the ANOVA F or it does not.

Test Statistics

Five procedures were evaluated and compared with the intention of determining one test that holds the most acceptable empirical Type I error rates combined with the highest power findings. The computational methods for each procedure are provided in Appendix A. 1) ANOVA F test. This test is included in this study as it is almost exclusively adopted by applied researchers, regardless of whether the assumptions of the test are violated; 2) Welch on trimmed means (Wilcox, 2003) using a Johnson transformation and Johansen interaction term (JW- J_t). The JW- J_t circumvents the problem of heterogeneous variances by unweighting the error term and the problem of non-normality by transforming and trimming the data. The Welch-James using trimmed means and Winsorized variances was found by Keselman, Kowalchuk, and Lix (1998) to be robust to heterogeneity and non-normality in non-orthogonal (unequal sample size) designs. Further, Luh & Guo (2001) recommended the use of this procedure

with a Johnson transformation. 3) Alexander-Govern with trimmed means and Johnson's transformation (JA- G_t). Luh & Guo (2001) found that the Alexander-Govern test with a combination of trimmed means and Johnson's transformation had acceptable empirical Type I error control under several conditions of non-normality and variance heterogeneity (Luh & Guo, 2001; Luh & Guo, 2004). 4) Welch on trimmed data (W_t). The Welch test on trimmed data is advantageous under heterogeneity of variance, as it unweights the pooled error term. In other words, the largest sample sizes no longer have the most influence on the pooled error term.

The final procedure investigated in this study is the Welch test, with the Johansen interaction procedure, on ranked data (W_r). Cribbie, Wilcox, Bewell & Keselman (2007) found that the Welch (1951) test on ranked data provided the best balance between Type I error control and power in one-way independent groups designs when both the assumptions of normality and variance homogeneity were violated. It is hypothesized in this study that the use of the W_r will also provide the best balance between Type I error control and power in 2 x 2 factorial independent groups designs. The use of a heteroscedastic test statistic in combination with ranked data is expected to simultaneously correct for violations of the assumptions of variance homogeneity and normality. Ranking data assigns the lowest score on the outcome variable a value of 1 and every other score a rank relative to that score, regardless of group membership. Thus, outlying data points become less distant and the problems associated with extreme data points are reduced. The W_r procedure is exactly as described for the Welch (see Appendix A), but because trimming and Winsorizing are unnecessary when using ranked data, the substitutions

$$d_{jk} = \frac{s^2}{n_{jk}}$$

for

$$d_{jk} = \frac{(n_{jk} - 1)s^2_{wjk}}{h_{jk}(h_{jk} - 1)}$$

and \bar{X}_{jk} for \bar{X}_{ijk} are made.

The Johansen test (see Appendix A) is used for evaluating the statistical significance of the interaction term.

Methodology

The current study aims to facilitate decision-making by applied researchers by discovering the one procedure which can offer the best balance of empirical Type I error control and power for 2 x 2 factorial designs. It is hypothesized that the W_r will be such a procedure, following the findings of Cribbie, et al. (2007) for one-way designs.

To test this hypothesis, a Monte Carlo study was conducted using 5000 simulations. R-project (Ihaka & Gentleman, 1996) and SAS/IML (SAS Institute Inc, 1989) software were used, with data generated using the rnorm and the RANNOR generators, respectively. The variables manipulated were: degree of sample size imbalance, variance inequality, pairings of unequal group sizes and variances (positive and negative), population distribution shape, and population means. The total sample size for the current study was set at 56 with specific individual cell sizes outlined below.

The procedures were tested with equal variances and with largest to smallest variance ratios of 4:1 and 8:1, respectively. This disparity was found by Keselman et al. (1998) to be common in psychological testing. The unequal variances were then reversed when sample sizes were unequal in order to test for both positive and negative pairings of unequal sample sizes and variances. The sample size and variance conditions investigated in this study are presented in Table 1.

Data were tested when population distribution shapes were normal and non-normal. The data were drawn from distributions defined by Hoaglin (1985) where both skewness (g) and kurtosis (h) can be manipulated to create varying levels of non-normality. In the current study, the distributions were set to normal ($g = 0, h = 0$), moderately skewed ($g = 0.5, h = 0$), and heavily skewed ($g = 1, h = 0$). Standard normal variates were generated with SAS RANNOR (SAS Institute, 1989) and R-project RNORM (Ihaka & Gentleman, 1996) and to

obtain data from a skewed g - and h - distribution, these variables were converted to:

$$\varepsilon = g^{-1} [\exp(gZ) - 1] \exp(hZ^2 / 2),$$

when $g = 0, \varepsilon = Z \exp(hZ^2 / 2)$. For $g > 0$, the mean of the g - and h - distribution

$$\mu_{gh} = \frac{\left(\exp \left\{ \frac{g^2}{[2(1-h)]} \right\} - 1 \right)}{\left[g(1-h)^{\frac{1}{2}} \right]}$$

was subtracted from each observation, and for trimmed data the population trimmed mean (μ_{gh}) was subtracted from each observation. In order to create cells with mean μ_{jk} and standard deviation σ_{jk} , the resulting ε_{ijk} were converted to $Y_{ijk} = \mu_{jk} + (\varepsilon_{ijk} \sigma_{jk})$. For the W_r , the population mean rank is not equal across cells when the distribution shapes are skewed and the variances are unequal. Therefore, for each condition of skewness and variance heterogeneity, we adjusted the distribution of the cells so that the population mean ranks were equal. Specifically, the empirically derived population mean rank for each cell was subtracted from Y_{ijk} .

Null Hypotheses

Given a 2 x 2 independent groups factorial design, the null hypotheses for the row and column main effects are: $H_0: \mu_1 = \mu_2$ where

$$\mu_j = \frac{\sum_k \mu_{jk}}{2} \text{ and } \mu_k = \frac{\sum_j \mu_{jk}}{2}.$$

When trimmed means are applied, as is the case for the W_t , the null hypotheses becomes $H_0: \mu_{t1} = \mu_{t2}$

$$\text{where } \mu_{tj} = \frac{\sum_k \mu_{tjk}}{2} \text{ and } \mu_{tk} = \frac{\sum_j \mu_{tjk}}{2}.$$

The null hypotheses for the interaction term can be expressed as $H_0: \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$ for the usual means and for trimmed means $H_0: \mu_{t11} - \mu_{t12} = \mu_{t21} - \mu_{t22}$. For ranked data, the null hypotheses for the main effects and interactions (without a heteroscedastic test statistic) relate to the population mean ranks (i.e., μ_{rjk}) only when the distributions are the same shape and variances are equal. Hence, an important part of this study

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

is to evaluate how the Welch on ranks performs when variances are unequal.

Results

Normal Distributions and Equal Variances

When distributions were normal and variances were equal, all tests produced acceptable empirical Type I error rates. The ANOVA F and the W_r held the highest power under these conditions, although differences among procedures were minimal. (Empirical Type I error and power rates are presented in Tables 2 - 6.)

Skewed Distributions and Equal Variances

When distributions were moderately skewed and variances were equal, empirical Type I error rates were within the acceptable range for all procedures. The power of the procedures was very similar in terms of main effects, but when interaction is present, the W_r is the most powerful.

When distributions were heavily skewed and had equal variances, the ANOVA F and the W_r maintained Type I error rates within the acceptable range while the other procedures were deflated relative to α . The W_r was more powerful than the ANOVA F (and all other procedures).

Heterogeneity and Normal Distributions

When unequal variances were combined with normal distributions, the ANOVA F had Type I error control that was deflated relative to α when the pairing of the unequal variances and sample sizes was positive and inflated relative to α when the pairing was negative. Type I error rates for the W_t slightly exceeded the robustness criteria when testing interactions with negatively paired sample sizes and variances, but all other procedures had Type I error rates within the acceptable range. Power findings were similar across all procedures, with the W_r slightly higher, particularly for interactions when there was a negative pairing of unequal sample sizes and variances.

Heterogeneity and Skewed Distributions

When distributions were moderately skewed and variances were unequal, the

ANOVA F and the W_t had unacceptable Type I error control. Specifically, the ANOVA F had inflated Type I error rates when sample sizes and variances were negatively skewed and deflated Type I error rates when sample sizes and variances were positively skewed, for both main effects and interactions. The W_t procedure had inflated Type I error rates when testing interactions with negatively paired sample sizes and variances. The W_r maintained much higher power than all other procedures, again particularly in the case of negative pairings. Finally, when distributions were heavily skewed with unequal variances, the W_r was the only procedure that maintained empirical Type I error rates within the acceptable range, and even when other procedures had acceptable Type I error rates the power of the W_r was generally superior.

Conclusion

Factorial designs are extremely common in psychological research. The method most commonly used for analyzing factorial designs, the ANOVA F statistic, is clearly a poor choice when the assumptions of homogeneity and normality are violated. The F test simply falls short of the expectations that researchers assign it. The goal of the current paper was to elucidate the problems with the popular ANOVA F test while at the same time offering a comparison of alternative procedures across numerous conditions of normality/non-normality and variance homogeneity/heterogeneity with respect to the balance between empirical Type I error control and statistical power.

It is strikingly clear that the most popular procedure, the ANOVA F , is also the most inappropriate test for factorial research unless data conform to the assumptions of normality and variance homogeneity. Empirical Type I error rates stray considerably from the nominal α , especially when variances are unequal or unequal variances are combined with non-normal distributions. When α is set at .05, the empirical Type I error rates for the ANOVA F can be as low as 1.8% or as high as 14% under the conditions used in the current study. Further, if the ratio of the largest to smallest variances exceeds 8:1 or more extreme sample size imbalance is present (both realities in real-world

data), the rates of Type I error become even more alarming (see Hsuing & Olejnik, 1996).

These results are troubling given that the assumptions of the ANOVA F are routinely violated. Micceri (1989) investigated the distribution shapes of over 400 sets of data from empirical studies and found that in psychometric and ability type scores about 70% were asymmetric and/or had heavy tails. In other words, most of the studies had distributions that could be considered non-normal. Further, Keselman, Kowalchuk, and Lix (1998) discuss the regular occurrence of unequal variances in psychology, and unbalanced cell sizes are the norm in psychological research.

The closer the data come to meeting assumptions, the more choices there are for researchers in terms of accuracy and power. As the data move farther from normality and variance homogeneity, the decision is made easier by elimination. The procedure that holds empirical Type I error rates closest to α and has the highest power is the Welch on ranked data using the Johansen procedure for interactions. Under all conditions, the procedure performed well in terms of Type I error control and power. The most exciting aspect of the findings in this project is that the Welch on ranked data worked well under the majority of conditions that were investigated for a 2 x 2 design, including equal variances and normal distributions. In other words, researchers don't need to sort through a confusing decision-making process. This procedure can easily fill the role that the ANOVA F now occupies by offering more accuracy and power when assumptions are violated while only losing a trivial amount of power when assumptions are met. Therefore, it is highly recommended that researchers routinely adopt the Welch procedure with ranked data when analyzing factorial designs.

With regard to limitations of the current study, Micceri (1989) notes that Monte-Carlo investigations don't necessarily replicate real-world data. With real-world data, researchers might experience different kinds of non-normality than the distribution shapes that were investigated in this study. Likewise, the degree of variance heterogeneity has innumerable possibilities while only five conditions were investigated in the current project. However, the

conditions investigated in the current project covered many of the most extreme assumption violations that researchers will encounter and thus if the procedure is robust under these conditions, it will likely be robust under most conditions encountered in applied research.

An obvious future direction for this procedure is to investigate the performance of the Welch on ranks in higher order factorial designs. Although it is expected that the results of this study will replicate in larger factorial designs, this hypothesis still needs to be evaluated, especially in light of the fact that Seaman, Walls, Wise, and Jaeger (1994) report that in designs larger than a 2 x 2 factorial that because rank transformations are nonlinear, the expected rank of an observation in one cell will depend nonlinearly on the original population means of the other cells.

It is expected that the complications that arise when utilizing ranks with traditional test statistics [e.g., the rank transform procedure suggested by Conover and Iman (1981)] will not have a significant effect on the Welch on ranks procedure because it utilizes heteroscedastic test statistics; however this is still to be demonstrated. Another important consideration in future research is the effect of between-cell distribution shape heterogeneity. In other words, the degree of skew might differ from group to group and exacerbate the effects of skewness beyond what was reported in this paper. In fact, Wilcox (2005) notes that skewness per se is not necessarily the problem, but the degree to which skewness varies from group to group raises cause for alarm.

As a result of the findings of the current study, it is strongly recommended that researchers discontinue the use of the ANOVA F procedure. Instead, it is suggested that researchers utilize the Welch on ranked data (with Johansen procedure for interactions) regularly for analyzing independent groups factorial designs.

References

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

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

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

Cribbie, R. A., Wilcox, P. R., Bewell, C. & Keselman, H. J. (2007). Tests for treatment group equality when data are non-normal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, 6, 117-132.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67, 85-93.

Hoaglin, D. C. (1985). Summarizing shape numerically: the *g*- and *h*-distributions. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes*. NY: Wiley.

Hsuang, T-H., & Olejnik, S. (1996). Type I error rates and statistical power for the James Second-order test and the univariate *F* test in two-way fixed-effects ANOVA models under heteroscedasticity and/or non-normality. *The Journal of Experimental Education*, 65, 57-71.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust non-orthogonal analysis revisited: An update based on trimmed means. *Psychometrika*, 63, 145-163.

Ihaka, R. & Gentleman, R. (1996). "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299-314.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350-386.

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.

Luh, W. M. (1999). Developing trimmed mean test statistics for two-way fixed-effects ANOVA models under variance heterogeneity and non-normality. *The Journal of Experimental Education*, 67, 243-264.

Luh, W. M., & Guo, J. H. (2001). Using Johnson's transformation and robust estimators with heteroscedastic test statistics: An examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design. *British Journal of Mathematical and Statistical Psychology*, 54, 79-94.

Luh, W. M., & Guo, J. H. (2004). Improved robust test statistic based on trimmed means and Hall's transformation for two-way ANOVA models under non-normality. *Journal of Applied Statistics*, 31, 623-643.

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

Pegg, P. O., Auerbach, S. M., Seel, R. T., Buenaver, L. F., Kiesler D. J., & Plybon, L. E. (2005). The impact of patient-centered information on patients' treatment satisfaction and outcomes in traumatic brain injury rehabilitation. *Rehabilitation Psychology*, 50, 366-374.

SAS Institute, Inc. (1989). *SAS/IML software: Usage and reference, Version 6 (1st Ed.)*. Cary, NC: Author.

SAS Institute, Inc. (1996). *SAS Basic software, Version 6 (12th Ed.)*. Cary, NC: Author.

Seaman, J. W., Jr, Walls, S. C., Wise, S. E., & Jaeger, R. G. (1994). Caveat emptor: rank transform methods and interaction. *Trends in Ecology and Evolution*, 9, 261-263.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336.

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics*, 14, 269-278.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. California: Academic Press.

Wilcox, R. R. (2005). New methods for comparing groups. *Current Directions in Psychological Science*, 14, 272-275

Wilcox, R. R. & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8, 254-274.

Table 1: Means, Sample Sizes and Variances Utilized in the Monte Carlo Study

Condition	Relevant Statistic			
	Means			
	μ_{11}	μ_{12}	μ_{21}	μ_{22}
No Main Effect or Interaction	0	0	0	0
Main Effect + Interaction	0	0	0	2
No Main Effect, Interaction	0	1	1	0
Sample Sizes				
	n_{11}	n_{12}	n_{21}	n_{22}
Equal Sample Sizes	14	14	14	14
Moderately Unequal Sample Sizes	11	14	14	17
Extremely Unequal Sample Sizes	7	10	18	21
Variances				
	σ_{11}	σ_{12}	σ_{21}	σ_{22}
Equal Variances	1	1	1	1
Moderately Unequal Variances (Positively Paired with Unequal Sample Sizes)	1	2	3	4
Moderately Unequal Variances (Negatively Paired with Unequal Sample Sizes)	4	3	2	1
Extremely Unequal Variances (Positively Paired with Unequal Sample Sizes)	1	3	5	8
Extremely Unequal Variances (Negatively Paired with Unequal Sample Sizes)	8	5	3	1

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

Table 2: Type I Error Rates for Main Effects with Normal and Skewed Distribution for $N = 56$

Distribution Shape	Variances	F	JA-G _t	JW-J _t	W _t	W _r
Normal	Equal	.050	.047	.048	.048	.050
	Positive Pair	.034	.046	.048	.048	.055
	Negative Pair	<u>.094</u>	.050	.048	.050	.054
Moderate Skew	Equal	.048	.043	.041	.046	.050
	Positive Pair	.037	.045	.045	.046	.053
	Negative Pair	<u>.095</u>	.049	.047	.049	.055
Heavy Skew	Equal	.042	.040	.037	.038	.052
	Positive Pair	.041	.041	.035	.037	.053
	Negative Pair	<u>.093</u>	.039	.035	.046	.054

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).

Bolded entries indicate conservative empirical Type I error rate. Bolded and underlined entries represent liberal Type I error rates

Table 3: Type I Error Rates for Interactions with Normal and Skewed Distribution for $N = 56$

Distribution Shape	Variances	F	JA-G _t	JW-J _t	W _t	W _r
Normal	Equal	.054	.048	.050	.056	.051
	Positive Pair	.035	.047	.049	.055	.054
	Negative Pair	<u>.095</u>	.051	.049	<u>.062</u>	.054
Moderate Skew	Equal	.050	.045	.042	.054	.052
	Positive Pair	.033	.045	.046	.053	.051
	Negative Pair	<u>.095</u>	.050	.045	<u>.064</u>	.054
Heavy Skew	Equal	.044	.039	.035	.052	.052
	Positive Pair	.029	.040	.035	.050	.051
	Negative Pair	<u>.079</u>	.042	.034	<u>.064</u>	.057

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).

Bolded entries indicate conservative empirical Type I error rate. Bolded and underlined entries represent liberal Type I error rates

Table 4: Power Findings for Main Effects with Normal and Skewed Distribution when both Main Effects and Interaction were Present for N = 56

Distribution Shape	Variances	<i>F</i>	JA-G _t	JW-J _t	W _t	W _r
Normal	Equal	.940	.893	.886	.897	.911
	Positive Pair	.468	.484	.490	.492	.467
	Negative Pair	.556	.367	.363	.377	.493
Moderate Skew	Equal	.838	.816	.859	.841	.847
	Positive Pair	.344	.487	.494	.459	.518
	Negative Pair	.480	.321	.345	.361	.402
Heavy Skew	Equal	.516	.680	.750	.708	.733
	Positive Pair	.134	.427	.422	.362	.556
	Negative Pair	.330	.266	.290	.310	.334

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).

Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2 α criteria.

Table 5: Power Findings for Interactions with Normal and Skewed Distribution when Both Main Effects and Interactions were Present for N = 56

Distribution Shape	Variances	<i>F</i>	JA-G _t	JW-J _t	W _t	W _r
Normal	Equal	.941	.898	.888	.946	.954
	Positive Pair	.470	.483	.491	.548	.468
	Negative Pair	.558	.365	.363	.504	.651
Moderate Skew	Equal	.832	.854	.874	.941	.962
	Positive Pair	.356	.432	.456	.522	.584
	Negative Pair	.449	.326	.328	.505	.577
Heavy Skew	Equal	.508	.712	.764	.886	.921
	Positive Pair	.188	.343	.382	.448	.726
	Negative Pair	.278	.240	.254	.472	.502

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), W_t (Welch on trimmed data), W_r (Welch on ranked data).

Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2 α criteria.

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

Table 6: Power Findings for Interactions with Normal and Skewed Distribution When Interaction Was Only Present for $N = 56$

Distribution Shape	Variances	F	JA-G _t	JW-J _t	W _t	W _r
Normal	Equal	.943	.895	.894	.910	.925
	Positive Pair	.469	.489	.494	.509	.560
	Negative Pair	.557	.370	.362	.409	.465
Moderate Skew	Equal	.832	.842	.843	.867	.925
	Positive Pair	.366	.466	.466	.477	.596
	Negative Pair	.473	.382	.371	.406	.482
Heavy Skew	Equal	.511	.724	.724	.752	.910
	Positive Pair	.204	.395	.383	.403	.668
	Negative Pair	.288	.357	.333	.365	.613

F (ANOVA F), JA-G_t (Alexander-Govern with trimmed means and Johnson transformation), JW-J_t (Welch-James with trimmed means and Johnson transformation), Welch (Welch on trimmed data), W_r (Welch on ranked data).

Greyed power findings indicate cases where empirical Type I error rate does not fall within +/- .2 α criteria.

Appendix A:
ANOVA F Procedure

The main effect of one factor (*A*) is a measure of the ratio of mean squared group variation to mean squared error and is defined as:

$$F_A = \frac{\frac{nK \sum (\bar{X}_{.k} - \bar{X}_{..})^2}{J-1}}{\frac{\sum (X - \bar{X}_{..})^2 - n \sum (\bar{X}_{jk} - \bar{X}_{..})^2}{JK(n-1)}}$$

where *n* = cell group size, *N* = total sample size, *j* = 1 ... *J* (number of levels for factor *A*), *k* = 1 ... *K* (number of levels for factor *B*), *X* is an observation, \bar{X}_{jk} is the mean of the cell at the *i*th row and the *j*th column, $\bar{X}_{..}$ is the grand mean, $\bar{X}_{.j}$ is the mean of the *j*th level of factor *A*, and $\bar{X}_{.k}$ is the mean of the *k*th level of factor *B*. The degrees of freedom for factor *A* are *J* - 1 and *JK*(*n* - 1).

The main effect for factor *B* is likewise defined, with the means of each level obtained across (and disregarding) all levels of Factor *A*. The equation is:

$$F_B = \frac{\frac{nJ \sum (\bar{X}_{j.} - \bar{X}_{..})^2}{K-1}}{\frac{\sum (X - \bar{X}_{..})^2 - n \sum (\bar{X}_{jk} - \bar{X}_{..})^2}{JK(n-1)}}$$

The degrees of freedom for the main effect of *B* are *K*-1 and *JK*(*n*-1). The interaction term for the ANOVA *F* test is a ratio of mean squared cell variation (less mean squared variance of both factors) to mean squared error and is defined as:

$$F_{AB} = \frac{\frac{n \sum (\bar{X}_{jk} - \bar{X}_{..})^2 - nK \sum (\bar{X}_{.j} - \bar{X}_{..})^2 - nJ \sum (\bar{X}_{.k} - \bar{X}_{..})^2}{(J-1)(K-1)}}{\frac{\sum (X - \bar{X}_{..})^2 - n \sum (\bar{X}_{jk} - \bar{X}_{..})^2}{JK(n-1)}}$$

The main effect for factor *B* is likewise defined, with the means of each level obtained across (and disregarding) all levels of Factor *A*. The equation is:

$$F_B = \frac{\frac{nJ \sum (\bar{X}_{j.} - \bar{X}_{..})^2}{K-1}}{\frac{\sum (X - \bar{X}_{..})^2 - n \sum (\bar{X}_{jk} - \bar{X}_{..})^2}{JK(n-1)}}$$

The degrees of freedom for the main effect of *B* are *K* - 1 and *JK*(*n* - 1). The interaction term for the ANOVA *F* test is a ratio of mean squared cell variation (less mean squared variance of both factors) to mean squared error. It is defined as:

$$F_{AB} = \frac{\frac{n \sum (\bar{X}_{jk} - \bar{X}_{..})^2 - nK \sum (\bar{X}_{.j} - \bar{X}_{..})^2 - nJ \sum (\bar{X}_{.k} - \bar{X}_{..})^2}{(J-1)(K-1)}}{\frac{\sum (X - \bar{X}_{..})^2 - n \sum (\bar{X}_{jk} - \bar{X}_{..})^2}{JK(n-1)}}$$

The degrees of freedom for the interaction term are (*J*-1)(*K*-1) and *JK*(*n* - 1).

Appendix B:

The Welch Procedure using Johansen Interaction Term

Wilcox (2003, p. 345) defines the Welch procedure using trimmed means and Winsorized variances. Winsorizing is a method by which trimmed scores are replaced with the remaining highest and lowest score in the data. This generates an appropriate estimate of variance when using a trimmed mean as opposed to estimating variance using only the scores left after trimming by accounting for the original sample size. The current study adopts these procedures for the Welch. Consider *X*₁, ..., *X*_{*n*}, a random sample from a single group, ordered from smallest to largest. Let *e* = [γ*n*], where γ is the proportion of symmetric trimming, set at .20 in this study, and [X] is the greatest integer less than or equal to *X*, and let *h*_{*jk*} = *n* - 2*e* be the effective sample size (i.e., sample size after trimming).

RANK-BASED APPROACH FOR 2 x 2 INDEPENDENT GROUPS DESIGNS

A trimmed mean can be expressed as $\bar{X}_t = \sum_{i=e+1}^{n-e} \frac{X_i}{h_{jk}}$. The main effect of Factor A first takes a measure of typical deviation for each cell:

$$d_{jk} = \frac{(n_{jk} - 1)s_{wj}^2}{h_{jk}(h_{jk} - 1)},$$

where $s_w^2 = \frac{\sum (Y_i - \bar{X}_w)}{n-1}$, and $\bar{X}_w = \frac{\sum Y_i}{n}$,

where $Y_i = X_{e+1}$ if $X_i \leq X_{e+1}$, X_i if $X_{e+1} < X_i < X_{n-e}$ and X_{n-e} if $X_i \geq X_{n-e}$.

A measure of row means is indicated by $R_j = \sum_{k=1}^K \bar{X}_{tjk} / k$, where t indicates trimmed cell means. Next, the inverse of the sum of the row deviations is $r_j = \frac{1}{\sum_k d_{jk}}$, and these two terms contribute to a measure of predicted variance for Factor A , defined by $\hat{R} = \frac{\sum r_j R_j}{\sum r_j}$.

Two final terms contribute to the Welch statistic:

$$\hat{v}_j = \frac{(\sum_k d_{jk})^2}{\sum_k d_{jk}^2 / (h_{jk} - 1)}$$

and

$$B_a = \sum_{j=1}^J \frac{1}{\hat{v}_j} \left(1 - \frac{r_j}{\sum r_j}\right)^2.$$

The main effect is defined as:

$$V_a = \frac{1}{(J-1) \left(1 + \frac{2(J-2)B_a}{(J^2-1)}\right)} \sum_{j=1}^J r_j (R_j - \hat{R})^2$$

The numerator degrees of freedom are $v_1 = J - 1$ and for the denominator, $v_2 = \frac{J^2 - 1}{3B_a}$. The

main effect of Factor B is similarly obtained, using $W_k = \sum_{j=1}^J \bar{X}_{tjk}$, $w_k = \frac{1}{\sum_j d_{jk}}$,

$$\hat{\omega}_k = \frac{(\sum_j d_{jk})^2}{\sum_j d_{jk}^2 / (h_{jk} - 1)}, \quad \hat{W} = \frac{\sum w_k W_k}{\sum w_k},$$

$$B_b = \sum_{k=1}^K \frac{1}{\hat{\omega}_k} \left(1 - \frac{w_k}{\sum w_k}\right)^2, \quad \text{and}$$

$$V_b = \frac{1}{(K-1) \left(1 + \frac{2(K-2)B_b}{K^2-1}\right)} \sum_{k=1}^K w_k (W_k - \hat{W})^2.$$

Degrees of freedom for Factor B are $v_1 = K - 1$

and $v_2 = \frac{K^2 - 1}{3B_b}$. To test for interactions,

Wilcox recommends the Johansen (1980) method. The inverse of the mean cell deviation is $D_{jk} = \frac{1}{d_{jk}}$ which are summed across each

factor and in total to determine (respectively)

$$D_{j.} = \sum_{k=1}^K D_{jk}, \quad D_{.k} = \sum_{j=1}^J D_{jk}, \quad \text{and}$$

$D_{..} = \sum D_{jk}$. The predicted values of the cell means are determined using:

$$\tilde{X}_{tjk} = \sum_{l=1}^J \frac{D_{lk} \bar{X}_{tlk}}{D_{.k}} + \sum_{m=1}^K \frac{D_{jm} \bar{X}_{tjm}}{D_{j.}} - \sum_{l=1}^J \sum_{m=1}^K \frac{D_{lm} \bar{X}_{tlm}}{D_{..}}.$$

The interaction is determined with a ratio of the cell mean residuals to cell mean deviation using

$$V_{ab} = \sum_{j=1}^J \sum_{k=1}^K D_{jk} \left(\bar{X}_{tjk} - \tilde{X}_{tjk} \right)^2 .$$

Using the example, the interaction is calculated as follows:

The critical value for the Johansen method is found by computing

$$A = \sum_j \sum_k \frac{1}{f_{jk}} \left\{ 1 - D_{jk} \left(\frac{1}{D_{.j}} + \frac{1}{D_{.k}} - \frac{1}{D_{..}} \right) \right\}^2 ,$$

where $f_{jk} = h_{jk} - 1$ and c is the cutoff value in the $1 - \alpha$ chi-square distribution, with

$$h(c) = \frac{c}{2(J-1)(K-1)} \left\{ 1 + \frac{3c}{(J-1)(K-1)+2} \right\} A.$$

Appendix C:

Alexander-Govern Procedure with Trimmed Means and Johnson Transformation

This procedure involves terms identical to those used for the Welch statistic: r_j , R_j , \hat{v}_j , & \hat{R} for the row effect and w_k , W_k , \hat{w}_k , & \hat{W} for the column effect, with $d_{jk} = \frac{s_{jk}^2}{n_{jk}}$ for both row and

column effects. The A-G then computes the row Z statistic using $T_j = \sqrt{r_j} (R_j - \hat{R})$, $A_j = \hat{v}_j - 0.5$,

$$a_j = 48A_j^2, \quad C_j = \left[A_j \ln \left(1 + \frac{T_j^2}{\hat{v}_j} \right) \right]^{\frac{1}{2}},$$

$$D_j = 4C_j^7 + 33C_j^5 + 240C_j^3 + 855C_j,$$

$$E_j = 10a_j^2 + 8a_j C_j^4 + 1000a_j,$$

$$Z_j = C_j + \frac{C_j^3 + 3C_j}{a_j} - \frac{D_j}{E_j}, \text{ and } AG = \sum Z_j^2.$$

This test statistic is compared to a χ^2 critical value at $1 - \alpha$ with $J - 1$ degrees of freedom. For the example, the critical value is 3.84 when $\alpha = .05$. To test for interactions, the Johansen method is recommended by Luh (1999), which

is the same as used by the Welch and so its definition will suffice.

For use in the transformation, the third central Winsorized moment is defined using

$$\hat{\mu}_3 = \frac{\sum (Y_i - \bar{X}_w)^3}{n}, \text{ where } Y_i \text{ are the observations in the cell of interest and } \bar{X}_w = \frac{\sum Y_i}{n} \text{ is the Winsorized mean,}$$

$$\hat{\sigma}_w^2 = \frac{(n-1)s_w^2}{(h_{jk} - 1)} \text{ is the squared standard error of}$$

the trimmed mean and $\hat{\mu}_w = \frac{n\hat{\mu}_3}{h_{jk}}$ is the third

central sample Winsorized moment. The transformation is executed in the residual computations for the T_t terms. These residuals are defined as

$$R_{tj} - \hat{R}_t = \sum_{k=1}^K \left\{ \left(\bar{X}_{tjk} - \hat{X}_{t.k} \right) + \frac{\hat{\mu}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{\mu}_{wjk} \left(\bar{X}_{tjk} - \hat{X}_{t.k} \right)^2}{3\hat{\sigma}_{wjk}^4} \right\}$$

for the row effect and

$$W_{tj} - \hat{W}_t = \sum_{k=1}^K \left\{ \left(\bar{X}_{tjk} - \hat{X}_{t.j} \right) + \frac{\hat{\mu}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{\mu}_{wjk} \left(\bar{X}_{tjk} - \hat{X}_{t.j} \right)^2}{3\hat{\sigma}_{wjk}^4} \right\}$$

for the column effect, where

$$\hat{X}_{t.j} = \frac{\sum_l w_{tl} \bar{X}_{tjl}}{\sum_l w_{tl}}$$

and

$$\hat{X}_{t.k} = \frac{\sum_i r_{ti} \bar{X}_{tik}}{\sum_i r_{ti}}.$$

Appendix D:

Welch-James with Trimmed Means and Johnson Transformation

C_1 are contrast matrices associated with either the main effect of factor A or B or AB . The cell means are: $\bar{Y}_{jk} = \sum_i Y_{ijk} / n_{jk}$. The matrix of cell means is: $\bar{Y}_j = (\bar{Y}_{j1}, \dots, \bar{Y}_{jK})$ and the $1 \times J$ matrix of cell means is thus, $\bar{Y} = (\bar{Y}_1', \dots, \bar{Y}_j')$. The sample variance matrix of Y is:

$$S = \text{diag} \left(\frac{s_{11}^2}{n_{11}}, \dots, \frac{s_{JK}^2}{n_{JK}} \right).$$

The test statistic is:

$$T_{WJ} = \frac{(C_1 \bar{Y})'(C_1 S C_1')^{-1} (C_1 \bar{Y})}{r + 2A - \frac{6A}{(r+2)}}$$

where

$$A = \sum_{jk} \frac{(1 - P_{jk,jk})^2}{n_{jk} - 1}$$

and $P_{jk,jk}$ = the jk, jk^{th} element of the matrix $I - S C_1' (C_1 S C_1')^{-1} C_1$. T_{WJ} has an approximate F distribution with degrees of freedom $f_1 = r$ and $f_2 = r(r+2)/(3/4)$.

The Johnson transformation applied to the $W-J_t$ is defined by Luh & Guo (2001) as follows \bar{X}_{tjk} is replaced by

$$\left(\bar{X}_{tjk} - \hat{X}_{t..} \right) + \frac{\hat{u}_{wjk}}{6\hat{\sigma}_{wjk}^2 f_{jk}} + \frac{\hat{u}_{wjk} (\bar{X}_{tjk} - \hat{X}_{t..})^2}{3\sigma_{wjk}^4}$$

where

$$\hat{X}_{t..} = \frac{\sum_{jk} f_{jk} \bar{X}_{tjk}}{\sum_{jk} f_{jk}},$$

$$\hat{\mu}_3 = \frac{\sum (Y_i - \bar{X}_w)^3}{n},$$

$$\hat{\sigma}_w^2 = \frac{(n-1)s_w^2}{(h_{jk} - 1)},$$

and

$$\hat{\mu}_w = \frac{n\hat{\mu}_3}{h_{jk}}.$$

A Comparison of Maximum Likelihood and Expected A Posteriori Estimation for Polychoric Correlation Using Monte Carlo Simulation

Jinsong Chen Jaehwa Choi
The George Washington University

This study aims to compare the maximum likelihood (ML) and expected a posteriori (EAP) estimation for polychoric correlation (PCC) under diverse conditions, especially when considering a sample size. As the ML is the classical solution to estimate PCC, the EAP is a new method based on Bayes' theorem. Different types of prior distributions are also adapted to investigate the sensitivity of prior distribution onto the PCC estimate for the EAP case. The Monte Carlo simulation is used for this comparison by a specialized program code in MATLAB.

Key words: Polychoric correlation, maximum likelihood, expected a posteriori.

Introduction

It is fairly common that observed variables are measured using ordinal scales, which represent categorizations of underlying constructs that are continuous. This scenario is especially relevant in psychological and educational measurement. As an estimate of the relation between the two continuous constructs underlying two such ordinal variables, the polychoric correlation (PCC) has been widely employed. For instance, PCC has been used in many confirmatory factor analysis (CFA) or structural equation model (SEM) scenarios recently (e.g., Flora, 2002; Flora & Curran, 2004; Rigdon & Ferguson, 1991). The estimation of PCC has been conducted using maximum likelihood (ML) methods (e.g., Olsson, 1979), which can be accomplished using several popular statistical applications such as PRELIS (Jöreskog, 2002-2005) or SAS PROC FREQ (SAS Institute Inc., 2004).

Regarding the ML estimation of PCC, research showed that it: 1) produces an unbiased estimate of the correlation between the original bivariate normal variables (Babakus & Ferguson, 1988; Olsson, 1979); 2) outperforms Pearson's product-moment correlation (PPMC), Spearman's rho, and Kendall's tau-b for ordinal data (Babakus & Ferguson, 1988); and 3) is rather robust to modest violation of the underlying normality assumptions (Quiroga, 1992).

Even though estimating PCC with the ML method has been quite satisfactory as stated above, empirical or simulation results from previous research are based on relatively large sample sizes. For instance, the sample size was 500 for Olsson (1979), and 200 or above for Quiroga (1992). In many situations however, the sample size could be much smaller (e.g., less than 100), and the performance of the ML estimator has not been studied yet in the case of smaller sample sizes. Furthermore, due to the properties of numerical procedure of ML (i.e., iterative hill-climbing method using gradients of the target function), the ML estimation method for PCC also has several disadvantages such as, local maxima and non-converged solution.

Recently, expected a posteriori (EAP) estimation for PCC was introduced (Choi, Chen, & Kim, in press). As the EAP method is based on Bayes' theorem (Bock & Aitken, 1981), the

Jinsong Chen is a doctoral candidate in the Graduate School of Education and Human Development. Email: cjs@gwmail.gwu.edu. Jaehwa Choi is an Assistant Professor of Educational Research in the Graduate School of Education and Human Development. Email: jaehwa.choi@yahoo.com.

estimation of PCC can incorporate prior information regarding the correlation. The EAP method has been spotlighted in social and behavioral methodologies, such as Item Response Theory (IRT) models (e.g., Mislevy & Stocking, 1989). Also, this estimator has been compared with the ML method in IRT models (see Chen, Hou, & Dodd, 1998 for a summary). Because both PCC and EAP are becoming increasingly popular in social science research, understanding the behavior of the newly developed EAP estimator for PCC from a systematic comparison with the ML estimator would be beneficial. In this article, a methodological framework of both ML and EAP estimators will be introduced and the performance of the two estimators under various conditions will be compared, especially in the case of small sample size, using a Monte Carlo simulation study.

Polychoric Correlation and the Maximum Likelihood (ML) Estimation

Traditionally, ML has been the only estimator for estimating PCC, and the procedures are summarized here (see Olsson, 1979 for details). Two ordinal variables are observed with r and s possible categories. Given that the two corresponding continuous latent constructs follow the crucial assumption of bivariate normal distribution, the log-likelihood function of any sample is:

$$\ln L = \ln C + \sum_{i=1}^s \sum_{j=1}^r n_{ij} \ln \pi_{ij} \quad (1.1)$$

where C is a constant and π_{ij} is the probability that a given observation falls into the contingency table cell (i, j) between two ordinal variables,

$$\pi_{ij} = \Phi_2(a_i, b_j) - \Phi_2(a_{i-1}, b_j) - \Phi_2(a_i, b_{j-1}) + \Phi_2(a_{i-1}, b_{j-1}) \quad (1.2)$$

where a and b are the threshold parameters for the categories $i = 1, 2, \dots, s$ and $j = 1, 2, \dots, r$, with $a_0 = b_0 = -\infty$ and $a_s = b_r = +\infty$, and Φ_2 is

the bivariate standard normal cumulative density function (CDF) with correlation ρ .

The threshold and correlation parameters can be estimated by: 1) taking partial derivatives of the log-likelihood function with respect to the parameters (thresholds and correlation), 2) setting these equations equal to zero, and 3) solving these equations for the parameters of interest using the numerical iterative procedure such as the Newton-Raphson method (Olsson, 1979). This method attempts to estimate all parameters of interest simultaneously, and was referred to as the “full ML method” or the “one-step ML method” by Olsson (1979).

Olsson also presented the two-step ML method for PCC, which estimates threshold values first:

$$a_i = \Phi_1^{-1}(P_{i.}) \text{ and } b_j = \Phi_1^{-1}(P_{.j}) \quad (1)$$

where P_{ij} is the observed proportion in cell (i, j) , $P_{i.}$ and $P_{.j}$ are observed cumulative marginal proportions of the contingency table, and Φ_1 is the univariate normal CDF.

These threshold values are then substituted into the log-likelihood function, Equation (1), and the correlation parameter is estimated similar to the one-step ML method illustrated above. Olsson (1979) further showed that the difference of estimation between the one-step and two-step ML methods is negligible. Therefore, the two-step method is used in this study for the purpose of computational convenience.

Several issues of the ML methods are worthwhile to be noted here. As mentioned earlier, ML methods are iteratively searching the maximum of the log-likelihood function using the gradients (the first and second derivative of the log-likelihood function). Therefore, in general, ML estimators present the following disadvantages: 1) it is possible to get a non-converged solution; 2) there is no guarantee of getting the global maximum; 3) consequently, the ML estimates depend on a starting value; 4) above disadvantages tend to get worse as sample sizes decrease. Because it is very common for one to analyze small sample sizes (e.g., less than 100) in social and behavioral applied research,

these disadvantages of ML estimation have occasionally frustrated researchers who want to estimate PCC over the last several decades.

Expected A Posteriori Estimation

The EAP estimation method for PCC proposed by Choi et al. (in press) also adopts the assumption of bivariate underlying normal distribution, and uses the same procedure, Equation (3), in the above two-step ML method to estimate threshold values. However, when it estimates PCC, it follows Bayes' theorem: posterior distribution \propto likelihood function \times prior distribution. In other words, subjective belief about what the true correlation is likely to be can be incorporated into the estimation procedure through the prior distribution. Here is a brief development of the EAP method for PCC (more details of the EAP estimator of IRT model are available in Bock & Aitken, 1981):

$$\Pr(\rho|x) = \frac{\Pr(\rho)\Pr(x|\rho)}{\int_{-1}^1 \Pr(\rho)\Pr(x|\rho)d\rho} \quad (2)$$

where $\Pr(\rho|x)$ is the posterior distribution of given x which is frequency data of two variables, $\Pr(\rho)$ is a prior distribution of ρ , and $\Pr(x|\rho)$ is the same likelihood function L in the ML method. Then, the EAP (i.e., the mean of the posterior distribution) estimate of PCC can be simply expressed as:

$$\rho_{EAP} = E[\Pr(\rho|x)] = \int_{-1}^1 \rho \Pr(\rho|x)d\rho. \quad (3)$$

For the purpose of numerical computation of the integration, the above two equations can be re-expressed as:

$$\Pr(\rho|x) = \frac{\Pr(\rho)\Pr(x|\rho)}{\sum_{i=1}^k \Pr(\rho_i)\Pr(x|\rho_i)} \quad (4)$$

and

$$\hat{\rho}_{EAP} = \sum_{i=1}^k \rho_i \Pr(\rho_i|x) \quad (5)$$

where k is the number of equally spaced quadrature points from -1 to +1.

As described in the above development, the EAP method is based on both Bayes' theorem and the non-iterative numerical integration. Consequently, the EAP method has the following advantages over the ML method: 1) there is no non-convergence issue (i.e., the estimates always exist); 2) there is no risk of a local maxima problem; 3) the estimates do not depend on a starting value; 4) the capability of including the a priori knowledge/belief on the parameter into the estimation process using prior distribution.

Methodology

Beyond the methodological advantages of the EAP over the ML illustrated above, it would be useful to investigate the empirical behavior of the two estimators over various conditions especially for a small sample sizes. In this study, a Monte Carlo simulation was used to examine the effect of sample size, population correlation magnitude, and number of categories on the PCC estimation for both EAP and ML estimators. The procedures can be summarized in the following way: (1) bivariate normal data was randomly generated from specific population correlation magnitude (ρ) and sample size (n); (2) the generated interval data was categorized over number of categories based on the threshold scheme and; (3) PCC was estimated by the EAP and ML estimation methods; and (4) the above procedures were repeated 1,000 times (i.e., iteration number $in = 1,000$).

Data Generation

The sample size variable was $n = 30, 50, 100,$ and 500 observations. These numbers were chosen to reflect from small to moderate sample size that might be commonly encountered in the social sciences. The population correlation variable was chosen with $\rho = 0, 0.1, 0.3, 0.5,$ and 0.7 magnitudes, ranging from null to moderate high.

The categorization rule (threshold scheme) used in this simulation study was the normal category option, which was also called the equal category width option within the range from -3 to 3 in standard normal distribution (Bollen & Barb, 1981). Therefore, the

ML AND EAP FOR POLYCHORIC CORRELATION

distribution of categorized data gets closer to normal as the number of categories is increased. The number of category for both ordinal variables was $r = s = 2, 3, 5,$ and 7 .

EAPPCC, a MATLAB subroutine (Choi et al., in press) was adopted for both EAP and ML estimators. Also, entire Monte Carlo simulation was implemented by a specialized program code in MATLAB (The MathWorks Inc., 2007), and the MVNRND function in MATLAB was used to generate bivariate normal data with specified population correlation (ρ) magnitude and sample size (n).

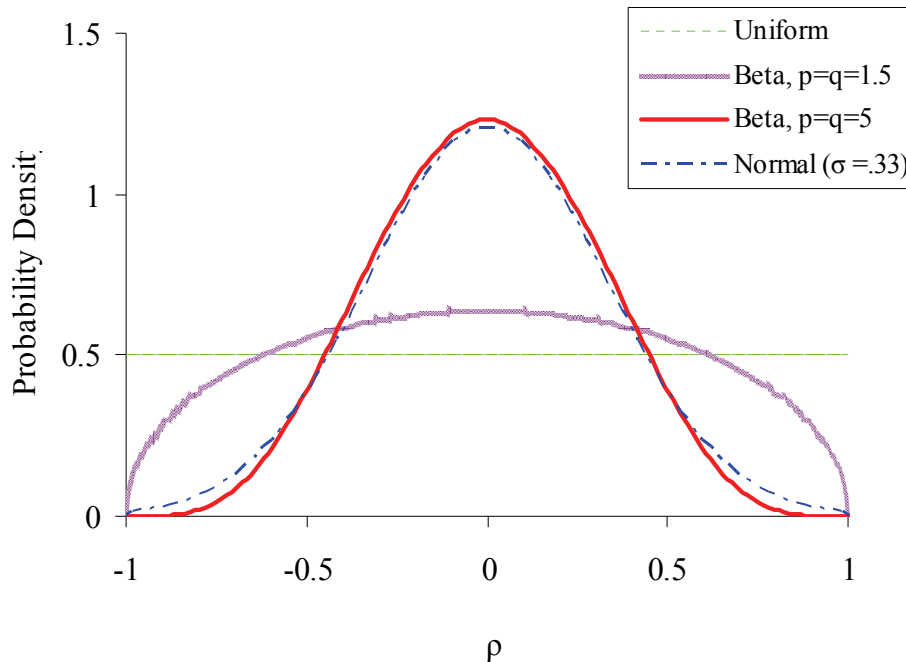
Estimation Options

In this study, the following estimators were considered: the ML, the EAP with Uniform $(-1, 1)$ prior (EAPU), the EAP with Normal $(0, 1/3)$ prior (EAPN), the EAP with Beta $(5, 5)$ prior (EAPB5), and the EAP with Beta $(1.5, 1.5)$ prior (EAPB1.5). The latter three were compared for prior sensitivity onto estimates.

Note that the range of Beta and Normal prior distribution have been adjusted to $[-1, 1]$ for the purpose of constructing the appropriate priors for the correlation. As shown in Figure 1, the shape of Beta $(5, 5)$ is very similar to that of Normal $(0, 1/3)$. From the perspective of informativeness of prior, above two priors are more informative than Beta $(1.5, 1.5)$ prior whereas the uniform distribution is least informative, specifically non-informative. Note that for comparing the performance of ML and EAP estimator, the EAPU (EAP with Uniform prior) was compared to the ML.

A two-step ML method was employed for the ML method and the EAP method adopted 100 quadrature points ($k = 100$). Chen et al. (1998) showed that any quadrature points of 20 or above were substantially the same. Therefore, 100 quadrature points would be a fair balance considering both accuracy requirement and computational load in this study.

Figure 1: Probability density functions for four prior distributions: Uniform $(-1, 1)$, Beta $(1.5, 1.5)$, Beta $(5, 5)$, Normal $(0, 1/3)$



First, the convergence of the ML estimator was examined using the convergence rate (*CR*):

$$CR = pcin / in \times 100 \% \quad (6)$$

where *pcin* was the properly converged iteration number, and *in* was the total iteration number attempted (i.e., *in* = 1,000).

Second, for evaluating the bias of the ML and EAPU (EAP with uniform distribution as prior) estimates, mean relative bias (*MRB*) was employed as the major statistics, and its general form was (Bandalos, 2006):

$$MRB = \sum_{i=1}^{pcin} \left[(\hat{\rho}_i - \rho) / \rho \right] / pcin \quad (7)$$

where $\hat{\rho}_i$ was the *i*th iteration PCC estimate. In case of $\rho = 0$, mean bias (*MB*) was used instead of *MRB*, to avoid the issue of dividing by zero:

$$MB = \sum_{i=1}^{pcin} \left[\hat{\rho}_i - \rho \right] / pcin . \quad (8)$$

In case average *MRB* values over different ρ magnitudes were needed (e.g., Figures 2 and 4), the *MB* values of $\rho = 0$ were excluded from averaging *MRB* values. Additionally, mean value (*M*) of estimates were also presented and analyzed:

$$M = \sum_{i=1}^{pcin} \hat{\rho}_i / pcin . \quad (9)$$

Third, for evaluating the variability of the ML and EAPU estimates, root mean squared error (*RMSE*) and standard deviation of mean values (*SD*) were examined with the following definitions:

$$RMSE = \left[\sum_{i=1}^{pcin} (\hat{\rho}_i - \rho)^2 / pcin \right]^{1/2}, \quad (10)$$

and

$$SD = \left[\sum_{i=1}^{pcin} (\hat{\rho}_i - M)^2 / pcin \right]^{1/2}. \quad (11)$$

MRB and *SD* were used for examining the prior sensitivity of EAP estimators.

Results

Convergence Rate for the ML Estimator

In this study, the convergence of both EAP and ML estimators was assessed; as shown in previous studies (e.g., Flora & Curran, 2004), not all iterations were converges for the ML estimator in this study as well. From the pilot study, it was found that the convergence rate was very low (< 30%) in most cases when a fixed initial value (0) was used. Therefore, PPMC was used with categorized data as the initial value for the ML method in this simulation study.

As indicated in Table 1, the average convergence rates were below 100% in all scenarios for the ML. Furthermore, as the sample size, number of categories, or ρ magnitude decrease, the rates tended to become worse. In contrast, as expected, EAP estimates could be obtained for all iterations for all conditions (i.e., *pcin* = 1,000 for the EAP).

Table 1: Convergence Rates of the ML Estimator

<i>n</i>	30	50	100	500	
%	98.1	98.7	99.1	99.4	
<i>r = s</i>	2	3	5	7	
%	97.5	99.1	99.5	99.4	
ρ	0	0.1	0.3	0.5	0.7
%	97.8	98.1	98.7	99.1	99.5

Note. Values were averaged over other conditions

Bias of ML and EAPU Estimates

Statistics regarding the bias of estimators (*M* and *MRB*) are presented in Table 2, and are also summarized and depicted in Figure 2.

First, for $\rho = 0$, the differences between the ML and EAP (specifically the EAPU) estimators were negligible, and *M* for both estimates were very close to zero (i.e., $|M| < 0.01$ for most cases).

ML AND EAP FOR POLYCHORIC CORRELATION

Table 2: Simulation Results of the ML and EAPU Estimates

	ML				EAPU			
	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>
<i>n</i>	$\rho = 0, r = s = 2$				$\rho = 0, r = s = 2$			
30	-0.013	0.295	-0.013	0.295	-0.011	0.249	-0.011	0.249
50	0.001	0.224	0.001	0.224	0.001	0.201	0.001	0.202
100	-0.003	0.161	-0.003	0.161	-0.003	0.152	-0.003	0.152
500	0.002	0.074	0.002	0.069	0.002	0.072	0.002	0.068
<i>n</i>	$\rho = 0, r = s = 5$				$\rho = 0, r = s = 5$			
30	0.007	0.231	0.007	0.231	0.006	0.202	0.006	0.202
50	0.007	0.165	0.007	0.165	0.007	0.150	0.007	0.150
100	0.000	0.111	0.000	0.110	0.000	0.105	0.000	0.105
500	0.004	0.050	0.004	0.051	0.004	0.050	0.004	0.050
<i>n</i>	$\rho = 0.1, r = s = 2$				$\rho = 0.1, r = s = 2$			
30	0.102	0.291	2.115	0.291	0.087	0.247	-13.291	0.247
50	0.096	0.224	-3.980	0.225	0.086	0.202	-13.581	0.203
100	0.091	0.150	-9.227	0.151	0.086	0.142	-14.056	0.143
500	0.103	0.073	3.334	0.070	0.102	0.071	1.924	0.069
<i>n</i>	$\rho = 0.1, r = s = 5$				$\rho = 0.1, r = s = 5$			
30	0.109	0.217	9.145	0.217	0.095	0.191	-4.611	0.191
50	0.097	0.163	-2.837	0.163	0.089	0.149	-11.277	0.150
100	0.102	0.112	1.488	0.112	0.096	0.107	-3.552	0.107
500	0.102	0.050	1.460	0.050	0.100	0.050	0.297	0.050
<i>n</i>	$\rho = 0.3, r = s = 2$				$\rho = 0.3, r = s = 2$			
30	0.282	0.265	-6.102	0.267	0.240	0.226	-20.096	0.234
50	0.300	0.205	0.080	0.205	0.271	0.186	-9.587	0.188
100	0.302	0.150	0.515	0.150	0.286	0.143	-4.631	0.143
500	0.299	0.069	-0.476	0.066	0.295	0.069	-1.761	0.065
<i>n</i>	$\rho = 0, r = s = 3$				$\rho = 0, r = s = 3$			
30	-0.016	0.275	-0.016	0.276	-0.014	0.236	-0.014	0.236
50	0.002	0.214	0.002	0.215	0.002	0.194	0.002	0.195
100	0.001	0.140	0.001	0.140	0.001	0.132	0.001	0.132
500	0.001	0.062	0.001	0.062	0.001	0.061	0.001	0.061
<i>n</i>	$\rho = 0, r = s = 7$				$\rho = 0, r = s = 7$			
30	0.002	0.208	0.002	0.208	0.002	0.182	0.002	0.182
50	-0.001	0.155	-0.001	0.155	-0.001	0.142	-0.001	0.141
100	0.003	0.113	0.003	0.113	0.003	0.108	0.003	0.107
500	0.002	0.048	0.002	0.048	0.002	0.048	0.002	0.048

^a Mean bias (*MB*) was used instead of *MRB* in case of $\rho = 0$

Table 2: Simulation Results of the ML and EAPU Estimates (continued)

<i>n</i>	ML				EAPU			
	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>
<i>n</i>	$\rho = 0.1, r = s = 3$				$\rho = 0.1, r = s = 3$			
30	0.112	0.270	11.719	0.271	0.096	0.231	-4.081	0.231
50	0.104	0.203	3.799	0.201	0.094	0.184	-6.248	0.182
100	0.108	0.136	7.582	0.137	0.102	0.129	1.688	0.129
500	0.103	0.058	2.509	0.058	0.101	0.058	1.211	0.058
<i>n</i>	$\rho = 0.1, r = s = 7$				$\rho = 0.1, r = s = 7$			
30	0.082	0.192	-18.201	0.193	0.071	0.168	-28.592	0.171
50	0.104	0.147	3.537	0.147	0.095	0.135	-5.346	0.135
100	0.106	0.111	5.876	0.111	0.101	0.105	0.714	0.105
500	0.099	0.047	-0.563	0.047	0.098	0.047	-1.691	0.047
<i>n</i>	$\rho = 0.3, r = s = 3$				$\rho = 0.3, r = s = 3$			
30	0.323	0.255	7.648	0.256	0.280	0.222	-6.835	0.223
50	0.316	0.184	5.209	0.185	0.287	0.170	-4.226	0.170
100	0.299	0.132	-0.243	0.132	0.284	0.127	-5.286	0.127
500	0.297	0.055	-1.025	0.055	0.294	0.055	-2.140	0.055
<i>n</i>	$\rho = 0.3, r = s = 5$				$\rho = 0.3, r = s = 5$			
30	0.301	0.200	0.462	0.200	0.266	0.180	-11.369	0.183
50	0.297	0.151	-1.169	0.151	0.273	0.142	-9.041	0.144
100	0.304	0.104	1.263	0.103	0.290	0.100	-3.199	0.100
500	0.300	0.046	-0.006	0.046	0.297	0.046	-0.982	0.046
<i>n</i>	$\rho = 0.5, r = s = 2$				$\rho = 0.5, r = s = 2$			
30	0.492	0.235	-1.602	0.235	0.422	0.206	-15.541	0.220
50	0.488	0.176	-2.329	0.176	0.444	0.163	-11.237	0.173
100	0.499	0.128	-0.161	0.128	0.475	0.123	-4.947	0.125
500	0.501	0.058	0.236	0.056	0.495	0.058	-0.988	0.056
<i>n</i>	$\rho = 0.5, r = s = 5$				$\rho = 0.5, r = s = 5$			
30	0.515	0.172	3.003	0.172	0.463	0.164	-7.332	0.168
50	0.501	0.130	0.235	0.130	0.468	0.127	-6.339	0.131
100	0.508	0.088	1.693	0.088	0.491	0.088	-1.778	0.088
500	0.499	0.041	-0.164	0.041	0.496	0.041	-0.884	0.041
<i>n</i>	$\rho = 0.7, r = s = 2$				$\rho = 0.7, r = s = 2$			
30	0.685	0.192	-2.217	0.193	0.598	0.178	-14.637	0.205
50	0.692	0.142	-1.180	0.143	0.636	0.138	-9.079	0.151
100	0.698	0.099	-0.241	0.099	0.669	0.098	-4.400	0.102
500	0.700	0.044	-0.009	0.043	0.693	0.044	-0.945	0.044

^a Mean bias (*MB*) was used instead of *MRB* in case of $\rho = 0$

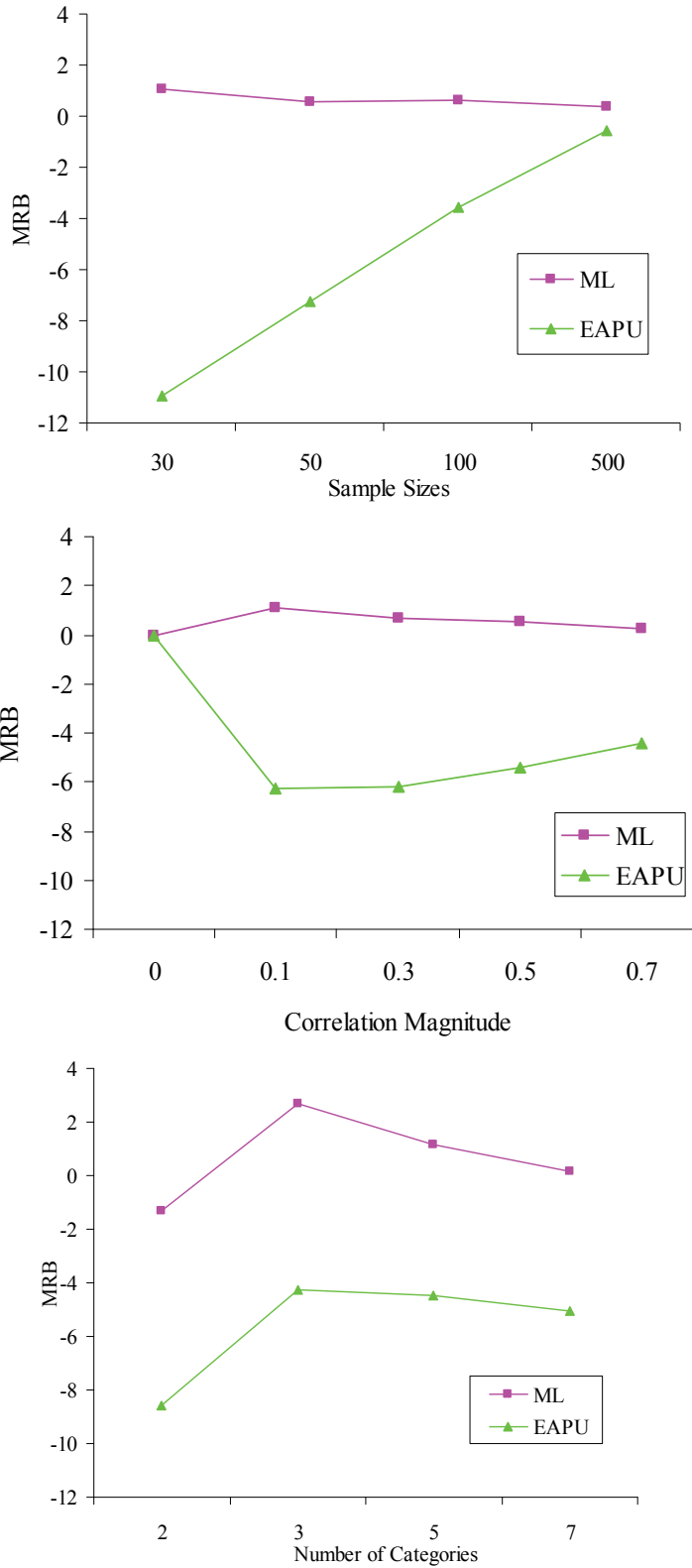
ML AND EAP FOR POLYCHORIC CORRELATION

Table 2: Simulation Results of the ML and EAPU Estimates (continued)

<i>n</i>	ML				EAPU			
	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>
<i>n</i>	$\rho = 0.7, r = s = 5$				$\rho = 0.7, r = s = 5$			
30	0.709	0.124	1.299	0.124	0.655	0.128	-6.458	0.136
50	0.711	0.092	1.622	0.092	0.679	0.095	-2.997	0.097
100	0.705	0.064	0.714	0.065	0.689	0.066	-1.559	0.067
500	0.701	0.028	0.119	0.028	0.698	0.029	-0.330	0.029
<i>n</i>	$\rho = 0.3, r = s = 7$				$\rho = 0.3, r = s = 7$			
30	0.310	0.187	3.472	0.188	0.275	0.170	-8.252	0.171
50	0.303	0.146	1.003	0.146	0.280	0.137	-6.707	0.138
100	0.301	0.102	0.353	0.102	0.288	0.099	-3.927	0.099
500	0.301	0.045	0.301	0.044	0.298	0.044	-0.644	0.044
<i>n</i>	$\rho = 0.5, r = s = 3$				$\rho = 0.5, r = s = 3$			
30	0.503	0.214	0.571	0.214	0.440	0.193	-12.068	0.202
50	0.507	0.162	1.445	0.162	0.466	0.154	-6.707	0.157
100	0.502	0.112	0.445	0.112	0.481	0.110	-3.861	0.112
500	0.500	0.051	0.049	0.051	0.496	0.051	-0.863	0.051
<i>n</i>	$\rho = 0.5, r = s = 7$				$\rho = 0.5, r = s = 7$			
30	0.513	0.156	2.530	0.157	0.464	0.151	-7.108	0.155
50	0.510	0.122	2.018	0.122	0.479	0.120	-4.156	0.122
100	0.502	0.083	0.341	0.083	0.486	0.083	-2.896	0.084
500	0.501	0.037	0.214	0.037	0.498	0.037	-0.456	0.037
<i>n</i>	$\rho = 0.7, r = s = 3$				$\rho = 0.7, r = s = 3$			
30	0.719	0.169	2.641	0.170	0.640	0.161	-8.573	0.172
50	0.706	0.122	0.791	0.122	0.659	0.121	-5.906	0.127
100	0.695	0.090	-0.665	0.090	0.672	0.090	-4.028	0.094
500	0.701	0.038	0.178	0.038	0.697	0.038	-0.491	0.039
<i>n</i>	$\rho = 0.7, r = s = 7$				$\rho = 0.7, r = s = 7$			
30	0.705	0.116	0.646	0.116	0.657	0.120	-6.143	0.128
50	0.703	0.089	0.366	0.089	0.674	0.093	-3.677	0.096
100	0.703	0.057	0.436	0.057	0.689	0.058	-1.542	0.059
500	0.700	0.026	0.057	0.026	0.698	0.026	-0.329	0.026

^a Mean bias (*MB*) was used instead of *MRB* in case of $\rho = 0$

Figure 2: *MRB* across Different Sample Sizes, Correlation Magnitude, and Number of Categories
 (Values averaged over other conditions; Mean Bias (*MB*) was Used in Case of $\rho = 0$)



ML AND EAP FOR POLYCHORIC CORRELATION

Second, as sample size increased, the difference between the two estimators disappeared, and both estimators performed better (i.e., smaller magnitude of *MRB*). Also, it seemed the EAP estimator was more sensitive to the change of sample size than the ML estimator according to Table 2 and Figure 2.

Third, for small sample sizes ($n = 30$ and 50) with non-zero ρ magnitude, *MRB* patterns of the two estimators substantially differed. For the ML case, *MRB* values can be largely positive or negative. Accordingly, the average *MRB* values (depicted in Figure 2) appeared to be smaller than most individual biases. Namely, the average *MRB* in Figure 2 were attenuated compared with the actual bias of the ML estimator in terms of *MRB*. For EAPU case, the *MRB* values tended to be largely negative for most cases, which supported the consistency between average and individual *MRB* values in Table 2 and Figure 2. These observations suggested the existence of systematic underestimation for the EAP estimator.

Fourth, the EAP estimator performed slightly better as ρ magnitude increased except in the null case. Also, the EAP estimator performed poorly for two categories as compared to higher numbers of categories, as shown in Figure 2. Again, the average *MRB* values for the ML in Figure 2 were less meaningful because they represent averages of the negative and positive values in each simulation cell *MRB* values.

Variability of ML and EAPU Estimates

Detailed *RMSE* and *SD* values of ML and EAP estimates are presented in Table 2, and are summarized and depicted in Figure 3.

First, in terms of *RMSE* and *SD*, the EAP outperformed the ML estimator in most cases. However, the differences in *RMSE* and *SD* among estimators are negligible for many cases.

Second, in small sample sizes ($n = 30$ and 50), the EAP estimator clearly outperformed the ML estimator (Figure 3). As sample size increased, the difference of *RMSE* values between two estimators disappeared, and the variability or fluctuation became increasingly

smaller, which suggested that both estimators were asymptotically efficient.

Third, a similar pattern was found over ρ magnitude. In small magnitude, the EAP estimator evidently outperformed the ML estimator. As the magnitude increased, the difference between the two estimators disappeared, and the variability of estimates decreased for both estimators. However, *RMSE* values were more sensitive to the change of sample size than that of ρ magnitude. This can be observed by comparing different charts in Figure 3.

Fourth, for number of categories, the EAP estimator also appeared to outperform the ML estimator in all cases. However, the differences in *RMSE* values between five and seven categories were very small for both estimators.

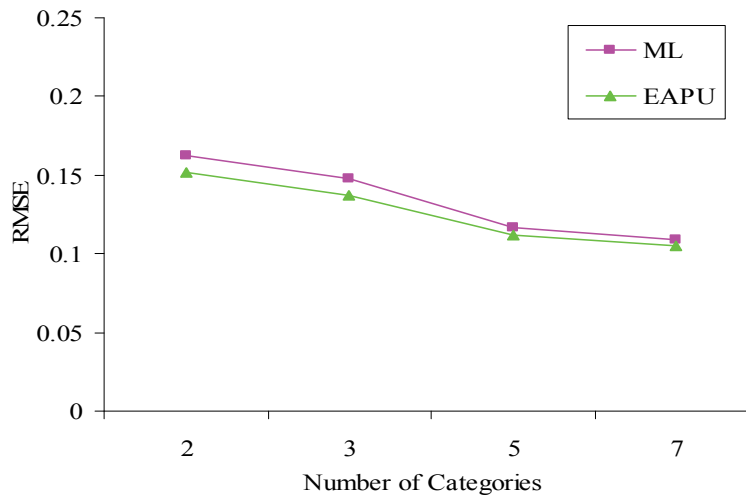
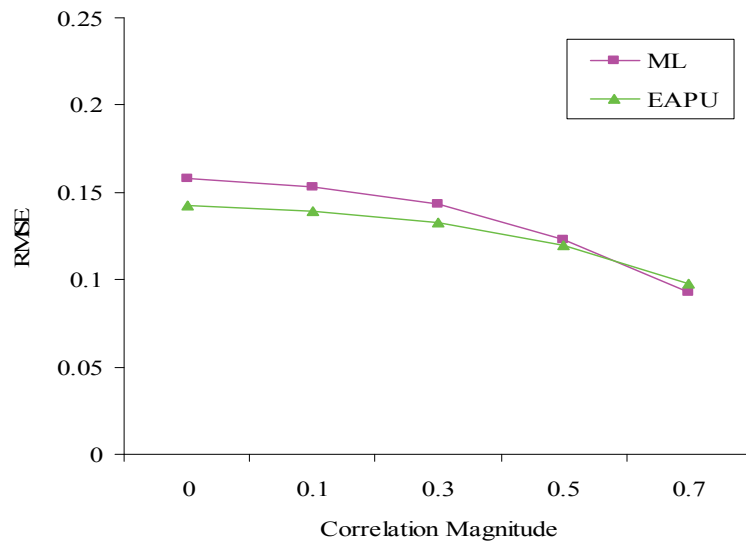
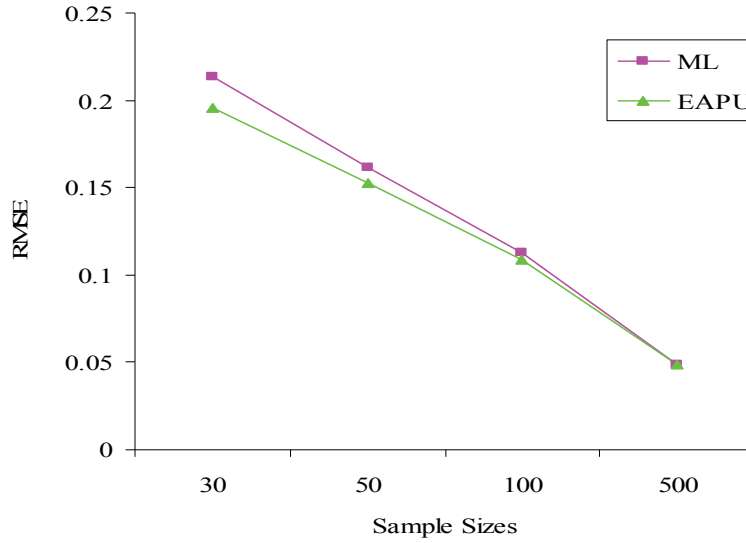
Prior Sensitivity for EAP with Different Prior Distributions

Statistics of EAP estimates with different prior distributions are presented in Table 3, and those are also summarized and depicted in Figures 4 through 5.

First, the EAP estimators whose prior distributions were more informative increasingly biased toward the mean of prior distribution. As shown in Table 3, the mean of estimates (*M*) for both EAPN and EAPB5 estimators, whose prior distributions are most informative (Figure 1), are very close to zero for most conditions. Also, *MRB* values for both EAPN and EAPB5 were extremely negatively biased (approximately -92) for all non zero population correlation cases. The *M* and *MRB* values for EAPB1.5, whose prior distribution is moderately informative in this study, are always between the above informative prior cases and the least informative prior case, EAPU. The *MRB* values in Figure 4 present essentially the same results.

Second, for estimators whose prior distributions were more informative, the variability of estimates in terms of *SD* was less. Both EAPN and EAPB5 cases showed the smallest *SD* values, whereas the *SD* values of the EAPB1.5 estimator were larger when compared with the above two estimators, but smaller when compared with the EAPU case (Figure 5). These results were also consistent with the

Figure 3: *RMSE* across Different Sample Sizes, Correlation Magnitude, and Number of Categories
(Values were averaged over other conditions.)



ML AND EAP FOR POLYCHORIC CORRELATION

Table 3: Simulation Results of EAPU Estimates with Different Prior Distributions

	EAPU				EAPB1.5			
	<i>M</i>	<i>SD</i>	<i>MRB^a</i>	<i>RMSE</i>	<i>M</i>	<i>SD</i>	<i>MRB^a</i>	<i>RMSE</i>
<i>n</i>	$\rho = 0$				$\rho = 0$			
30	-0.004	0.217	-0.004	0.217	-0.001	0.089	-0.001	0.089
50	0.002	0.172	0.002	0.172	0.001	0.069	0.001	0.069
100	0.000	0.124	0.000	0.124	0.000	0.049	0.000	0.049
500	0.002	0.058	0.002	0.057	0.001	0.022	0.001	0.022
<i>n</i>	$\rho = 0.1$				$\rho = 0.1$			
30	0.087	0.209	-12.644	0.210	0.036	0.085	-63.600	0.107
50	0.091	0.168	-9.113	0.168	0.038	0.067	-62.341	0.092
100	0.096	0.121	-3.801	0.121	0.039	0.048	-61.026	0.077
500	0.100	0.056	0.435	0.056	0.040	0.022	-60.496	0.065
<i>n</i>	$\rho = 0.3$				$\rho = 0.3$			
30	0.265	0.200	-11.638	0.203	0.112	0.082	-62.550	0.205
50	0.278	0.159	-7.390	0.160	0.115	0.065	-61.684	0.196
100	0.287	0.117	-4.261	0.117	0.116	0.047	-61.302	0.190
500	0.296	0.053	-1.382	0.052	0.118	0.022	-60.675	0.184
<i>n</i>	$\rho = 0.5$				$\rho = 0.5$			
30	0.447	0.179	-10.512	0.186	0.191	0.077	-61.865	0.319
50	0.464	0.141	-7.110	0.146	0.194	0.061	-61.179	0.312
100	0.483	0.101	-3.370	0.102	0.199	0.044	-60.226	0.304
500	0.496	0.047	-0.798	0.046	0.203	0.024	-59.395	0.300
<i>n</i>	$\rho = 0.7$				$\rho = 0.7$			
30	0.637	0.147	-8.967	0.160	0.275	0.071	-60.667	0.430
50	0.662	0.112	-5.415	0.118	0.284	0.055	-59.471	0.419
100	0.680	0.078	-2.882	0.081	0.289	0.039	-58.658	0.412
500	0.697	0.034	-0.524	0.034	0.297	0.023	-57.632	0.406

Note. Values were averaged across different number of categories

Table 3: Simulation Results of EAPU Estimates with Different Prior Distributions (continued)

	EAPB5				EAPN			
	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>	<i>M</i>	<i>SD</i>	<i>MRB</i> ^a	<i>RMSE</i>
<i>n</i>	$\rho = 0$				$\rho = 0$			
30	-0.001	0.017	-0.001	0.017	-0.001	0.016	-0.001	0.016
50	0.000	0.013	0.000	0.013	0.000	0.012	0.000	0.012
100	0.000	0.010	0.000	0.010	0.000	0.009	0.000	0.009
500	0.000	0.004	0.000	0.004	0.000	0.004	0.000	0.004
<i>n</i>	$\rho = 0.1$				$\rho = 0.1$			
30	0.007	0.016	-92.914	0.095	0.006	0.015	-93.578	0.095
50	0.007	0.013	-92.636	0.094	0.007	0.012	-93.352	0.094
100	0.007	0.009	-92.400	0.093	0.007	0.009	-93.164	0.093
500	0.007	0.005	-92.328	0.092	0.007	0.004	-93.114	0.093
<i>n</i>	$\rho = 0.3$				$\rho = 0.3$			
30	0.022	0.016	-92.712	0.278	0.020	0.014	-93.393	0.280
50	0.022	0.012	-92.585	0.278	0.020	0.011	-93.306	0.280
100	0.022	0.009	-92.538	0.278	0.020	0.008	-93.286	0.280
500	0.023	0.005	-92.409	0.277	0.021	0.004	-93.185	0.280
<i>n</i>	$\rho = 0.5$				$\rho = 0.5$			
30	0.037	0.014	-92.688	0.464	0.033	0.013	-93.368	0.467
50	0.037	0.011	-92.603	0.463	0.033	0.010	-93.319	0.467
100	0.038	0.008	-92.453	0.463	0.034	0.007	-93.205	0.466
500	0.039	0.005	-92.293	0.462	0.035	0.004	-93.076	0.466
<i>n</i>	$\rho = 0.7$				$\rho = 0.7$			
30	0.052	0.013	-92.623	0.648	0.047	0.011	-93.304	0.653
50	0.053	0.009	-92.470	0.647	0.048	0.009	-93.192	0.652
100	0.053	0.007	-92.370	0.647	0.048	0.006	-93.123	0.652
500	0.054	0.004	-92.213	0.646	0.049	0.004	-92.998	0.651

Note. Values were averaged across different number of categories

ML AND EAP FOR POLYCHORIC CORRELATION

Figure 4: *MRB* of the EAP Estimator across Different Sample Sizes and Correlation Magnitude
(Values were averaged over other conditions)

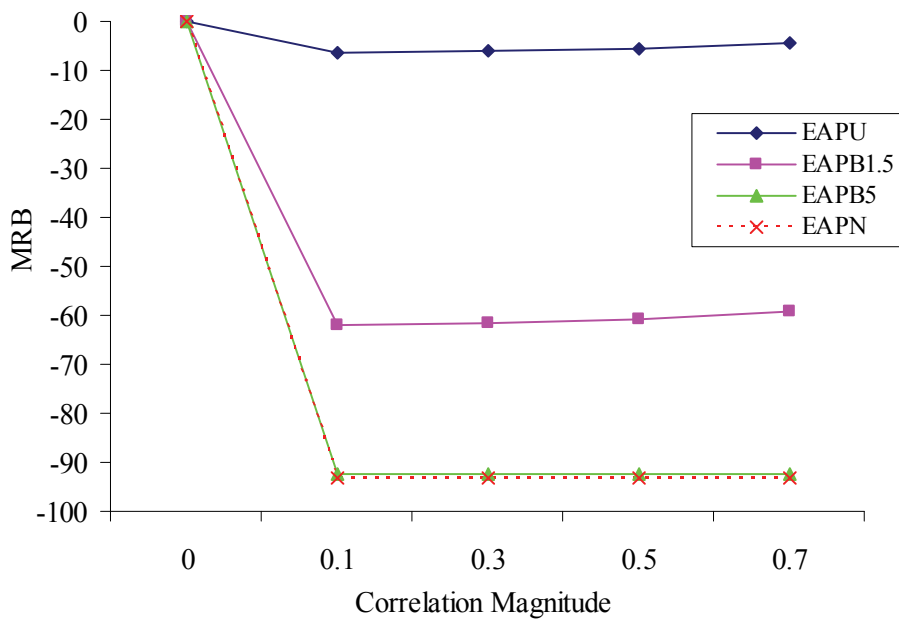
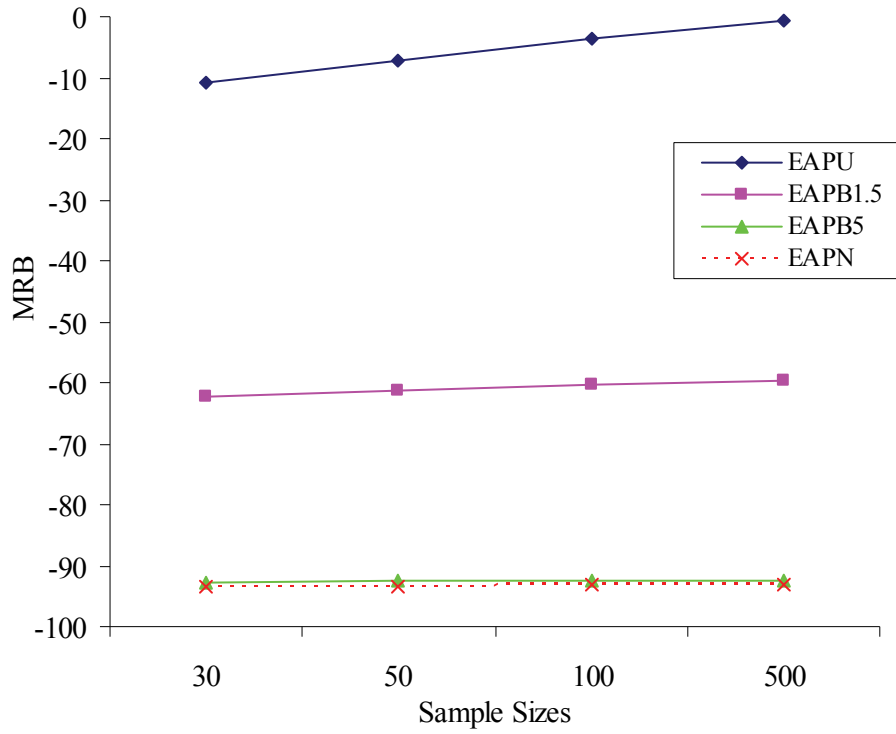
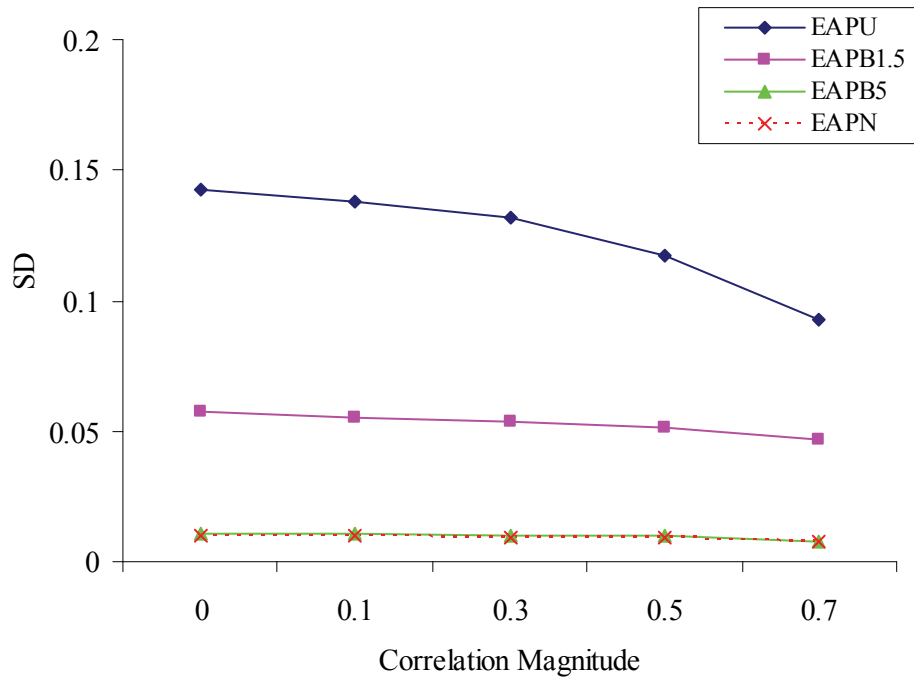
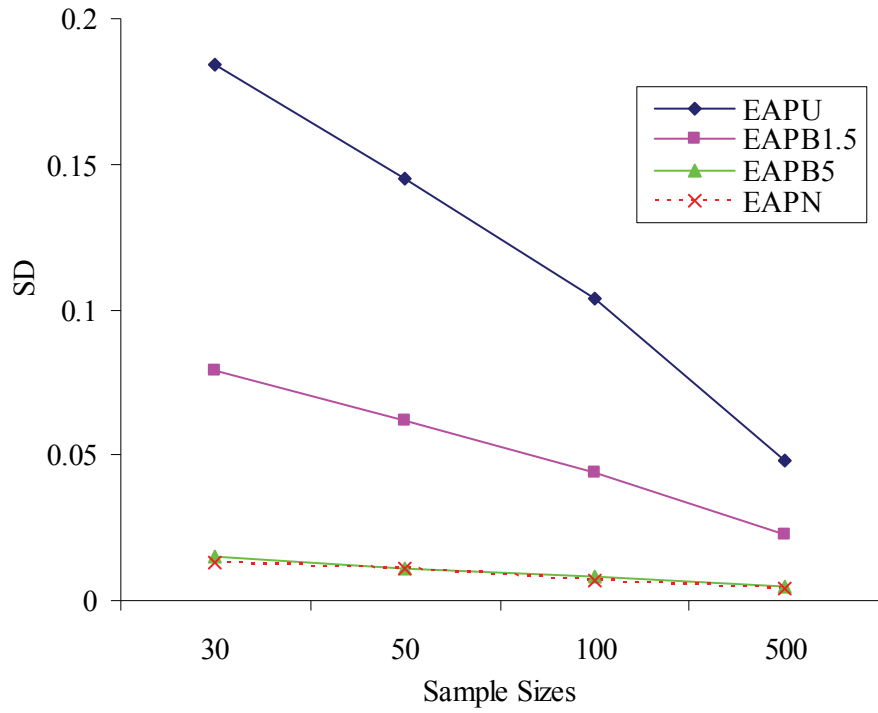


Figure 5: *SD* of the EAP Estimator across Different Sample Sizes and Correlation Magnitude
(Values were averaged over other conditions)



informativeness order of different prior distributions (Figure 1).

Third, the performance of the EAP estimators in terms of biases (*MRB*) and variability (*SD*) got better as sample size or ρ increase. However, as the prior distribution got more informative, both *MRB* and *SD* are less sensitive to the change of sample size or ρ magnitude, whereas the EAPU estimates were most sensitive to those changes (Figures 4 through 5). Furthermore, for $\rho = 0$ magnitude, biases of all estimators disappeared.

Conclusion

This study attempted to evaluate and compare the behaviors of ML and EAP estimators for PCC focusing on small sample size cases. The convergence rate of the ML estimator improves when using the PPMC for categorized data as the initial value rather than a fixed number, 0. However, non-converged cases for the ML remain an issue. In contrast, for the EAPU estimator, there is no non-convergent issue. These results could prove promising and useful to the applied researcher who is planning to estimate PCC and perhaps already arrived at a non-converged solution from the ML estimator.

For small sample sizes, the ML estimator can substantially underestimate or overestimate the ρ magnitude, whereas the EAP estimator always underestimates the ρ magnitude as shown in Figure 6. This, shrinkage effect, is a well known problem of EAP, similar to other Bayesian estimates. Because of the nature of EAP, a weighted average of posterior distribution which is a function of prior, EAP estimates is generally biased toward the mean of the prior distribution. Because all priors used in this research have zero as a mean, EAP estimates were biased toward zero for all cases. Important issues to be addressed include 1) the shrinkage effect is more apparent as the ρ magnitude increased or sample size decreased; and 2) the shrinkage effect disappears when $\rho = 0$.

Although this shrinkage effect is obviously a negative aspect of the EAP estimator, the systematic underestimation pattern could be wisely utilized to arrive at a conservative estimate of the true value. As

shown in Figure 2 and Table 2, the true value is most likely 0% to 15% higher than the EAPU estimate. Applied researchers should note that the ML estimator cannot provide such information.

For the variability of estimates in terms of *RMSE* or *SD*, the EAP estimator generally outperforms the ML estimator. The results are more apparent when the sample size, ρ magnitude, or number of categories is small. Because the EAP estimator is a weighted average over a prior distribution, it tends to provide more stable estimates than the ML estimator in those conditions.

For the EAP estimator with small sample sizes, the use of two categories would not be recommended due to relatively large bias and variability of estimates. Meanwhile, the use of three, five, or seven categories does not provide much difference in *MRB* values. Although *RMSE* gets smaller as the number of categories increases, the difference in *RMSE* between two categories and higher numbers of categories is not as imminent as that of *MRB* (i.e., the average *RMSE* with two categories – the average *RMSE* with seven categories < 0.05). In sum, especially when sample sizes are small (50 or below) and number of categories are not large (five or below), the EAP estimator can be recommended over the ML because the EAP is free from the convergence issue and provides smaller estimate variability. Also, as the sample size increases, both the shrinkage effect and the difference between the ML and EAP estimators disappear.

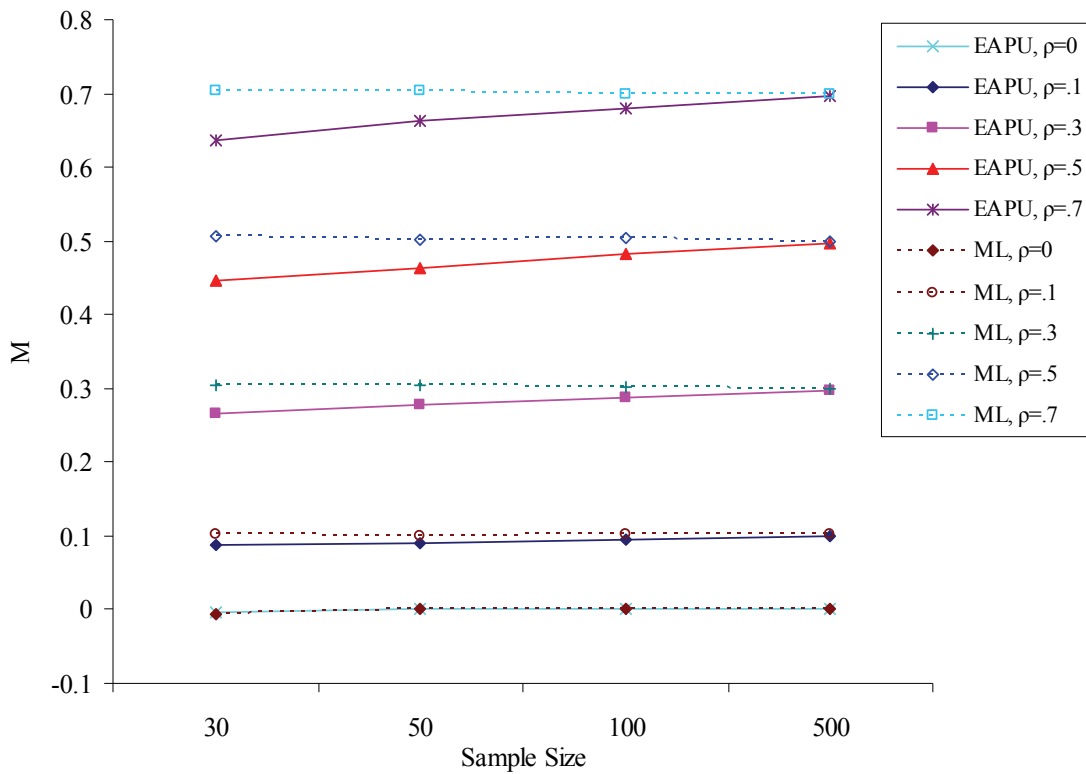
EAP estimators with more informative prior distribution could result in stronger bias toward the mean of the prior distribution, and provide less variation of estimates in terms of *SD*. For EAP estimators with relatively informative prior distributions like EAPN or EAPB5, both bias and variation of estimates are insensitive to the change of the sample size or ρ magnitude. For applied researchers with a strong a priori belief of the true correlation, EAP can provide a more stable estimate. Researchers can also include such information into an estimation procedure by adopting an informative prior distribution in the EAP estimator. This advantage of the EAP is beneficial especially

when the research involves small sample sizes in practice.

The following suggestions for future research on the estimation of PCC are based on the findings from this study. First, other Bayesian estimate, e.g., maximum a posteriori (MAP) estimate might be considered for analysis. As it relies on the mode, rather than the mean, of the posterior distribution, MAP could have some advantages against either the EAP or ML estimates. Second, the violation of

normality under different settings should be considered. As it has been shown that modest violation to normality is not critical for the ML estimator under relatively large sample size (Flora & Curran, 2004; Olsson, 1979; Quiroga, 1992), the situations for small sample size or for other estimators are not fully understood yet. Third, as this study focused on the point estimate, investigations on the interval estimates over different estimators would be needed, and should be addressed in future research.

Figure 6: M across Different Sample Sizes
(Values were averaged over number of categories)



ML AND EAP FOR POLYCHORIC CORRELATION

References

- Babakus, E., & Ferguson, C. E. (1988). On choosing the appropriate measure of association when analyzing rating scale data. *Academy of Marketing Science*, 16(1), 95-103.
- Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385-427). Charlotte, NC: Information Age Publishing Inc.
- Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, 46, 232-239.
- Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd Ed.)*. Pacific Grove, CA: Duxbury.
- Chen, S.-K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, 58(4), 569-595.
- Choi, J., Chen, J., & Kim, S. (in press). EAPPCC: A Matlab subroutine for estimating polychoric correlation matrices using an expected a posteriori estimation method. *Applied Psychological Measurement*.
- Flora, D. B. (2002). *Evaluation of categorical variable methodology for confirmatory factor analysis with Likert-type data*. Unpublished Doctoral dissertation, University of North Carolina at Chapel Hill, Chapel Hill.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466-491.
- Jöreskog, K. G. (2002-2005). *Structural equation modeling with ordinal variables using LISREL*. Retrieved 10/07/2007, from <http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf>.
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13(1), 57-75.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44, 443-460.
- Quiroga, A. M. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables*. Unpublished Doctoral dissertation, Acta Universitatis Upsaliensis.
- Rigdon, E. E., & Ferguson, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, 28(4), 491-497.
- SAS Institute Inc. (2004). *SAS/STAT 9.1 user's guide*. Cary, NC: SAS Institute Inc.
- The MathWorks Inc. (2007). Documentation for MathWorks Products, R2007b [Electronic Version]. Retrieved 10/07/2007 from <http://www.mathworks.com/access/helpdesk/help/helpdesk.html>.

It's Back!

Design and Analysis of Time-Series Experiments

(with a new Introduction by the first author)

Gene V Glass, Arizona State University

Victor L. Willson, Texas A&M University

John M. Gottman, The Gottman Institute, Seattle, Washington

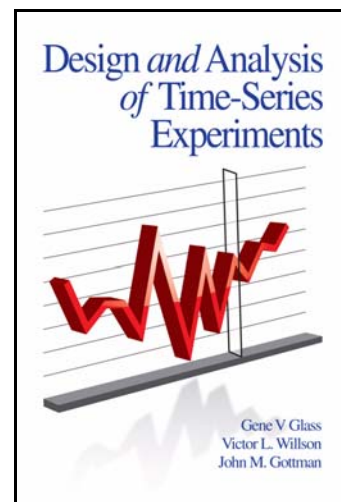
Hailed as a landmark in the development of experimental methods when it appeared in 1975, *Design and Analysis of Time-Series Experiments* is available again after several years of being out of print.

Gene V Glass, Victor L. Willson and John M. Gottman have carried forward the design and analysis of perhaps the most powerful and useful quasi-experimental design identified by their mentors in the classic Campbell & Stanley text *Experimental and Quasi-experimental Design for Research* (1966). In an era when governments seek to resolve questions of experimental validity by fiat and the label "Scientifically Based Research" is appropriated for only certain privileged experimental designs, nothing could be more appropriate than to bring back the classic text that challenges doctrinaire opinions of proper causal analysis.

Glass, Willson & Gottman introduce and illustrate an armamentarium of interrupted time-series experimental designs that offer some of the most powerful tools for discovering and validating causal relationships in social and education policy analysis. Drawing on the ground-breaking statistical analytic tools of Box & Jenkins, the authors extend the comprehensive autoregressive-integrated-moving-averages (ARIMA) model to accommodate significance testing and estimation of the effects of interventions into real world time-series. Designs and full statistical analyses are richly illustrated with actual examples from education, behavioral psychology, and sociology.

"...this book will come to be viewed as a true landmark. ... [It] should stand the test of time exceedingly well." ~ James A. Walsh (*Educational & Psychological Measurement*, 1975)

"Ordinary least squares estimation is usually inapplicable because of autoregressive error.... Glass, Willson, and Gottman have assembled the best approach." ~Donald T. Campbell



Publication Date:

Winter 2009

ISBN's:

Paperback: 978-1-59311-980-5

Hardcover: N/A

Price:

Paperback: \$39.99

Hardcover:

Trim Size: 6 X 9

Subject:

Education, Statistics

Special Price: \$25.99 paperbacks plus s/h

Book URL: <http://www.infoagepub.com/products/content/p489c9049a428d.php>

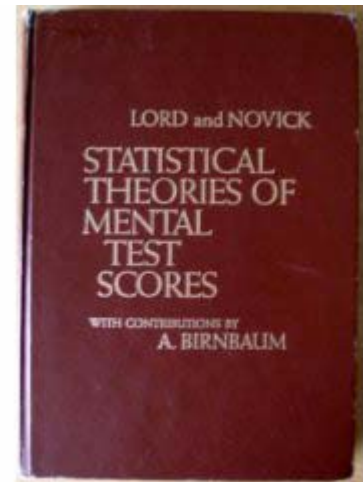
It's Back in Paperback!

Statistical Theories of Mental Test Scores

by **Frederic M. Lord** and **Melvin R. Novick**

A classic returns.

- ...pioneering work...
- ...comprehensive...
- ...classic...
- ...definitive...
- ...unquestioned status and authority...



Tatsuoka was right:

"This comprehensive and authoritative work is a major contribution to the literature of test theory. Without doubt it is destined to become a classic in the field." ~ Maurice Tatsuoka (1971)

One of the most important books in the history of psychometrics has been virtually unavailable to scholars and students for decades. A gap in the archives of modern test theory is now being filled by the release in paperback for the first time of the classic text, *Statistical Theories of Mental Test Scores*, by the late and honored statisticians and psychometricians, Frederic M. Lord and Melvin R. Novick. No single book since 1968 when Lord & Novick first appeared has had a comparable impact on the practice of testing and assessment.

Information Age Publishing is proud to make this classic text available to a new generation of scholars and researchers.

<http://www.infoagepub.com/products/content/p4810c9a0891af.php>

Publication Date:

Spring 2008

ISBN's:

Paperback: 978-1-59311-934-8

Price:

Paperback: \$59.99

Trim Size: 6 X 9

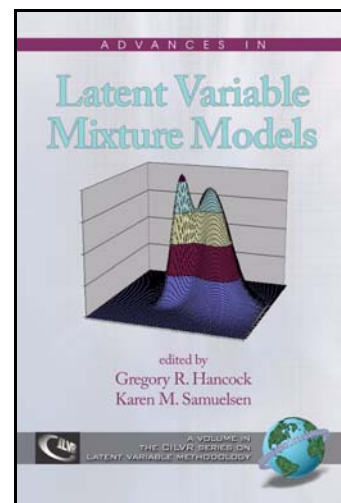
Subject:

Education, Statistics

Advances in Latent Variable Mixture Models

Edited by **Gregory R. Hancock**, *University of Maryland, College Park*, and **Karen M. Samuelsen**, *University of Georgia*

The current volume, *Advances in Latent Variable Mixture Models*, contains chapters by all of the speakers who participated in the 2006 CILVR conference, providing not just a snapshot of the event, but more importantly chronicling the state of the art in latent variable mixture model research. The volume starts with an overview chapter by the CILVR conference keynote speaker, Bengt Muthén, offering a “lay of the land” for latent variable mixture models before the volume moves to more specific constellations of topics. Part I, *Multilevel and Longitudinal Systems*, deals with mixtures for data that are hierarchical in nature either due to the data’s sampling structure or to the repetition of measures (of varied types) over time. Part II, *Models for Assessment and Diagnosis*, addresses scenarios for making judgments about individuals’ state of knowledge or development, and about the instruments used for making such judgments. Finally, Part III, *Challenges in Model Evaluation*, focuses on some of the methodological issues associated with the selection of models most accurately representing the processes and populations under investigation. It should be stated that this volume is not intended to be a first exposure to latent variable methods. Readers lacking such foundational knowledge are encouraged to consult primary and/or secondary didactic resources in order to get the most from the chapters in this volume. Once armed with that basic understanding of latent variable methods, we believe readers will find this volume incredibly exciting.



CONTENTS: Editors’ Introduction, *Gregory R. Hancock and Karen M. Samuelsen*. Acknowledgments. Latent Variable Hybrids: Overview of Old and New Models, *Bengt Muthén*. **PART I: Multilevel and Longitudinal Systems.** Multilevel Mixture Models, *Tihomir Asparouhov and Bengt Muthén*. Longitudinal Modeling of Population Heterogeneity: Methodological Challenges to the Analysis of Empirically Derived Criminal Trajectory Profiles, *Frauke Kreuter and Bengt Muthén*. Examining Contingent Discrete Change Over Time with Associative Latent Transition Analysis, *Brian P. Flaherty*. Modeling Measurement Error in Event Occurrence for Single, Non-Recurring Events in Discrete-Time Survival Analysis, *Katherine E. Masyn*. **PART II: Models for Assessment and Diagnosis.** Evidentiary Foundations of Mixture Item Response Theory Models, *Robert J. Mislevy, Roy Levy, Marc Kroopnick, and Daisy Rutstein*. Examining Differential Item Functioning from a Latent Mixture Perspective, *Karen M. Samuelsen*. Mixture Models in a Developmental Context, *Karen Draney, Mark Wilson, Judith Glück, and Christiane Spiel*. Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment, *Amy Soller and Ron Stevens*. The Mixture General Diagnostic Model, *Matthias von Davier*. **PART III: Challenges in Model Evaluation.** Categories or Continua? The Correspondence Between Mixture Models and Factor Models, *Eric Loken and Peter Molenaar*. Applications and Extensions of the Two-Point Mixture Index of Model Fit, *C. Mitchell Dayton*. Identifying the Correct Number of Classes in Growth Mixture Models, *Davood Tofghi and Craig K. Enders*. Choosing a “Correct” Factor Mixture Model: Power, Limitations, and Graphical Data Exploration, *Gitta H. Lubke and Jeffrey R. Spies*. About the Contributors.

Books of Related Interest:

Structural Equation Modeling: A Second Course

<http://www.infoagepub.com/products/content/1-59311-015-4.php>

2006

Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.95

Real Data Analysis

<http://infoagepub.com/products/content/978-1-59311-565-4.php>

2007

Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.95

Publication Date:

Fall 2007

ISBN’s:

Paperback: 978-1-59311-847-1

Hardcover: 978-1-59311-848-8

Price:

Paperback: \$39.99

Hardcover: \$73.99

Trim Size: 6 X 9

Subject:

Education

Structural Equation Modeling: A Second Course

Edited by **Gregory R. Hancock**, *University of Maryland*
and **Ralph O. Mueller**, *The George Washington University*

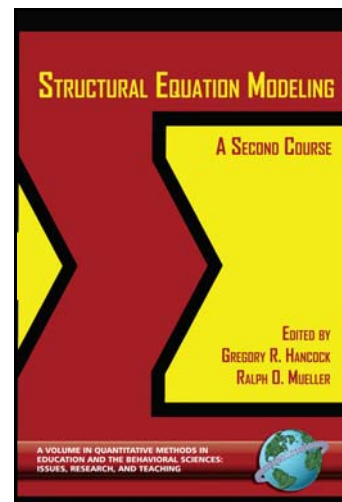
A volume in **Quantitative Methods in Education and the Behavioral Sciences:
Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)

"I believe that this volume represents a vital contribution to the field of SEM beyond the introductory level."

From the Preface by
Richard G. Lomax, *The University of Alabama*



This volume is intended to serve as a didactically-oriented resource covering a broad range of advanced topics often not discussed in introductory courses on structural equation modeling (SEM). Such topics are important in furthering the understanding of foundations and assumptions underlying SEM as well as in exploring SEM as a potential tool to address new types of research questions that might not have arisen during a first course. Chapters focus on the clear explanation and application of topics, rather than on analytical derivations, and contain syntax and partial output files from popular SEM software.

CONTENTS: Introduction to Series, *Ronald C. Serlin*. Preface, *Richard G. Lomax*. Dedication. Acknowledgements. Introduction, *Gregory R. Hancock & Ralph O. Mueller*. **Part I: Foundations.** The Problem of Equivalent Structural Models, *Scott L. Hershberger*. Formative Measurement and Feedback Loops, *Rex B. Kline*. Power Analysis in Covariance Structure Modeling, *Gregory R. Hancock*. **Part II: Extensions.** Evaluating Between-Group Differences in Latent Variable Means, *Marilyn S. Thompson & Samuel B. Green*. Using Latent Growth Models to Evaluate Longitudinal Change, *Gregory R. Hancock & Frank R. Lawrence*. Mean and Covariance Structure Mixture Models, *Phill Gagné*. Structural Equation Models of Latent Interaction and Quadratic Effects, *Herbert W. Marsh, Zhonglin Wen, & Kit-Tai Hau*. **Part III: Assumptions.** Nonnormal and Categorical Data in Structural Equation Modeling, *Sara J. Finney & Christine DiStefano*. Analyzing Structural Equation Models with Missing Data, *Craig K. Enders*. Using Multilevel Structural Equation Modeling Techniques with Complex Sample Data, *Laura M. Stapleton*. The Use of Monte Carlo Studies in Structural Equation Modeling Research, *Deborah L. Bandalos*. About the Authors.

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 \$39.99 Hardcover ISBN: 978-1-59311-685-9 \$73.99

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.99

Publication Date:
2005

ISBN's:
Paperback: 1-59311-014-6
Hardcover: 1-59311-015-4

Price:
Paperback: \$39.99
Hardcover: \$73.99

Subject:
Education, Statistics

Series URL: <http://www.infoagepub.com/products/series/serlin.html>



New Book Information

Multilevel Modeling of Educational Data

Edited by **Ann A. C'Connell**, *Ohio State University*
and **D. Betsy McCoach**, *University of Connecticut*

A volume in **Quantitative Methods in Education and the Behavioral Sciences:
Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the *Educational Statisticians, SIG*)

Multilevel Modeling of Educational Data, co-edited by Ann A. O'Connell, Ed.D., and D. Betsy McCoach, Ph.D., is the next volume in the series: *Quantitative Methods in Education and the Behavioral Sciences: Issues, Research and Teaching* (Information Age Publishing), sponsored by the Educational Statisticians' Special Interest Group (Ed-Stat SIG) of the American Educational Research Association. The use of multilevel analyses to examine effects of groups or contexts on individual outcomes has burgeoned over the past few decades. Multilevel modeling techniques allow educational researchers to more appropriately model data that occur within multiple hierarchies (i.e.- the classroom, the school, and/or the district). Examples of multilevel research problems involving schools include establishing trajectories of academic achievement for children within diverse classrooms or schools or studying school-level characteristics on the incidence of bullying. Multilevel models provide an improvement over traditional single-level approaches to working with clustered or hierarchical data; however, multilevel data present complex and interesting methodological challenges for the applied education research community.

In keeping with the pedagogical focus for this book series, the papers this volume emphasize applications of multilevel models using educational data, with chapter topics ranging from basic to advanced. This book represents a comprehensive and instructional resource text on multilevel modeling for quantitative researchers who plan to use multilevel techniques in their work, as well as for professors and students of quantitative methods courses focusing on multilevel analysis. Through the contributions of experienced researchers and teachers of multilevel modeling, this volume provides an accessible and practical treatment of methods appropriate for use in a first and/or second course in multilevel analysis. A supporting website links chapter examples to actual data, creating an opportunity for readers to reinforce their knowledge through hands-on data analysis. This book serves as a guide for designing multilevel studies and applying multilevel modeling techniques in educational and behavioral research, thus contributing to a better understanding of and solution for the challenges posed by multilevel systems and data.

CONTENTS: **Series Introduction**, *Ronald C. Serlin*. **Acknowledgements**. Part I: **Design Contexts for Multilevel Models**. Introduction, *Ann A. O'Connell and D. Betsy McCoach*. The Use of National Datasets for Teaching and Research, *Laura M. Stapleton and Scott L. Thomas*. Using Multilevel Modeling to Investigate School Effects, *Xin Ma, Lingling Ma, and Kelly D. Bradley*. Modeling Growth Using Multilevel and Alternative Approaches, *Janet K. Holt*. Cross-Classified Random Effects Models, *S. Natasha Beretvas*. Multilevel Logistic Models for Dichotomous and Ordinal Data, *Ann A. O'Connell, Jessica Goldstein, H. Jane Rogers, and C. Y. Joanne Peng*. Part II: **Planning and Evaluating Multilevel Models**. Evaluation of Model Fit and Adequacy, *D. Betsy McCoach and Anne C. Black*. Power, Sample Size, and Design, *Jessica Spybrook*. Part III: **Extending the Multilevel Framework**. Multilevel Methods for Meta-Analysis, *Sema A. Kalaian and Rafa M. Kasim*. Multilevel Measurement Modeling, *Kihito Kamata, Daniel J. Bauer, and Yasuo Miyazaki*. Part IV: **Mastering the Technique**. Reporting Results from Multilevel Analyses, *John M. Ferron, Kristin Y. Hogarty, Robert F. Dedrick, Melinda R. Hess, John D. Niles, and Jeffrey D. Kromrey*. Software Options for Multilevel Models, *J. Kyle Roberts and Patrick McLeod*. Estimation Procedures for Hierarchical Linear Models, *Hariharan Swaminathan and H. Jane Rogers*.

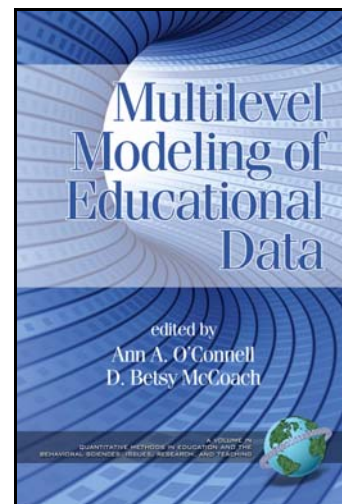
Also Available:

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.95

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.95



Publication Date:
Spring 2008

ISBN's:
Paperback: 978-1-59311-684-2
Hardcover: 978-1-59311-685-9

Price:
Paperback: \$39.99
Hardcover: \$73.99

Subject:
Education, Statistics

Series URL: <http://www.infoagepub.com/products/series/serlin.html>

Book URL: <http://www.infoagepub.com/products/content/p478cb9504908a.php>



IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

Real Data Analysis

Edited by **Shlomo S. Sawilowsky**, *Wayne State University*

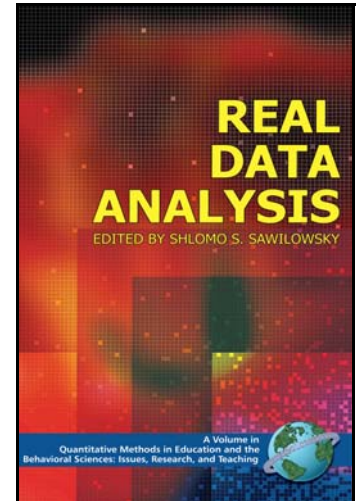
A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)

The invited authors of this edited volume have been prolific in the arena of Real Data Analysis (RDA) as it applies to the social and behavioral sciences, especially in the disciplines of education and psychology. Combined, this brain trust represents 3,247 articles in refereed journals, 127 books published, US \$45.3 Million in extramural research funding, 34 teaching and 92 research awards, serve(d) as Editor/Assistant Editor/Editorial Board Member for 95 peer reviewed journals, and provide(d) ad hoc reviews for 362 journals. Their enormous footprint on real data analysis is showcased for professors, researchers, educators, administrators, and graduate students in the second text in the AERA/SIG ES Quantitative Methods series.

CONTENTS: Preface. *Shlomo S. Sawilowsky*. **PART I: FOUNDATIONS.** The Co-Evolution of Statistics and Hz, *Joseph M. Hilbe*. Effective Sample Size: A Crucial Concept, *Thomas R. Knapp*. Advances in Missing Data Methods and Implications for Educational Research, *Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, Lee H. Ehman*. Methods for Simulating Real World Data for the Psycho-Educational Sciences, *Todd Christopher Headrick*. How and Why I Use Real, Messy Data to Investigate Theory and Inform Decision Making, *Ted Micceri*. **PART II: STATISTICAL METHODS.** Using E-Mail Messages to Help Students Prepare for a Statistics Exam, *Schuyler Huck*. Randomization Tests: Statistical Tools for Assessing the Effects of Educational Interventions When Resources are Scarce, *Joel R. Levin*. A Skipped Multivariate Measure of Location: One- And Two-Sample Hypothesis Testing, *Rand R. Wilcox, H. J. Keselman*. Robust Step-Down Tests for Multivariate Group Differences, *Lisa M. Lix, Ian Clara, Aynslie Hinds, Charles Bernstein*. Dunn-Sidak Critical Values and *p* Values, *Roger E. Kirk, Joel Hetzer*. Controlling Experiment-wise Type I Errors: Good Advice for Simultaneous and Sequential Hypothesis Testing, *Shlomo S. Sawilowsky, Patric R. Spence*. Robustness and Power of Ordinal *d* for Paired Data, *Du Feng*. Factorial ANOVA in SPSS: Fixed-, Random-, and Mixed-Effects Models, *Richard G. Lomax, Stacy Hughey Surman*. ANOVA: Effect Sizes, Simulating Interaction vs. Main Effects, and a Modified ANOVA Table, *Shlomo S. Sawilowsky*. ANCOVA and Quasi-Experimental Design: The Legacy of Campbell and Stanley, *Shlomo S. Sawilowsky*. **PART III: MEASUREMENT:** Thinking About Item Response Theory from a Logistic Regression Perspective: A Focus on Polytomous Models, *Amery D. Wu, Bruno D. Zumbo*. Some Practical Uses of Item Response Time to Improve the Quality of Low-Stakes Achievement Test Data, *Steven L. Wise, Xiaojing Kong*. Using Moving Averages to Detect Exposed Test Items in Computer-Based Testing, *Ning Han, Ronald K. Hambleton*. An Empirical Calibration of the Effects of Multiple Sources of Measurement Error on Reliability Estimates for Individual Differences Measures, *Frank L. Schmidt, Huy Ahn Le*. Latent Structure of Attitudes toward Abortion, *C. Mitchell Dayton*. **PART IV: DATA ANALYSIS.** Hierarchical Linear Models and the Estimation of Students' Mathematics Achievement, *Kathrin A. Parks, Dudley L. Poston, Jr.* Grade Inflation: An Examination at the Institutional Level, *Sharon L. Weinberg*. Using Discrete-Time Survival Analysis to Study Gender Differences in Leaving Mathematics, *Suzanne E. Graham, Judith D. Singer*. Nonparametric procedures for testing for dropout rates on University courses with application to an Italian case study, *Rosa Arboretti Giancristofaro, Fortunato Pesarin, Luigi Salmaso, Aldo Solari*. Nonparametric Approaches for Multivariate Testing with Mixed Variables and for Ranking on Ordered Categorical Variables with an Application to the Evaluation of Ph. D. Programs, *Rosa Arboretti Giancristofaro, Fortunato Pesarin, Luigi Salmaso*. Randomized Replicated Single-case Experiments: Treatment of Pain-related Fear by Graded Exposure *In Vivo*, *Patrick Onghena, Johan W. S. Vlaeyen, Jeroen de Jong*. Whole Brain Correlations: Examining Similarity Across Conditions of Overall Patterns of Neural Activation in fMRI, *Arthur Aron, Susan Whitfield, Wemara Lichty*. Principal Component Analysis of Senate Voting Patterns. *Jan de Leeuw*



Publication Date:

2007

ISBN's:

Paperback: 978-1-59311-564-7

Hardcover: 978-1-59311-565-4

Price:

Paperback: \$39.99

Hardcover: \$73.99

Subject:

Education, Statistics

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 \$39.99 Hardcover ISBN: 978-1-59311-685-9 \$73.99

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.99

Series URL: <http://www.infoagepub.com/products/series/serlin.html>

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.
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 professionals are US \$95 per year; for graduate students are US \$47.50 per year; and for libraries, universities, and corporations are 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.

The easy way to find open access journals

DOAJ DIRECTORY OF
OPEN ACCESS
JOURNALS

www.doaj.org

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

Aims

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

Scope

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

In DOAJ browse by subject

Agriculture and Food Sciences
Biology and Life Sciences
Chemistry
General Works
History and Archaeology
Law and Political Science
Philosophy and Religion
Social Sciences

Arts and Architecture
Business and Economics
Earth and Environmental Sciences
Health Sciences
Languages and Literatures
Mathematics and statistics
Physics and Astronomy
Technology and Engineering

Contact

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

Funded by



www.soros.org

Hosted by



LUND
UNIVERSITY
www.lu.se