Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 27

5-1-2006

Vol. 5, No. 1 (Full Issue)

JMASM Editors

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2-282

1538 – 9472/06/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2-282

1538 – 9472/06/$95.00

Editorial Board

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2-282

1538 – 9472/06/$95.00

Journal Of Modern Applied Statistical Methods

Invited Articles

2 – 13	**James Algina**, **H. J. Keselman**, **Randal D. Penfield**	Confidence Intervals for an Effect Size When Variances Are Not Equal
14 – 21	**Rand Wilcox**	ANCOVA: A Robust Omnibus Test Based on Selected Design Points

Regular Articles

22 – 40	**Bruce R. Fay**	The Effect on Type I Error and Power of Various Methods of Resolving Ties for Six Distribution-Free Tests of Location
41 – 43	**Phillip I. Good**, **Cliff Lunneborg***	Limitations of the Analysis of Variance
44 – 65	**Rhonda Kowalchuk**, **H. J. Keselman**, **Rand R. Wilcox**, **James Algina**	Multiple Comparison Procedures, Trimmed Means, and Transformed Statistics
66 – 72	**M. Masoom Ali**, **J. S. Cho**, **Munni Begum**	Nonparametric Bayesian Multiple Comparisons for Dependence Parameter in Bivariate Exponential Populations
73 – 94	**Joel I. Deichmann**, **Abdolreza Eshghi**, **Dominique Haughton**, **Selin Sayek**, **Nicholas Teebagy***, **Heikki Topi**	Understanding Eurasian Convergence: Application of Kohonen Self-Organizing Maps
95 – 106	**Stan Lipovetsky**	Entropy Criterion in Logistic Regression and Shapley Value of Predictors
107 – 117	**Samir Safi**, **Alexander White**	The Efficiency of OLS in the Presence of Auto-Correlated Disturbances in Regression Models
118 – 129	**Terry Dielman**	Choosing Smoothing Parameters for Exponential Smoothing: Minimizing Sums of Squared versus Sums of Absolute Errors

*****_Posthumously_

130 – 139 **Hua Liang** Penalized Splines for Longitudinal Data with an Application in AIDS Studies

140 – 151 **V. Karioti**, **C. Caroni** Properties of the GAR(1) Model for Time Series of Counts

152 – 170 **Debasis Kundu**, **Avijit Joarder** Analysis of Type-II Progressively Hybrid Censored Competing Risks Data

171 – 180 **Ayman Baklizi**, **Mohammed Ebrahem** Comparison of Some Simple Estimators of the Lognormal Parameters Based on Censored Samples

181 – 193 **Edward J. Mascha**, **Jeffrey M. Albert** Properties of Bound Estimators on Treatment Effect Heterogeneity for Binary Outcomes

194 – 216 **Matthew E. Elam** Two New Unbiased Point Estimates of a Population Variance

217 – 229 **Shenghai Zhang**, **Mary E. Thompson** Variance Estimation and Construction of Confidence Intervals for GEE Estimator

230 – 247 **Show-Mann Liou,** **Joanne Peng** The Use of Hierarchical ANCOVA in Curriculum Studies

248 – 257 **Michael B.C. Khoo,** **S. H. Quah**, **C. K. Ch'ng** A Combined Individuals and Moving Range Control Chart

Early Scholars

258 – 265 **K. Thangavel**, **D. Ashok Kumar** A Combined Standard Deviation Based Data Clustering Algorithm

Algorithms and Code

266 – 272 **Debasis Kundu**, **Rameshwar Gupta**, **Anubhav Manglick** JMASM22: A Convenient Way of Generating Normal Random Variables Using Generalized Exponential Distribution

273 – 280 **Andrés M. Alonso** JMASM23: Cluster Analysis in Epidemiological Data (Matlab)

Errata

281 **Juliet Popper Shaffer** Confidence Intervals on Subsets May be Misleading

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2-282

1538 – 9472/06/$95.00

End Matter

282 – 283 **Author** Statistical Pronouncements V

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

Three areas are appropriate for *JMASM*: (1) development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) development or study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods. Problems may arise from applied statistics and data analysis; experimental and nonexperimental research design; psychometry, testing, and measurement; and quantitative or qualitative evaluation. They should relate to the social and behavioral sciences, especially education and psychology.

Editorial Assistant: **John Cuzzocrea**

Production Staff: **Christina Gase, Jack Sawilowsky**

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

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2- 13

1538 – 9472/06/$95.00

Invited Articles

Confidence Intervals For An Effect Size When Variances Are Not Equal

James Algina
University of Florida

H. J. Keselman
University of Manitoba

Randall D. Penfield
University of Miami

Confidence intervals must be robust in having nominal and actual probability coverage in close agreement. This article examined two ways of computing an effect size in a two-group problem: (a) the classic approach which divides the mean difference by a single standard deviation and (b) a variant of a method which replaces least squares values with robust trimmed means and a Winsorized variance. Confidence intervals were determined with theoretical and bootstrap critical values. Only the method that used robust estimators and a bootstrap critical value provided generally accurate probability coverage under conditions of nonnormality and variance heterogeneity in balanced as well as unbalanced designs.

Key words: Effect size, confidence interval, trimmed means, Winsorized variance, noncentral distribution

Introduction

Estimating effect size (ES) and setting intervals for such estimates has become a requirement in many scientific journals as a result of the American Psychological Association's (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999). Indeed, according to Thompson (2003, personal communication) at least 23 journals require authors to follow the recommendation put forth by the task force.

Not surprisingly, there has been a renewed interest in ES estimates and accompanying confidence intervals (CIs). See, for example, Algina and Keselman (2003), Bird (2002), Cumming and Finch (2001), and Steiger and Fouladi (1997).

Glass (1976) used a control group standard deviation (in a two-group problem) to standardize the difference between the group means. However, other values have been used to standardize the mean difference. For example, Hedges (1981) used the square root of the pooled variance, which is referred to as the pooled standard deviation. If the variance equality assumption is not met, then the standard deviation for either one of the groups could be used as the standardizer. In the context of comparing an experimental and control treatment, Glass, McGaw, and Smith (1981) recommended using the standard deviation for the control group, but pointed out that the experimental group standard deviation could be used. Glass et al. (1981) presented an example demonstrating that the value of the ES estimate

James Algina (algina@ufl.edu) is Professor of Educational Psychology. His research interests are in applied statistics and psychometrics. H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology. His research interests are in applied statistics. Randall D. Penfield (penfield@miami.edu) is Assistant Professor of Education. His research interests are in educational measurement and psychometrics.

can vary depending on which group's standard deviation is used as the standardizer. As well, they point out that both ES estimates would be correct. As Glass et al. (1981) noted, "These facts are not contradictory; they are two distinct features of a finding which cannot be expressed by one number" (p 107).

Thus, Olejnik and Algina (2000) noted that when the equality of variance assumption is violated, the researcher will have to select one standard deviation that expresses the contrast (i.e., the effect) on the scale the researcher imagines is most important, or will have to report the mean difference standardized by several standard deviations and discuss the implications of these ESs. Before turning to methods that can be used when variances appear to be heterogeneous, it is important to point out that heterogeneity of variance can occur due to some additional factor in the data that is not modeled in the analysis. It is better to model such factors than to uncritically use methods that are appropriate for heterogeneous variances.

When the population variances are assumed to be equal for the two levels of the factor, the population ES (PES) is

$$\delta_{Pooled} = \frac{\mu_2 - \mu_1}{\sigma}$$

where μ_j is the population mean for level j and σ is the population standard deviation, which is assumed to be equal for the two levels of the factor. The PES can be estimated by

$$\hat{\delta}_{Pooled} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{Pooled}}$$

where $\bar{Y}_j$ $(j = 1, 2)$ is a treatment level group mean, n_j $(n_1 + n_2 = N)$ is the sample size for the jth group, and S_{Pooled} is the pooled standard deviation.

According to Steiger and Fouladi (1997), a CI for the PES, which is exact under the assumptions for the independent samples t test, can be derived by using the noncentral t distribution with N – 2 degrees of freedom. First, a CI for the noncentrality parameter

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\mu_2 - \mu_1}{\sigma}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\delta_{Pooled}$$

is obtained. Then, by multiplying the limits of the interval for λ by the inverse of

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

a CI for δ_{Pooled} is obtained. The lower limit of the CI for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\bar{Y}_2 - \bar{Y}_1}{S_{Pooled}}\right)$$

is the $1-\alpha/2$ quantile. For example, if $t = 2.131$ and $N - 2 = 15$, the lower limit of the 95% CI for λ is zero, because 2.131 is the .975 quantile of the t distribution with a noncentrality parameter equal to zero. The upper limit of the $100(1-\alpha/2)\%$ interval for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic is the $\alpha/2$ quantile of the distribution (See Steiger & Fouladi, 1997).

The PES based on the standard deviation for the jth group is

$$\delta_j = \frac{\mu_2 - \mu_1}{\sigma_j}$$

and can be estimated by

$$\hat{\delta}_j = \frac{\bar{Y}_2 - \bar{Y}_1}{S_j}$$

where S_j is the square root of the usual unbiased sample variance. With this ES, the noncentral t-based interval for δ is no longer correct. However, under the assumptions that the data in each group are normally distributed and all data are distributed independently, a noncentral t-based approximate CI for δ_{j} can be derived. Thus, the CI does not assume equal variances, but the interval is based on normal distribution theory. This normality assumption is likely to be problematic because $\bar{Y}_2 - \bar{Y}_1$ and S_j are not distributed independently when the distribution is skewed for the *j*th treatment. For example, if the distribution is positively skewed for the first treatment, the sampling correlation between $\bar{Y}_2 - \bar{Y}_1$ and S_1 will be negative.

Therefore, large values for $\bar{Y}_2 - \bar{Y}_1$ will tend to be associated with small values for S_1 and $\hat{\delta}_1$ will tend to be positively biased. Moreover, the distribution theory used in deriving the CI will no longer apply. As a result the CI may not have the correct probability coverage. In fact, in an investigation of CIs for ESs in *dependent* samples designs, Algina, Keselman, and Penfield (2005a) showed that nonnormality has a negative impact on coverage probability for a noncentral t based approximate CI for δ_j.

Purposes of this article

Therefore, one purpose of the research was to investigate coverage probability for the noncentral t-based CI for δ_{j} when data are sampled in an *independent* samples design from a nonnormal distribution. Considering the prediction that the noncentral t-based CI for δ_j is likely to be negatively impacted by nonnormality, a second purpose of the article was to investigate alternatives to the interval.

One reasonable alternative is to use the percentile bootstrap to construct a CI for δ_j. A second alternative is to replace the least squares estimates in $\hat{\delta}_j$ with robust estimates. This approach was recommended by Algina et al. (2005a) in the context of CIs for δ_j in repeated measures designs and by Algina, Keselman, and Penfield (2005b) in the context of CIs for δ in independent samples and is consistent with the observation in Wilcox and Keselman (2003) that the common population definition and sample estimate of ES (i.e., δ_{Pooled} and $\hat{\delta}_{Pooled}$ or δ_j and $\hat{\delta}_j$ for the two-group problem), based on least squares estimators, are not robust to distribution shape. That is, skewed distributions and distributions containing outliers can cause the PES value and its estimate to be grossly misleading (Wilcox, 2003, Sec 8.11). Accordingly, in place of $\hat{\delta}_j$, the following is used

$$\hat{\delta}_{R_j} = .642\left(\frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{S_{W_j}}\right) \tag{1}$$

where $\bar{Y}_{tj}$ is the 20% trimmed mean for the *j*th group $(j = 1, 2)$ and $S^2_{W_j}$ is the 20% Winsorized variance for group *j*. Twenty percent refers to the percentage trimmed from each tail. The constant .642 is the population value for the Winsorized standard deviation for a standard normal distribution for 20% trimming. (See Wilcox, 2003, for a justification of 20% trimming and computational definitions of the trimmed mean and Winsorized variance). For a normal distribution, both $\hat{\delta}_{R_j}$ and $\hat{\delta}_j$ converge to δ_j as the sample sizes increase. Probability coverage for a noncentral t-based CI and for a percentile bootstrap CI for δ_{R_j} was investigated (defined later in equation (2)).

A Noncentral t-Based CI for δ_j

If the variances are unequal, in a two-group independent samples design, the population and sample ES is defined as

$$\delta_1 = \frac{\mu_2 - \mu_1}{\sigma_1}$$

and

$$\hat{\delta}_1 = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1},$$

respectively. (The standard deviation for the second group could also be used. Glass et al. (1981) pointed out that these ESs provide different information.)

It is well known that if $U \sim N(\mu,1)$, $V \sim \chi^2(k)$, and U and V are independently distributed, then

$$\frac{U}{\sqrt{\frac{V}{k}}} \sim t(k,\mu)$$

where $t(k,\mu)$ is the noncentral t distribution with degrees of freedom k and noncentrality parameter μ. Using this result with

$$U = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and

$$V = \frac{(n_1 - 1) S_i^2}{\sigma_1^2}$$

then

$$\frac{\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{S_1^2}{\sigma_1^2}}} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1\sqrt{\frac{1}{n_1} + \frac{\sigma_2^2}{n_2\sigma_1^2}}} \sim t(n_1 - 1, \lambda)$$

where

$$\lambda = \frac{\mu_2 - \mu_1}{\sigma_1\sqrt{\frac{1}{n_1} + \frac{\sigma_2^2}{n_2\sigma_1^2}}}.$$

If the estimate of λ is calculated as

$$\hat{\lambda} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}} = \frac{\hat{\delta}_1}{\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}}$$

the noncentral t distribution, with $n_1 - 1$ degrees of freedom, can be used to find a CI on λ. Specifically, the upper limit of a $100(1-\alpha)\%$ interval for λ is the noncentrality parameter for the noncentral t distribution with $n_1 - 1$ degrees of freedom in which $\hat{\lambda}$ is the $\alpha/2$ quantile of the distribution; the lower limit is the noncentrality parameter for the noncentral t distribution in which $\hat{\lambda}$ is the $(1-\alpha/2)$ quantile. Then, multiplying the lower and upper limit by $\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}$, an approximate CI for δ_1 is obtained. The interval is approximate because the limits of the CI for λ are multiplied by a random variable.

To obtain an estimate of the robust ES, let $[.2n_j]$ indicate that $.2n_j$ is rounded down to the nearest integer, $g_j = [.2n_j]$, $h_j = n_j - 2g_j$, and then let

$$\tilde{S}_j^2 = \frac{(n_j - 1) S_{W_j}^2}{h_j - 1}$$

and

$$\tilde{\sigma}_j^2 = \frac{(n_j - 1)\sigma_{W_j}^2}{h_j - 1}$$

where $\sigma_{W_j}^2$ is the population Winsorized variance for treatment j. To obtain a CI for

$$\delta_{R_1} = .642\left(\frac{\mu_{t2} - \mu_{t1}}{\sigma_{w_1}}\right) \qquad (2)$$

define

$$\lambda_R = \frac{\mu_{t2} - \mu_{t1}}{\tilde{\sigma}_1\sqrt{\frac{1}{h_1} + \frac{\tilde{\sigma}_2^2}{h_2\tilde{\sigma}_1^2}}} = \frac{\delta_{R_1}}{.642\sqrt{\frac{n_1 - 1}{h_1 - 1}\left(\frac{1}{h_1} + \frac{\tilde{\sigma}_2^2}{h_2\tilde{\sigma}_1^2}\right)}} \qquad (3)$$

where μ_{tj} is the population trimmed mean. Also define

$$\hat{\lambda}_R = \frac{\bar{Y}_{t2}-\bar{Y}_{t1}}{\tilde{S}_1\sqrt{\frac{1}{h_1}+\frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}}} = \frac{\hat{\delta}_{R_1}}{.642\sqrt{\frac{n_1-1}{h_1-1}\left(\frac{1}{h_1}+\frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}\right)}} . \quad (4)$$

The upper limit of a $100(1-\alpha)\%$ interval for λ_R is the noncentrality parameter for the noncentral *t* distribution, with h_1-1 degrees of freedom, in which $\hat{\lambda}_R$ is the $\alpha/2$ quantile of the distribution; the lower limit is the noncentrality parameter for the noncentral *t* distribution in which $\hat{\lambda}_R$ is the $(1-\alpha/2)$ quantile. An approximate CI for δ_{R_1} is obtained by multiplying the lower and upper limit by

$$.642\sqrt{\left(\frac{n_1-1}{h_1-1}\right)\left(\frac{1}{h_1}+\frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}\right)} .$$

The interval is approximate for two reasons. First, when trimmed means and Winsorized variances are used, there is no guarantee that the noncentral t distribution is the appropriate distribution for calculating a CI for λ_R. Second, the interval is approximate because the limits of the CI for λ_R are multiplied by a random variable.

The investigations of these intervals were carried out in three studies.

Study 1

Methodology

Probability coverage of CIs for δ_1 and δ_{R_1} based on the noncentral t distribution were investigated. It is important to recognize that δ_1 and δ_{R_1} are different parameters. When applied to normal distributions, the parameters will be equal, but otherwise will most likely be unequal. Thus, there is no attempt to compare the interval estimates of the δ_1 and δ_{R_1} .

Probability coverage was investigated for all combinations of the following three factors: $n_1 = n_2 = 20$ to 100 in steps of 20, PESs $\left(\delta_1 \text{ and } \delta_{R_1}\right)$ ranging from 0 to 1.6 in steps of .4, and population distribution (four cases from the family of g and h distributions). The nominal confidence level for all intervals was .95 and each condition was replicated 5000 times.

The data were generated from the g and h distribution (Hoaglin, 1985). Specifically, four g and h distributions were chosen for investigation: (a) $g = h = 0$, a standard normal distribution; (b) $g = .76$ and $h = -.098$, a distribution with skew and kurtosis equal to that for an exponential distribution $(\gamma_1 = 2,\ \gamma_2 = 6)$; (c) $\mathsf{g} = 0$ and $h = .225$, a long-tailed symmetric distribution $(\gamma_1 = 0 \text{ and } \gamma_2 = 154.84)$; and (d) $g = .225$ and $h = .225$, a long-tailed skewed distribution ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$). To generate data from a g and h distribution, standard unit normal variables Z_{ij} were converted to g and h distributed random variables via

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

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

$$Y_{ij} = Z_{ij}\exp\left(\frac{hZ_{ij}^2}{2}\right).$$

Z_{ij} scores were generated by using RANNOR in SAS (SAS, 1999). For simulees in treatment 2, the Y_{i2} scores were transformed to

$$\sqrt{PVR}\left(Y_{i2}-\mu_2\right)+\mu_2+\sigma_1\times\delta_1 \qquad (5)$$

where PVR is the ratio of the population variance for the transformed Y_{i2} scores to the variance of the Y_{i1} scores and was set equal to 4 for all conditions in Study 1. The scores generated by using equation (5) were used in the

CI for δ_1. Additional levels of PVR were planned for investigation. Because the results for $PVR = 4$ indicated poor probability coverage in some conditions and the focus should be to find intervals that work well in a wide variety of conditions, the intervals being estimated were dismissed.

To facilitate reporting of results for the CI for δ_{R_1}, the Y_{i2} scores were transformed to

$$\sqrt{PVR}\left(Y_{i2} - \mu_{t2}\right) + \mu_{t2} + \frac{\sigma_{W_1}}{.642}\delta_1 . \qquad (6)$$

This method of generating the scores in treatment 2 results in $\delta_1 = \delta_{R_1}$. The CI for δ_{R_1} was also investigated using equation (5) to generate Y_{i2} scores, $\delta_1 \neq \delta_{R_1}$. The general pattern of results was the same in the two sets of conditions.

Results

Estimated coverage probability for the two CIs are reported in Table 1 for the four g and h distributions, all sample size values, and all values of the PES (The CI for δ_{R_1} is based on Y_{i2} generated by using equation (6)). The results show that both CIs had estimated probability coverage near the nominal confidence level when the data were normally distributed $\left(g = h = 0\right)$, but both could have poor probability coverage when the data were nonnormal. As the PES increased, both CIs had increasingly worse coverage probability. Coverage probability appeared to be largely unaffected by sample size.

Study 2

Both noncentral *t*-based CIs had good coverage probability when the data were normal despite the fact that both CIs are only approximately correct. However, both could have poor coverage probability when the data were nonnormal. Therefore, the use of a percentile bootstrap CI to construct an interval on δ_1 was investigated.

Methodology

Probability coverage of a percentile bootstrap CI for all combinations of the following $n_1 = n_2 = 20$ to 100 in steps of 20, population distribution (four cases from the family of g and h distributions), and δ_1 ranging from 0 to 1.6 in steps of .4 was investigated. In all conditions, $PVR = 4$. The distributions from Study 1 were investigated and the data was generated by using the procedure described for Study 1. Because a CI for δ_1 was being investigated, the data for treatment 2 were generated by using Equation (5). As in Study 1, 5000 replications were conducted for each condition combination. 600 bootstrap replications were used. In all conditions, the nominal confidence level was .95.

Results

Estimated coverage probability for the bootstrap CI for δ_1 is reported in Table 2 for all sample size values and all levels of PES. The results show that the percentile CI for δ_1 can have poor coverage probability and therefore should not be used. These intervals were particularly poor when the sample size was small and δ_1 was large.

Study 3

The results indicate that each of the noncentral *t*-based and percentile bootstrap CIs for δ_1 and the noncentral t-based CI for δ_{R_1} can have poor coverage probability with nonnormal data. Therefore, coverage probability for a percentile bootstrap interval for δ_{R_1} was investigated.

Table 1. Estimated Coverage Probabilities for Noncentral t Distribution-Based CIs for δ_1 and δ_{R_1}

		$g = .000$, $h = .000$		$g = .000$, $h = .225$		$g = .760$, $h = -.098$		$g = .225$, $h = .225$	
		δ_1	δ_{R_1}	δ_1	δ_{R_1}	δ_1	δ_{R_1}	δ_1	δ_{R_1}
0.00	20	.954	.955	.954	.954	.943	.949	.956	.962
	40	.959	.955	.955	.954	.948	.951	.957	.957
	60	.954	.957	.956	.955	.947	.950	.954	.958
	80	.953	.954	.952	.948	.949	.953	.951	.953
	100	.954	.951	.955	.952	.948	.948	.952	.949
0.40	20	.948	.950	.955	.955	.924	.932	.940	.954
	40	.955	.952	.949	.951	.920	.925	.932	.952
	60	.957	.953	.943	.951	.928	.928	.931	.943
	80	.945	.943	.937	.952	.930	.932	.921	.948
	100	.948	.946	.937	.953	.920	.926	.918	.944
0.80	20	.949	.949	.936	.948	.900	.913	.906	.937
	40	.948	.947	.927	.948	.894	.907	.891	.927
	60	.952	.951	.919	.949	.895	.911	.874	.933
	80	.949	.943	.915	.951	.895	.915	.872	.931
	100	.953	.948	.913	.948	.893	.902	.859	.934
1.20	20	.951	.943	.914	.940	.871	.890	.876	.925
	40	.953	.943	.893	.941	.867	.892	.843	.925
	60	.953	.948	.885	.940	.858	.894	.825	.922
	80	.950	.939	.877	.938	.859	.887	.809	.920
	100	.946	.940	.871	.933	.858	.886	.799	.914
1.60	20	.956	.949	.883	.931	.836	.866	.837	.915
	40	.948	.941	.862	.920	.836	.872	.802	.911
	60	.953	.945	.843	.932	.831	.875	.773	.909
	80	.948	.939	.836	.933	.823	.860	.764	.915
	100	.947	.941	.834	.928	.830	.865	.749	.917

Note: $PVR = 4$.

Table 2. Estimated Coverage Probabilities for the Bootstrap Percentile CI for δ_1

δ_1	$n_1 = n_2$	$g = .000$,	$g = .000$,	$g = .760$,	$g = .225,$
		$h = .000$	$h = .225$	$h = -.098$	$h = .225$
0.0	20	.936	.929	.920	.921
	40	.942	.937	.939	.935
	60	.939	.935	.935	.938
	80	.948	.946	.935	.940
	100	.945	.939	.940	.941
0.4	20	.934	.922	.926	.915
	40	.939	.929	.930	.928
	60	.942	.935	.937	.932
	80	.950	.941	.940	.933
	100	.948	.936	.947	.931
0.8	20	.931	.904	.915	.900
	40	.934	.921	.928	.904
	60	.943	.921	.933	.916
	80	.945	.933	.940	.907
	100	.944	.929	.938	.916
1.2	20	.929	.882	.905	.862
	40	.937	.901	.922	.874
	60	.943	.905	.925	.884
	80	.938	.918	.930	.880
	100	.949	.913	.934	.892
1.6	20	.926	.861	.883	.824
	40	.940	.881	.911	.838
	60	.945	.889	.908	.850
	80	.943	.895	.927	.850
	100	.942	.893	.927	.848

Note: $PVR = 4$

Methodology

Probability coverage was investigated for all combinations of: sample size n_1 = 20, 40, and 60 in combination with $n_2 = n_1$ and $n_2 = n_1 + 20$; population distribution (four cases from the family of g and h distributions), various PESs, $\delta_{R_1} = .00$, .40, .80, 1.20 and 1.60, and $PVR = .25$, .5, 1, 4, and 8. As in Study 2, $g = h = 0$, $g = .76$ and $h = -.098$, $g = 0$ and $h = .225$, and $g = .225$ and $h = .225$ were investigated. Because a CI for δ_{R_1} was being investigated, the data for treatment 2 were generated by using Equation (6). In all conditions the nominal confidence level was .95. As in the previous study, 5,000 replications and 600 bootstrap replications were used.

Results

Table 3 contains estimated coverage probabilities for the percentile bootstrap CI for all conditions with $PVR = 8$. Estimated coverage

Table 3. Estimated Coverage Probabilities for the Percentile Bootstrap CI for δ_{R_1}

n_1, n_2	δ_{R_1}	$g = .000$, $h = .000$	$g = .000$, $h = .225$	$g = .760$, $h = -.098$	$g = .225$, $h = .225$
20, 20	.00	.943	.945	.945	.950
	.40	.950	.956	.954	.951
	.80	.948	.955	.952	.954
	1.20	.961	.964	.957	.966
	1.60	.960	.966	.962	.960
20, 40	.00	.949	.957	.949	.952
	.40	.951	.954	.956	.958
	.80	.953	.959	.951	.961
	1.20	.967	.964	.958	.965
	1.60	.959	.969	.957	.963
60, 60	.00	.949	.947	.947	.948
	.40	.953	.944	.943	.952
	.80	.949	.950	.948	.957
	1.20	.952	.951	.952	.949
	1.60	.947	.959	.954	.958
60 80	.00	.945	.952	.944	.950
	.40	.952	.949	.946	.951
	.80	.949	.959	.951	.959
	1.20	.955	.954	.953	.956
	1.60	.955	.961	.954	.953
100,100	.00	.950	.948	.949	.947
	.40	.947	.948	.953	.951
	.80	.950	.946	.949	.957
	1.20	.951	.953	.951	.952
	1.60	.953	.956	.953	.956
100,120	.00	.948	.955	.947	.948
	.40	.939	.951	.948	.948
	.80	.955	.949	.950	.948
	1.20	.951	.947	.955	.955
	1.60	.956	.960	.959	.959

Note. $PVR = 8$.

probabilities for other values of PVR were not noticeably different from those in Table 3. Over the 120 conditions reported in Table 3, empirical coverage ranged from .939 to .969, with an average coverage value of .953. The results suggest coverage probability increased as δ_{R_1} increased, but was largely unaffected by the sampled distribution and whether the sample sizes were equal.

Conclusion

Estimating the magnitude of a treatment effect has become a required mode of analysis for many scientific journals in the social and behavioral sciences as a result of recommendations made by the APA Task Force regarding statistical inference. Not surprisingly, issues related to estimating the magnitude of an effect have become of paramount interest to applied researchers. One issue is what standard deviation to use in the denominator of the ES statistic. That is, since Glass's (1976), which used the control group's standard deviation to standardize the mean difference, other approaches have been recommended. Hedges (1981) recommended using the pooled standard deviation when the variances are homogeneous. Glass et al. (1981) recognized that if homogeneity of variances is not a reasonable assumption, the standard deviation for either group could be used as the denominator. This applies regardless of whether one of the treatment groups is a control group.

A second issue is how to use the ES measures to construct a CI. It is well known that when the pooled standard deviation is used in the denominator, CIs can be constructed by using the noncentral t distribution and will be exact when the scores are independently drawn from normal distributions and with equal variances. As shown in this article, an alternative interval based on the noncentral t distribution can be used when the standard deviation for one of the groups is used in the denominator, as would be done if Glass's (1976) ES were used or if the recommendation of Glass et al. (1981) were used when the variances are not homogeneous. However, the theory underlying this interval assumes data that are normal in form, which implies that the numerator and denominator of the ES are independently distributed. Independence does not hold when the data for the group that contributes the standard deviation are skewed. Accordingly, the interval could not be recommended without first examining its operating characteristics under nonnormality

As Wilcox and Keselman (2003) indicated, ES measures can be inaccurate when the data are drawn from nonnormal distributions because of the effects of nonnormality on means and standard deviations. Therefore, CIs calculated from a robust effect size $\left(\hat{\delta}_{R_1}\right)$ in which trimmed means replace means and the square root of the Winsorized variance replaces the standard deviation were also investigated. An additional issue considered was whether one could obtain accurate probability coverage for CIs for ES when coverage was based on theoretically obtained critical values (i.e., based on the noncentral *t* distribution) or obtained through a bootstrapping method. This was an important issue because others have demonstrated the benefits of using bootstrapping methodology (See, e.g., Keselman et al., 2002).

It this article, it was found that: (1) the classical approach, which divides the mean difference by a standard deviation from one group $\left(\text{i.e., } \hat{\delta}_1\right)$ in combination with the interval based on the noncentral *t* distribution had poor probability coverage when data were skewed, (2) the robust approach, which divides the difference of the trimmed means by the square root of the Winsorized variance from one group $\left(\text{i.e., } \hat{\delta}_{R_1}\right)$ in combination with the interval based on the noncentral *t* distribution also had poor probability coverage when data were nonnormal, (3) bootstrap CIs for δ_1 can perform poorly, and (4) the percentile bootstrap interval for δ_{R_1} was very little affected by nonnormality, providing a very good interval for δ_{R_1}.

An emphasis must be placed on the belief that it is important to estimate a robust parameter, that is, the robust PES, rather than the usual parameter of ES, when data are nonnormal. Researchers should be interested in estimates of a parameter that is robust to conditions of skewness and outlying values. Inferences pertaining to robust parameters may be more valid than inferences pertaining to the least squares derived parameters when dealing with populations that are nonnormal (e.g., Hample, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990). Hogg (1974, p. 919) maintained that most distributions are skewed in practice, and Tukey (1960) argued that most distributions will have heavy tails. Therefore, according to this perspective, the justification for (testing hypotheses and) setting robust intervals for robust parameters is that (testing the usual hypotheses and) setting intervals around the usual parameters is a mistake or at least shortsighted when other robust methods are available, methods that are not generally affected by a relatively few data points in a distribution or some minor characteristic of the distribution, points and characteristics that need not affect the quantity researchers are interested in.

As well, it was found that the natural sample estimate of the robust parameter, one based on trimmed means and a Winsorized variance, provides probability coverage that is fairly close to the target value of .95, when upper and lower critical values for the interval were obtained through a percentile bootstrap method. Despite the preference for a robust parameter, others may feel that, given a hypothesis about the least square means (which is not recommended with nonnormal data), δ is the appropriate effect size measure. These researchers must face the fact that neither the noncentral *t* distribution-based CI nor the percentile bootstrap CI will necessarily have coverage probability near the nominal value.

References

Algina, J. & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, *63*, 537-553.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005a). Effect sizes and their intervals: The two-level repeated measures case. *Educational and Psychological Measurement*, *65*, 241-258.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005b). An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*, 317-328.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, *62*, 197-226.

Cumming G. & Finch S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hoaglin, D. C. (1983). Summarizing shape numerically: The g-and h distributions. In D. C. Hoaglin, F. Mosteller, & Tukey, J. W. (Eds.), *Data analysis for tables, trends, and shapes: Robust and exploratory techniques*. New York: Wiley.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, *69*, 909-927.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Keselman, H. J., Wilcox, R., R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and non normality. *Journal of Modern Applied Statistical Methods, 1(2)*, 288-309.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286.

SAS Institute Inc. (1999). *SAS/IML user's guide, version 8*. Cary, NC: Author.

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

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.) *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

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

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254-274.

Wilkinson, L. & the Task force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 14-21

ANCOVA: A Robust Omnibus Test Based On Selected Design Points

Rand R. Wilcox
Dept of Psychology
University of Southern California

Many robust analogs of the classic analysis of covariance method have been proposed. One approach, when comparing two independent groups, uses selected design points and then compares the groups at each design point using some robust method for comparing measures of location. So, if K design points are of interest, K tests are performed. There are rather obvious ways of performing, instead, an omnibus test that for all K points, no differences between the groups exist. One of the main results here is that several variations of these methods can perform very poorly in simulations. An alternative approach, based in part on the usual sample median, is suggested and found to perform reasonably well in simulations. It is noted that when using other robust measures of location, the method can be unsatisfactory.

Key words: ANCOVA, bootstrap methods, measures of depth, smoothers

Introduction

The analysis of covariance (ANCOVA) problem is to compare two independent groups based on some outcome of interest, Y , in a manner that takes into account some covariate, X. A classic and well-known approach assumes that the error term of the usual linear regression model is homoscedastic and has a normal distribution, the regression lines associated with each group are parallel, and the variances associated with the error terms for each group are assumed to be identical. More formally, if for the jth group *(j* = 1, 2), then there are n_j randomly sampled pairs of observations, say *(X_{ij}, Y_{ij}), i* = 1, . . . n_j, the classic assumption is that for the jth group,

$$Y_{ij} = \beta X_{ij} + \beta_{oj} + \varepsilon_{ij} \qquad (1)$$

where ε_{ij} has variance $\sigma_j^2, \sigma_1^2 = \sigma_2^2$, and ε_{ij} is independent of X_{ij}. So by implication, for each group, the conditional variance of Y , given X, does not vary with X, and each group has the same slope.

It is known that violating one or more of these assumptions can result in serious practical problems. Concerns about the robustness of the method date back to at least Atiqullah (1964) who concluded that non-normality is a practical problem. Another obvious concern is the assumption that the regression lines are parallel. There are several robust methods for testing this assumption (e.g., Wilcox, 2003, 2005), but it remains unclear when such tests have enough power to detect situations where having non-parallel lines is a practical concern. Yet another concern about equation (1) is the assumption that the association between Y and X is linear.

Rand R. Wilcox is Professor of Psychology at the University of Southern California, Los Angeles. Email: rwilcox@usc.edu.

Of course, in some situations this is a reasonable approximation, but this is not always the case. Many alternative methods have been derived that eliminate the assumption that the association is linear (e.g. Bowman & Young, 1996; Delgado, 1993; Dette & Neumeyer, 2001; Hall, Huber, & Speckman, 1997; Kulasekera, 1995; Kulasekera & Wang, 1997; Munk & Dette, 1998; Neumeyer & Dette, 2003; Young & Bowman, 1995; Wilcox, 2003). However, some of these methods require homoscedasticity and for most there are few if any simulation results that support their use with small to moderate sample sizes.

A simple and very flexible approach to ANCOVA is described in Wilcox (2003, section 14.8). It allows the regression lines to be non-linear, it allows heteroscedasticity, it performs well in simulations, and in the event standard assumptions are met, all indications are that it has nearly the same amount of power as the classic ANCOVA method (e.g., Wilcox, 2005, p. 526). Roughly, the method is based on multiple comparisons. Examination of the method suggests a simple and rather obvious approach to performing an omnibus test instead. But results reported here make it clear that several variations of this approach perform very poorly in simulations. (Details are given later in the article). The main result in this article is that an alternative approach, based in part on the usual sample median and the depth of the null vector in a bootstrap cloud, nearly eliminates this problem. The main exception is a situation where, simultaneously, the conditional distribution of Y is discrete, skewed, and the possible values for Y are relatively small in number.

Considered and Discarded Methods

It helps to describe the first general method that was considered and discarded and then suggest a related approach that gives more satisfactory results. It is assumed that for the jth group, Y and X are related through some unknown function, m_j. More formally, it is assumed that

$$Y_{ij} = m_j(X_{ij}) + \varepsilon_{ij}$$

where ε_{ij} has a median of zero, variance σ_{ij}^2, and is independent of X_{ij}. Let $m_j(x)$ be the population median of Y for the jth group, given that the covariate of the jth group is $X_j = x$. (Comments on using other location estimators are given later in the article). Let $x_1, ..., x_K$ be K values of X that are of interest. The method in Wilcox (2003, section 14.8) includes as a special case the problem of testing

$$H_0 : m_1(x_k) = m_2(x_k), k = 1, ..., K,$$

for each k. That is, K tests are to be performed. Let $\delta(x_k) = m_1(x_k) - m_2(x_k)$. The goal here is to test

$$H_0 : \delta(x_1) = \cdots = \delta(x_K) = 0$$

(2)

Here, it is assumed that K = 5 and that the choices for $x_1, ..., x_5$ are made empirically in a manner about to be described. Of course, it is not being suggested that other choices for the design points or K are inappropriate. For example, a researcher might have interest in K specific design points, rather than points determined as is done here. The idea is to provide a data-driven method for checking whether the regression lines differ, paying particular attention to design points where valid inferences about the medians of the Y values can be made.

The choice of the five design points stems in part from what is called a running interval smoother. To describe the details, attention is temporarily focused on a single group of subjects. The basic strategy is to find all X_i values close to x and estimate $m(x)$ with the median of the corresponding Y values. The method begins by computing the median absolute deviation statistic:

$$MAD = median\{| X_1 - M |, ..., | X_n - M |\},$$

where M is the usual sample median of the X values. Let MADN = MAD/.6745. The only reason for rescaling MAD is that under normality, MADN estimates σ. This rescaling helps describe the running interval smoother in terms of familiar concepts, but ultimately it is not important. Then X_i is said to be close to x if

$$| X_i - x | \le f \times MADN,$$

where f is some constant, called the span. Here, following Wilcox (2003), f = 1 is used. Let $\overline{m}_j = \Sigma m_j(x_k)/K$. A seemingly natural alternative to (2) is to test

$$H_0 : \overline{m}_1 = \overline{m}_2 \qquad (3)$$

That is, view the problem in the context of a 2 by K ANOVA and test the hypothesis that there is no main effect for the first factor. Many robust methods for testing this hypothesis have been proposed (Wilcox, 2005), which include various bootstrap techniques. But when checking the ability of this approach to control the probability of a Type I error for the problem at hand, poor results were obtained in situations described later in the article. Included were non-bootstrap methods for 20% trimmed means and medians (Wilcox, 2003, sections 10.3 & 10.5) plus bootstrap variations of these methods described in Wilcox (2005). In particular, it was found that in some situations, when testing at the .05 level, the actual Type I error probability was estimated to exceed .2.

Description of the Recommended Method

The one method that performed well in simulations is based on testing (2) rather than (3). The general strategy is to generate bootstrap samples, yielding bootstrap estimates of δ_k, and then determine how deeply the null vector is nested within this bootstrap cloud. Two approaches to measuring the depth of the null vector are considered. General theoretical results related to this approach are reported in Liu and Singh (1997).

To elaborate, momentarily assume that the xk values have been chosen and let Y_{ijk} $(i = 1..., n_{jk}; k = 1,..., K)$ be the Y_{ij} values such that

$$| X_{ij} - x_k | \le f \times MADN. \qquad (4)$$

For fixed k and j, generate a bootstrap sample by randomly sampling with replacement n_{jk} values from Y_{ijk} yielding $Y^*_{ijk}, (i = 1,..., n_{jk})$. Let M^*_{jk} be the usual sample median based on the Y^*_{ijk} values and let $\delta^*_k = M^*_{1k} - M^*_{2k}$. Repeat this process B times yielding $\delta^*_{bk}, b = 1,..., B$. So, there are B vectors of bootstrap δ^*_{bk} values, each vector having length K. Then roughly, the null hypothesis is rejected depending on how deeply the null vector $(0,...,0)$ is nested within this bootstrap cloud.

The problem of choosing the x_k values is approached as follows. Let $N_j(x)$ be the number of points in the jth group that are considered close to x based on (4). For notational convenience, assume that for fixed j, the X_{ij} values are in ascending order. That is, $X_{1j} \le \cdots \le X_{njJ}$. The regression lines are said to be comparable at x if simultaneously $N_j(x) \ge 12$ for both j = 1 and 2. The value 12 is chosen simply to reflect a sample of points large enough so as to expect reasonable control over the probability of a Type I error, but obviously some other (larger) value could be used if desired.

Suppose x_1 is taken to be the smallest X_{i1} value for which the regression lines are comparable. That is, search the first group for the smallest X_{i1} such that $N_1(X_{i1}) \ge 12$. If $N_2(X_{il}) \ge 12$, the two regression lines are considered comparable at X_{i1} and $x_1 = X_{i1}$ is set. If $N_2(x_{il}) < 12$, consider the next largest X_{i1} value and continue until it is simultaneously true that $N_1(X_{i1}) \ge 12$ and $N2(Xi1) \ge 12$. $K = 5$ is used, but again some other value is certainly reasonable. Let x_5 be the largest X_{i1} value in the first group for which the regression lines are comparable. That is, x_5 is the largest X_{i1} value such that $N_1(x_5) \geq 12$ and $N_2(x_5) \geq 12$. Let i_5 be the corresponding value of i. The other three design points are chosen as follows. Let $i_3 = (i_1 + i_5)/2$, $i_2 = i_1 + i_3/2$, and $i_4 = (i_3 + i_5)/2$. Round i_2, i_3, and i_4 down to the nearest integer and set $x_2 = X_{i_2 1}$, $x_3 = X_{i_3 1}$, and $x_4 = X_{i_4 1}$.

There are various ways of measuring how deeply a point is nested within a multivariate cloud of data (e.g., Liu & Singh, 1997, Wilcox, 2005). The simplest is based on Mahalanobis distances and is the first of the two methods considered here. However, the most obvious estimate of the covariance matrix associated with the bootstrap vectors is not used. Rather, it is estimated with

$$s_{km} = \frac{1}{B-1}\sum_{b=1}^{B}(\delta^*_{bk} - \delta_k)(\delta^*_{bm} - \delta_m).$$

That is, for fixed k, rather than use $\Sigma\delta^*_{bk}/B$ as the estimate of the center of the bootstrap cloud, use δ_k instead. Put another way, there is no need to estimate the center of the bootstrap cloud, it is already known and given by the vector $(\delta_1,\ldots,\delta_K)$. Indeed, if it is estimated with $\Sigma\delta^*_{bk}/B$, control over the probability of a Type I error deteriorates, consistent with a variety of other methods surveyed by Wilcox (2005). Let $S = (s_{km})$ be the corresponding covariance matrix, in which case the distance of the bth bootstrap vector from the center is given by

$$d_b = \sqrt{(\delta^*_{b1} - \delta_1,\ldots,\delta^*_{bK} - \delta_K)S^{-1}(\delta^*_{b1} - \delta_1,\ldots,\delta^*_{bK} - \delta_K)'}.$$

Let

$$D = \sqrt{(\delta_1 - 0,\ldots,\delta_K - 0)S^{-1}(\delta_1 - 0,\ldots,\delta_K - 0)'},$$

which is the distance of the null vector from the center of the bootstrap cloud. The (generalized) p-value is

$$\hat{p}^* = \frac{1}{B}\Sigma I(D \leq d_b),$$

where $I(D \leq d_b) = 1$ if $D \leq d_b$ and $I(D \leq d_b) = 0$ if $D > d_b$. This will be called method M.

The second method considered here for measuring the depth of a point in the bootstrap cloud is a projection-type method given in Wilcox (2005, section 6.2.5); it represents a slight variation of a method discussed by Donoho and Gasko (1992) and has been found to perform well in connection with other methods described in Wilcox (2005). The computational details are relegated to an appendix. This will be called method P.

A Simulation Study

Simulations were used to assess the small-sample properties of the method just described. Observations were generated according to the models

$$Y = \varepsilon$$

$$Y = X + \varepsilon$$

and

$$Y = X^2 + \varepsilon,$$

where X has a standard normal distribution and ε has one of four g-and-h distributions (Hoaglin, 1985), which contain the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$W = \begin{cases} \dfrac{\exp(gZ)-1}{g}\exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

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

g	h	k_1	k_2
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	1.75	8.9
0.2	0.2	2.81	155.99

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 a symmetric distribution with heavy tails ($g = h = 0.2$). In Table 1, the theoretical skewness and kurtosis for each distribution is considered. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

A general concern about methods aimed at comparing population medians, based on the usual sample median, is that for discrete data where tied values can occur, control over the probability of a Type I error can be poor. This is the case when using the method proposed by Bonett and Price (2002) as well as a related method in Wilcox (2003, section 8.7.1). In a paper submitted for publication, the author has found that certain bootstrap methods correct this problem while others do not. The main point here is that considering discrete distributions where tied values are likely is crucial for the problem at hand. Accordingly, additional simulations were run by generating ε from a beta-binomial distribution:

$$P(X = x) = \frac{B(m-x+r, x+s)}{(m+1)B(m-x+1, x+1)B(r,s)},$$

where B is the complete beta function. Here m = 10, 12 and 20 were considered. With m = 12, for example, the possible values for X are the integers $0,1,\ldots,12$. The values for r and s were taken to be $r = s = 4$, as well as $r = 1$ and $r = 9$. For $r = s = 4$ the distribution is bell-shaped and symmetric with mean $m/2$. In Figure 1, the probability function when $r = 1$, $s = 9$ and $m = 12$ is exhibited.

In Table 2, the estimated probability of a Type I error when testing at the .05 level and $n_1 = n_2 = 40$ is exhibited. The estimates are based on 1,000 replications with $B = 600$. (From Robey & Barcikowski, (1992), 1,000 replications is sufficient from a power point of view. More specifically, if the hypothesis that the actual Type I error rate is .05 is tested, and if power is to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required.) The results for $Y = X + \varepsilon$ did not reveal any new insights, and so for brevity they are not reported. To get some idea of the effect of homoscedasticity, additional simulations were run where values in the first group were multiplied by $\sigma_1 = 4$. The g-and-h distribution has a median of zero, so the null hypothesis remains true. For the beta-binomial distributions, the data were shifted to have a median of zero before multiplying by $\sigma_1 = 4$. The top portion of Table 2 are the results when there is homoscedasticity $(\sigma_1 = 1)$.

Figure 1: The beta-binomial probability function with $m = 12$, $r = 1$ and $s = 9$

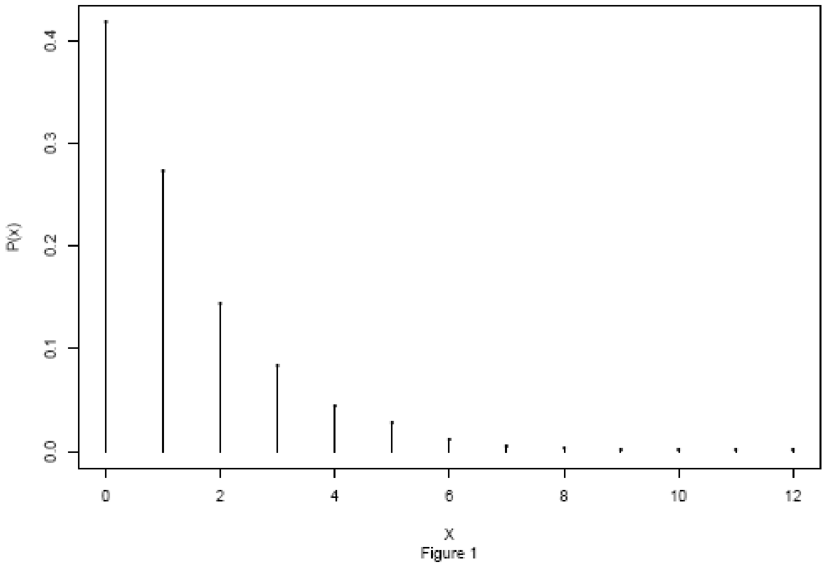


Table 2: Estimated Type I error probabilities

		$\sigma_1 = 1$			
ϵ		$Y = \epsilon$		$Y = X_1^2 + \epsilon$	
g	h	P	M	P	M
0.0	0.0	.064	.059	.052	.054
0.0	0.2	.039	.065	.047	.057
0.2	0.0	.061	.073	.050	.061
0.2	0.2	.041	.064	.048	.054
m	$r,\ s$				
12	1, 9	.464	.071	.053	.066
20	1, 9	.428	.027	.055	.066
10	4, 4	.152	.058	.063	.068
		$\sigma_1 = 4$			
g	h	P	M	P	M
0.0	0.0	.048	.088	.061	.089
0.0	0.2	.042	.076	.052	.076
0.2	0.0	.047	.089	.060	.087
0.2	0.2	.038	.077	.053	.076
m	$r,\ s$				
12	1, 9	.267	.142	.096	.146
20	1, 9	.133	.081	.077	.101
10	4, 4	.062	.055	.055	.073

First, consider the homoscedastic case with continuous g-and-h distributions. Both methods P and M perform reasonably well. To avoid an estimated Type I error probability greater than .07, method P is preferable. Under heteroscedasticity, method M can be unsatisfactory, with estimates exceeding .08, while again method P gives fairly satisfactory results. But when tied values occur, method P can be disastrous and should not be used. Method M now performs well under homoscedasticity $(\sigma_1 = 1)$, but under heteroscedasticity, it breaks down as well with estimates exceeding .1.

All simulations were repeated with $n_1 = n_2 = 60$, no new insights were found, so the results are not reported.

Conclusion

A positive result is that when tied values occur with probability zero, method P performs fairly well in terms of Type I errors, even when there is heteroscedasticity. However, when tied values are likely, it can be unsatisfactory. If tied values are likely and there is homoscedasticity, method M performs reasonably well, but it can break down when there is heteroscedasicity. So a possible argument in favor of method M is that when the (conditional) distributions of Y do not differ, it provides good control over the probability of a Type I error. But a negative feature is that it is sensitive to more than one feature of the data. That is, it does not isolate the reason for rejecting, which could be due to differences between medians or heteroscedasticity.

Some additional simulations were run with *m* = 2 0,*r* = 2 and *s* = 9. The ability of method P to control the probability of a Type I error improved substantially versus the situation where *r* = 1, but the estimated probability of a Type I error for the model $Y = \varepsilon$ was .099. So it seems that some tied values can probably be tolerated when using method P, but it is difficult to know when this is the case.

A criticism of the sample median is that under normality, or when sampling from a light-tailed distribution, it is relatively inefficient. By trimming less, say 20%, good efficiency is obtained under normality and some protection against low efficiency due to heavy-tailed distributions is obtained. (Note that the usual sample median belongs to the class of trimmed means with the maximum amount of trimming.) However, replacing the usual sample median with a 20% trimmed mean, the methods studied here are unsatisfactory in terms of estimated Type I errors, at least for the situations considered. Consideration was given to estimating the population median with the Harrell and Davis (1982) estimator with the goal of achieving better efficiency under normality, but again control over the probability of a Type I error was no longer satisfactory.

References

Atiqullah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika, 51,* 365–372.

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.

Bowman, A. & Young, S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics, 45,* 83–98.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters, 17,* 199–204.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics, 27,* 1012–1040.

Dette, H. & Neumeyer, N. (2001). Nonparametric analysis of covariance. *Annals of Statistics, 29,* 1361–1400.

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics, 20,* 1803–1827.

Hall, P., Huber, C., & Speckman, P. L. (1997). Covariate-matched one-sided tests for the difference between functional means. *Journal of the American Statistical Association, 92,* 1074–1083.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69,* 635–640.

Hoaglin, D. C. (1985). *Summarizing shape numerically: The g-and-h distribution*. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) Exploring Data Tables Trends and Shapes. New York: Wiley.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association, 90*, 1085–1093.

Kulasekera, K. B. & Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association, 92*, 500–511.

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

Munk, A. & Dette, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. *Annals of Statistics, 26*, 2339–2368.

Neumeyer, N. & Dette, H. (2003). Nonparametric comparison of regression curves: An empirical process approach. *Annals of Statistics, 31*, 880–920.

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.

Young, S. G. & Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics, 51*, 920–931.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques Testing*. San Diego CA: Academic Press.

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

Appendix

For notational convenience, projection distance is described in terms of a sample of *n* vectors from some multivariate distribution. The sample is denoted by $X_i, i = 1, ..., n$. Let ξ be some multivariate measure of location. Here, ξ is taken to be the W-estimator stemming from the minimum volume ellipsoid estimator. (For a detailed discussion of the minimum volume ellipsoid estimator, see Rousseeuw & Leroy, 1987). The outlier detection method in Rousseeuw and van Zomeren (1990) is applied, any points flagged as outliers are removed, and ξ is taken to be the mean of the remaining vectors. For any *i*, let

$$U_i = X_i - \xi ,$$

$$\begin{aligned} B_i &= U_i U_i' \\ &= \Sigma_{k=1}^{p} U_{ik}^2 \end{aligned}$$

and for any *j* let (j=1,…,n) let

$$W_{ij} = \sum_{k=1}^{p} U_{ik} U_{jk} ,$$

and

$$T_{ij} = \frac{W_{ij}}{B_i}(U_{i1}, ..., U_{ip}) \qquad (5)$$

The distance between ξ and the projection of X_j (when projecting onto the line connecting X_i and ξ) is

$$V_{ij} = \| T_{ij} \| ,$$

where $\| T_{ij} \|$ is the Euclidean norm associated with the vector T_{ij}. Let

$$d_{ij} = \frac{V_{ij}}{q_2 - q_1} , \qquad (6)$$

where for fixed i, q_2 and q_1 are estimates of the upper and lower quartiles, respectively, of the V_{ij} values. (Here, the ideal fourths based on the values $V_{i1}, ...V_{in}$ were used; see, for example, Wilcox, 2004.) The projection distance associated with X_j say D_j, is the maximum value of d_{ij}, the maximum being taken over $i = 1, ..., n$.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 22-40

REGULAR ARTICLES

The Effect On Type I Error And Power Of Various Methods Of Resolving Ties For Six Distribution-Free Tests Of Location

Bruce R. Fay
Wayne County Regional Educational Service Agency, Michigan

The impact on Type I error robustness and power for nine different methods of resolving ties was assessed for six distribution-free statistics with four empirical data sets using Monte Carlo techniques. These statistics share an underlying assumption of population continuity such that samples are assumed to have no equal data values (no zero difference–scores, no tied ranks). The best results across all tests and combinations of simulation parameters were obtained by randomly resolving ties, although there were exceptions. The method of dropping ties and reducing the sample size performed poorly.

Key words: Distribution-free, ties, location-shift, Monte Carlo, Rosenbaum's test, Tukey's quick test, Kolmogorov-Smirnov test, Wilcoxon rank-sum test, Kruskal-Wallis test, Terpstra-Jonckheere test.

Introduction

Distribution-free tests are important in the context of social and behavioral science research because they have less stringent assumptions than parametric statistics. Micceri (1986, 1989) showed that many variables studied in the social and behavioral sciences clearly do not meet distributional assumptions of parametric tests, such as normality or homoscedasticity.

In terms of hypotheses of a pure shift in location parameter combined with a violation of the normality assumption, nonparametric statistics are much more powerful than their parametric counterparts. In many layouts, these advantages are evident with very small samples and improve dramatically as sample sizes increase (Blair & Higgins, 1980, van den Brink & van den Brink, 1989, Sawilowsky, 1990, Sawilowsky & Blair, 1992, Kelley, Sawilowsky, & Blair, 1994, MacDonald, 1999).

Many distribution-free statistics lose efficiency when there is a violation of their underlying assumption of population continuity. In practice, this means the samples are assumed to have no equal data values (no zero difference–scores, no tied ranks), either within groups or between groups. Data in the social and behavioral sciences almost never meet this assumption either because of the inherently discrete nature of the data (Micceri, 1986, 1989) or because of a lack of precision in measurement (Cliff, 1996a, 1996b).

Sparks (1967) conducted one of the few empirical studies to have specifically examined violation of continuity. He investigated Student's *t*-test (Student, 1908) and the Wilcoxon Rank-sum (Mann-Whitney *U*) test (Wilcoxon, 1945, Mann & Whitney, 1947) using discrete approximations to the normal, rectangular, and exponential distributions. Results were similar for both Student's *t*-test and the Wilcoxon-Mann-Whitney test when ties were randomly resolved. The Wilcoxon-Mann-Whitney test, however, produced very conservative results when ties were resolved using mid-ranks.

The practical consequence of violating the assumption of population continuity is that samples will contain equal data values resulting

Dr. Fay is an Assessment Consultant in Wayne County, Michigan where he works with the state and local education agencies in the areas of school improvement, accountability, accreditation, and assessment. His research interests include the study of the properties of statistics through computer-intensive Monte Carlo methods using Fortran.

in zero difference–scores or tied ranks. A useful distinction can be made, however, between consequential (critical, meaningful) and inconsequential (non-critical) ties. Ties can occur in such a way that regardless of how they are resolved they have no effect on the calculation of the test statistic or the resulting inference. Such ties are clearly inconsequential. Ties that occur only within a group, when looking for between group effects, are often of this type. By definition, inconsequential ties may be resolved by any simple procedure that maintains the integrity of the ranks, such as arbitrary assignment in sequence of the set of ranks for which the group of scores is tied. Other ties occur in such a way that different resolutions result in different values of the statistic that may, in turn, result in different inferential decisions. Such ties are clearly consequential.

Purpose of the Study

Even though the less stringent underlying assumptions of distribution-free tests are rarely met in practice, the effects of violation of assumptions on robustness of Type I error rates and power have not been studied extensively. Given the potentially deleterious effects of ties on these tests, and the necessity of dealing with them in some way, a careful investigation of the impact of different methods of resolution is warranted. This is especially true given the subtle nature of robustness (Bradley, 1978, Wilcox, 1998). Therefore, nine methods were used, as applicable, to resolve consequential ties prior to the computation of six statistics.

Fahoome (1999, 2002) studied the Type I error properties of large-sample approximation formulas for twenty nonparametric and/or distribution-free statistics, including the six presented here, using the theoretical standard Normal distribution and four of the Micceri (1986) data sets. Ties, however, were either ignored or resolved in one specific way on a test-by-test basis. These same data sets served as pseudo-population models for the present study.

Tests

The following distribution-free tests were investigated:

1. Kolmogorov-Smirnov Test of General Differences for Two Independent Samples (Kolmogorov, 1933).

2. Rosenbaum's Test of Location for Two Independent Samples (Rosenbaum, 1953, 1954, 1965).

3. Tukey's Quick Test of Location for Two Independent Samples (Tukey, 1959).

4. Wilcoxon-Mann-Whitney Test for Two Independent Samples (Wilcoxon, 1945, Mann & Whitney, 1947, Kruskal, 1957).

5. Kruskal-Wallis Test for *k* Independent Samples (*k* = 3 to 6) (Kruskal, 1952, Kruskal & Wallis, 1952).

6. Terpstra-Jonckheere Test of an Ordered Alternative Hypothesis for *k* Independent Samples (*k* = 3 to 6) (Terpstra, 1952, Jonckheere, 1954).

Resolution of Ties

The nine methods for dealing with consequential ties (zero difference–scores or tied ranks) were:

1. (M-1) Resolve consequential ties in the manner least favorable to rejection of the null hypothesis and in the manner most favorable to rejection of the null hypothesis, calculate the statistic for each of these resolutions, and then calculate the mid-range (mean) value of these two statistics and use it to conduct the test.

2. (M-2) Count ties as 1/2 (Rosenbaum's Test and Tukey's Quick Test only).

3. (M-3) Alternately resolve each set of tied-for ranks.

4. (M-4) Randomly resolve each set of tied-for ranks.

5. (M-5) Delayed increment (Kolmogorov-Smirnov Test of General Differences only).

6. (M-6) Assign the mid-rank of a set of tied ranks to each score without further correction.

7. (M-7) Weighted average of all possible resolutions (Rosenbaum's Test only).

8. (M-8) Drop matching tied-for ranks and reduce *N* accordingly.

9. (M-9) Drop all tied-for ranks (if possible) and reduce *N* accordingly.

Methods 3, 4, 6, and 9 were described by Bradley (1968) as well as Gibbons and Chakraborti (1992). Methods 1, 2, 5, and 7 were described by Neave and Worthington (1988). Method 1 is related to a method described by Bradley (1968). Method 9 is widely mentioned in textbooks. Method 8 was not encountered in the literature but was added to the study as a variation of Method 9 that preserved equal sample sizes when dropping tied values.

Bradley (1968) also described methods involving calculation of statistics for all possible resolutions of consequential ties, the results being used to establish probability bounds for the test or to calculate a mean probability. Although theoretically attractive, these methods are often impractical, requiring the calculation of very large numbers of statistics and/or the availability of the probabilities (see, however, Fay, 2002, for a discussion of methods for generating critical values and associated probabilities for some of these tests). For many tests, the calculation of an average statistic, based on all possible resolutions of ties, turns out to be equivalent to resolving each set of tied-for ranks using the mid-rank (Neave & Worthington, 1988). Bradley (1968) warned, however, that under some circumstances the use of mid-ranks might give a statistic something closer to its minimum or maximum value rather a median or mean value. This might account for the results in Sparks (1967).

Many of the methods involve schemes for eliminating ties, either by: (a) breaking them, that is, by somehow assigning the available ranks to the tied observations, or (b) dropping them. Other methods, such as mid-ranks, result in modified samples that still contain duplicate (and perhaps non-integer) ranks, even though this cannot happen when all assumptions of the test are met. Averaging the statistics from the least and most likely to reject resolutions can also result in non-integer values of statistics that are normally integer-valued. Such statistics were still referred to a standard table of critical values, for example, Neave (1981), as the performance when used in this manner was a major point of this study. The test/method combinations investigated are shown in Table 1.

Data Sets

A theoretical distribution and four empirical data sets were used as sources of samples. The theoretical standard Normal distribution ($\mu = 0$, $\sigma = 1$) did not produce samples with significant numbers of duplicate data values and thus served as a baseline for the performance of these tests under conditions meeting their underlying continuity assumption. The four empirical data sets, due to Micceri (1986), were (a) Extreme Asymmetric (EA), (b) Extreme Bi-modal (EB), (c) Multi-modal Lumpy (ML), and (d) Smooth Symmetric (SS).

The four Micceri (1986) data sets are inherently discrete and decidedly non-normal (see Appendix, Figures A1 through A4). They were also discussed in Micceri (1989), Sawilowsky, Blair and Micceri (1990), Sawilowsky and Blair (1992), and Fahoome (1999, 2002). With regard to the extreme bimodal data set, Fahoome (1999) concluded:

> [B]ecause of the small number (6) of data points, there were an extremely large number of ties, even for relatively small sample sizes. This data is Likert-type data. The performance by most tests was extremely poor. Most of the tests had inflated Type I error rates, some as high as 0.99999. A few had very low Type I error rates. (p. 462)

In spite of this finding, the extreme bimodal data set was retained for this study because of the widespread existence of such data. Properties of these data sets are given in Table 2.

Table 1. Tests and Applicable Methods of Resolving Ties

	Test					
Method	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-1[g]	X	X	X	X	X	X
M-2[h]	na	X	X	na	na	na
M-3[i]	X	X	X	X	X	X
M-4[j]	X	X	X	X	X	X
M-5[k]	X	na	na	na	na	na
M-6[l]	na	na	na	X	X	X
M-7[m]	na	X	na	na	na	na
M-8[n]	X	X	X	X	X	X
M-9[o]	X	X	X	X	X	X

Note: Cells marked ‘na’ indicate that the method does not apply to the test.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum’s Test, [c]TQ = Tukey’s Quick Test,
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test,
[f]T-J = Terpstra-Jonckheere Test, [g]M-1 = Average of least and most likely to reject,
[h]M-2 = Count ties as ½, [i]M-3 = Alternating, [j]M-4 = Random, [k]M-5 = Delayed Increment,
[l]M-6 = Mid-ranks, [m]M-7 = Weighted average, [n]M-8 = Drop matching, [o]M-9 = Drop all.

Table 2. Properties of Selected Micceri (1986,1989) Data Sets

	Parameter				
Data Set	μ[a]	ϕ[b]	σ[c]	γ_3[d]	γ_4[e]
Extreme Asymmetric	24.50	27.00	5.79	-1.33	4.11
Extreme Bi-modal	2.97	4.00	1.69	-0.08	1.30
Multi-modal Lumpy	21.15	18.00	1.90	0.19	1.80
Smooth Symmetric	13.19	13.00	4.91	0.01	2.66

Note: Excerpted from “A more realistic look at robustness and type II error properties of the t test to departures from population normality,” by S. S. Sawilowsky & R. C. Blair, 1992, *Psychological Bulletin*, *111*(2), 352-360, Table I, p. 353, copyright 1992 by Psychological Bulletin. Adapted with permission.
[a]μ = mean, [b]ϕ = median; [c]σ = variance, [d]γ_3 = skewness, [e]γ_4 = kurtosis.

Methodology

The simulations were programmed in Fortran 90/95. A main program was built for each of the six tests to conduct both the Type I error and power studies by controlling the combinations of simulation parameters and making calls to the appropriate modules. For each unique combination of distribution, sample size, number of groups (for *k*-sample tests only), and effect size (for power studies only), 1 million samples were drawn. For each sample one- and two-sided tests where conducted at both nominal alpha .01 and .05 for each applicable method of resolving ties (Table 1). Counts were maintained of significant and non-significant results, as well as un-testable trials, until the end of the simulation cycle when they were converted to proportions and written to output files.

Separate programs were written for each of the six tests to conduct the simulations for the drop ties and reduce N methods of resolving ties as these methods often led to tests on unequal sample sizes for which the test statistic could either not be computed or for which critical values were unavailable. This necessitated a modified approach to the simulations in which un-testable samples were discarded and additional samples were drawn until: (a) 10,000 testable samples were obtained, or (b) the program reached its 10,000,000th cycle, whichever came first.

All sample sizes from 3 to 30 [3(1)30] were examined, limited only by the availability of critical values. Because the method of dropping ties and reducing *N* often resulted in unequal sample sizes, this method was only studied for tests where tables of critical values for unequal sample sizes were available (Neave, 1981, Neave & Worthington, 1988) or could be generated (Fay, 2002). Power studies were conducted for equal initial per-group sample sizes of 3(3)30 if Type I error results were satisfactory and critical values were available.

One of the most widely suggested methods for dealing with (consequential) ties is to resolve them in all possible ways, obtaining a value of the statistic (or its associated probability) for each resolution. A mean value of the statistic is then obtained and tested, or a mean value of the probability established. This method was only implemented for Rosenbaum's test as there was a practical method for doing so. It was not otherwise used in this study because of the practical difficulties involved in implementing it for even moderate sample sizes when there are numerous ties at several different values. Also, comprehensive tables of exact probabilities are even more difficult to obtain than critical value tables for some of these tests.

Bradley (1978) recommended conservative bounds for robust Type I error of nominal alpha ± 10% and liberal bounds of nominal alpha ± 50%. Many distribution-free tests, however, cannot achieve nominal alpha at small sample sizes. The entries in critical value tables are typically best conservative values that may fall below Bradley's recommended 10% lower bound. As the main interest in the Type I error studies was the ability of each test to resist inflation of Type I error rate the conservative and liberal criteria were combined such that Type I error rates were considered acceptable if they fell in the range of $.5\alpha$ to 1.1α or were no more conservative than the results obtained when sampling from the standard Normal distribution.

The power of a test was of no interest if the Type I error rate was not robust to violations of assumptions. A priori, it was expected that those combinations of test conditions that produced Type I error rates well below nominal alpha would also have attenuated power.

For the power studies, a one-sided test was made in the direction of the simulated effect, while significant results in the wrong tail constituted Type III errors (MacDonald, 1999). Pure shift-effects of known size were simulated by shifting one or more of the groups relative to a base group. Nominal effect size multipliers of 0.2, 0.5, 0.8 and 1.2 were planned following Cohen (1988) and Sawilowsky and Blair (1992). Because of the necessity of generating integral shifts with the empirical data sets in order to obtain between-group ties, actual effect size multipliers for each empirical data set differed slightly from these targets, as shown in Table 3. The performance of the six tests with respect to the various methods of resolving ties, when used with such data, was of primary interest in this study.

Statistical Tests

All six tests share the assumptions of: (a) random and independent sampling of continuous populations, with sufficient precision of measurement to avoid tied observations (Bradley, 1968, Conover, 1999), (b) independence of sample observations both within and between groups (Hollander & Wolfe, 1999). All the tests have null hypotheses that assume all samples are drawn from identical populations. Assumptions about the populations under the alternative hypothesis differ for each test. The tests can be used successfully with discrete populations, but become approximate with the tabled critical values generally providing best conservative estimates.

Kolmogorov-Smirnov Test

Background. Neave and Worthington (1988) and Conover (1999) identified this as Smirnov's (1939) application of Kolmogorov's (1933) goodness-of-fit test. Everitt (1998) described it as "A distribution free method that tests for any difference between two population probability distributions. The test is based on the absolute maximum difference between the cumulative distribution functions of the samples from each population" (p. 179). The maximum distance referred to is the vertical distance between the cumulative probability distributions.

Hypotheses. The null hypothesis for the two-sided test is that the two sampled populations have identical distributions. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way. In the case of a one-sided test, the alternative hypothesis is that one population is stochastically greater than the other. Neave (1981) suggested that the test only be used in the two-sided situation, the Wilcoxon–Mann–Whitney test being more powerful for the directional hypothesis.

Table 3. Actual Shifts and Effect Sizes for Nominal Effect Sizes

	Nominal Effect Size			
Data Set (σ[a])	S[b](.2σ)	M[c](.5σ)	L[d](.8σ)	VL[e](1.2σ)
Extreme Asymmetric (5.79)				
NS[f]	1.158	2.895	4.632	6.948
AS[g]	1	3	5	7
AES[h]	0.173σ	0.518σ	0.864σ	1.209σ
Extreme Bi-modal (1.69)				
NS	0.338	0.845	1.352	2.028
AS	n/a	1	n/a	2
AES	n/a	0.592σ	n/a	1.183σ
Multi-modal Lumpy (11.90)				
NS	2.380	5.950	9.520	14.280
AS	2	6	10	14
AES	0.168σ	0.504σ	0.840σ	1.176σ
Smooth Symmetric (4.91)				
NS	0.982	2.455	3.982	5.892
AS	1	2	4	6
AES	0.204σ	0.407σ	0.815σ	1.222σ
Standard Normal (1.00)				
NS	0.200	0.500	0.800	1.200
AS	0.200	0.500	0.800	1.200
AES	0.200σ	0.500σ	0.800σ	1.200σ

Note: Developed based on Cohen (1988) and Sawilowsky and Blair (1992).
[a]σ = Standard deviation, [b]S = Small, [c]M = Medium, [d]L = Large, [e]VL = Very Large.
[f]NS = Nominal Shift, [g]AS = Actual Shift, [h]AES = Actual Effect Size (rounded).

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). Let there be $N = n_A + n_B$ ranked observations, each designated as an A or B. For the A observations, maintain a count above the letter sequence, starting from zero and incremented by n_B each time an A is encountered. For the B observations, maintain a count below the letter sequence, starting from zero and incremented by n_A each time a B is encountered. The final count for both A's and B's should be $M = n_A \times n_B$. Compute the differences, d_i = B$_i$ – A_i, by subtracting the A counts from the B counts for each letter position. Find the absolute value of these differences. For the two-sided test, take D^* = max$|d_i|$. For a one-sided test take $D_+^* = \max|pos(d_i)|$ or $D_-^* = \max|neg(d_i)|$ depending on what is expected under H$_1$. Conover (1999) defined the test statistic, T, in terms of two empirical distribution functions, S_A and S_B, using the supremum. For the two-sided test, $T = \sup_x |S_A(x) - S_B(x)|$. For the one-sided test that A < B (stochastically), $T^+ = \sup_x [S_A(x) - S_B(x)]$. Thus, for the one-sided test that A > B (stochastically), $T^- = \sup_x [S_B(x) - S_A(x)]$.

Rejection Region.

Critical regions are usually tabulated as $D^* \geq \mathit{critical\ value}$. Note that $D^* = n_A n_B D$, where D is the statistic derived from a direct comparison of the sample cdf's, is more convenient to work with as it takes only integer values (Neave & Worthington, 1988).

Rosenbaum's Test

Background.

This test first appeared in its current form in Rosenbaum (1954), which was based on Rosenbaum (1953). In both articles, Rosenbaum cited Wilks (1942) as the original source of the formulas for deriving the critical value tables. Rosenbaum (1965) reiterated this earlier work. The test is classified as a runs test. It is a quick and easy test, but is not routinely included in textbooks on nonparametric statistics. Neave and Worthington (1988) presented it as a test for general differences between two sampled populations where spread tends to increase with an increase in the mean, consistent with Rosenbaum (1954). They claimed that under the conditions of an increase in spread with an increase in the median tests such as the Wilcoxon-Mann-Whitney test and Tukey's Quick test have almost no power because of the change in spread. Likewise, tests for spread, such as the Siegel-Tukey test (Siegel & Tukey, 1960), have little or no power because of the change in location. If more general differences were suspected, or needed to be protected against, the Kolmogorov-Smirnov test was suggested as a better choice. Processes that are known to be exponential or Poisson in nature, where the standard deviation is related to the mean, would be excellent candidates for analysis by Rosenbaum's test. Thus, Rosenbaum's test appears to occupy a somewhat unique place among its better-known peers.

Hypotheses.

The null hypothesis is that there is no difference in the two sampled populations. The alternative hypothesis can be two-sided or one-sided. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way. In the case of a one-sided test, the alternative hypothesis is that one population is stochastically greater than the other.

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). For the two-sided test, determine which sample has the overall greatest value and then count the number of observations in that sample that are greater than the greatest value in the other sample and let this be the test statistic *R*. For the one-sided test, determine if the greatest overall value comes from the sample whose population is hypothesized under H_1 to have the greater mean. If it does, proceed as for the two-sided test, if not, set $R = 0$.

Rejection Region.

Critical regions are of the form $R \geq$ *critical value*. The table of critical values must be entered with n_1 as the size of the sample from which R is calculated and n_2 as the size of the other sample (Neave & Worthington, 1988).

Tukey's Quick Test

Background.

This test first appeared in Tukey (1959). It is a two-sample test constructed according to Duckworth's (1958) portability specifications. It is a quick test because it only requires a few of the sample observations to be ordered. It is also compact, in the sense that tables of critical values are not generally needed for most applications, as only a limited number of critical values occur in practice. These two characteristics combine to make the test portable. Like Rosenbaum's test, Tukey's Quick test is based on extreme runs and is not routinely included in applied textbooks.

Hypotheses.

The test is primarily a test for differences in location of the medians of the two sampled populations and is most appropriate when there is reason to believe that the sampled populations have the same spread, or better, the same shape (Neave & Worthington, 1988). The null hypothesis is that there is no difference in the two sampled populations or no difference in the medians of the populations. The alternative hypothesis can be two-sided or one-sided. The two-sided alternative hypothesis is simply that the two sampled populations are different in some way, or have different medians. In the case of a one-sided test the alternative hypothesis is that one population is stochastically greater than the other, or that there is a directional difference in the medians.

Procedure and Test Statistic.

The following procedure was described in Neave and Worthington (1988). It begins by arranging the sample observations in a single combined array from least to greatest, keeping track of original sample membership, say A and B, and then ranking them. For a two-sided test, if the minimum and maximum observed values come from the same sample then the test statistic is $T_y = 0$. If the minimum and maximum observed values come from different samples, then the test statistic is the sum of the extreme runs, that is, if the minimum value comes from sample A and the maximum from sample B, then count the number of A's from the beginning of the array until the first B is reached, say C_L, and count the number of B's from the end of the array back until the first A is reached, say C_U, and set $T_y = C_L + C_U$. For a one-sided test, if the minimum and maximum observed values come from the same sample, set $T_y = 0$. If the minimum and maximum observed values come from different samples, determine if the maximum observation comes from the sample that is expected to have the greater median. If not, set $T_y = 0$. If so, calculate T_y just as for the two-sided.

Rejection Region.

Critical regions are of the form $T_y \geq$ *critical value* and tables are available in Neave and Worthington (1988). However, for one-sided tests with sample sizes that are not too small and not too dissimilar, the .05 and .01 critical values are generally 6 and 9, respectively. For a two-sided test under the same conditions, the .05 and .01 critical values are generally 7 and 10, respectively. These critical values are reported to work well for ratios of sample sizes from 1 to 1.5. Equal sample sizes are not required, although tables of critical values should be employed when the ratio of larger to smaller sample exceeds 1.5 (Tukey, 1959).

Wilcoxon-Mann-Whitney Test

Background.

Wilcoxon (1945) introduced the rank-sum version of this test for equal sample sizes in the same article as the signed-rank test, while Mann and Whitney (1947) independently developed the Mann–Whitney *U* test. The two versions are procedurally different but mathematically equivalent and are often referred to jointly in the literature as the Wilcoxon-Mann-Whitney test (Sprent & Smeeton, 2001). The test is applied to ordinal data. Tables of critical values are more commonly available for the Mann-Whitney version of the test. In either form this is one of the better-known distribution-free tests, and is the one that corresponds most directly to Student's *t*-test for two independent samples (Student, 1908). It is also a powerful test, with an asymptotic relative efficiency that never falls below 0.864 with respect to the *t*-test (Lehmann, 1998), although it is often much more powerful under conditions that violate the assumptions of the *t*-test, yet respect its own assumptions (Blair & Higgins, 1980).

The Wilcoxon-Mann-Whitney test is generally regarded as a test of whether two independent samples represent the same population versus populations that differ in location, either of their medians or with respect to the rank ordering of their scores (Sheskin, 1997). Bergmann, Ludbrook, and Spooren (2000) described it as a test of group mean ranks or, equivalently, rank sums, for testing two different hypotheses: (a) a shift in otherwise identical populations, or (b) a difference in mean ranks between randomized groups. A detailed theoretical treatment of the test was given in Lehmann (1998). Kruskal (1957) detailed the history of the test from 1941 to 1957.

Hypotheses.

The alternative hypothesis under the population model assumes that the populations have identical probability distributions other than a constant shift (Sheskin, 1997), also known as a translation, or location–shift, model. If *F* and *G* are the population distribution functions, the location-shift model requires $G(x) = F(x - \Delta), \forall x$. The null hypothesis is then H_0: $[\Delta = 0]$ (Hollander & Wolfe, 1999). The null hypothesis can also be stated as no difference in the medians of the populations, or H$_0$: [$\phi_1 = \phi_2$] (Neave & Worthington, 1988). With equal sample sizes, this is equivalent to the hypothesis that the sum of ranks for each group is the same, or H$_0$: [$\sum R_1 = \sum R_2$]. For unequal sample sizes this generalizes as the mean rank of the groups being equal, or H$_0$: [$\overline{R}_1 = \overline{R}_2$] (Sheskin, 1997). The parallel to Student's *t*-test is most evident in this form.

The test can be one-sided or two-sided. The two-sided alternative hypothesis for shift is H$_1$: [$\Delta \neq 0$] (Hollander & Wolfe, 1999) and the alternative hypothesis for medians is H$_1$: [$\phi_1 \neq \phi_2$] (Neave & Worthington, 1988). The alternative hypotheses for ranks are H$_1$: [$\sum R_1 \neq \sum R_2$] or H$_1$: [$\overline{R}_1 \neq \overline{R}_2$] (Sheskin, 1997). For a one-sided test, the alternative hypotheses for shift are either H$_1$: [$\Delta < 0$], or H$_1$: [$\Delta > 0$] (Hollander & Wolfe, 1999). The alternative hypotheses for medians are H$_1$: [$\phi_1 < \phi_2$] or H$_1$: [$\phi_1 > \phi_2$] (Neave & Worthington, 1988). The alternative hypotheses for ranks are H$_1$: [$\sum R_1 < \sum R_2$], H$_1$: [$\sum R_1 > \sum R_2$], H$_1$: [$\overline{R}_1 < \overline{R}_2$] or H$_1$: [$\overline{R}_1 > \overline{R}_2$] (Sheskin, 1997).

Procedure and Test Statistic.

Siegel and Castellan (1988) and Neave and Worthington (1988) described the Wilcoxon version of the test. Given two samples, A and B, with $N = n_A + n_B$, combine the observations in a single array, keeping track of original sample membership, and then rank them from 1 to *N*. Compute R_A as the sum of the ranks of the observations from sample A and R_B as the sum of the ranks of the observations from sample B. The test statistic, *W*, is the rank sum that would be expected to be smaller if H$_1$ were true.

Rejection Region.

Tables of critical values are usually given for the Mann-Whitney *U* test (Neave & Worthington, 1988, Sheskin, 1997), with critical regions of the form $U_{min} \leq$ *critical value* representing best conservative values. The test can be applied to unequal sample sizes with appropriate critical value tables. Because they are mathematically equivalent, the results of the Wilcoxon procedure can be converted to values of *U*. Neave and Worthington (1988) gave the conversion for a two-sided test as: $U = \min\left[U_A,\ U_B\right]$, with $U_A = R_A - \frac{1}{2}n_A\left(n_A + 1\right)$ and $U_B = n_A n_B - U_A = R_B - \frac{1}{2}n_B\left(n_B + 1\right)$. For a one-sided test, use either U_A or U_B according to which one is expected to have the smaller value under H$_1$. Converting to values of *U* also accounts for the effect of unequal sample sizes.

Kruskal-Wallis Test

Background.

This test was introduced in Kruskal (1952) and Kruskal and Wallis (1952). Vogt (1999) described it as, "A nonparametric test of statistical significance used when testing more than two independent samples. It is an extension of the Mann-Whitney *U* test, and of the Wilcoxon [rank-sum test], to three or more independent samples. It is a nonparametric one-way ANOVA for rank order data" (p. 151).

Everitt (1998) described the test as a "distribution free method that is the analogue of the analysis of variance of a one-way design. It tests whether the groups to be compared have the same population median" (p. 180). The test is applied to ordinal (rank ordered) data (Sheskin, 1997). Power comparisons with the *F*-test are very favorable. Conover (1999) gave the following asymptotic relative efficiencies for the Kruskal-Wallis test relative to the *F*-test: (a) For distributions that differ only in their means, never less than 0.864, but as high as infinity, (b) for Normal populations, 0.955, (c) for uniform distributions, 1.0, and (d) for exponential distributions, 1.5.

Hypotheses.

For *k* groups, the population distribution functions, $F_1, \ldots, F_k$ are assumed to have the relationship $F_j\left(x\right) = F\left(x - \tau_j\right), -\infty < j < \infty$ over all *j* (*j* = 1 to *k*) where *F* is a continuous distribution function with unknown median and τ_j is the unknown treatment effect for the *j*th population (Hollander & Wolfe, 1999). The null hypothesis can be stated as no difference in the medians of the populations, H$_0$: [$\phi_1 = \phi_2 = \ldots = \phi_n$] (Neave & Worthington, 1988, Siegel & Castellan, 1988), identical populations, H$_0$: [All of the k population distribution functions are identical] (Conover, 1999) or identical treatment effects, H$_0$: [$\tau_1 = \tau_2 = \cdots = \tau_k$] (Hollander & Wolfe, 1999). The alternative hypothesis assumes that the populations differ only in location (Sprent & Smeeton, 2001) and that at least one of the populations, medians or treatment effects is different from the others.

Vargha and Delaney (1998) took exception to the use of the Kruskal-Wallis test with the foregoing assumptions on the grounds that the attendant hypotheses, while mathematically correct, were too narrow to be of practical value to researchers. They claimed that the Kruskal-Wallis test "cannot detect with consistently increasing power any alternative other than exceptions to stochastic homogeneity" (p.170). This, in turn, is mathematically equivalent to the "equality of expected values of the rank sample means" (p.170). They argued that the requirement for identical distributions under H$_0$ is too strict, and that only variance homogeneity is needed. Further, they asserted that the H$_1$ to which the test is actually sensitive is "the tendency for observations in at least one of the populations to be larger (or smaller) than all the remaining populations together" (p. 186).

The test is two-sided with an omnibus alternative hypothesis for shift of H$_1$: [$\tau_1, \ldots, \tau_k$ not all equal] (Hollander & Wolfe, 1999), H$_1$: [not all of ϕ_1, ϕ_2, ..., ϕ_k are equal] (Neave & Worthington, 1988, Siegel & Castellan, 1988) or H$_1$: [At least one of the populations tends to yield larger observations than at least one of the other populations] (Conover, 1999). All of these hypotheses can be formulated in terms of rank-sums (for the equal sample size case) or mean ranks (for the general case) as H$_0$: [$\Sigma R_1 = \Sigma R_2 = \ldots = \Sigma R_k$] or H$_0$: [$\bar{R}_1 = \bar{R}_2 = \ldots = \bar{R}_k$], with the alternative hypothesis of H$_1$: [not H$_0$] (Sheskin, 1997). The alternative hypothesis is stated in this way because it only requires that some pair of groups be different, not that all groups are different, consistent with Conover (1999).

Procedure and Test Statistic.

The general procedure, which does not assume equal sample sizes, is to combine the samples and rank the observations while keeping track of original group membership. For each of the k groups, let the number of observations be n_i $(i = 1, 2, \ldots, k)$ such that the total number of observations is $N = \sum_{i=1}^{k} n_i$. Calculate the rank-sum for each group as $s_i = \sum_{j=1}^{n_i} r_{ij}$, where r_{ij} is the rank assigned to the jth observation in the ith group. The sum of the mean squared ranks is calculated as $S_k = \sum_{i=1}^{k}\left(\frac{s_i^2}{n_i}\right)$. The statistic is then calculated as $H = \frac{12}{N(N+1)} S_k - 3(N+1)$. This is the common computational formulation (Sprent & Smeeton, 2001, Neave & Worthington, 1988, Feir-Walsh & Toothaker, 1974, Siegel & Castellan, 1988, Conover, 1999).

Conover (1999) defined the test statistic as $T = \frac{1}{S^2}\left(S_k - \frac{N(N+1)^2}{4}\right)$ where S_k and N are as defined above and $S^2 = \frac{1}{N-1}\left(\sum_{\substack{all \\ ranks}} R\left(X_{ij}\right)^2 - N\frac{(N+1)^2}{4}\right)$. He noted that S^2 simplified to $N(N+1)/12$ in the absence of ties such that $T = H$ as defined above. H can also be defined as $H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \bar{R}\right)^2$, where n_i is as above, $\bar{R}_i$ is the mean rank of group i, and $\bar{R}$ is the overall mean rank of the N total observations (Neave & Worthington, 1988, Siegel & Castellan, 1988). In this form it can be seen most clearly that the statistic is a weighted sum of squared deviations. Post-hoc procedures using pairwise comparisons are available (Conover, 1999, Sheskin, 1997, Siegel & Castellan, 1988), but are not considered further here.

Rejection Region.

Critical regions are of the form $H \geq \mathit{critical\ value}$ (Neave & Worthington, 1988). Approximate critical values can be obtained from a chi-squared distribution with *k* – 1 degrees-of-freedom, but see Fahoome (1999, 2002). The test will work with unequal sample sizes since the calculation of the statistic involves a weighted sum of squares of differences between group mean ranks and the overall mean rank, although critical value tables tend to be limited (Neave, 1981).

Terpstra-Jonckheere Test

Background.

The Terpstra-Jonckheere test was developed independently by Terpstra (1952) and Jonckheere (1954). Like the Kruskal-Wallis test, it is an extension of the Wilcoxon-Mann-Whitney test on ranks for the one-way design. It differs from the Kruskal-Wallis test in that it postulates a specific ordering of the groups under the alternative hypothesis based on prior knowledge, that is, that the situation being tested supports an a priori expectation of a specific, identifiable order of the population medians based on the experimental design, not on the observed data (Hollander & Wolfe, 1999, Siegel & Castellan, 1988). A general assumption is that all of the possible assignments of joint ranks are equally possible (Hollander & Wolfe, 1999).

Hypotheses.

For *k* groups, the population distribution functions, $F_1,\ldots,F_k$ are assumed to have the relationship $F_j(x) = F(x - \tau_j), -\infty < x < \infty$ over all *j*, (*j* = 1 to *k*), where *F* is a continuous distribution function with unknown median and τ_j is the unknown treatment effect for the *j*th population (Hollander & Wolfe, 1999). The null hypothesis can be stated in terms of medians as H_0: [$\phi_1 = \phi_2 = \ldots = \phi_k$] (Neave & Worthington, 1988, Siegel & Castellan, 1988), identical populations as H_0: [$F_1(x) = F_2(x) = \cdots = F_k(x), \forall x$] (Sprent & Smeeton, 2001), or treatment effects as H_0: [$\tau_1 = \tau_2 = \cdots = \tau_k$] (Hollander & Wolfe, 1999). If the *k* groups are numbered to correspond to the expected order, the alternative hypothesis is one-sided and given by H_1: [$\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$, with at least one strict inequality] (Hollander & Wolfe, 1999), H_1: [$F_1(x) \leq F_2(x) \leq \cdots \leq F_k(x)$, at least one inequality strict for some x] (Sprent & Smeeton, 2001), or H_1: [$\phi_1 \leq \phi_2 \leq \ldots \leq \phi_k$, at least one of the inequalities is strict] (Neave & Worthington, 1988, Siegel & Castellan, 1988).

Procedure and Test Statistic.

The procedure calculates the Mann-Whitney *U* statistic for all pairs of samples and then combines the results. If the Wilcoxon rank-sum procedure is used the resulting statistics must be converted to Mann-Whitney *U* statistics before being combined. For the alternative hypothesis, as stated above, the test statistic was given by Neave and Worthington (1988) as

$$\begin{aligned} J &= U_{21} + U_{31} + \ldots + U_{k1} + U_{32} + \ldots + U_{ij} + \ldots + U_{k(k-1)} \\ &= \sum_{j=1}^{k-1} \sum_{i=j+1}^{k} U_{ij} \end{aligned},$$

where U_{ij} represents the Mann-Whitney *U* statistic for each pair of samples, computed in the order dictated by H_1 to give the least value of each U_{ij}. This is consistent with Siegel and Castellan (1988) and others. To the extent that H_1 tends to be true, each of the U_{ij} will tend to be small and thus their sum will tend to be small.

For *k* groups there will be *k*(*k* - 1)/2 values of *U*. Hollander and Wolfe (1999) gave the Mann-Whitney procedure for calculating the values of *U* directly, including an adjustment for ties (equivalent to using mid-ranks in the Wilcoxon version of the procedure) as

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi^*(X_{iu}, X_{jv}), 1 \leq u < v \leq k,$$

where

$$\phi^*(a,b) = \begin{cases} 1 \text{ if } a < b \\ \dfrac{1}{2} \text{ if } a = b \\ 0 \text{ if } a > b \end{cases}.$$

This is consistent with Siegel and Castellan

Table 4. Test / Method Combinations with Acceptable Type I Error Results

	Test					
Method	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-1[g]	EA,--	EA,--	--,--	EA,EB	--,--	EA,EB
	ML,SS	ML,SS	ML,--	ML,SS	ML,(SS)	ML,SS
M-2[h]	na	EA,--	--,--	na	na	na
		ML,SS	ML,--			
M-3[i]	EA,--	--,--	--,--	--,EB	--,EB	EA,EB
	ML,SS	ML,SS	ML,--	ML,SS	ML,SS	ML,SS
M-4[j]	EA,EB	EA,EB	EA,--	EA,EB	EA,EB	EA,EB
	ML,SS	ML,SS	ML,--	ML,SS	ML,SS	ML,SS
M-5[k]	--,--	na	na	na	na	na
	ML,--					
M-6[l]	na	na	na	EA,EB	EA,EB	EA,EB
				ML,SS	ML,SS	ML,SS
M-7[m]	na	--,--	na	na	na	na
		ML,SS				

Note. EA = Extreme Asymmetric Data Set, EB = Extreme Bi-modal Data Set, ML = Multi-modal Lumpy Data Set, SS = Smooth Symmetric Data Set.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test,
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test,
[f]T-J = Terpstra-Jonckheere Test.
[g]M-1 = Average of least and most likely to reject, [h]M-2 = Count ties as ½, [i]M-3 = Alternating,
[j]M-4 = Random, [k]M-5 = Delayed Increment, [l]M-6 = Mid-ranks, [m]M-7 = Weighted average.

(1988). The test is approximate when ties are present.

Rejection Region.

Critical regions are of the form $J \leq critical\ value$. The test supports unequal samples sizes and more extensive critical value tables are available as Table *R* in Neave and Worthington (1988). As the sample size increases, the null distribution of *J* becomes asymptotically normal. Formulas exist for obtaining approximate critical values (Neave & Worthington, 1988, Siegel & Castellan, 1988), but see Fahoome (1999, 2002).

Results

Type I Error Results

Question 1: For samples drawn from the same population, is the Type I error rate maintained between $.5\alpha$ and 1.1α for each combination of test, method, number of groups, directionality, sample size, and distribution?

Combinations of tests, methods and Micceri (1986) data sets that demonstrated acceptable Type I Error rates are shown in Table 4. Results for the theoretical standard Normal distribution are not shown, as it did not produce ties. Note, however, that the performance of these tests with the theoretical Normal distribution was not always acceptable due to the

Table 5. Preferred Methods[k, l, m, n, o, p] by Test and Micceri (1986) Data Set

	Test					
Data Set	K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
EA[g]	M-4, M-1	M-1/ M-2/ M-4	na	M-4	M-4, M-6	M-4
EB[h]	na	na	na	M-4	M-4/ M-6	M-4
ML[i]	M-4	M-3	M-4	M-3	M-4/ M-6, M-1	M-4
SS[j]	M-4	M-3	M-4	M-4, M-3	M-4/ M-6, M-1	M-4

Note. A/B indicates very similar results, A, B indicates A better than B.
[a]K-S = Kolmogorov-Smirnov Test, [b]R = Rosenbaum's Test, [c]TQ = Tukey's Quick Test.
[d]W-M-W = Wilcoxon-Mann-Whitney Test, [e]K-W = Kruskal-Wallis Test.
[f]T-J = Terpstra-Jonckheere Test.
[g]EA = Extreme Asymmetric Data Set, [h]EB = Extreme Bi-modal Data Set,
[i]ML = Multi-modal Lumpy Data Set, [j]SS = Smooth Symmetric Data Set.
[k]M-1 = Average of least and most likely to reject, [l]M-2 = Count ties as ½, [m]M-3 = Alternating,
[n]M-4 = Random, [o]M-5 = Delayed Increment, [p]M-6 = Mid-ranks.

discrete nature of the statistics and the use of best conservative critical values whose probabilities were sometimes less than 0.5α. Following Bradley (1978), Type I error performance was judged to be acceptable if it was not inflated beyond 1.1α and was not more conservative than the corresponding performance with the theoretical Normal distribution. As shown in Table 4, the random method provided acceptable Type I error rates for the largest combination of tests and distributions. Most of the other methods provided acceptable results for specific combinations of test and data set with the exception of Methods 8 and 9 (not shown).

Method 9, the drop all ties and reduce N method, is one of the most widely recommended, especially in textbooks, for situations where there are not too many ties. But how many is too many? Methods 8 and 9 are absent from Table 4 because the Type I error results were unacceptable across all combinations of tests and simulation parameters.

Power Results

The remaining research questions were only studied for those combinations of test, method, number of groups, directionality, sample size and distribution for which Question 1 was answered in the affirmative as shown in Table 4. In order to answer the 3rd and 4th research questions it was necessary to analyze the power results from a large number of simulation runs in a manner that might permit determination of the order of preference of methods across various combinations of simulation parameters for each test.

Table 6. Best Method[g, h, i, j] By Test Across Distributions

K-S[a]	R[b]	TQ[c]	W-M-W[d]	K-W[e]	T-J[f]
M-4[i]	M-3[h]	M-4	M-4	M-4, M-3	M-4, M-6[j], M-1[g]

Note. A, B indicates A better than B.
[a] K-S = Kolmogorov-Smirnov Test, [b] R = Rosenbaum's Test, [c] TQ = Tukey's Quick Test,
[d] W-M-W = Wilcoxon-Mann-Whitney Test, [e] K-W = Kruskal-Wallis Test,
[f] T-J = Terpstra-Jonckheere Test.
[g] M-1 = Average of least and most likely to reject, [h] M-3 = Alternating, [i] M-4 = Random.
[j] M-6 = Mid-ranks.

Question 2: For samples drawn from populations differing only in location, is there a preferred method of resolving tied ranks for each combination of test and data set, irrespective of the number of groups, directionality, and sample size?

As shown in Table 5, the random method was the preferred method (13 of 20), or tied for first (4 of 20), for the vast majority of combinations of test and data set (17 of 20). The method of analysis employed for this purpose involved ranking the power results across methods for each specific combination of test, number of groups, nominal alpha level and distribution at each combination of nominal effect size multiplier and initial sample size. Mean ranks were then calculated in three ways: (a) by summing across nominal effect size multipliers at each initial sample size, (b) by summing across initial sample sizes at each nominal effect size multiplier, and (c) by summing across both nominal effect size multipliers and initial sample sizes.

Question 3: For samples drawn from populations differing only in location, is there a preferred method of resolving tied ranks for each test, irrespective of the number of groups, directionality, sample size, and data set?

This question requires a conclusion to be drawn about the relative behavior of the methods across data sets. The results of the preceding analysis were used to determine the number of first place finishes for each test for each combination of method and distribution across nominal alpha and number of groups. If a particular method consistently had the most first place finishes for a particular test, across data sets, then it could in some sense be considered the best method for that test/data set combination. As shown in Table 6, random resolution of ties was clearly superior for four of the six tests, and a close second for another.

Question 4: Is there a best method for resolving ties across all tests and data sets in the study?

Given the results presented in Tables 4, 5, and 6, random resolution of ties performs best across the set of tests, data sets and methods examined in this study.

Conclusion

This study examined various methods of resolving equal data values (tied ranks) in a set of distribution-free statistical tests of location or general difference for *k* independent samples using Monte Carlo simulations with theoretical Normal and discrete, non-normal data. These tests were all based on the assumption of continuity in the underlying population. As such, the presence of ties—which occurred frequently with the discrete, non-normal data sets—and the efficacy of various methods of resolving them were of theoretical and practical interest.

Of the methods investigated for resolving ties, random resolution seemed to perform best, in the sense of guarding against inflation of Type I error rates while maintaining power, for the majority of combinations of simulation parameters, but not all. This is of interest both theoretically and practically. First, although random resolution might be expected to produce the best results on theoretical grounds, it does not always do so. There are also strong objections in practice to resolving ties at random as the outcome of any particular test then depends on a secondary random event. But what are the consequences of the alternatives if random resolution is rejected on these grounds? How well do the common alternatives, such as mid-ranks or dropping tied values, work?

The often-recommended method of dropping tied values and reducing the sample size performed very poorly across all combinations of simulation parameters. Based on the results of this study, this method should not be used. All of these tests and methods also performed poorly with Likert scale data (i.e., Micceri, 1986, Extreme Bi-modal data set). They should not be used with discrete population data sets that contain relatively few distinct values.

References

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician, 54*(1), 72-77.

Blair, R. C., & Higgins, J. J (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational Statistics, 5*(4), 309-335.

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

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

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*(3), 331-350.

Cliff, N. (1996b). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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

Duckworth, W. E. & Wyatt, J. V. (1958). Rapid statistical techniques for operations research workers. *Operations Research Quarterly, 9*, 218-233.

Everitt, B. S. (1998). *The Cambridge dictionary of statistics*. Cambridge, England: Cambridge University Press.

Fahoome, G. F. (1999). A Monte Carlo study of twenty-one nonparametric statistics with normal and nonnormal data. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Fahoome, G. F. (2002). Review of twenty nonparametric statistics and their large sample approximations. *Journal of Modern Applied Statistical Methods, 1*(2), 248-268.

Fay, B. R. (2002). JMASM4: Critical values for four nonparametric and/or distribution-free tests of location for two independent samples. *Journal of Modern Applied Statistical Methods, 1*(2), 489-517.

Feir-Walsh, B. J. & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34*(4), 789-799.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed. Vol. 131). New York: Marcel Dekker, Inc.

Hollander, M., & Wolfe, D. (1999). *Nonparametric statistical methods* (2nd ed.). New York: John Wiley and Sons, Inc.

Jonckheere, A. R. (1954). A distribution free k-sample test against ordered alternatives. *Biometrika, 41*, 133-145.

Kelley, D. L., Sawilowsky, S. S., & Blair, R. C. (1994, October). Comparison of ANOVA, McSweeney, Bradley, Harwell-Serlin, and Blair-Sawilowsky tests in the balanced 2x2x2 layout. Paper presented at the annual meeting of the Midwestern Educational Research Association, Chicago, IL.

Killian, C. R., & Hoover, H. D. (1974, April). An investigation of selected two-sample hypothesis testing procedures when sampling from empirically based test score models. Paper presented at the 59th annual meeting of the American Educational Research Association, Chicago, IL.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari, 4*, 83-91.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics, 23*, 525-540.

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of American Statistical Association, 52*, 356-360.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks on one-criterion analysis of variance. *Journal of American Statistical Association, 47*, 583-621.

Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks* (1st (Revised) ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education, 67*(4), 369-379.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50-60.

Micceri, T. (1986, November). A futile search for that statistical chimera of normality. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166.

Neave, H. R. (1981). *Elementary statistics tables*. London: Unwin Hyman Ltd.

Neave, H. R. & Worthington, P. L. B. (1988). *Distribution-free tests*. London: Unwin Hyman Ltd.

Rosenbaum, S. (1953). Tables for a nonparametric test of dispersion. *Annals of Mathematical Statistics*, *24*, 663-668.

Rosenbaum, S. (1954). Tables for a nonparametric test of location. *Annals of Mathematical Statistics, 25*, 146-150.

Rosenbaum, S. (1965). On some two-sample non-parametric tests. *Journal of American Statistical Association, 60,* 1118-1126.

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

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*(2), 352-360.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika, 55*(4), 729.

Sheskin, D. J. (1997). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, Fl: CRC Press.

Siegal, S. & Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of American Statistical Association*, *55*, 429-445.

Siegel, S. & Castellan, N. J., Jr., (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Boston: McGraw-Hill.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of Mathematics University of Moscow, 2*(2), 3-14.

Sparks, J. N. (1967). Effects of Inapplicability of the continuity condition upon the probability distributions of selected statistics and their implications for research in education (Final report for project no. BR-6-8467 PA-24). Pennsylvania State Univ. (RIE SYN71840) (ERIC Document Reproduction Service No. ED021317)

Sprent, P. & Smeeton, N. C. (2001). *Applied nonparametric statistical methods* (3rd ed.). Boca Raton, Fl: Chapman and Hall / CRC.

Student [W. S. Gosset], (1908). The probable error of a mean. *Biometrika, 6*, 1-25.

Terpstra, T. J. (1952). A non-parametric k sample test and its connection with the H-test (S-92, VP2). Amsterdam: Mathematical Center.

Tukey, J. W. (1959). A quick, compact, two-sample test to Duckworth's specifications. *Technometrics, 1,* 31-48.

van den Brink, W. P. & van den Brink, S. G. (1989). A comparison of the power of the t test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *British Journal of Mathematical and Statistical Psychology, 42*(2), 183-189.

Vargha, A. & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics, 23*(2), 170-192.

Vogt, P. W. (1999). Dictionary of statistics and methodology: *A nontechnical guide for the social sciences* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology, 51,* 1-39.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80-83.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Annals of Mathematical Statistics, 13,* 400-409.

Appendix

Micceri (1986) data sets (see Sawilowsky & Blair, 1992):

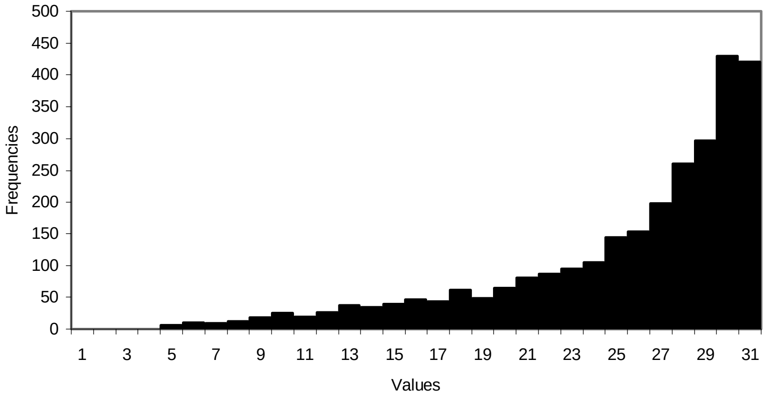


Figure A1. Micceri (1986) extreme asymmetric data set. See Sawilowsky & Blair (1992).

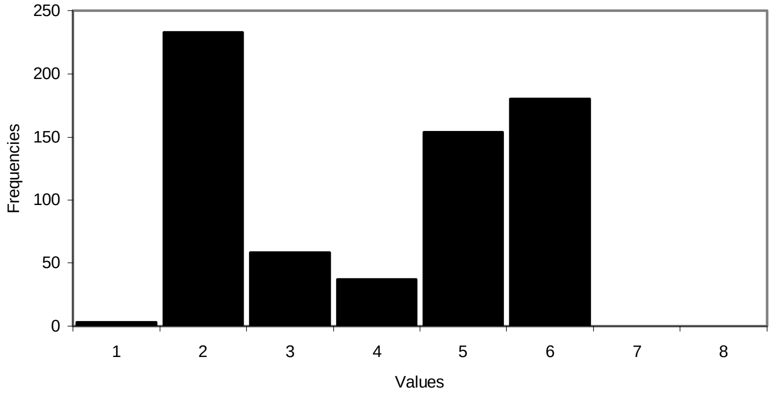


Figure A2. Micceri (1986) extreme bi-modal data set. See Sawilowsky & Blair (1992).

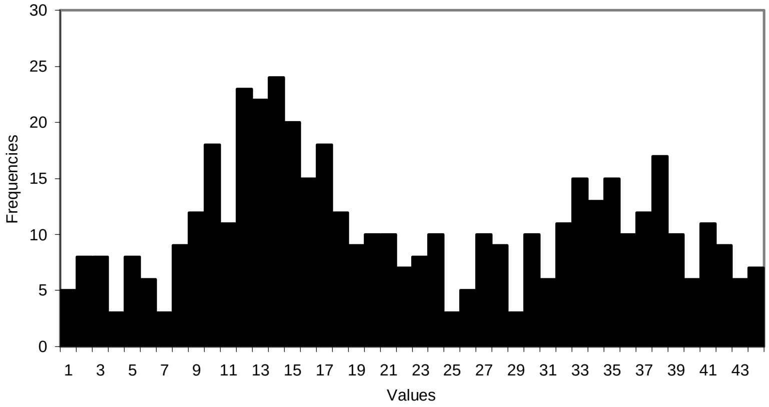


Figure A3. Micceri (1986) multi-modal lumpy data set. See Sawilowsky & Blair (1992).

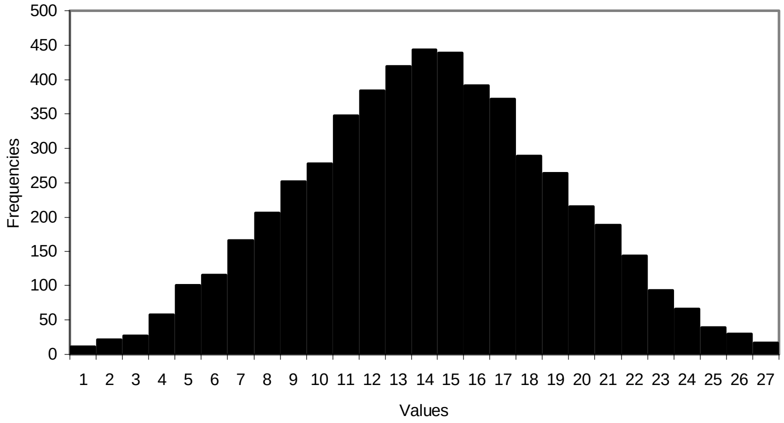


Figure A4. Micceri (1986) smooth symmetric data set. See Sawilowsky & Blair (1992).

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 41-43

1538 – 9472/06/$95.00

Limitations Of The Analysis Of Variance

Phillip I. Good
Information Research
Huntington Beach, C.A.

Cliff Lunneborg
Department of Statistics
University of Washington

Conditions under which the analysis of variance will yield inexact p-values or would be inferior in power to a permutation test are investigated. The findings for the one-way design are consistent with and extend those of Miller (1980).

Key words: Analysis of variance, permutation tests, exact tests, robust tests, one-way designs, k-sample designs.

Introduction

The analysis of variance has three major limitations:

1. It is designed to test against any and all alternatives to the null hypothesis and thus may be suboptimal for testing against a specific hypothesis.

2. It is optimal when losses are proportional to the square of the differences among the unknown population means, but may not be optimal otherwise. For example, when losses are proportional to the absolute values of the differences among the unknown population means, expected losses would be minimized via a test that makes use of the absolute values of the differences among the sample means; see, for example, Good (2005).

3. It is designed for use when the observations are drawn from a normal distribution and though it is remarkably robust, it may not yield exact p-values when the observations come from distributions that are heavier in the tails than the normal. Even in cases when the analysis of variance yields almost exact p-values, it may be less powerful than the corresponding permutation test when the observations are drawn from non-normal distributions under the alternative.

The use of the F-distribution for deriving p-values for the analysis of variance is based upon the assumption of normality; see, for example, the derivation in Lehmann (1986). Nevertheless, Jagers (1980) shows that the F-ratio is almost exact in many non-normal situations.

The purpose of the present note is to explore the conditions under which a distribution would be sufficiently non-normal that the analysis of variance applied to observations from that distribution would be either inexact or less powerful than a permutation test.

Findings: General Hypotheses

When the form of the distribution is known explicitly, one often can transform the observations to normally-distributed ones and then apply the analysis of variance; see, Lehman (1986) for a list of citations. Consequently, the

Philip Good is a statistical consultant. He authored numerous books that include, *Introduction to Statistics via Resampling Methods and R/S-PLUS* and *Common Errors in Statistics and How to Avoid Them*. Email: pigood@verizon.net. The late Cliff Lunneborg was Professor Emeritus, Statistics & Psychology and author of *Modeling Experimental and Observational Data* and *Data Analysis by Resampling: Concepts and Applications.*

present investigation is limited to the study of observations drawn from contaminated normal distributions, both because such distributions are common in practice and because they cannot be readily transformed.

In R, examples of samples such distributions would include the following:

```
rnorm(n,2*rbinom(n,1,0.3))
ifelse(rbinom(n,1,0.3),rnorm(n,0.5),
  rnorm(n,1.5,1.5))
```

for both of which the analysis of variance was exact in 1000 simulations of an unbalanced 1x3 design with 3, 4, and 5 observations per cell.

Regardless of the underlying distribution, providing the observations are exchangeable under the null hypothesis, one can always make use of the permutation distribution of a test statistic to obtain an exact test. Let X_{ij} denote the jth observation in the ith cell of a one-way design. Eliminating factors from the F-ratio that are invariant under rearrangement of the observations between cells, such as the within sum of squares that forms its denominator, a permutation test based on the F-ratio reduces to a test based on the sum $\sum_i(\sum_j X_{ij})^2$. It was this test that was used in head-to-head comparisons with the one-way analysis of variance.

When a 1x3 design was formed using the following code

```
s1=rnorm(size[1],rbinom(size[1],1,0.3))
s2=ifelse(rbinom(size[2],1,0.3),
  rnorm(size[2],0.5),rnorm(size[2],1.5,1.5))
s3=ifelse(rbinom(size[2],1,0.3),
  rnorm(size[3],1),rnorm(size[3],2,2))
```

the power of the analysis of variance and the permutation test based upon 1000 simulations were comparable for a balanced design with as few as three observations per cell (α=10%, β=22%). But for an unbalanced design with 3, 4, and 5 observations per cell, the permutation test was more powerful at the 10% level with β=30%, compared to 18% for the analysis of variance.

When a 1x4 design was formed using the following code:

```
s0=rnorm(size[1],rbinom(size[1],1,0.5))
s1=rnorm(size[2],rbinom(size[2],1,0.5))
s2=rnorm(size[3],rbinom(size[3],1,0.5))
s3=rnorm(size[4],2 + rbinom(size[4],1,0.5))
```

the power of the analysis of variance and the permutation test were comparable for a balanced design with as few as three observations per cell (α=10%, β=57%). However, for an unbalanced design with 2, 3, 3, and 4 observations per cell, the permutation test was more powerful at the 10% level with β=86%, compared with 65% for the analysis of variance.

If the designs are balanced, the simulations support Jagers (1980) result, that the analysis of variance is both exact and powerful, whether observations are drawn from a contaminated normal distribution, a distorted normal distribution (z=2*z if z>0), a censored normal distribution (z = -0.5 if z< -0.5), or a discrete distribution such as would arise from a survey on a five-point Likert scale. When the design is unbalanced, Jagers' result does not apply, and the permutation test has superior power. The results confirm and extend the findings of Miller (1986).

Findings: Specific Hypotheses

When testing for an ordered dose response, the Pearson's product moment correlation coefficient is usually employed as a test statistic with p-values obtained from a t distribution. Alternatively, the exact permutation procedure due to Pitman (1937) could be employed. In the simulations with contaminated normal distributions, it was found that the parametric procedure for testing for an ordered dose response was both exact (to within the simulation error) and as powerful as the permutation method.

For testing other specific hypotheses, the permutation method may be preferable, simply because no well-tabulated parametric distribution exists. An example would be the alternative that exactly one of the k-populations from which the samples are drawn is different from the others for which an exact test based on the distribution of $\max_k |\bar{X}_. - \bar{X}_k|$ is readily obtained by permutation means.

To further explore the possibilities, a copy of the code along with a complete listing of the simulation results is provided at mysite.verizon.net/res7sf1o/AnovPower.txt. (A manuscript assessing the robustness of the two-way analysis of variance is in preparation.)

References

Good, P. (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (3rd Ed.). New York: Springer.

Jagers, P. (1980). Invariance in the linear model: An argument for chi-square and F in nonnormal situations. *Mathematische Operationsforschung und Statistik, 11*, 455-464.

Lehmann, E. L. (1986). *Testing statistical hypotheses*. (2nd ed.). New York: Wiley.

Miller, R. G. (1986). *Beyond ANOVA, basics of applied statistics*. New York: Wiley.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society, 4*, 225-232.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 44-65

1538 – 9472/06/$95.00

Multiple Comparison Procedures, Trimmed Means And Transformed Statistics

Rhonda K. Kowalchuk
Southern Illinois University Carbondale

H. J. Keselman
University of Manitoba

Rand R. Wilcox
University of Southern California

James Algina
University of Florida

A modification to testing pairwise comparisons that may provide better control of Type I errors in the presence of non-normality is to use a preliminary test for symmetry which determines whether data should be trimmed symmetrically or asymmetrically. Several pairwise MCPs were investigated, employing a test of symmetry with a number of heteroscedastic test statistics that used trimmed means and Winsorized variances. Results showed improved Type I error control than competing robust statistics.

Key words: Multiple comparison procedures, trimmed estimators, symmetric and asymmetric trimming, heteroscedastic test statistic, nonnormality, variance heterogeneity.

Introduction

Pairwise multiple comparison procedures (MCPs) are adversely affected by nonnormality, particularly when variances are heterogeneous and group sizes are unequal (Keselman, Lix, & Kowalchuk, 1998). Specifically, Type I errors are liberal, resulting in spurious rejections of null hypotheses. The deleterious effects of nonnormality on rates of Type I error are, for the most part, attributable to asymmetry of distributions, that is, to skewness (Westfall & Young, 1993). These results are predictable on theoretical grounds. Cressie and Whitford (1986) showed that Student's two-sample *t* test is not asymptotically correct when the group distributions have unequal third cumulants and sample sizes are unequal; therefore, Type I error inflation is expected. In the one-way independent groups problem, Keselman, Lix, et al. (1998) found Type I error rates for popular pairwise MCPs approached .21 ($\alpha = .05$) when data were obtained from skewed distributions where group variances and sample sizes were unequal and negatively paired with one another.

One potential solution to this Type I error inflation is to replace the usual least squares estimators with estimates which are less influenced by the effects of nonnormality. Indeed, many investigators have shown that better results can be obtained by using statistics designed for heterogeneity combined with robust estimators of central tendency and variability (see Keselman, Kowalchuk, & Lix, 1998; Lix & Keselman, 1998; Wilcox, Keselman, & Kowalchuk, 1998; Yuen, 1974). For example, Keselman, Lix et al. (1998) found that the methods due to Ryan (1960), Welsch (1977), Peritz (1970), Shaffer (1979; 1986), Hayter (1986), and Hochberg (1988) provided much better Type I error control, typically having rates less than .075 when based on a heteroscedastic statistic with trimmed means and Winsorized variances. Though rates improved, these methods were, nonetheless, still occasionally affected when distributions were nonnormal, variances were heterogeneous, and group sizes

Rhonda K. Kowalchuk (rkowal@siu.edu) is an Assistant Professor of Educational Psychology. H. J. Keselman (kesel@ms.umanitoba.ca) is a Professor of Psychology. Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology. James Algina (algina@ufl.edu) is a Professor of Educational Psychology.

were unequal. That is, rates occasionally exceeded .075.

An approach that may provide improved Type I error control for tests of trimmed mean equality (pairwise) is to use a preliminary test for symmetry which determines whether data should be trimmed symmetrically or asymmetrically. Keselman, Wilcox, Othman, and Fradette (2002) found that by using a test for symmetry in conjunction with a test for equality of trimmed means, Type I error rates were well controlled when data were extremely heterogeneous and nonnormal in a one-way independent groups design. The test of symmetry investigated was first proposed by Hogg, Fisher, and Randles (1975) and later modified by Babu, Padmanabhan and Puri (1999). Specifically, two indices are computed, one that determines tail thickness and the other symmetry of the underlying distribution. The calculations determine whether a test of mean equality is based on symmetrically or asymmetrically trimmed means (see Othman, Keselman, Wilcox, Fradette, & Padmanabhan, 2002, for details of the test of symmetry).

Keselman, Lix, et al. (1998) symmetrically trimmed 20% of the data per group and used an approximate degrees of freedom Welch (1938) test statistic for the pairwise comparisons. Although, 20% symmetric trimming is recommended (Wilcox, 1995), theory would imply that asymmetric trimming would be more appropriate when data are skewed (Keselman et al., 2002; Othman et al., 2002). The rationale behind asymmetric trimming is to remove more of the offending data (i.e., data that does not represent the bulk of the observations, that is, the 'typical' score) from the tail containing more of the outlying values. Keselman et al. (2002) found other percentages of trimming, either symmetrically or asymmetrically, resulted in better Type I error control than uniformly adopting 20% symmetric trimming. For example, 15% symmetric trimming or 15% asymmetric trimming resulted in fewer non-robust values compared to always adopting 20% symmetric trimming.

In addition, Keselman et al. (2002) found that transformations (i.e., Johnson, 1978; Hall, 1992) of the Welch-James heteroscedastic statistic improved Type I error control. The Johnson and Hall transformations are intended to remove the bias due to skewness. This is consistent with Guo and Luh (2000) and Luh and Guo (1999) who found that transformations of the Welch-James statistic improved its performance when trimmed means were used and distributions were skewed and heavy-tailed. As well, Keselman et al. (2002) found improved Type I error control when the transformed heteroscedastic statistics were preceded by a test of symmetry under extreme conditions of nonnormality and variance heterogeneity in a one-way independent groups design. Thus, the purpose of this article was to investigate whether these procedures would be beneficial in the pairwise multiple comparison problem.

Test of Symmetry

Othman et al. (2002) provided the details for the test of symmetry, a test based on the work of Hogg et al. (1975) and Babu et al. (1999). Essentially, two indices are computed, one index (Q_2) determines tail-weight (light or heavy) while the other index (Q_1) determines the symmetry of an underlying distribution. The value of the Q_2 index classifies a distribution as normal-tailed, heavy-tailed, or very heavy-tailed which then determines the number of sample points to be used in the computation of the Q_1 index. If the distribution is determined to be (a) normal-tailed, then all sample points are used, (b) heavy-tailed, then the top and bottom 10% of sample points are trimmed, or (c) very heavy-tailed, then the top and bottom 20% of sample points are trimmed. That is, the value of the Q_1 index determines the symmetry/asymmetry of a distribution (i.e., left skewed, symmetric or right skewed) which then determines the type of trimming (symmetric vs asymmetric). Keselman et al. (2002) provided a SAS/IML (1989) program to compute the test of symmetry.

Robust Estimation

Robust estimates of central tendency and variability were applied to heteroscedastic statistics. Specifically, trimmed means and Winsorized variances were used in order to test the hypothesis of the equality of population trimmed means in the pairwise multiple comparison problem. Let

$$Y_{(1)j} \le Y_{(2)j} \le \ldots \le Y_{(n_j)j}$$

represent the ordered observations associated with the jth (j=1,…,J) group, where n_j is the sample size in the jth group. Let

$$g_j = \left[\gamma n_j\right]$$

where γ represents the proportion of observations to be trimmed in each tail of the distribution and [x] is the greatest integer ≤ x. The effective sample size for the jth group becomes $h_j = n_j - 2g_j$. The jth sample trimmed mean is

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

The sample Winsorized mean is necessary in order to compute the Winsorized variance. The jth sample Winsorized mean is

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

where

$$X_{ij} = \begin{cases} Y_{(g_j+1)j} & \text{if } Y_{ij} \le Y_{(g_j+1)j} \\ Y_{ij} & \text{if } Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ Y_{(n_j-g_j)j} & \text{if } Y_{ij} \ge Y_{(n_j-g_j)j}. \end{cases}$$

The sample Winsorized variance is required in order to get a valid estimate of the standard error of a trimmed mean. The sample Winsorized variance for the jth group is

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

and the estimated standard error of the trimmed mean is

$$\sqrt{\left(n_j-1\right)\hat{\sigma}_{wj}^2 \Big/ \left[h_j\left(h_j-1\right)\right]}.$$

Under asymmetric trimming, and assuming that the distribution is positively (right) skewed so that observations in the upper tail of the distribution are trimmed, the effective sample size for the jth group becomes $h_j = n_j - g_j$. The jth sample trimmed mean is

$$\hat{\mu}_{tj} = \frac{1}{h_j}\sum_{i=1}^{n_j-g_j} Y_{(i)j} \qquad (3)$$

and the jth sample Winsorized mean is

$$\hat{\mu}_{wj} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}, \qquad (4)$$

where

$$X_{ij} = \begin{cases} Y_{ij} & \text{if } Y_{ij} < Y_{(n_j-g_j)j} \\ Y_{(n_j-g_j)j} & \text{if } Y_{ij} \ge Y_{(n_j-g_j)j}. \end{cases}$$

The sample Winsorized variance is computed based on the previous equation with the new definition of $\hat{\mu}_{wj}$ and the estimated standard error of the trimmed mean is also computed based on the previous equation with the new definitions of h_j and $\hat{\sigma}_{wj}^2$.

Definitions of the Heteroscedastic Statistics

Johanson's (1980) Welch-James (WJ)-type heteroscedastic statistic (see Lix & Keselman, 1995) with robust estimators has been found to obtain better Type I error control than the WJ statistic with least squares estimators in independent groups designs under nonnormality and variance heterogeneity (see Guh & Luh, 2000; Keselman, Kowalchuk, et al., 1998; Keselman, Lix, et al., 1998; Lix & Keselman, 1998; Luh & Guo, 1999; Wilcox et al. 1998). Guo and Luh (2000) found that two transformations of the WJ statistic combined with the use of trimmed means and Winsorized variances resulted in better Type I error control than the WJ statistic with trimmed means and without a transformation for various skewed and heavy-tailed distributions. Specifically, Johnson's (1978) or Hall's (1992) transformations of the WJ statistic are intended to remove skewness. Hence, the transformations contend with skewness, trimmed means contend with heavy tails, and a heteroscedastic statistic contends with variance heterogeneity (Luh & Guo, 1999).

In the present study, both transformations of the WJ statistic for removing skewness were investigated along with the nontransformed WJ statistic. Let $\hat{\mu}_{tj}, \hat{\mu}_{wj}, \hat{\sigma}^2_{wj}$ and h_j be the trimmed mean, Winsorized mean, Winsorized variance, and trimmed sample size, respectively, for group j. The third central Winsorized moment of the jth group is

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

Let

$$\tilde{\sigma}^2_{wj} = \frac{\left(n_j - 1\right)}{\left(h_j - 1\right)}\hat{\sigma}^2_{wj},\ \tilde{\mu}_{wj} = \frac{n_j}{h_j}\hat{\mu}_{3j},$$

$$q_j = \frac{\tilde{\sigma}^2_{wj}}{h_j},\ w_{tj} = \frac{1}{q_j},\ U_t = \sum_{j=1}^{J} w_{tj},$$

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

Luh and Guo (1999) defined Johnson's (1978) transformed trimmed mean statistic as

$$T_{(Johnson)j} = \left(\hat{\mu}_{tj} - \hat{\mu}_t\right) + \frac{\tilde{u}_{wj}}{6\tilde{\sigma}^2_{wj}h_j} + \frac{\tilde{u}_{wj}}{3\tilde{\sigma}^4_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^2 . \qquad (5)$$

From Guo and Luh (2000), Hall's (1992) transformed trimmed mean statistic can be defined as:

$$T_{(Hall)j} = \left(\hat{\mu}_{tj} - \hat{\mu}_t\right) + \frac{\tilde{u}_{wj}}{6\tilde{\sigma}^2_{wj}h_j} + \frac{\tilde{u}_{wj}}{3\tilde{\sigma}^4_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^2 + \frac{\tilde{u}_{wj}}{27\tilde{\sigma}^8_{wj}}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^3 . \qquad (6)$$

Keselman, Wilcox, and Lix (2003) indicated that sample trimmed means, sample Winsorized variances, and trimmed sample sizes can be used to compute the WJ statistic. That is,

$$WJ = \sum_{j=1}^{J} w_{tj}\left(\hat{\mu}_{tj} - \hat{\mu}_t\right)^2, \qquad (7)$$

which, when divided by c, is distributed as an *F* variable with degrees of freedom equal to J – 1 and

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

where

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

Thus, the transformed WJ statistics may be defined as

$$JWJ = \sum_{j=1}^{J} w_{tj}\left(T_{(Johnson)j}\right)^2 \qquad (8)$$

and

$$HWJ = \sum_{j=1}^{J} w_{tj}\left(T_{(Hall)j}\right)^2 . \qquad (9)$$

When Johnson's transformed WJ statistic (JWJ) and Hall's transformed WJ statistic (HWJ) are divided by c, they are also distributed as *F* variates with no change in degrees of freedom.

The WJ, JWJ, and HWJ statistics were used not only for the omnibus test, if one was required, but for the pairwise tests for each of the MCPs investigated.

Multiple Comparison Methods

The MCPs investigated, adopt stepwise testing for controlling the overall (familywise) rate of Type I error. Specifically, the MCPs examined were the: (a) Ryan (1960)-Welsch (1977) multiple range procedure, (b) Peritz (1970) procedure, (c) Shaffer (1986) sequentially rejective Bonferroni procedure, (d) Shaffer (1986) sequentially rejective Bonferroni procedure that begins with an omnibus test, (e) Hochberg (1988) step-up sequentially acceptive Bonferroni procedure, (f) multiple range procedure that begins with an omnibus test (see Shaffer 1979; 1986), and (g) Hayter (1986) two-stage modified least significant difference (LSD) procedure. These MCPs were previously investigated by Keselman, Lix, et al. (1998).

The Ryan (1960) and Welsch (1977) multiple range procedure begins by examining the J range, and steps down to examine successively smaller ranges only when a larger range test is declared significant. The designation q is used to denote this MCP. According to Ryan and Welsch, the overall rate of Type I error is controlled at α (when assumptions are satisfied) for a set of p (p = 2,…, J) means if each test is assessed for significance at a level equal to

$$\alpha_p = 1-(1-\alpha)^{p/J} \quad [2 \le p \le J-2],$$
$$\alpha_{J-1} = \alpha_J = \alpha.$$

The Peritz (1970) procedure follows the same step-down logic of the usual range procedure, but assesses significance with Newman (1939), Keuls (1952), and/or Ryan-Welsch critical values. This MCP is designated PER. Shaffer’s (1986) sequentially rejective Bonferroni procedure uses probability (*p*) values in assessing the pairwise hypotheses taking into account the number of hypotheses rejected at earlier stages in the sequence of testing in arriving at decisions regarding significance. The abbreviation for this MCP is SRB.

Shaffer’s (1986) sequentially rejective Bonferroni procedure begins with an omnibus test (i.e., WJ, JWJ, HWJ), and if rejected, assesses significance of the pairwise comparisons by taking into account the number of true pairwise hypotheses remaining given previous rejections. Because three omnibus statistics are being investigated, there are three SRB MCPs and they are designated as WJ/SRB, JWJ/SRB, and HWJ/SRB.

Hochberg’s (1988) step-up sequentially acceptive Bonferroni procedure uses the *p* values associated with the pairwise tests to arrive at accept-reject decisions; these are determined sequentially and hypotheses can be rejected by implication. Hochberg’s MCP is designated as HOCH. Another set of MCPs were based on the modified range procedure due to Shaffer (1979; 1986), which starts with an omnibus test and only upon rejection, moves on to test range hypotheses with Ryan-Welsch critical values, modifying the J-range critical value to one based on J-1 means. The abbreviations of these three (stage 1 omnibus) Shaffer MCPs are WJ/q, JWJ/q, and HWJ/q. Lastly, Hayter’s (1986) modified LSD begins with an omnibus test, which if rejected leads to the Stage 2 tests of the pairwise comparisons using a Studentized range critical value for J-1 means. The three MCPs based on Hayter’s method are designated: WJ/HAY, JWJ/HAY, and HWJ/HAY. Detailed descriptions of all the pairwise MCPs can be found in the original references.

Methodology

Seven pairwise MCPs were compared in terms of Type I error control under conditions of nonnormality and variance heterogeneity in one-way independent groups designs. Variables that were examined by Keselman, Lix, et al. (1998) were chosen for investigation. Eight variables were manipulated in the present study: (a) number of groups (3 and 6), (b) sample size (equal or not equal), (c) degree/pattern of variance heterogeneity [moderate and large/all (mostly) unequal and all but one equal], (d) pairing of groups sizes and variances, (e) type of nonnormal population distribution, (f) method of computing a test of symmetry, (g) percentage of trimming, and (h) type of heteroscedastic statistic.

One-way independent groups designs containing three and six groups to evaluate the effect of number of pairwise comparisons on Type I error were chosen for investigation. That is, for the former case, only three pairwise comparisons were tested, whereas, in the latter case, 15 pairwise comparisons were tested.

The sample sizes in each of the groups were either equal or unequal. When equal, C = 0, and when unequal, C = .163 and .327, where C denotes a coefficient of group size variation defined as

$$\left(\sum_j \left(n_j - \bar{n}\right)^2 / J\right)^{1/2} \Big/ \bar{n}, \text{ where } \bar{n}$$

is the average group size. When equal, group sizes were set at 20 in both the J = 3 and J = 6 designs. When unequal, and for the J = 3 design, the two cases of group size inequality were 16, 20, 24 (C = .163) and 12, 20, 28 (C = .327), while for the J = 6 design, the group sizes were 16, 16, 20, 20, 24, 24 (C = .163) and 12, 12, 20, 20, 28, 28 (C = .327).

Two patterns of variance heterogeneity were examined: (a) all (most) variances unequal (Pattern 1) and (b) all variances equal but one (Pattern 2). When J = 3, Pattern 1 was 1, 9, 16 and Pattern 2 was 1, 1, 16. The patterns for J = 6 were, respectively, 1, 1, 4, 9, 9, 16, and 1, 1, 1, 1, 1, 16.

Seven cases of group sizes and variances pairings were investigated. Group sizes were both equal and unequal and paired with equal and unequal variances. Specifically, the combinations were: (a) equal n_j; equal σ_j^2, (b/b') equal n_j; unequal σ_j^2, (c/c') unequal n_j; unequal σ_j^2 (positively paired), (d/d') unequal n_j; unequal σ_j^2 (negatively paired). The b/c/d notation represents the Pattern 1 variance conditions, whereas the b'/c'/d' notation represents the Pattern 2 variance conditions. Considering the group size and variance inequalities, there were a total of eleven combinations.

To examine distributional shape, four nonnormal distributions with varying degrees of skewness (γ_1) and kurtosis (γ_2) were chosen for investigation. A chi-square (χ^2) distribution and three g- and h-distributions (Hoaglin, 1985) were selected. Specifically, the four nonnormal distributions were: (a) $\chi^2_{(3)}$ distribution (γ_1 = 1.63, γ_2 = 4.00); (b) g = .5 and h = 0 distribution (γ_1 = 1.75, γ_2 = 8.9); (c) g = 1 and h = 0 distribution (γ_1 = 6.2, γ_2 = 114); and (d) g = .25 and h = .25 distribution (γ_1 and γ_2 undefined). The three g- and h- distributions are hereafter notated as (g = .5, h = 0), (g = 1, h = 0), and (g = .25, h = .25), respectively. These nonnormal distributions were selected because educational and psychological research data are typically skewed and/or heavy-tailed (Micceri, 1989; Wilcox, 1990).

To generate pseudorandom variates having a chi-square (χ^2) distribution with 3 degrees of freedom, three standard normal variates were squared and summed. The variates were transformed to $\chi^2_{(3)}$ variates having mean μ_{tj} (population trimmed mean) and σ_j^2 (see Hastings & Peacock, 1975, p. 46-51, for further details). To generate data from a g- and h-distribution, standard unit normal variables (Z) were converted to the random variable

$$X_{ij} = \frac{\exp\left(gZ_{ij}\right) - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right), \quad (10)$$

according to the values of g and h selected for investigation. μ_{tj} was subtracted from each observation. To obtain a distribution with standard deviation σ_j, each transformed X_{ij} (j = 1, …, J) was then multiplied by a value of σ_j. The standard deviation of a g- and h-distribution is not equal to one, and thus the values for the variances/standard deviations reflect the ratio of the variances/standard deviations between the groups (see Wilcox, 1994). Each population distribution was empirically generated and the indices of tail weight and symmetry were computed in order to determine whether the population trimmed mean used for centering should be based on symmetric or asymmetric (e.g., right tailed) trimming for the percentage of trimming cases investigated.

Three approaches to computing the test of symmetry were examined by calculating the indices (Q_1 and Q_2) within each group and then: (a) using a weighted mean of the indices across all groups to determine the type of trimming for every group (average estimate; see Othman et al. 2002); (b) using the value for each particular group to determine the type of trimming for that group (individual estimate), and (c) using a weighted mean of the indices across two groups to determine the type of trimming for the groups involved in each particular comparison (pairwise estimate). The test of symmetry based on pairwise estimates could not be applied to an omnibus test, so only the MCPs that do not require an omnibus test were considered for this approach. In addition, the pairing of groups had to be predetermined in order to compute the weighted mean of the indices across the two groups in each pairwise comparison and this prevented the use of the approach with the range MCPs. Thus, the third approach was applied to only the SRB and HOCH procedures. The three approaches to symmetric/asymmetric trimming were compared to always adopting symmetric trimming. The Q_1 and Q_2 indices determine whether symmetrically/asymmetrically trimmed means for each group were used in the pairwise MCPs. For those MCPs that require an omnibus test, the same approach to trimming (i.e., average estimate, individual estimate or symmetric trimming) was adopted for the omnibus and the pairwise tests.

The following combinations of symmetric and asymmetric trimming percentages were investigated: (a) either 10% symmetric or 20% asymmetric trimming (10/20), (b) either 15% symmetric or 30% asymmetric trimming (15/30), (c) either 20% symmetric or 40% asymmetric trimming (20/40), (d) either 10% symmetric or 10% asymmetric trimming (10/10), (e) either 15% symmetric or 15% asymmetric trimming (15/15), and (f) either 20% symmetric or 20% asymmetric trimming (20/20). As well, symmetrically trimming 10%, 15%, and 20% of the data was investigated. Hence, the various combinations of trimming percentages were chosen to evaluate whether there would be an optimal proportion of trimming.

Three heteroscedastic statistics were examined: (a) Welch-James statistic (WJ), (b) Johnson's (1978) transformation of WJ (JWJ), and (c) Hall's (1992) transformation of WJ (HWJ) (see Guo & Luh, 2000; Keselman et al. 2002; Luh & Guo, 1999). The seven pairwise MCPs were computed with each of the heteroscedastic statistics, resulting in a total of 21 pairwise MCPs.

Type I error rates were based on five thousand replications using a .05 level of significance for the complete null hypothesis.

Results

Bradley's (1978) liberal criterion of robustness to assess Type I error rates was chosen. That is, if an empirical estimate of Type I error ($\hat{\alpha}$) was contained within the interval of $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$, then the procedure was considered robust. For a significance level of .05, the interval is $.025 \leq \hat{\alpha} \leq .075$. If the Type I error was not contained in this interval, then a procedure was considered nonrobust for that particular condition. In the tables, bold entries correspond to these latter values.

Because of the large number of MCPs investigated and the form of assumption violations examined, only the mean Type I error rates (percentages), averaging across the eleven combinations of group sizes, and variances were tabled. Plus and minus symbols next to the tabled error rates are used to identify whether the minimum to maximum range of Type I error rates across the eleven combinations contained a conservative (-) value, a liberal (+) value, or both conservative and liberal (±) values. A conservative value is defined as an error rate below Bradley's lower limit (2.50%) and a liberal value is defined as an error rate above Bradley's upper limit (7.50%). Because of space considerations and the similar pattern of results for the chi-square and (g = .5, h = 0) distributions, only the latter are tabled.

J = 3

Tables 1 through 3 contain the summary percentages for the (g = .5, h = 0), (g = 1, h = 0), and (g = .25, h = .25) distributions, respectively. When the number of groups is equal to three, a few of the MCPs investigated are identical. Specifically, the Hayter (1986) two-stage and Shaffer (1986) sequentially rejective Bonferroni procedure that begins with an omnibus test are identical (denoted as WJ/*, JWJ/*, and HWJ/* in Tables 1 through 3). Additionally, the Ryan (1960)–Welsch (1977) and Peritz (1970) procedures are identical (denoted as q / PER in Tables 1 through 3).

g = .5 and h = 0 Distribution

When data were obtained from this particular nonnormal distribution, all MCPs were robust when preceded by the symmetry test with 10/10 symmetric/asymmetric trimming where the indices of tail weight and symmetry were averaged over all groups and under the 10%, 15%, and 20% symmetric trimming cases (see Table 1). The chi-square distribution had a similar pattern of results, however all MCPs were also robust under the 15/15 and 20/20 symmetric/asymmetric trimming where the indices were averaged over all groups. MCPs preceded by the test of symmetry generally had mean Type I error rates closer to the nominal 5% level compared to the strategy of always adopting symmetric trimming. For the symmetry test based on averaging (tail-weight and symmetry) indices across all groups, the mean error rates across robust MCPs were 4.83%, 4.80%, and 5.22% for the 10/10, 15/15, and 20/20 trimming cases, respectively and for the symmetry test based on the indices taken per group, the mean error rate across robust MCPs was 5.32% for the 10/10 trimming case. For the symmetric trimming conditions of 10%, 15%, and 20%, the mean error rates across MCPs were 4.75%, 4.68%, and 4.80%, respectively. In addition, the general pattern for MCPs preceded by a test for symmetry was for error rates to increase as the proportion of trimming increased (i.e., from 10/20 to 15/30 to 20/40 and from 10/10 to 15/15 to 20/20).

The MCPs based on the WJ statistic generally had more conservative error rates than the same MCPs based on the modified WJ statistics (i.e., JWJ and HWJ), when preceded by a test of symmetry, a pattern opposite to that observed for the symmetric trimming cases. For example, under the 10/10 trimming case preceded by the test of symmetry based on indices (tail weight and symmetry) averaged across all groups, the mean error rates for the MCPs based on the WJ, JWJ, and HWJ statistics were equal to 4.70%, 4.87%, and 4.91%, respectively. However, when adopting 20% symmetric trimming, the mean error rates across MCPs based on the WJ, JWJ, and HWJ statistics were equal to 4.94%, 4.73%, and 4.74%, respectively. For the chi-square distribution, regardless of whether the MCPs were preceded by a test of symmetry, the MCPs based on the JWJ and HWJ statistics generally had more conservative Type I error rates than the corresponding MCPs based on the WJ statistic.

The mean error rates for the SRB and HOCH procedures based on symmetric trimming were more conservative than when the MCPs were preceded by a test of symmetry. When the test of symmetry was based on individual group estimates of tail weight and symmetry, the MCP's mean error rates were highest, and decreased when the test was based on pairwise estimates and further decreased when the symmetry test was based on average estimates across groups (a result consistent with that obtained for the chi-square distribution). Noteworthy is that the error rates for the SRB and HOCH MCPs fell within Bradley's (1978) limits for the 10/10 trimming percentage regardless of the method of computing the test for symmetry; a result consistent with that obtained for the chi-square distribution. An optimal strategy is to use a test of symmetry with either pairwise estimates or average estimates across groups with either 10/10 or 15/15 symmetric/asymmetric trimming.

g = 1 and h = 0 Distribution

The use of the test of symmetry resulted in improved Type I error control when data were obtained from the (g = 1, h = 0) nonnormal distribution (see Table 2). That is, the MCPs with conservative and/or liberal error rates based on symmetric trimming became either robust or closer to Bradley's (1978) limits when preceded

Table 1. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 3; g = .5, h = 0 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
q / PER (WJ)	5.45+	7.29+	**9.56+**	4.07	4.50	5.00	**11.58+**	**19.96+**	**29.43+**	4.90	6.58+	**9.10+**
WJ / q (WJ)	6.37+	**8.50+**	**11.03+**	4.64	5.06	5.64	**13.28+**	**23.15+**	**33.42+**	5.57	**7.51+**	**10.43+**
WJ / * (WJ)	7.48+	**9.73+**	**12.42+**	5.70	6.10+	6.73+	**15.50+**	**26.68+**	**38.03+**	6.72+	**8.99+**	**12.33+**
SRB (WJ)	5.65+	7.39+	**9.50+**	4.46	4.73	5.26	**12.31+**	**21.63+**	**31.96+**	5.24	6.95+	**9.61+**
HOCH (WJ)	5.87+	**7.67+**	**9.83+**	4.62	4.90	5.43	**12.68+**	**22.17+**	**32.65+**	5.43	7.19+	**9.91+**
q / PER (JWJ)	5.72+	7.31+	**9.28+**	4.25	4.43	4.87	**12.58+**	**21.57+**	**31.46+**	5.05	6.67+	**9.30+**
JWJ / q (JWJ)	6.55+	**8.48+**	**10.69+**	4.80	5.09	5.44	**14.28+**	**24.71+**	**35.19+**	5.70+	**7.63+**	**10.56+**
JWJ / * (JWJ)	**7.66+**	**9.71+**	**12.15+**	5.83	6.15+	6.52+	**16.62+**	**28.38+**	**40.00+**	6.80+	**9.14+**	**12.51+**
SRB (JWJ)	5.86+	7.43+	**9.43+**	4.66	4.72	5.08	**13.53+**	**23.41+**	**34.29+**	5.36	7.13+	**9.80+**
HOCH (JWJ)	6.06+	**7.68+**	**9.74+**	4.82	4.89	5.25	**13.87+**	**23.97+**	**35.03+**	5.54	7.34+	**10.08+**
q / PER (HWJ)	5.75+	7.36+	**9.37+**	4.29	4.46	4.89	**12.62+**	**21.63+**	**31.53+**	5.10	6.71+	**9.33+**
HWJ / q (HWJ)	6.58+	**8.52+**	**10.79+**	4.83	5.11	5.46	**14.31+**	**24.77+**	**35.30+**	5.75+	**7.67+**	**10.58+**
HWJ / * (HWJ)	**7.69+**	**9.76+**	**12.25+**	5.87	6.17+	6.54+	**16.66+**	**28.44+**	**40.12+**	6.87+	**9.17+**	**12.53+**
SRB (HWJ)	5.90+	7.49+	**9.55+**	4.70	4.76	5.10	**13.58+**	**23.47+**	**34.40+**	5.42	7.17+	**9.83+**
HOCH (HWJ)	6.09+	**7.73+**	**9.86+**	4.86	4.92	5.27	**13.90+**	**24.01+**	**35.12+**	5.59	7.39+	**10.10+**
Pairwise Estimate												
SRB (WJ)	6.39+	**9.03+**	**12.41+**	4.46	4.91	5.70						
HOCH (WJ)	6.59+	**9.24+**	**12.68+**	4.62	5.08	5.86						
SRB (JWJ)	6.85+	**9.31+**	**12.53+**	4.68	4.96	5.62						
HOCH (JWJ)	7.00+	**9.51+**	**12.79+**	4.83	5.10	5.77						
SRB (HWJ)	6.89+	**9.37+**	**12.63+**	4.73	4.99	5.64						
HOCH(HWJ)	7.03+	**9.57+**	**12.88+**	4.88	5.13	5.79						
No Preliminary Test (symmetric trimming)												
	10	15	20									
q / PER (WJ)	4.24	4.26	4.43									
WJ / q (WJ)	4.65	4.65	4.80									
WJ / * (WJ)	5.60	5.59	5.84									
SRB (WJ)	4.62	4.65	4.75									
HOCH (WJ)	4.74	4.77	4.87									
q / PER (JWJ)	4.23	4.09	4.29									
JWJ / q (JWJ)	4.54	4.47	4.54									
JWJ / * (JWJ)	5.52	5.43	5.56									
SRB (JWJ)	4.55	4.46	4.57									
HOCH (JWJ)	4.71	4.58	4.69									
q / PER (HWJ)	4.26	4.11	4.30									
HWJ / q (HWJ)	4.60	4.50	4.55									
HWJ / * (HWJ)	5.59	5.47	5.57									
SRB (HWJ)	4.61	4.49	4.59									
HOCH (HWJ)	4.77	4.61	4.71									

Notes: 10/20 = 10% symmetric/20% asymmetric trimming; 15/30 = 15% symmetric/30% asymmetric trimming; 20/40 = 20% symmetric/40% asymmetric trimming; 10/10 = 10% symmetric/10% asymmetric trimming; 15/15 = 15% symmetric/15% asymmetric trimming; 20/20 = 20% symmetric/20% asymmetric trimming; q/PER indicates that q and Peritz procedures are equivalent; /* indicates that the SRB and Hayter procedures are equivalent; HOCH is the Hochberg procedure; 10 = 10% symmetric trimming; 15 = 15% symmetric trimming; 20 = 20% symmetric trimming; bold entries indicate values that exceeded Bradley's (1978) lower and upper limits; + indicates a liberal value, - indicates a conservative value, and ± indicates both conservative and liberal values in the minimum to maximum range of error rates.

Table 2. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 3; g = 1, h = 0 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
q / PER (WJ)	4.46	4.55	4.83	4.33	4.09	4.31	6.77+	**10.28+**	**14.18+**	4.25	4.47	5.34
WJ / q (WJ)	5.01	5.14	5.42	4.97	4.78	4.84	**7.55+**	**11.35+**	**15.64+**	4.82	5.14	6.06+
WJ / * (WJ)	6.26+	6.31+	6.55+	6.46+	6.12+	6.06+	**9.11+**	**13.36+**	**18.05+**	6.16+	6.46+	7.46+
SRB (WJ)	5.06-	5.04	5.22	5.30±	4.85-	4.93-	7.42+	**10.99+**	**15.02+**	5.05	5.14	5.89+
HOCH (WJ)	5.17-	5.18	5.38	5.42±	4.97-	5.03-	**7.63+**	**11.32+**	**15.49+**	5.15	5.29	6.06+
q / PER (JWJ)	4.42	4.22	4.29	4.74	4.18	4.20	7.24+	**10.92+**	**14.85+**	4.61	4.61	5.33
JWJ / q (JWJ)	5.01	4.78	4.78	5.23	4.87	4.76	**8.03+**	**11.96+**	**16.08+**	5.07	5.30	6.12+
JWJ / * (JWJ)	6.22+	5.95+	6.02+	6.70+	6.22+	5.94+	**9.66+**	**14.12+**	**18.68+**	6.43	6.62+	7.48+
SRB (JWJ)	4.97-	4.81-	4.90-	5.53	4.89	4.78-	**7.94+**	**11.82+**	**15.92+**	5.25	5.15	5.89+
HOCH (JWJ)	5.11-	4.92-	5.04-	5.67	5.01	4.91-	**8.12+**	**12.16+**	**16.36+**	5.39	5.28	6.06+
q / PER (HWJ)	4.43	4.24	4.35	4.83	4.23	4.21	7.26+	**10.96+**	**14.92+**	4.70	4.66	5.34
HWJ / q (HWJ)	5.03	4.82	4.86	5.33	4.90	4.78	**8.06+**	**12.03+**	**16.17+**	5.15	5.34+	6.14+
HWJ / * (HWJ)	6.24+	5.98+	6.11+	6.81+	6.25+	5.97+	**9.69+**	**14.15+**	**18.75+**	6.53+	6.66+	7.50+
SRB (HWJ)	4.99-	4.84-	4.97-	5.64	4.93	4.80-	**7.97+**	**11.86+**	**16.00+**	5.36	5.20	5.90+
HOCH (HWJ)	5.14-	4.95-	5.11-	5.78	5.05	4.94-	**8.16+**	**12.20+**	**16.44+**	5.51	5.32	6.09+
Pairwise Estimate												
SRB (WJ)	5.22	5.39+	5.82+	5.25±	4.80-	4.93						
HOCH (WJ)	5.33	5.53+	5.97+	5.35±	4.91-	5.02						
SRB (JWJ)	5.31	5.31±	5.62±	5.50	4.84	4.81-						
HOCH (JWJ)	5.43	5.41±	5.74±	5.62	4.95	4.93-						
SRB (HWJ)	5.33	5.34±	5.68±	5.61	4.89	4.83-						
HOCH(HWJ)	5.46	5.44±	5.81±	5.74	5.00	4.96-						
No Preliminary Test (symmetric trimming)												
	10	15	20									
q / PER (WJ)	4.61	4.41	4.60-									
WJ / q (WJ)	5.22	5.00	5.00									
WJ / * (WJ)	6.64+	6.31+	6.31+									
SRB (WJ)	5.57±	5.24-	5.31-									
HOCH (WJ)	5.66±	5.36-	5.40±									
q / PER (JWJ)	4.80	4.41	4.50									
JWJ / q (JWJ)	5.24	4.94	4.89									
JWJ / * (JWJ)	6.66+	6.18+	6.09+									
SRB (JWJ)	5.59	5.11	5.15-									
HOCH (JWJ)	5.71+	5.23	5.24									
q / PER (HWJ)	4.91	4.45	4.52									
HWJ / q (HWJ)	5.34	5.01	4.91									
HWJ / * (HWJ)	6.78+	6.24+	6.12+									
SRB (HWJ)	5.68+	5.17	5.17-									
HOCH (HWJ)	5.82+	5.30	5.26									

Note. See note from Table 1

Table 3. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 3; g = .25, h = .25 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
q / PER (WJ)	4.72	6.11+	**7.64+**	3.66	4.11-	4.69	6.62+	**11.73+**	**17.69+**	3.36-	4.49	6.38+
WJ / q (WJ)	5.17	6.83+	**8.62+**	3.94	4.48	5.09	7.28+	**12.92+**	**19.32+**	3.71	4.95	6.88+
WJ / * (WJ)	6.11+	**7.92+**	**9.93+**	4.77	5.40	6.07+	**8.63+**	**15.25+**	**22.51+**	4.47	5.90	**8.23+**
SRB (WJ)	4.92	6.29+	**7.86+**	3.91	4.35	4.88	7.15+	**13.03+**	**19.75+**	3.59	4.82	6.83+
HOCH (WJ)	5.07	6.45+	**8.05+**	4.03	4.48	5.03	7.32+	**13.30+**	**20.12+**	3.70	4.95	7.03+
q / PER (JWJ)	5.95+	**7.87+**	**10.06+**	4.47	4.81	5.47	**8.20+**	**15.19+**	**22.75+**	3.95	5.29	**7.71+**
JWJ / q (JWJ)	6.49+	**8.57+**	**10.90+**	4.84	5.22	5.95+	**8.85+**	**16.47+**	**24.37+**	4.23	5.72	**8.18+**
JWJ / * (JWJ)	**7.57+**	**9.68+**	**12.16+**	5.84	6.22+	6.99+	**10.54+**	**19.24+**	**28.16+**	5.11	6.77+	**9.75+**
SRB (JWJ)	6.31+	**8.11+**	**10.40+**	4.79	5.10	5.74+	**9.06+**	**16.97+**	**25.65+**	4.15	5.64	**8.40+**
HOCH (JWJ)	6.45+	**8.29+**	**10.58+**	4.92	5.25	5.89+	**9.25+**	**17.21+**	**25.99+**	4.28	5.80	**8.61+**
q / PER (HWJ)	6.02+	**7.97+**	**10.21+**	4.51	4.85	5.53	**8.29+**	**15.37+**	**23.00+**	3.98	5.33	**7.80+**
HWJ / q (HWJ)	6.57+	**8.66+**	**11.02+**	4.88	5.26	5.98+	**8.96+**	**16.65+**	**24.60+**	4.27	5.76+	**8.26+**
HWJ / * (HWJ)	**7.67+**	**9.80+**	**12.28+**	5.91	6.27+	7.02+	**10.66+**	**19.46+**	**28.42+**	5.16	6.82+	**9.84+**
SRB (HWJ)	6.38+	**8.23+**	**10.57+**	4.85	5.17	5.80+	**9.17+**	**17.22+**	**26.01+**	4.19	5.69	**8.51+**
HOCH (HWJ)	6.52+	**8.41+**	**10.74+**	4.97	5.32	5.95+	**9.36+**	**17.48+**	**26.32+**	4.33	5.85	**8.70+**
Pairwise Estimate												
SRB (WJ)	5.33	7.24+	**9.42+**	3.99	4.55	5.20						
HOCH (WJ)	5.44	7.36+	**9.57+**	4.10	4.66	5.31						
SRB (JWJ)	6.88+	**9.71+**	**13.14+**	4.93	5.42	6.24+						
HOCH (JWJ)	7.01+	**9.86+**	**13.28+**	5.05	5.54	6.36+						
SRB (HWJ)	6.98+	**9.86+**	**13.45+**	4.99	5.47	6.31+						
HOCH(HWJ)	7.11+	**10.00+**	**13.58+**	5.10	5.60	6.42+						
No Preliminary Test (symmetric trimming)												
	10	15	20									
q / PER (WJ)	3.42-	3.50-	3.72									
WJ / q (WJ)	3.64	3.82	3.98									
WJ / * (WJ)	4.42	4.62	4.84									
SRB (WJ)	3.61	3.72	3.94									
HOCH (WJ)	3.72	3.81	4.05									
q / PER (JWJ)	4.07	3.81	3.84									
JWJ / q (JWJ)	4.39	4.15	4.11									
JWJ / * (JWJ)	5.32	5.00	4.98									
SRB (JWJ)	4.29	4.04	4.03									
HOCH (JWJ)	4.43	4.16	4.14									
q / PER (HWJ)	4.10	3.83	3.85									
HWJ / q (HWJ)	4.44	4.17	4.11									
HWJ / * (HWJ)	5.37	5.03	5.00									
SRB (HWJ)	4.34	4.06	4.05									
HOCH (HWJ)	4.47	4.18	4.15									

Note. See note from Table 1

by a test of symmetry, particularly for the MCPs based on the modified WJ statistic (i.e., JWJ or HWJ).

Specifically, all the MCPs based on the 10/10 trimming case with the test of symmetry based on individual group estimates of tail weight and symmetry had rates of Type I error within Bradley's (1978) limits except the Hayter (1986) two-stage and Shaffer (1986) sequentially rejective Bonferroni procedure that begins with an omnibus test utilizing the WJ statistic (denoted WJ/*) and the HWJ statistic (denoted HWJ/*) with liberal rates of 8.28% and 7.52%, respectively. Interestingly, this particular condition had the largest number of MCPs that fell within Bradley's lower and upper limits. The mean error rates across robust MCPs based on the JWJ and HWJ heteroscedastic statistics for the 10/10 and 15/15 trimming cases were 5.34% and 4.76%, respectively for the test of symmetry based on average estimates across groups and 5.27% and 5.07%, respectively for the test of symmetry based on individual group estimates.

The MCPs based on the WJ statistic generally had more conservative error rates than the same MCPs based on the modified WJ statistic (i.e., JWJ and HWJ) when preceded by a test of symmetry except under the 10/20, 15/30, 20/40, and 20/20 trimming cases for the test of symmetry based on average estimates across groups where the opposite pattern was observed (i.e., WJ based MCPs had higher mean error rates). Additionally, Type I error rates for the MCPs tended to decrease with an increase in the percentage of trimming (i.e., from 10/20 to 15/30 to 20/40 and from 10/10 to 15/15 to 20/20), except for the MCPs preceded by a test of symmetry based on individual group estimates where the pattern was reversed, that is, error rates tended to increase as the proportion of trimming increased.

The mean error rates for the SRB and HOCH procedures indicate that an optimal strategy is to use a test of symmetry based either on indices of tail weight and symmetry averaged across the pairwise comparisons or averaged across all groups with 15/15 symmetric/asymmetric trimming (i.e., mean error rates closer to the nominal 5% level). A result consistent with the (g = .5, h = 0) distribution.

g = .25 and h = .25 Distribution

When nonnormal data were obtained from the (g = .25, h = .25) distribution, the use of the symmetry test based on the individual group indices resulted in all MCPs having liberal Type I error rates, for the 10/20, 15/30, 20/40, and 20/20 trimming cases (see Table 3). However, improved Type I error control was obtained when the test of symmetry was based on indices averaged across all groups or averaged across the two groups defining the pairwise comparison. Interestingly, all MCPs had rates below Bradley's (1978) upper limit for the 10/10 trimming case when preceded by the preliminary test of symmetry, regardless of the method of computing the test. In addition, all MCPs had rates of Type I error below Bradley's upper limit when always adopting 10%, 15%, or 20% symmetric trimming.

The use of the averaged over all groups tail weight and symmetry indices resulted in Type I error rates closer to the nominal level compared to always adopting symmetric trimming. For example, the 10/10 and 15/15 trimming cases had mean rates of Type I error across non-liberal MCPS equal to 4.69% and 4.91%, respectively, whereas the 10%, 15%, and 20% symmetric trimming cases had mean error rates, across MCPs equal to 4.27%, 4.13%, and 4.19%, respectively.

The MCPs based on the JWJ or HWJ heteroscedastic statistics had rates of Type I error closer to the nominal level compared to MCPs based on the WJ statistic. For example, (a) with the symmetry test based on average estimates across groups, the mean rates of Type I error across all five MCPs when based on the WJ, JWJ, and HWJ test statistics for the 10/10 trimming condition equaled 4.06%, 4.97%, and 5.02%, respectively, (b) with the symmetry test based on individual group estimates, the mean error rates for the 10/10 trimming condition equaled 3.77%, 4.34%, and 4.39%, respectively, and (c) with symmetric trimming, the mean rates for 20% trimming equaled 4.11%, 4.22%, and 4.23%, respectively.

Mean rates of Type I error for the SRB and HOCH procedures, when preceded by a test of symmetry based on tail weight and symmetry estimates from the two groups forming the pairwise comparison, were higher than when the symmetry test was based on the average estimate of the indices across all groups for a given trimming condition, with the highest rates occurring when individual group indices of tail weight and symmetry were used. The optimal level of trimming occurs under the 10/10 symmetric/asymmetric trimming case when the MPCs were based on the JWJ or HWJ statistics (i.e., mean error rates closest to the nominal 5% level).

J = 6

Tables 4 through 6 contain the summary percentages of Type I error for the MPCs for the (g = .5, h = 0), (g = 1, h = 0), and (g = .25, h = .25) distributions, respectively. The SRB and HOCH procedures had identical error rates across the eleven pairings of groups sizes and variances, thus they have been combined into one row in the tables (denoted as SRB/HOCH).

g = .5 and h = 0 Distribution

All MCPs had Type I error rates below Bradley's (1978) upper limit (i.e., 7.50%) when based on the test of symmetry with indices of tail weight and symmetry averaged over groups except Hayter's (1986) two-stage and Shaffer's (1986) sequentially rejective Bonferroni procedure that begins with an omnibus test (i.e., WJ/HAY, JWJ/HAY, HWJ/HAY, WJ/SRB, JWJ/SRB, HWJ/SRB) under the 20/40 symmetric/asymmetric trimming case (see Table 4). Unlike when J = 3, some MCPs had error rates below Bradley's lower limit (i.e., 2.50%). Specifically, the effected MCPs were the range procedures [(PER (WJ), q (WJ), WJ/q, PER (JWJ), q (JWJ), JWJ/q, PER (HWJ), q (HWJ), and HWJ/q)] when they were based on the test of symmetry using an average estimate of tail weight and symmetry across all of the groups and symmetric trimming (a result consistent with that obtained for the chi-square distribution).

The mean error rate across MCPs for the 10/20, 15/30, 10/10, 15/15, and 20/20 trimming cases when preceded by the test of symmetry based on average estimates of tail weight and symmetry across all groups was equal to 3.57%, 3.94%, 3.39%, 3.31%, and 3.39%, respectively and for the 10/10 trimming case, when preceded by the test of symmetry based on individual group estimates of tail weight and symmetry, the mean error rate was equal to 4.17%. Thus, an optimal strategy and level of trimming is to use 10/10 symmetric/asymmetric trimming with the test of symmetry based on individual group estimates (a result consistent with that obtained for the chi-square distribution).

The pattern of error rates differed with the type of heteroscedastic statistic. Error rates tended to increase as the proportion of trimming increased for the 10/20, 15/30 and 20/40 trimming cases and for the 10/10, 15/15, and 20/20 trimming cases. However, MCPs based on the JWJ and HWJ statistics, had rates that tended to decrease as the proportion of trimming increased for the 10/10, 15/15, and 20/20 conditions with the test of symmetry based on average group estimates (a result consistent with that obtained for the chi-square distribution).

The MCPs based on the WJ statistic generally had more conservative rates of error than the same MCPs based on the modified WJ statistics (i.e., JWJ and HWJ), when preceded by a test of symmetry based on individual group estimates or pairwise estimates of tail weight and symmetry, a pattern opposite to that observed for the symmetry test based on average estimates across groups (except under the 10/10 trimming case) or when always adopting symmetric trimming.

For example, under the 10/10 trimming case with the test of symmetry based on indices (tail weight and symmetry) for individual groups, the mean error rates for the MCPs based on the WJ, JWJ, and HWJ statistics were equal to 4.07%, 4.20%, and 4.25%, respectively and when based on average indices across groups, the mean error rates were equal to 3.29%, 3.43%, and 3.46%, respectively. On the other hand, when adopting 20% symmetric trimming the mean error rates across MCPs based on the WJ, JWJ, and HWJ statistics were equal to 3.63%, 3.45%, and 3.47%, respectively. This pattern is consistent with the results obtained for the chi-square distribution.

Table 4. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 6; g =.5, h = 0 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
PER (WJ)	3.05-	3.59-	4.43-	2.67-	2.77-	2.97-	**9.35+**	**17.72+**	**27.66+**	3.29-	4.83±	7.00+
q (WJ)	2.97-	3.47-	4.23-	2.58-	2.69-	2.84-	**9.06+**	**17.20+**	**26.59+**	3.18-	4.70±	6.64+
WJ / q (WJ)	2.61-	3.13-	4.04-	**2.26-**	**2.41-**	2.50-	**8.90+**	**17.70+**	**27.83+**	2.89-	4.34±	6.41+
WJ / SRB (WJ)	4.33	4.84	6.00+	3.97	4.06	4.25	**14.31+**	**28.11+**	**44.63+**	4.95	7.20+	**11.11+**
WJ / HAY (WJ)	5.13	5.69	7.18+	4.67	4.73	5.01	**16.69+**	**32.57+**	**50.30+**	5.79	**8.41+**	**13.06+**
SRB/HOCH (WJ)	3.84	4.19	5.07	3.57	3.53	3.79	**12.09+**	**23.83+**	**38.70+**	4.33	6.19	**9.42+**
PER (JWJ)	2.86-	3.21-	3.98-	2.82-	2.71-	2.67-	**10.74+**	**20.23+**	**30.90+**	3.50-	4.98±	7.36+
q (JWJ)	2.75-	3.08-	3.78-	2.75-	2.63-	2.56-	**10.45+**	**19.63+**	**29.87+**	3.39-	4.83±	7.02+
JWJ / q (JWJ)	2.52-	2.80-	3.45-	**2.39-**	**2.29-**	**2.29-**	**10.32+**	**20.13+**	**31.07+**	2.99-	4.45+	6.73+
JWJ / SRB (JWJ)	4.22	4.45	5.49+	4.11	3.95	3.99	**16.08+**	**31.60+**	**49.35+**	4.99	7.29+	**11.52+**
JWJ / HAY (JWJ)	5.01	5.30	6.51+	4.76	4.57	4.71	**18.72+**	**36.12+**	**54.93+**	5.83	**8.56+**	**13.53+**
SRB/HOCH (JWJ)	3.70	3.91	4.88	3.74	3.46	3.54	**13.87+**	**27.27+**	**43.70+**	4.48	6.37	**9.83+**
PER (HWJ)	2.91-	3.27-	4.09-	2.85-	2.73-	2.70-	**10.80+**	**20.32+**	**31.03+**	3.55-	5.03+	7.40+
q (HWJ)	2.78-	3.14-	3.91-	2.77-	2.66-	2.58-	**10.51+**	**19.70+**	**30.01+**	3.43-	4.85±	7.04+
HWJ / q (HWJ)	2.56-	2.85-	3.58-	**2.41-**	**2.33-**	**2.31-**	**10.38+**	**20.21+**	**31.22+**	3.03-	4.48+	6.78+
HWJ / SRB (HWJ)	4.28	4.55	5.68+	4.15	3.99	4.03	**16.18+**	**31.76+**	**49.57+**	5.04	7.35+	**11.58+**
HWJ / HAY (HWJ)	5.06	5.40	6.71+	4.80	4.61	4.76	**18.82+**	**36.24+**	**55.13+**	5.89	**8.61+**	**13.59+**
SRB/HOCH (HWJ)	3.75	4.00	5.03	3.79	3.49	3.58	**13.97+**	**27.42+**	**43.96+**	4.54	6.42	**9.89+**
Pairwise Estimate												
SRB/HOCH (WJ)	4.58	6.21+	**8.98+**	3.46	3.72	4.35						
SRB/HOCH (JWJ)	4.97	6.33+	**9.16+**	3.77	3.75	4.23						
SRB/HOCH (HWJ)	5.01	6.42+	**9.29+**	3.81	3.78	4.25						
No Preliminary Test (symmetric trimming)												
	10	15	20									
PER (WJ)	2.69-	2.81-	2.89-									
q (WJ)	2.60-	2.71-	2.69-									
WJ / q (WJ)	**2.36-**	**2.38-**	**2.43-**									
WJ / SRB (WJ)	4.14	4.23	4.53									
WJ / HAY (WJ)	4.81	4.86	5.21									
SRB/HOCH (WJ)	3.66	3.77	4.01									
PER (JWJ)	2.85-	2.77-	2.81-									
q (JWJ)	2.76-	2.67-	2.58-									
JWJ / q (JWJ)	**2.35-**	**2.28-**	**2.27-**									
JWJ / SRB (JWJ)	4.11	4.06	4.25									
JWJ / HAY (JWJ)	4.72	4.67	4.97									
SRB/HOCH (JWJ)	3.78	3.62	3.81									
PER (HWJ)	2.90-	2.79-	2.82-									
q (HWJ)	2.80-	2.70-	2.60-									
HWJ / q (HWJ)	**2.38-**	**2.30-**	**2.28-**									
HWJ / SRB (HWJ)	4.19	4.08	4.27									
HWJ / HAY (HWJ)	4.77	4.70	5.00									
SRB/HOCH (HWJ)	3.81	3.67	3.85									

Notes: 10/20 = 10% symmetric/20% asymmetric trimming; 15/30 = 15% symmetric/30% asymmetric trimming; 20/40 = 20% symmetric/40% asymmetric trimming; 10/10 = 10% symmetric/10% asymmetric trimming; 15/15 = 15% symmetric/15% asymmetric trimming; 20/20 = 20% symmetric/20% asymmetric trimming; PER is the Peritz procedure; HAY is the Hayter procedure; SRB/HOCH indicates that SRB and Hochberg procedures had equivalent rates; 10 = 10% symmetric trimming; 15 = 15% symmetric trimming; 20 = 20% symmetric trimming; bold entries indicate values that exceeded Bradley's (1978) lower and upper limits; + indicates a liberal value, **-** indicates a conservative value, and ± indicates both conservative and liberal values in the minimum to maximum range of error rates.

Table 5. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 6; g =1, h=0 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
PER (WJ)	**2.43-**	**2.43-**	2.53-	**2.39-**	**2.27-**	**2.43-**	5.05	**8.49+**	**12.78+**	**2.35-**	2.74-	3.45-
q (WJ)	**2.33-**	**2.32-**	**2.38-**	**2.32-**	**2.23-**	**2.32-**	4.88	**8.16+**	**12.13+**	**2.29-**	2.67-	3.30-
WJ / q (WJ)	**2.17-**	**2.06-**	**2.17-**	**2.18-**	**2.11-**	**2.17-**	4.58	**7.91+**	**12.02+**	**2.14-**	2.55-	3.15-
WJ / SRB (WJ)	4.46-	3.99	4.28	4.77±	4.36-	4.45-	**7.85+**	**12.51+**	**18.40+**	4.52±	4.72	5.88+
WJ / HAY (WJ)	5.11+	4.71	5.07	5.44±	5.01	5.08±	**9.16+**	**14.66+**	**21.50+**	5.21+	5.59	6.92+
SRB/HOCH (WJ)	3.90-	3.53-	3.68-	4.14±	3.84-	3.88-	6.76+	**10.57+**	**15.68+**	3.87-	4.07-	4.96
PER (JWJ)	**2.36-**	**1.89-**	**1.88-**	2.93-	2.50-	**2.34-**	5.84+	**9.52+**	**13.99+**	2.81-	3.00-	3.50-
q (JWJ)	**2.25-**	**1.82-**	**1.76-**	2.85-	**2.44-**	**2.23-**	5.62+	**9.19+**	**13.32+**	2.73-	2.93-	3.30-
JWJ / q (JWJ)	**2.10-**	**1.66-**	**1.56-**	2.51-	**2.26-**	**2.08-**	5.31+	**8.99+**	**13.18+**	**2.40-**	2.67-	3.19-
JWJ / SRB (JWJ)	4.28-	3.56-	3.65-	5.08+	4.37-	4.24-	**8.80+**	**13.92+**	**19.98+**	4.76	4.77	5.78+
JWJ / HAY (JWJ)	4.93-	4.13-	4.24-	5.86+	5.07	4.91-	**10.20+**	**16.19+**	**22.90+**	5.54	5.64	6.90+
SRB/HOCH (JWJ)	3.79-	3.10-	3.22-	4.72±	3.84-	3.77-	**7.82+**	**12.02+**	**17.40+**	4.32-	4.11-	4.98
PER (HWJ)	**2.37-**	**1.93-**	**1.94-**	3.00-	2.52-	**2.35-**	5.89+	**9.58+**	**14.10+**	2.87-	3.03-	3.52-
q (HWJ)	**2.27-**	**1.85-**	**1.82-**	2.92-	**2.47-**	**2.25-**	5.68+	**9.27+**	**13.43+**	2.80-	2.96-	3.32-
HWJ / q (HWJ)	**2.11-**	**1.68-**	**1.63-**	2.58-	**2.29-**	**2.09-**	5.36+	**9.04+**	**13.28+**	**2.47-**	2.70-	3.22-
HWJ / SRB (HWJ)	4.30-	3.59-	3.74-	5.21+	4.41-	4.26-	**8.87+**	**14.03+**	**20.11+**	4.88	4.82	5.81+
HWJ / HAY (HWJ)	4.96-	4.17-	4.36-	6.01+	5.12	4.93-	**10.27+**	**16.29+**	**23.07+**	5.68+	5.70	6.94+
SRB/HOCH (HWJ)	3.80-	3.13-	3.31-	4.83±	3.88-	3.77-	**7.87+**	**12.09+**	**17.51+**	4.41-	4.16-	5.00
Pairwise Estimate												
SRB/HOCH (WJ)	3.99-	3.88-	4.30-	3.95±	3.65-	3.83-						
SRB/HOCH (JWJ)	4.10-	3.60-	3.99-	4.55±	3.73-	3.78-						
SRB/HOCH (HWJ)	4.12-	3.64-	4.07-	4.66±	3.77-	3.79-						
No Preliminary Test (symmetric trimming)												
	10	15	20									
PER (WJ)	2.52-	**2.44-**	2.49-									
q (WJ)	**2.43-**	**2.37-**	**2.34-**									
WJ / q (WJ)	**2.28-**	**2.18-**	**2.17-**									
WJ / SRB (WJ)	4.84±	4.53-	4.68-									
WJ / HAY (WJ)	5.51±	5.14±	5.42+									
SRB/HOCH (WJ)	4.23±	3.95-	4.12-									
PER (JWJ)	3.00-	2.61-	2.54-									
q (JWJ)	2.90-	2.52-	**2.37-**									
JWJ / q (JWJ)	2.52-	**2.25-**	**2.10-**									
JWJ / SRB (JWJ)	5.01+	4.37-	4.48-									
JWJ / HAY (JWJ)	5.78+	5.08	5.22+									
SRB/HOCH (JWJ)	4.65±	3.99-	4.03-									
PER (HWJ)	3.07-	2.65-	2.56-									
q (HWJ)	2.96-	2.56-	**2.37-**									
HWJ / q (HWJ)	2.60-	**2.28-**	**2.12-**									
HWJ / SRB (HWJ)	5.13+	4.45-	4.53									
HWJ / HAY (HWJ)	5.92+	5.15	5.25+									
SRB/HOCH (HWJ)	4.78±	4.05-	4.07-									

Note. See note from Table 4

Table 6. Summary Percentages of Type I Error for Multiple Comparison Procedures (J = 6; g =.25, h=.25 Distribution)

	Average Estimate						**Individual Estimate**					
	10/20	15/30	20/40	10/10	15/15	20/20	10/20	15/30	20/40	10/10	15/15	20/20
PER (WJ)	2.57-	3.25-	3.99-	**2.17-**	**2.43-**	2.62-	4.84	**10.70+**	**17.47+**	**2.16-**	3.12-	4.55
q (WJ)	**2.48-**	3.13-	3.79-	**2.08-**	**2.33-**	**2.47-**	4.66	**10.33+**	**16.82+**	**2.07-**	3.00-	4.30
WJ / q (WJ)	**2.16-**	2.85-	3.60-	**1.74-**	**1.96-**	**2.17-**	4.25	**9.86+**	**16.52+**	**1.69-**	2.52-	3.87
WJ / SRB (WJ)	3.57	4.46	5.55+	3.04	3.37	3.81	7.14+	**16.38+**	**27.48+**	3.00-	4.42	6.81+
WJ / HAY (WJ)	4.21	5.29	6.67+	3.58	3.97	4.54	**8.25+**	**18.37+**	**30.36+**	3.45	5.13	**7.91+**
SRB/HOCH (WJ)	3.04	3.83	4.71	2.66-	2.95-	3.31-	6.40+	**14.60+**	**24.66+**	2.67-	3.93	6.11
PER (JWJ)	3.74-	4.79±	6.06+	3.00-	3.16-	3.29-	7.08+	**16.04+**	**25.79+**	2.63-	3.96-	6.44+
q (JWJ)	3.63-	4.65±	5.81+	2.90-	3.05-	3.11-	6.88+	**15.71+**	**25.15+**	2.54-	3.83-	6.12+
JWJ / q (JWJ)	3.16-	4.24-	5.56+	**2.44-**	2.57-	2.72-	6.30+	**14.98+**	**24.86+**	**2.11-**	3.28-	5.52+
JWJ / SRB (JWJ)	5.10	6.36+	**7.95+**	4.25	4.32	4.58	**10.40+**	**24.27+**	**39.41+**	3.69	5.66	**9.44+**
JWJ / HAY (JWJ)	5.85	7.26+	**9.05+**	4.89	4.96	5.35	**11.63+**	**26.25+**	**41.98+**	4.20	6.40+	**10.61+**
SRB/HOCH (JWJ)	4.56	5.65+	7.14+	3.85	3.81	4.12	**9.58+**	**22.47+**	**36.87+**	3.36	5.19	**8.81+**
PER (HWJ)	3.80-	4.92+	6.24+	3.04-	3.21-	3.33-	7.25+	**16.46+**	**26.33+**	2.68-	4.02-	6.58+
q (HWJ)	3.69-	4.76±	5.99+	2.93-	3.10-	3.17-	7.04+	**16.08+**	**25.70+**	2.59-	3.90-	6.25+
HWJ / q (HWJ)	3.24-	4.33-	5.74+	**2.49-**	2.63-	2.79-	6.42+	**15.35+**	**25.38+**	**2.13-**	3.34-	5.64+
HWJ / SRB (HWJ)	5.23	6.51+	**8.17+**	4.34	4.39	4.66	**10.65+**	**24.80+**	**40.10+**	3.73	5.74	**9.62+**
HWJ / HAY (HWJ)	5.70+	7.39+	**9.26+**	4.98	5.03	5.46	**11.86+**	**26.74+**	**42.66+**	4.24	6.50+	**10.81+**
SRB/HOCH (HWJ)	4.66	5.80+	7.35+	3.92	3.89	4.19	**9.82+**	**23.13+**	**37.70+**	3.42	5.31	**9.06+**
Pairwise Estimate												
SRB/HOCH (WJ)	3.86	5.50+	7.42+	2.84	3.30	3.84						
SRB/HOCH (JWJ)	6.14+	**9.45+**	**13.92+**	4.14	4.56	5.36						
SRB/HOCH (HWJ)	6.31+	**9.82+**	**14.51+**	4.22	4.69	5.51+						

No Preliminary Test (symmetric trimming)			
	10	15	20
PER (WJ)	**2.06-**	**2.18-**	**2.24-**
q (WJ)	**1.97-**	**2.10-**	**2.10-**
WJ / q (WJ)	**1.65-**	**1.74-**	**1.81-**
WJ / SRB (WJ)	2.90	3.06	3.35
WJ / HAY (WJ)	3.41	3.62	3.99
SRB/HOCH (WJ)	2.53-	2.70-	2.95-
PER (JWJ)	2.67-	2.53-	**2.39-**
q (JWJ)	2.58-	**2.44-**	**2.21-**
JWJ / q (JWJ)	**2.19-**	**2.03-**	**1.91-**
JWJ / SRB (JWJ)	3.87	3.60	3.52
JWJ / HAY (JWJ)	4.44	4.17	4.15
SRB/HOCH (JWJ)	3.46	3.17	3.15
PER (HWJ)	2.70-	2.56-	**2.40-**
q (HWJ)	2.61-	**2.46-**	**2.23-**
HWJ / q (HWJ)	**2.22-**	**2.05-**	**1.92-**
HWJ / SRB (HWJ)	3.92	3.63	3.54
HWJ / HAY (HWJ)	4.50	4.21	4.18
SRB/HOCH (HWJ)	3.50	3.21	3.17

Note. See note from Table 4

The SRB and HOCH methods had mean error rates closest to the nominal level when preceded by the test of symmetry based on average group estimates for the 20/40 trimming case or pairwise estimates for the 10/20 trimming case. Specifically, the mean error rates for the procedures based on the WJ, JWJ, and HWJ statistics were 5.07%, 4.88%, and 5.03%, respectively when using the average group estimates of tail weight and symmetry and 4.58%, 4.97%, and 5.01%, respectively for the pairwise estimate indices. It is worth noting that under the 20/40 trimming case, the SRB/HOCH procedures were the only MCP to have robust error rates when preceded by a test of symmetry.

g = 1 and h = 0 Distribution

All MCPs had rates of Type I error below Bradley's (1978) upper limit when preceded by a test of symmetry based on indices averaged across all groups for the 15/30, 20/40, and 15/15 trimming conditions and when the test of symmetry was based on individual group indices for the 15/15 trimming condition (see Table 5). Few trimming conditions resulted in MCPs with error rates within Bradley's limits. The condition with the most robust MCPs occurred with a test of symmetry based on tail weight and symmetry estimates from the individual groups with 15/15 symmetric/asymmetric trimming. For this particular trimming condition, the mean error rates were closer to the nominal 5% level for MCPs preceded with the symmetry test based on the individual group estimates (average rate equal to 3.82%) compared to MCPs preceded with the test of symmetry based on average estimates across all groups (average rate equal to 3.39%). Furthermore, MCPs based on the JWJ and HWJ statistics generally had error rates closer to the nominal level compared to MCPs based on the WJ statistic. For example, under the 15/15 trimming case with the test of symmetry based on tail weight and symmetry estimates from individual groups, the mean error rates across the MCPs based on WJ, JWJ, and HWJ statistics were equal to 3.72%, 3.85%, and 3.90%, respectively.

Noteworthy is that the form of the heteroscedastic statistic had an influence on Type I error rates regardless of whether a test of symmetry was used. For example, under the 15% symmetric trimming condition, the liberal error rate for the Hayter (1986) procedure based on the WJ statistic became nonliberal when based on the JWJ or HWJ statistic. This follows the general pattern that error rates tended to be smaller (more conservative) for MCPs based on the JWJ or HWJ statistics compared to when the MCPs were based on the WJ statistic. However, under the 10/10 and 15/15 symmetric/ asymmetric trimming cases when preceded by the test of symmetry, the opposite pattern was obersed, that is, the MCPs based on the WJ statistic were more conservative than the same MCPs based on the modified WJ statistics (i.e., JWJ and HWJ), a result consistent with the (g = .5, h = 0) distribution under the 10/10 trimming case. In addition, Type I error rates for the MCPs tended to decrease with an increase in the proportion of trimming cases (i.e., from 10/20 to 15/30 to 20/40 and from 10/10 to 15/15 to 20/20), except for the MCPs preceded by a test of symmetry based on individual group estimates where the pattern was reversed, that is, error rates tended to increase as the proportion of trimming increased (i.e., a pattern consistent with the results for J = 3).

Type I error rates for the SRB and HOCH procedures indicated that a test of symmetry based on the individual group indices provided mean error rates closer to the nominal 5% level compared to always adopting symmetric trimming or trimming symmetrically/asymmetrically based on the pairwise or across all groups average indices. For example, the mean error rates for SRB/HOCH, based on the WJ, JWJ, and HWJ statistics, were 4.96%, 4.98%, and 5.00%, respectively, under the 20/20 trimming case when using individual group indices of tail weight and symmetry, and were 4.12%, 4.03%, and 4.07%, respectively, for the 20% symmetric trimming case.

g = .25 and h = .25 Distribution

All MCPs had rates of Type I error below Bradley's (1978) upper limit for the 10/10, 15/15, and 20/20 trimming cases when preceded by the test of symmetry with average estimates across groups and the 10/10 trimming case when preceded by the test of symmetry with individual group estimates (see Table 6). Under the 10%, 15%, and 20% symmetric trimming cases, all MCPs had non-liberal error rates. The MCPs based on the range statistic tended to have conservative error rates, whereas under these trimming cases, the MCPs with rates within Bradley's limits were the WJ/SRB, WJ/HAY, JWJ/SRB, JWJ/HAY, SRB/HOCH (JWJ), HWJ/SRB, HWJ/HAY, and SRB/HOCH (HWJ).

The mean error rates, however, were more conservative under the symmetric trimming cases compared to the rates obtained for the MCPs when a symmetric/asymmetric strategy based on indices of tail weight and symmetry was adopted. Specifically, the mean error rates across non-liberal MCPs for the 10/10, 15/15, and 20/20 trimming cases when preceded by the test of symmetry with average group estimates were equal to 3.24%, 3.40%, and 3.65%, respectively and the mean rate for the 10/10 and 15/15 trimming case when preceded by the test of symmetry with individual group estimates were equal to 2.91% and 4.15%, respectively. Whereas, under the 10%, 15%, and 20% symmetric trimming cases, the mean error rates across MCPs were equal to 2.95%, 2.86%, and 2.85%, respectively.

MCPs based on the WJ statistic tended to have more conservative rates than when based on the JWJ or HWJ statistic. For example, under the 20/20 trimming case with the test of symmetry based on average group estimates, the mean error rates for the MCPs based on the WJ, JWJ, and HWJ statistics were 3.15%, 3.86%, and 3.93%, respectively and under the 15/15 trimming case with the test of symmetry based on individual group estimates, the mean error rates for non-liberal MCPs based on the WJ, JWJ, and HWJ statistics were 3.69%, 4.38%, and 4.46%, respectively. The general pattern was for error rates to increase as the proportion of trimming increases when the MCPs were preceded by a test of symmetry. However, this pattern only occurred for the MCPs based on a WJ statistic when always adopting symmetric trimming.

The SRB and HOCH procedures had higher mean error rates when based on symmetric/asymmetric trimming obtained from pairwise estimates than when based on indices obtained from all the groups. For example, liberal rates under the 10/20 trimming case based on pairwise estimates became robust when symmetric/asymmetric trimming was based on indices of tail weight and symmetry averaged over all groups. The data suggests that an optimal strategy was 10/20 symmetric/asymmetric trimming based on Q_1 and Q_2 obtained from all groups in the design. Specifically, the mean error rates for the SRB/HOCH procedures, based on the JWJ and HWJ statistics, were 4.56% and 4.66%, respectively.

Conclusion

In the present study, the strategy of computing a test of symmetry in order to determine whether to trim nonnormal data symmetrically (from both tails of the empirical distributions) or asymmetrically (from one tail of the empirical distributions) was compared to always utilizing an *a priori* symmetric trimming strategy, an approach previously investigated by Keselman, Lix et al. (1998) and typically recommended in the empirical literature (e.g., see Wilcox, 2003). We investigated the utility of testing for symmetry within the context of pairwise multiple comparison testing in a one-way independent groups design.

Three variations of a test of symmetry were investigated, each utilizing indices of tail weight and symmetry. The first variation obtains the indices of tail weight and symmetry by computing them within each group of a one-way completely randomized layout and then averages these values across the groups to obtain a summary measure of tail weight and symmetry. A second variation also takes an average of group indices, but only from the two groups comprising a particular pairwise comparison. The third variation, does no averaging across groups but measures tail weight and symmetry within each group of the pairwise comparison, using this information to determine whether data should be trimmed symmetrically or asymmetrically within each particular group.

The rationale behind all three approaches is to obtain an estimate of the typical score, that is, an estimate that represents the bulk of the observations, and accordingly outlying values are not wanted, found in the tail(s) of the nonnormal distributions, to adversely affect the score to be selected as typical – selecting a score that is not central to the distribution (e.g., the usual mean can be very far away from the central portion of a distribution of scores for skewed data). Though the rationale is the same for these three approaches, they respond to the need to obtain a good representation of the typical score in different ways.

The first method uses all of the data, across groups, to measure symmetry in the data and applies the results across all groups, that is, trims in a consistent fashion across all groups. The second and third approaches measure symmetry, or the lack there of, by only looking at the data involved in the pairwise comparison. The logic here is to ignore the type of nonsymmetry that may exist in groups that are not involved in a particular comparison. This rationale is similar to the approach of using a nonpooled error term, rather than a pooled error term, in order to avoid the biasing effects of variance heterogeneity in tests of mean equality. The third approach takes this rationale to its logical completion by finding the typical score in each group of the pairwise comparison by assessing symmetry/asymmetry within each individual group, rather than averaging over the two groups and applying the same form of trimming to both groups. That is, with this approach we are comparing the typical score from one group with the typical score from a second group, even though these typical scores were developed through different methods of trimming.

In addition to the use of a test of symmetry, the type of heteroscedastic statistic used in the computation of the MCPs was also investigated. The WJ statistic was investigated by Keselman, Lix et al. (1998) and the Johnson (1978) and Hall (1992) transformed WJ statistics investigated by Keselman et al. (2002). The MCPs with transformed WJ statistics [i.e., Hall (1992) or Johnson (1978)] based on a test of symmetry provided better Type I error control when distributions were nonnormal in form and had heterogeneous variances compared to the use of the WJ statistic with 20% symmetric trimming, the approach investigated by Keselman, Lix et al. (1998) and generally recommended in the literature.

Specifically, MCPs showed improved Type I error control, that is, nonrobust MCPs became robust and mean Type I error rates were closer to the nominal 5% level when data were first checked for symmetry and the MCPs were computed based on modified WJ statistics (i.e., JWJ or HWJ). A test of symmetry based on each individual group's indices of tail weight and symmetry generally provided mean Type I error rates closer to the nominal level for the MCPs than when the symmetry test was based on indices averaged over all groups in the design or just the groups in a particular pairwise comparison, particularly for the more extreme non-normal distributions. Across all nonnormal distributions investigated, optimal percentages of trimming in terms of controlling Type I error rates within Bradley's (1978) limits were the 10/10 and 15/15 symmetric/asymmetric trimming conditions. Interestingly, these proportions are less than the recommended 20% symmetric trimming.

The magnitude of Type I error rates changed as the pattern and percentage of trimming changed. Across the nonnormal distributions investigated, Type I error rates generally increased for the MCPs as the proportion of trimming increased over the 10/20, 15/30, and 20/40 trimming cases and for the 10/10, 15/15, and 20/20 trimming cases when preceded by a test of symmetry. However, under the following conditions the opposite pattern occurred when the MCPs were preceded by a symmetry test where the indices of tail weight and symmetry were obtained by averaging across the indices within each group of the design (a) for the chi-square distribution, Type I error rates decreased as the proportion of trimming increased (10/10, 15/15, and 20/20) for MCPs based on the JWJ and HWJ statistics, (b) for the ($g = .5$, $h = 0$) distribution, Type I error rates decreased as the proportion of trimming increased (10/10, 15/15, and 20/20) for MCPs based on the JWJ and HWJ statistics only for $J = 6$, and (c) for the ($g = 1$, $h = 0$) distribution, Type I error rates generally decreased as the proportion of trimming increased (from 10/20 to 15/30 to 20/40 and from 10/10 to 15/15 to 20/20).

The Type I error rates for the MCPs based on the JWJ or HWJ statistics were generally more conservative than the same MCPs based on the WJ statistic for the chi-square distribution. However, as the degree of nonnormality increased, this pattern reversed itself, firstly for the J = 3 condition and smaller percent trimming condition (10/10) for J = 6 for the (g = .5, h = 0) distribution, the smaller percent trimming conditions (10/10 and 15/15) for the (g = 1, h = 0) distribution, and across all trimming cases for the most extreme non-normal distribution (g = .25, h = .25) investigated. As the population distribution became more non-normal (e.g., skewed), the advantage of the transformed WJ statistics in terms of providing more robust MCPs was evident. This is not surprising given that the JWJ and HWJ statistics were developed to deal with the skewness bias. The Type I error rates for MCPs based on the JWJ statistic were slightly smaller (i.e., more conservative) than the rates for the same MCPs based on the HWJ statistic across the non-normal distributions investigated.

Taking into consideration the trimming cases that resulted in non-liberal error rates across most MCPs preceded by a test of symmetry with the pattern of error rates across trimming percentages and the generally superior performance of the MCPs with either the JWJ or HWJ statistics, the following general recommendations are provided for a strategy to achieve good Type I error control in a one-way independent groups design: (a) for distributions with skewness less than 2, adopt the 10% symmetric or 10% asymmetric trimming condition based on a test of symmetry where the indices of tail weight and symmetry are obtained by averaging over all groups when J = 3, whereas for J = 6, use a test of symmetry based on individual group indices of tail weight and symmetry and (b) for distributions with skewness greater than 2, adopt the 15% symmetric or 15% asymmetric trimming condition based on a test of symmetry using individual group indices of tail weight and symmetry.

As an overall recommendation, researchers may adopt any one of the MCPs with either the JWJ or HWJ statistic with trimmed means and Winsorized variances preceded by a test of symmetry in order to deal with nonnormal data and heterogeneous variances, conditions likely to be encountered in applied research. The importance of this finding is that educational researchers will be assured that the method will provide good Type I error control with generally more modest amounts of trimming compared to the generally recommended strategy of uniformly adopting 20% symmetric trimming.

References

Babu, J. G., Padmanabhan, A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal, 41*, 321-339.

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

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two-sample *t*-test. *Biometrical Journal, 28*, 131-148.

Guo, J. H., & Luh, W. M. (2000). An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. *Statistics & Probability Letters, 49*, 1-7.

Hall, P. (1992). On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B, 54*, 221-228.

Hastings, N. A. J., & Peacock, J. B. (1975). *Statistical distributions: A handbook for students and practitioners*. New York: Wiley.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. *Journal of the American Statistical Association, 81*, 1000-1004.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h-distributions. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables, trends, and shapes*, (pp. 461-513). New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-802.

Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association, 70*, 656-661.

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

Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, *73*, 536-544.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisted: An update based on trimmed means. *Psychometrika*, *63*, 145-163.

Kesleman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, *3*, 123-141.

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.

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, *1*, 288-309.

Keuls, M. (1952). The use of the "Studentized range" in conjunction with an analysis of variance. *Euphytica*, *1*, 112-122.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, *117*, 547-560.

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

Luh, W. M., & Guo, J. H. (1999). A powerful transformation trimmed mean method for one-way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*, *52*, 303-320.

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

Newman, D. (1939). The distribution of the range in samples from a normal population expressed in terms of an independent estimate of standard deviation. *Biometrika*, *31*, 20-30.

Othman, A. R., Keselman, H. J., Wilcox, R. R., Fradette, K., & Padmanabhan, A. R. (2002). A test of symmetry. *Journal of Modern Applied Statistical Methods*, *1*, 310-315.

Peritz, E. (1970). *A note on multiple comparisons*. Unpublished manuscript, Hebrew University, Jerusalem, Israel.

Ryan, T. A. (1960). Significance tests for multiple comparison proportions, variances and other statistics. *Psychological Bulletin*, *57*, 318-328.

SAS Institute Inc. (1989). *SAS/IML software: Usage and reference, version 6* (1st ed.). Cary, NC: Author.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. *Journal of Educational Statistics*, *4*, 14-23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826-831.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, *72*, 566-575.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*. New York: Wiley.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal*, *32*, 771-780.

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*, 289-306.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65*, 51-77.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*, *51*, 123-134.

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

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 66-72

Nonparametric Bayesian Multiple Comparisons for Dependence Parameter in Bivariate Exponential Populations

M. Masoom Ali
Dept. of Math. Sciences
Ball State University

J. S. Cho
Dept. of Inform. Statistics
Kyungsung University

Munni Begum
Dept. of Math. Sciences
Ball State University

A nonparametric Bayesian multiple comparisons problem (MCP) for dependence parameters in *I* bivariate exponential populations is studied. A simple method for pairwise comparisons of these parameters is also suggested. The methodology by Gopalan and Berry (1998) is extended using Dirichlet process priors, applied in the form of baseline prior and likelihood combination to provide the comparisons. Computation of the posterior probabilities of all possible hypotheses are carried out through a Markov Chain Monte Carlo, Gibbs sampling, due to the intractability of analytic evaluation. The process of MCP for the dependent parameters of bivariate exponential populations is illustrated with a numerical example.

Key words: Bivariate exponential population; Dirichlet process prior; Gibbs sampler; mixture of Dirichlet processes; multiple comparison; nonparametric Bayes.

Introduction

In reliability studies of mechanical components, dependence between two components occurs quite often. A system, which functions as long as at least one of the two identical components functions, has a functional correlation between the system components. Initially, let the two components be independently on test with life distributions that are exponential with parameter λ, denoted as $exp(\lambda)$. Failure of one changes the life distribution of the other to $exp(\lambda\theta)$, $\theta > 0$. When $\theta = 1$, the two components function independently. For $\theta > 1$, the workload of the remaining component is increased, thereby decreasing the mean life. Here θ is called the dependence parameter. Weier (1981) provided the Bayes estimators of the parameters and reliability using a conjugate prior for such problems.

The multiple comparison problem (MCP) for *I* bivariate exponential populations with dependence parameters $\boldsymbol{\theta} = (\theta_1, \ldots\ldots, \theta_I)$ can be viewed as making inferences concerning relationships among the θ's based on observations. This is tantamount to testing the following hypothesis,

$$H_0 : \theta_1 = \ldots\ldots. = \theta_I \text{ vs. } H_1 : \text{not } H_0.$$

For bivariate exponential populations, the frequentist approach of multiple comparison is not very straightforward. This is partly due to the difficulty in handling the distributional aspects and associated computations. The multiple comparison problem using nonparametric priors in a Bayesian inferential setup was studied by Gopalan and Berry (1998) providing specific applications to the Binomial and Normal populations. Following similar approach, the MCP for a set of geometric and negative binomial populations (Masoom, Cho, & Begum, 2005) was studied. In this article, the MCP for the dependence parameters of a set of

M. Masoom Ali is George & Frances Ball Distinguished Professor of Statistics. Email him at mali@bsu.edu. His research interests are in order statistics, Bayesian statistics, statistical inference and distribution problems. J. C. Cho is Associate Professor of Informational Statistics. His research interests are in Bayesian Statistics. Munni Begum is Assistant Professor of Statistics. Her research interests are in Bayesian Statistics, Statistical modeling and Biostatistical methods.

bivariate exponential populations along the same line was studied.

In a Bayesian approach, the posterior probabilities of respective hypotheses in MCP can be calculated with moderate effort. The prior information on the unknown parameters has to be quantified as a distribution. However, the selection of the prior distribution could be tricky. One of the criticisms Bayesian inferential methods often face is the subjectivity in prior specification. In real data analysis prior specification could be based on scientific knowledge about the parameters. Non-informative prior specification is optimal in cases when there is little known about the background information. It is very important that prior distributions be as objective as possible while doing Bayesian inference. A typical objective prior distribution is the Dirichlet process prior (DPP) that leads to nonparametric Bayesian inference.

The DPP is a prior distribution on the family of distributions that is dense in the space of distribution functions. The family of DPPs was introduced by Ferguson (1973) and was extended to mixtures of DPP by Antoniak (1974) in order to treat problems including the estimation of a mixing distribution, bio-assay, empirical Bayes problems and discrimination problems. Escobar (1988) started the application of Markov chain Monte Carlo (MCMC) methods in nonparametric Bayesian modeling. Novel computational techniques and developments of MCMC schemes, including key contributions by Doss (1994), Bush and MacEachern (1996), Escobar and West (1997), MacEachern and Müller (1998), West, Müller and Escobar (1994) made it possible to study nonparametric Bayesian methods widely.

The focus was on the Bayesian approach to the multiple comparisons problem for *I* bivariate exponential populations based on the nonparametric Dirichlet process priors in this article. The MCMC techniques, in particular Gibbs sampling, is adopted here to evaluate the posterior probabilities of the hypotheses.

Preliminaries

Let *(X, Y)* denote the lifetimes of the two components that have a bivariate exponential model. The joint probability density function of *(X, Y)* can be written as,

$$f(x,y)\mid\lambda,\theta) = 2\theta\lambda^{2}\exp\left(-2\lambda x-\lambda\theta y\right),$$

$$x,y>0,\quad \lambda,\theta>0 \tag{1}$$

with θ as the dependence parameter.

It is assumed that $(\boldsymbol{x},\boldsymbol{y}) = \{(\boldsymbol{x}_1,\boldsymbol{y}_1), (\boldsymbol{x}_2,\boldsymbol{y}_2), \ldots, (\boldsymbol{x}_I,\boldsymbol{y}_I)\}$ be a set of observations available on *I* populations, where $(\boldsymbol{x}_i,\boldsymbol{y}_i)=\{(x_{i1},y_{i1}),\ldots\ldots,(x_{in_i},y_{in_i})\}$ is an $n_i \times 1$ vector of conditionally independent observations on population i, $i=1,2,\ldots\ldots,I$; $j=1,2,\ldots\ldots,n_i$ and $\sum_{i=1}^{I} n_i = n$. Then the probability density function of (x_{ij},y_{ij}) is,

$$f\left(x_{ij},y_{ij}\mid\lambda_i,\theta_i\right)=2\theta_i\lambda_i^{2}\times\exp\left(-2\lambda_i x_{ij}-\lambda_i\theta_i y_{ij}\right),$$

$$x_{ij},y_{ij}>0,\quad \lambda_i,\theta_i>0. \tag{2}$$

Now a distribution function $G_0(.)$ and a positive scalar precision parameter α together determine the Dirichlet process prior *G*. Here $G_0(.)$ that defines the location of the DPP is sometimes called prior guess or baseline prior. The precision parameter α determines the concentration of the prior for *G* around the prior guess G_0, and therefore measures the strength of belief in G_0. The DPP is usually denoted by $G \sim D(G \mid G_0, \alpha)$. For large values of α, *G* is very likely to be close to G_0, while for small values of α, *G* is likely to put most of its probability mass on just a few atoms.

It is assumed that the θ_i's come from *G*, and that $G \sim D(G \mid G_0, \alpha)$ as stated above. This structure results in a posterior distribution which is a mixture of Dirichlet processes (Antoniak 1974). Now following the Polya urn representation of the Dirichlet process (Blackwell & MacQueen, 1973), the joint posterior distribution can be written as,

$$\theta_i \mid \mathbf{x},\mathbf{y} \propto \prod_{i=1}^{I} f(\mathbf{x}_i,\mathbf{y}_i \mid \theta_i) \times \frac{\alpha G_0(\theta_i)+\sum_{k<i}\delta(\theta_i \mid \theta_k)}{\alpha+i-1}, \tag{3}$$

where $\delta\ (\theta_i \mid \theta_k)$ is the distribution putting a point mass on θ_k. For each i =1,….. I, the conditional posterior distribution of θ_i is given by,

$$\theta_i \mid \theta_k, k \neq i, \mathbf{x},\mathbf{y} \propto q_0 G_b\left(\theta_i \mid \mathbf{x}_i,\mathbf{y}_i\right) + \sum_{k\neq i} q_k \delta\left(\theta_i \mid \theta_k\right), \tag{4}$$

where $G_b(\theta_\mathrm{i} \mid \boldsymbol{x}_\mathrm{i}, \boldsymbol{y}_\mathrm{i})$ is the baseline posterior distribution, $q_0 \propto \alpha \int f\left(\mathbf{x}_i,\mathbf{y}_i \mid \theta\right) dG_0\left(\theta_i\right)$, $q_k \propto f\left(\mathbf{x}_i,\mathbf{y}_i \mid \theta_k\right)$, and $1 = q_0 + \sum_{k\neq i} q_k$. Let $\mathbf{\Theta} = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots\ldots, \theta_I) : \theta_i \in R, i=1,2, \ldots\ldots, I\}$ be the I-dimensional parameter space. Equality and inequality relationships among θ's induce statistical hypotheses that are subsets of $\mathbf{\Theta}$. Thus, the MCP becomes testing the following hypotheses.

$H_0 : \boldsymbol{\theta}_0 = \{\theta_i: \theta_1 = \theta_2 = \ldots\ldots. = \theta_I\}$,
$H_1 : \boldsymbol{\theta}_1 = \{\theta i: \theta_1 \neq \theta_2, \theta_2 = \theta_3 = \ldots\ldots= \theta_I\}, \ldots\ldots H_N : \boldsymbol{\theta}_N = \{\theta_i: \theta_1 \neq \theta_2 \neq \theta_3 \neq \ldots\ldots \neq \theta_K\}$.

The hypotheses $H_r : \boldsymbol{\theta}_r$, r = 0,1,2, …….., N, are disjoint, and $\bigcup_{r=0}^{n} \boldsymbol{\theta}_r = \mathbf{\Theta}$.

The elements of $\mathbf{\Theta}$ themselves behave as described by (3) and so with positive probability, they will reduce to some $p < I$ distinct values. Let superscript * denote distinct values of the parameters. Then, any realization of I parameters θ_i generated from G lies in a set of $p < I$ distinct values, denoted by ($\boldsymbol{\theta}^* = \theta_1^*, \theta_2^*, \ldots.., \theta_p^*$). The computation of posterior probabilities for different hypotheses through Gibbs algorithm becomes manageable using the notion of *configuration* as termed by Gopalan and Berry (1998). Their definition of *configuration* is restated here:

Definition (*Configuration*): The set of indices $S = \{S_1,\ldots..,S_I\}$ determines a classification of the data $\boldsymbol{\Theta}=\{\theta_1,\ldots\ldots.,\theta_I\}$ into I^* distinct groups or clusters; the $n_j= \#\{S_{i=j}\}$ observations in group j share the common parameter value θ_j^*. Now, define I_j as the set of indices of observations in group j; That is, $I_j=\{i: S_i =j\}$. Let $(X,Y)_{(j)} = \{(X_i,Y_i): S_i = j\}$ be the corresponding group of $n_{I_j} = \sum_{i\in I_j} n_i$ observations. Thus, a one-to-one correspondence between hypotheses and configurations follows and the required computations are reduced by the fact that the distinct θ_i's are typically reduced to fewer than I due to the clustering of the θ_i's inherent in the Dirichlet process. Hence, (4) can be rewritten as:

$$\theta_i \mid \theta_k, k \neq i, \mathbf{x},\mathbf{y} \propto q_0 G_b\left(\theta_i \mid \mathbf{x}_i,\mathbf{y}_i\right) + \sum n_k q_k^* \delta\left(\theta_i \mid \theta_k^*\right), \tag{5}$$

with $q_k^* \propto f\left(\mathbf{x}_i,\mathbf{y}_i \mid \theta^{*}{}_{k}\right)$, and $1 = q_0 + \sum_{k\neq i} n_k q_k^*$. In addition to the simplification of notations, the cluster structure of the θ_i also improves the efficiency of the algorithm.

Posterior Sampling In Dirichlet Process Mixtures

A gamma distribution with parameters $(\alpha_{0i}, \beta_{0i})$ is considered as baseline prior G_0. This implies that $\theta_1, \theta_2,\ldots\ldots, \theta_I$ are *i.i.d.* from G_0. Then, a hierarchical set up for the Dirichlet process analysis as outlined above becomes,

$$\mathbf{x}_i,\mathbf{y}_i \mid \theta_i \sim BVE\left(\mathbf{x}_i,\mathbf{y}_i \mid \lambda_i, \theta_i\right), \tag{6}$$

$$\theta_i \mid G \sim G\left(\theta_i\right), \tag{7}$$

$$G \mid G_0, \alpha \sim D\left(G \mid G_0, \alpha\right), \tag{8}$$

$$G_0 \mid \alpha_{0i}, \beta_{0i} \sim Gam\left(\alpha_{0i}, \beta_{0i}\right), \tag{9}$$

$$\lambda_i \mid \alpha_{1i}, \beta_{1i} \sim Gam(\alpha_{1i}, \beta_{1i}), \tag{10}$$

BVE and *Gam* stand for bivariate exponential and gamma distributions, respectively. Now, the choice of the precision parameter α in Dirichlet process is extremely important for the model. A gamma prior for α with a shape parameter *a* and scale parameter *b* is considered, that is, *α ~ Gam(a,b).* Thus, the *Gam(a,b)* becomes the reference prior if *a → 0* and b → 0 and one has access to a neat data augmentation device for sampling α by Escobar and West (1995).

The configuration notation is more convenient to use in describing the Gibbs sampling algorithm as the full conditionals can be written in closed form as under:

$$(\theta_i \mid \mathbf{x}, \mathbf{y}, \theta_k, k \neq i, \alpha) \sim$$

$$q_0 Gam\left(n_i + \alpha_{0i}, \lambda_i \sum_{j=1}^{n_i} y_{ij} + \beta_{0i}\right) + \sum_{k \neq i} q_k \delta(d\theta_i \mid \theta_k), \tag{11}$$

$$(\lambda_i \mid \mathbf{x}, \mathbf{y}, \theta_i, \alpha) \sim$$

$$Gam\left(2n_i + \alpha_{1i}, 2\sum_{j=1}^{n_i} x_{ij} + \theta_i \sum_{j=1}^{n_i} y_{ij} + \beta_{1i}\right), \tag{12}$$

$$(\theta_j^* \mid \mathbf{x}, \mathbf{y}, \mathrm{S}) \sim$$

$$Gam\left(\sum_{i=1}^{I^*} n_i + \alpha_{0i}^*, \lambda_i \sum_{i=1}^{I^*} \sum_{j=1}^{n_i} y_{ij} + \beta_{0j}^*\right), \tag{13}$$

$$(\alpha \mid \eta, I^*) \sim$$

$$\pi_\eta Gam\left(a + I^*, b - \log(\eta)\right) +$$

$$(1 - \pi_\eta) Gam\left(a + I^* - 1, b - \log(\eta)\right), \tag{14}$$

$$(\eta \mid \alpha, I^*) \sim Beta(\alpha + 1, I^*), \tag{15}$$

where

$$q_0 \propto \alpha \lambda_i^{2n_i + \alpha_{0i} - 1} \exp\left(-2\lambda_i \sum_{j=1}^{n_i} x_{ij}\right) \frac{\Gamma(n_i + \alpha_{0i})}{\left[\lambda_i \left(\sum_{j=1}^{n_i} y_{ij} + \beta_{0i}\right)\right]^{n_i + \alpha_{0i}}},$$

$$q_k \propto \theta_k^{n_i} \lambda_k^{2n_i} \exp\left(-2\lambda_k \sum_{j=1}^{n_i} x_{ij} - \theta_k \lambda_k \sum_{j=1}^{n_i} y_{ij}\right)$$

Gibbs sampling proceeds by simply iterating through (11) - (15) in order, sampling at each stage based on the current values of all the conditioning variables.

The configuration induces the equality and inequality relationships among the *θ*'s that corresponds to the partitions on the parameter space **Θ** and in turn to the hypotheses of interest. In order to estimate the posterior probability of a hypothesis H_r from a large number (*L*) of sample draws, one takes

$$P(H_r \mid \mathbf{X}, \mathbf{Y}) \approx \frac{1}{L} \sum_{l=1}^{L} \delta_{S_l}(H_r), \tag{16}$$

where $\delta_{S_l}(H_r)$ denotes unit point mass for the case where *l* th draw of *S*, S_0 corresponds to H_r. The probability of equality for any two *θ*'s can be calculated from the posterior distributions on hypotheses, *P(H*$_r$ *|* ***X,Y****), r =1,2, ……., N.* This can be achieved by adding probabilities of those hypotheses in which the two θ_i and θ_j are equal. That is

$$P(\theta_i = \theta_j \mid \mathbf{X}, \mathbf{Y}) \approx \frac{1}{L} \sum_{l=1}^{L} \delta_{S_l}(\theta_i = \theta_j) =$$

$$\sum_{r=1}^{N} P(H_r \mid \mathbf{X}, \mathbf{Y}) \delta_{H_r}(\theta_i - \theta_j), i \neq j,$$

where $\delta_{S_l}(\theta_i - \theta_j)$ and $\delta_{H_r}(\theta_i = \theta_j)$ denote unit point mass for the case where S_l and H_r indicate $\theta_i = \theta_j$.

Illustrative Example

A numerical illustration of the multiple comparisons for the dependence parameters in bivariate exponential populations is presented in this section using simulated data. Four bivariate exponential populations each with size n_i*=20* are considered. Then, the numbers of possible hypotheses for multiple comparisons are 15. The observed summary statistics for these data are given in Table 1.

It follows from Table 1, that the true hypothesis may be $H_{true} : \theta_1 = \theta_2 \neq \theta_3 = \theta_4$. For the precision parameter α, one considers three Gamma priors with parameters *(a,b)*=(1.0, 1.0), (0.1, 0.1) and (0.01, 0.01) in order to have equal mean 1 and different variances 1, 10, and 100, respectively. This also facilitates that the latter prior be fairly non-informative, giving reasonable mass to both high and low values of α. As well, each θ_i, i=1,……, 4 were set *a priori* following a gamma distribution with parameters $\alpha_{0i} = \alpha_{1i} = 2.0$ and $\beta_{0i} = \beta_{1i} = 0.001$ to reflect vagueness of the prior knowledge.

The posterior probabilities for all possible hypotheses are approximated by the Gibbs sampling algorithm using 20,000 iterations with 10,000 burn-ins and 5 replications and are presented in Table 2. It is to be noted that the hypothesis $\theta_1 = \theta_2 \neq \theta_3 = \theta_4$ has the largest posterior probabilities 0.7883, 0.7274 and 0.7410 for all priors of the precision parameter α. Thus, the data lend greatest support to equalities for $\theta_1 = \theta_2$ and $\theta_3 = \theta_4$ being different from the others.

Table 3 presents the pairwise posterior probabilities for the equalities in pairs of θ's. The equalities of $(\theta_1 = \theta_2)$ and $(\theta_3 = \theta_4)$ have the largest posterior probabilities (0.9943, 0.9903, 0.9729) and (1.0000, 1.0000, 1.0000) for three cases of *(a, b)* respectively. This suggests that there is strong evidence in the equality $(\theta_1 = \theta_2)$ and $(\theta_3 = \theta_4)$.

The Bayesian approach using nonparametric Dirichlet process priors facilitates studying the problem of multiple comparisons in a number of different distributions. So far, the MCP was carried out for a univariate distribution. Here, it has been shown that the method can be extended to a bivariate distribution as well, with moderate effort. As an alternative to a formal Bayesian analysis of a mixture model that usually leads to intractable calculations, the DPP is used to provide a nonparametric Bayesian method for obtaining posterior probabilities for various hypotheses of equality among the dependence parameters of bivariate exponential populations.

Table 1 The observed summary statistics for each populations

Populations	1	2	3	4
$X_i = \sum_{j=1}^{n_i} X_{ij}$	1.500	1.560	0.700	0.720
$Y_i = \sum_{j=1}^{n_i} Y_{ij}$	6.500	6.000	1.300	1.130
$\hat{\theta}_{MLE}$	0.462	0.520	1.077	1.274

Table 2 Calculated posterior probabilities for each hypothesis with three cases of (a,b)

Hypothesis	(1.0, 1.0)	(0.1, 0.1)	(0.01, 0.01)
$\theta1 = \theta2 = \theta3 = \theta4$	.2059	.2629	.2320
$\theta1 = \theta2 = \theta3 \neq \theta4$	.0000	.0000	.0000
$\theta1 = \theta2 = \theta4 \neq \theta3$	.0000	.0000	.0000
$\theta1 = \theta2 \neq \theta3 = \theta4$	.7883	.7274	.7410
$\theta1 = \theta2 \neq \theta3 \neq \theta4$	.0000	.0000	.0000
$\theta1 = \theta3 = \theta4 \neq \theta2$	.0036	.0038	.0030
$\theta1 = \theta3 \neq \theta2 = \theta4$	.0000	.0000	.0000
$\theta1 = \theta3 \neq \theta2 \neq \theta4$	.0000	.0000	.0000
$\theta1 = \theta4 \neq \theta2 = \theta3$	.0000	.0000	.0000
$\theta1 = \theta4 \neq \theta2 \neq \theta3$	.0000	.0000	.0000
$\theta1 \neq \theta2 = \theta3 = \theta4$	.0003	.0007	.0015
$\theta1 \neq \theta2 = \theta3 \neq \theta4$	.0000	.0000	.0000
$\theta1 \neq \theta2 = \theta4 \neq \theta3$	.0000	.0000	.0000
$\theta1 \neq \theta2 \neq \theta3 = \theta4$	.0018	.0052	.0226
$\theta1 \neq \theta2 \neq \theta3 \neq \theta4$	.0000	.0000	.0000

Table 3 Pairwise Posterior Probabilities with three cases of *(a, b)*

Hypothesis	(1.0, 1.0)	(0.1, 0.1)	(0.01, 0.01)
$\theta_1 = \theta_2$	.9943	.9903	.9729
$\theta_1 = \theta_3$	.2096	.2667	.2349
$\theta_1 = \theta_4$	.2096	.2667	.2349
$\theta_2 = \theta_3$	.2062	.2636	.2334
$\theta_2 = \theta_4$	.2062	.2636	.2334
$\theta_3 = \theta_4$	1.0000	1.0000	1.0000

References

Ali, M. Masoom, Cho, J. S., Begum, M. (2005). Nonparametric Bayesian multiple comparisons for geometric populations. *Journal of the Korean Data and Information Science Society*, *16*(4), 1129-1140.

Ali, M. Masoom., Cho, J. S., Begum, M. (2006). Bayesian multiple comparisons with nonparametric dirichlet process priors for negative binomial populations. *Pakistan Journal of Statistics*, *22*(2), 88-98.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, *2*, 1152-1174.

Bush, C. A. & MacEachern, S. N. (1996). A semi-parametric bayesian model for randomized block designs. *Biometrika*, *83*, 275-285.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, *22*, 1763-1786.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. *Unpublished dissertation*, Yale University.

Escobar, M. D. & West, M. (1997). Computing nonparametric hierarchical models. *ISDS Discussion Paper #97-15*, Duke University.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Gopalan, R. & Berry, D. A. (1998). Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association*, *90*, 1130 - 1139.

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, *7*, 223-239.

Weier, D. R. (1981). Bayes estimation for a bivariate survival model based on exponential distributions. *Communication in Statistics-Theory and Methods*, *10*, 1415-1427.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 73-94

1538 – 9472/06/$95.00

Understanding Eurasian Convergence: Application Of Kohonen Self-Organizing Maps

Joel I. Deichmann Abdolreza Eshghi Dominique Haughton
Selin Sayek Nicholas Teebagy Heikki Topi

Data Analytics Research Team, Bentley College

Kohonen self-organizing maps (SOMs) are employed to examine economic and social convergence of Eurasian countries based on a set of twenty-eight socio-economic measures. A core of European Union states is identified that provides a benchmark against which convergence of post-socialist transition economies may be judged. The Central European Visegrád countries and Baltics show the greatest economic convergence to Western Europe, while other states form clusters that lag behind. Initial conditions on the social dimension can either facilitate or constrain economic convergence, as discovered in Central Europe vis-à-vis the Central Asian Republics. Disquiet in the convergence literature is resolved by providing an analysis of the Eurasian states over time.

Introduction

The definition of what constitutes the entity of Europe is debated widely (Almström, 2000; Jordan, 2002). In particular, no satisfactory physiographic barriers exist to distinguish Europe from neighboring Asia. Many scholars approximate the border as the Ural Mountains, the Volga River, or the Bosporus Strait, dividing Russia and Turkey between Europe and Asia (Jordan, 2002). Others conveniently define Europe according to the membership of the fifteen EU member states, but this definition leaves out Norway, Switzerland, and several wealthy micro-states, as well as (until 2004) the Central European candidate states. Jordan (2002) defined Europe in terms of the people who live there, identifying the cultural traits that define the source of Western civilization, in addition to ten secondary socio-economic characteristics that most European states share.

The collapse of the Soviet empire in 1989, coupled with the deepening and widening debate within the EU, has fueled an unprecedented movement toward a unified Europe. The post-socialist countries of Central and Eastern Europe have embarked upon a daunting task of instituting a series of dramatic economic and social reforms to create western-style market economies with the objective of becoming full-fledged members of the EU as quickly as possible. As noted by the Economic Analysis Division of the United Nations Economic Commission for Europe:

Joel Deichmann is Associate Professor of Geography. His research interests include Foreign Direct Investment. Abdolreza Eshghi is Professor of Marketing and Director for the Center for Marketing Technology. His research interests focus on customer relationship management. Dominique Haughton is Professor of Mathematical Sciences. Her areas of research include analyses of living standards in developing countries. Selin Sayek is Professor of Economics, Bilkent University, Turkey. Her research focuses on foreign direct investment and aid flows. Nicholas Teebagy, in memoriam, was Associate Professor of Mathematical Sciences. His research interests are in spline analyses of ozone data and finance. Dr. Heikki Topi is Associate Professor of Computer Information Systems and Director of the MSIT Program. His research focuses on human factors and usability issues in enterprise systems. Kohonen maps appear in this article in gray scale. See the *JMASM* web site (http://tbf.coe.wayne.edu/jmasm) for renditions of these maps in color.

> One of the strategic goals of the transition economies is to achieve sustained and high rates of economic growth that would enable them to catch up with – to converge upon – the living standards of the developed economies of Western Europe. And many of them regard EU membership as instrumental to promote this process. (United Nations Economic Commission for Europe, 2000)

In their efforts to join the EU, Central and Eastern European countries have opted for a wide variety of transition paths to treat their unique set of initial conditions, in turn leading to a correspondingly heterogeneous set of results. While some have either regained (e.g., Poland, Slovakia, and Slovenia) or are close to regaining (Czech Republic and Hungary) their pre-transition GDP levels, others (notably Georgia, Ukraine, and Moldova) continue to struggle with their transformational recession (United Nations Economic Commission for Europe, 2000).

The question of post-socialist convergence has been the subject of extensive scholarly research from various perspectives (see Szalkowski & Jankowicz, 1999; Genov, 1998; Bartlett, 1997; Brabant, 1998; Lang, 2003; Graham & Hart, 1999, to name just a few). However, there is no consensus on the extent of convergence and the factors that have led to highly heterogeneous outcomes. The research presented here is intended to address these issues by analyzing a comprehensive set of socio-economic variables for all of the Western European and post-Communist countries for which data are available.

More specifically, the purpose of this article is to map the progress of post-socialist countries in catching up with, or converging upon the advanced Western European economies over the past decade. In particular, not only is the overall convergence mapped, but the macroeconomic, social, and institutional factors that are responsible for the convergence, or lack thereof, are identified. In this context, the role played by economic factors versus social factors in catching up and converging with the EU is discussed. A secondary purpose of this research is to extend previous Kohonen analysis on transition economies (Deichmann et al., 2003) to include the existing EU members plus Norway, Switzerland, the USA (see note 1 in Fig.9), and Turkey. In so doing, it is hoped that the extent to which this broader group of Eurasian states clusters geographically when all reference to location is absent will be determined, and use the changes in the clusters over time to observe whether or not patterns of convergence exist among these groups of economies.

Post-Socialist Heterogeneity

A significant body of literature has documented the differential levels of convergence throughout Eurasia during the first ten years after the Iron Curtain fell. Using data through 1998, Estrin, Urga, and Lazarova (2001) examine average (GDP) growth rates for transition economies leading up to and following the abrupt changes that began in the early 1990s. Focusing upon twenty-six countries over twenty-seven years, the level of pre-transition convergence was examined since 1991. Among the twenty-six states, they found that Hungary, Poland, Slovenia, Estonia, and Armenia are the only states with positive average growth rates since the transition and only Armenia, Slovenia, and Hungary have sustained growth that might eventually allow their economies to catch up with those of Western Europe.

Also citing disparities in growth prior to the 1990s transition, the authors highlight the failure of reallocation mechanisms within the Soviet bloc, with the possible exception of the former Yugoslavia, which was only loosely affiliated with Moscow. They concluded that the failure of Soviet-led central planning to ameliorate regional disparities within the socialist bloc is likely to have facilitated the demise of supranational affiliations within the region. Unfortunately, the authors also find little evidence for convergence during the first decade of individual state policies.

Kočenda (2001) modeled the time-path of several macroeconomic variables to evaluate convergence of Central European and Baltic states. Variables under investigation include industrial output, prices, money (capital), and interest rates. Among these countries, there are dramatically differing initial conditions that favor the Czech Republic and its neighbors, while putting the Baltic states at a comparative disadvantage; for example, the former enjoy an earlier 1989 starting point, while the latter became independent in 1991 and have only recently introduced their own new currencies. Despite the countries' unique initial conditions, Kočenda (2001) found considerable evidence of convergence by these otherwise similar countries through the natural process of increased international trade and through the institutional processes of coordination to satisfy EU pre-accession requirements. However, Kutan and Yigit (2004) emphasized the importance of model specification and how it changes the results of Kočenda (2001). They showed that when heterogeneity is taken into account the within-group convergence is not as evident as suggested by Kočenda.

Brada and Kutan (2001) examined the extent of convergence of monetary policy of EU candidate and non-candidate transition states to the German monetary policy, which is viewed to be broadly representative of the European Central Bank. They concluded that the transition states (both candidate and non-candidate) lag far behind the non-transition EU candidates (Cyprus, Malta, and Turkey), revealing deeply rooted disadvantages of central planning that endure in transition countries. They contended that Hungary and Poland, which have pursued independent monetary policies throughout the 1990s, have the best prospects of converging to EU fiscal policies.

Brada, Kutan, and Zhou (2002) employed a rolling cointegration technique to evaluate the convergence of base money, broader money (M2), the consumer price index (CPI), and industrial output in five leading EU candidate countries: the Czech Republic, Estonia, Hungary, Poland, and Slovenia. Brada et al. (2002) argue that adequate convergence has yet to occur in the areas of monetary policy and industrial output, but that consumer prices and M2 are comparable to those in the EU, confirming earlier findings (Brada & Kutan, 2001) with a wider frame of inquiry. They concluded that considerable time will be necessary following accession and before the candidates join the Euro zone.

Wagner and Hlouskova (2001) focused on convergence in the real (vis-à-vis nominal) dimension, mainly economic growth. In doing so, they study the correlation between the initial level of GDP of ten Central and Eastern European economies and their average growth rates over the 1990s and find evidence for convergence only after 1998. They applied the distributional dynamics technique, formulating a statistical model to describe the evolution of the joint distribution of real per capita GDP of the CEE and EU economies. This method allows for the investigation of the mobility of each economy within the cross country income distribution over time. They concluded that their evidence reveals high persistence in the data combined with a low probability of an economy changing its location in the distribution. Therefore, neither of their methods suggested evidence of convergence among the CEE and EU economies through 1998.

Theoretical Explanation

The issue of heterogeneity in economic convergence among post-socialist countries can be explained with reference to a number of theoretical and conceptual arguments. First, theoretical models in development economics (Barro & Sala-i Martin, 1992) posited that economies with low initial GDP levels should grow faster than those with higher initial GDP levels, and eventually catch up with these more developed economies. This is certainly the case among post-socialist countries. These countries began the journey toward a free-market economy with varying initial economic conditions.

A second explanation was offered by Romer (1986) who argued that the characteristics inherent in technology prevent convergence from occurring because increasing returns to scale cause the rich countries to become richer while the poor countries fall further behind. A related argument is that convergence will only occur among countries with a well-developed human capital base, which allows for such countries to benefit from modern technology.

Third, Barro (1991) and Barro and Sala-i Martin (1992) suggested that absolute convergence does not exist as all countries have different long-run per capita income levels that prevent such convergence. However, they showed that each country converges to these differing long-run equilibria, and they labeled this phenomenon as conditional convergence.

Fourth, Sachs and Warner (1995) argued that unconditional convergence depends upon the policy choices of the respective economies, and that countries that pursue market-based economic policies, liberal trade policies, and respect private property rights show strong tendencies to convergence. Sachs and Warner (1995) found that the transition economies that have undertaken significant economic reforms show convergence signs to the European Union, while those that have not converged show persistence in their economic position.

Fifth, the specific manner of implementing economic reforms is also believed to be responsible for heterogeneous patterns of convergence. Some countries opted primarily for a top-down approach by privatizing the state-owned enterprises, whereas others (mainly the Central and Eastern European economies) generally favored a bottom-up approach by encouraging the establishment of new start-up enterprises and development of existing private firms (Brezinski & Fritsch, 1996; Woo, 1998). Ellman (1997) argued that experience from the past decade demonstrates that the development of new private firms is more important for the resumption of economic growth than is rapid privatization.

Another factor that may have influenced the convergence outcome is the pace at which reforms were implemented. Some countries implemented drastic macroeconomic stabilization policies known as the shock therapy approach, whereas others insisted upon a policy of gradualism, which entails structural and institutional reforms as a pre-condition to introducing macroeconomic stabilization reforms (Popov, 2000).

Finally and perhaps most importantly, the success of economic and social reforms is not only contingent upon their contents but upon the social and historical context in which they are implemented (Rosenbaum, 2001). In other words, market reforms presuppose societal values and norms that are consistent with democracy and a free-market economy. Some post-socialist countries have been more successful in implementing market reforms due to their historical and cultural ties with Western Europe. Rosenbaum's (2001) review of the economic history of Central and Eastern Europe indicates that the development of a secular civil society in Western and Central Europe resulted from conflict between the state and the princes on the one hand and the church on the other. Consequently, the intellectuals gained the opportunity to play off competing authorities against one another, giving rise to new philosophical and political ideas that led to the overthrow of the autocratic and feudal order and relegated the church to just one of many interest groups. By contrast, the church and the political authority remained in one hand in the East under Orthodoxy, which tended to block individualistic tendencies and the introduction of new ideas such as private property.

As also noted in Rosenbaum (2001, p. 895), whereas Christianized Poles, Czechs, and Hungarians adopted the institutional order of the West and became part of Western culture, Russia and much of the Balkan region remained insulated from the infusion of new ideas, leading to consolidation of power in the hands of the state. As a consequence, Orthodox cultures tend to accept the dominant role of the state in society and economy as fait accompli. Clearly, the historical experiences of post-socialist countries have far reaching implications for the role of the individual in determining her/his economic destiny. In short, when the historical and cultural experiences are consistent with free market values and norms, substantial progress toward convergence is observed over a relatively short period of time. However, when there is a mismatch between the historical and cultural experiences and the free market values and norms, the transition is likely to be slow and painful.

Methodology

Kohonen Self-organizing maps were used (SOMs) to examine post-socialist convergence in Eurasian countries. Kohonen maps were pioneered during the 1980s and have been used as a method of visualizing non-spatial data (Kohonen, 1982). Techniques for creating and interpreting Kohonen maps have been refined and reviewed by their namesake in a series of subsequent volume editions (Kohonen, 2001). SOMs have been employed in many contexts, for example in mapping non-geographic data ranging from text documents (Kohonen, 1999) to conference abstracts (Skupin, 2002; Kloptchenko et al., 2003).

The application of Kohonen maps continues to grow in a variety of disciplines (Deboeck, 1998; Oja & Kaski, 1999). One application that is particularly relevant here is the work of Costea, Kloptchenko, and Back (1998). They compared the relative advantages of SOMs and cluster analysis in evaluating the economic status of six transition economies: Russia, Ukraine, Romania, Poland, Slovenia, and Latvia. They introduced a very insightful way of depicting statistical trends in data over time: each observation corresponds to a country at a specific point in time, which facilitated a clear understanding of how countries migrate across the map over time.

The Kohonen Algorithm

The Kohonen algorithm can be briefly described as follows (see for example Kaski and Kohonen 1996): the algorithm assigns to each position *i* in a grid an arbitrary (random) vector $m_i(0)$ with as many components as input variables. At each time *t* the vector of variables *x(t)* corresponding to one of the observations updates the current vectors $m_i(t)$ according to the formula $m_i(t+1) = m_i(t) + h_{ci}(t)(x(t) - m_i(t))$, where $c = \arg\min_i(\| x - m_i \|)$ and $h_{ij}(t)$ is a function of *t* and of the geometric distance on the lattice between position *i* and position *j*. Typically $h_{ij} \to 0$ with increasing distance between *i* and *j* and increasing time. So the vector *x(t)* is allowed to update the vector $m_c(t)$ it is closest to as well as some neighboring vectors $m_i(t)$. When the algorithm converges, the m_i tend to be ordered along the lattice in a meaningful way (see note 2 in Fig. 9).

Data Issues

Due to data restrictions, the analysis is limited to the period 1992-2000. The breakups of the Soviet Union, Yugoslavia, and Czechoslovakia all resulted in missing values for the resulting new states during the early years of our analysis. These were addressed by entering the unions' values for each state (for example, the Czech Republic and Slovakia were both assigned the 1992 value for Czechoslovakia). After that point, any missing entries were replaced with the value estimated by regressing each variable on time for each country. Finally, many missing values for the year 2001 limited the analysis to the years through 2000.

Description of Variables

Procuring accurate, complete, and current socioeconomic data for the transition states is a formidable challenge (Costea, Kloptchenko, & Back, 2001). Most of the data were collected by national authorities and reported by the World Bank Development Indicators CD-ROM (2002) for the years 1992-2000.

The list of variables under consideration is presented in Table 1. The variables include economic, social, and political measures. The measures were chosen to capture each country's preconditions as well as subsequent measures (both absolute numbers and rates of change). The economic variables can be sub-grouped into real and nominal variables. The real variables encompass indicators of economic development, the role of government and fiscal policy in the economy, the level of physical infrastructure, the depth of financial markets, and international openness measures. The nominal variables include indicators regarding the domestic price of goods and the foreign currency price of the domestic currency, the inflation rate and the real exchange rate respectively, and the real interest rates. Explicit reform variables, as addressed by Sach and Warner (1995), are available only for transition states, and are therefore unsuitable for this analysis that spans the EU and other wealthy states as well.

The social infrastructure measures, which include variables that impact the development of human capital, such as education and health measures, are covered in the social dimension of the analysis. In addition to such social infrastructure measures several physical infrastructure measures are also included in this group, as they also contribute more to the development of social infrastructure than anything else. These measures include efficiency in electricity distribution and access to communication means such as telephone and the internet. Finally, the social indicators also include measures of extent of political rights and civil liberties.

TABLE 1. LIST OF ECONOMIC, SOCIAL, AND POLITICAL VARIABLES

Variable	Description
Prscgdp	Private sector credit as share of GDP
Electricyt	Electric power transmission and distribution losses
Kgdp	Gross capital formation as share of GDP
Infl	Inflation (GDP deflator based)
Growth	Real GDP per capita growth
Tradegdp	Trade as a share of GDP
Figdp	FDI as a share of GDP
Reserves	Reserves, months of import coverage
Cagdp	Current account balance as a share of GDP
Gdppc	GDP per capita (in real 1995 USD)
Fiscgdp	Overall fiscal balance including grants (share of GDP)
Rer	Real exchange rate
Rir	Real interest rate
Bankresliq	Bank reserves to liquid assets
Tellines	Telephone lines (per 1000)
Stuteach	Student to teacher ratio
Schoolenroll	Secondary school enrollment (gross)
Immunmeasl	Immunization against measles
Lifeexp	Life expectancy
Nodoctors	Number of physicians (per 1000)
Immunization	Immunization against DPT
Agedepend	Age dependency ratio (dependents to working-age population)
Healthpub	Public health expenditures (share of GDP)
Healthpr	Private health expenditures (share of GDP)
Healthsum	Total health expenditures (share of GDP)
Internet	Internet users as share of population
Civlib	Score for civil liberties (1=lowest, 7=highest)*
Polrights	Score for political rights (1=lowest, 7= highest)*

Data Source: World Bank (2002) except for *, which were obtained from Freedom House (2003)

Analysis

As in Costea, Kloptchenko, and Back (2001), all countries under investigation for each individual year are first plotted on a single map to monitor movements over time throughout the lattice on the basis of all available variables. The variables are then subdivided into social/political and economic measures in an effort to examine the role they play in convergence.

Analysis of Aggregate Maps

Figure 1 represents a self-organizing map of all country-year pairs (such as Moldova 1992, for example) over 1992-2000, constructed on the basis of all variables in Table 1 for all countries under investigation. The largest group of countries can be thought of as a European core—composed of mainly EU states located in the center-top of the figure (such as France, Germany, Ireland, Luxembourg, the Netherlands, the UK). Outside this core, several noteworthy peripheries exist, in addition to several distinct groups of laggard transition countries. As was observed in past work (Kaski & Kohonen, 1996; Deichmann et al., 2003), an outstanding feature of this first U-matrix is the preservation of many geographic relationships

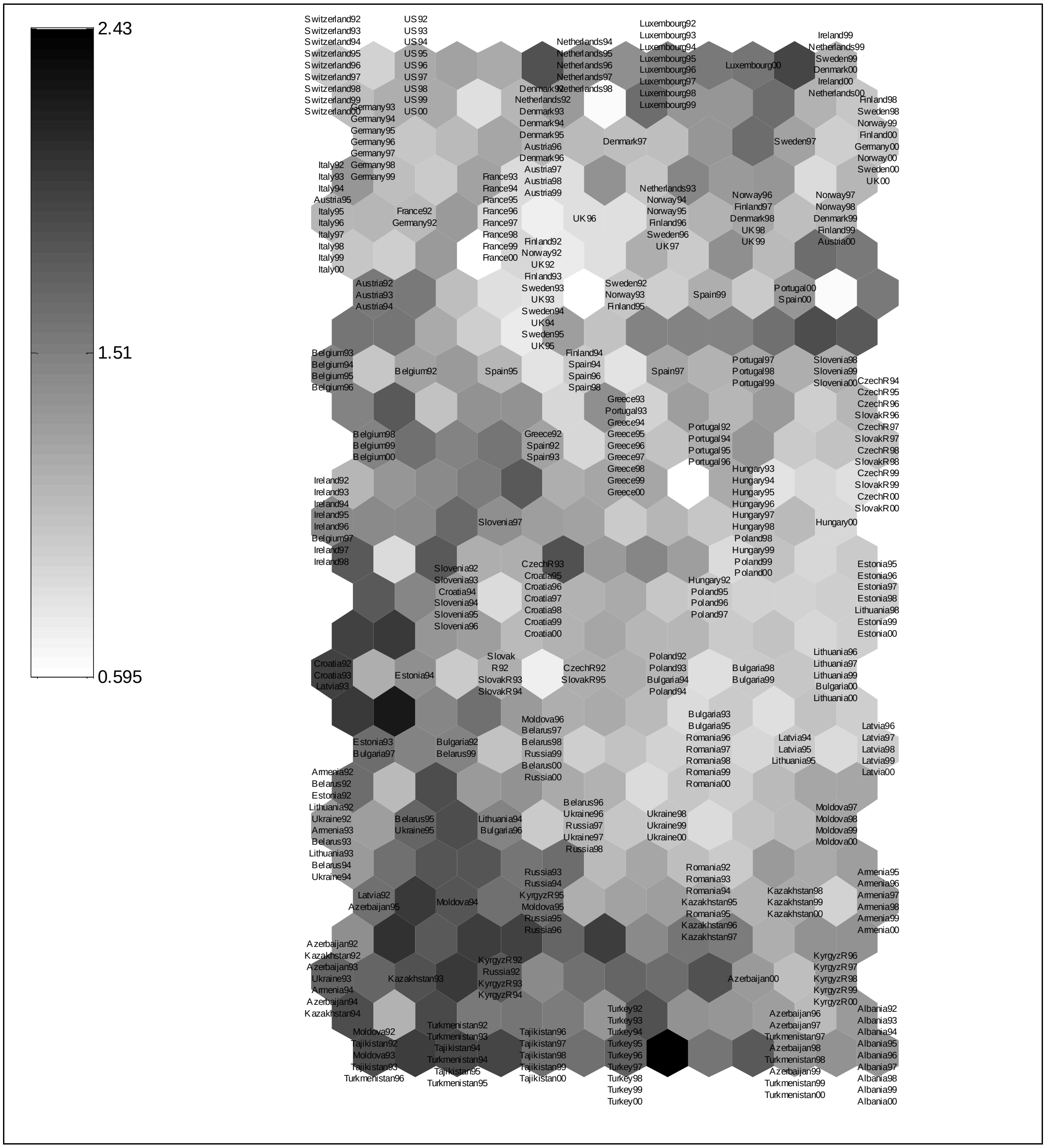


FIGURE 1. U-MATRIX OF COUNTRY MOVEMENTS FROM 1992-2000.
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

in the absence of explicitly geographic variables. This is clearly the case in Figure 1 and in subsequent figures.

Figure 2 provides estimated (by the Kohonen algorithm) values of the input variables at each grid position in the U-matrix. For example, it may be seen that estimated values of private health expenditures are high at the US (for all years) map position (top left of the U- matrix). Note that the U-matrix, in addition to actual grid positions, includes slots between grid positions which are colored to represent how close the grid positions are to one another. The color on an actual grid position represents how close the position is to its neighbors. For example, it is known that the positions of Switzerland and the US (at the top left of the map) are very close in terms of estimated variable values because the hexagon between them is dark blue (very light grey in grey scale format). Conversely, it may be seen that the position occupied by Croatia 92/93 and Latvia 93 (about two thirds of the way down on the left of the map), is distant from its neighbors because it is colored orange (a large distance color, as indicated by the color legend), dark grey in grey scale format.

A study of Figure 2 yields an interpretation - presented on Figure 3 – of the vertical and horizontal dimensions on the map. Together, the visual tools presented in Figures 1-3 facilitate an overall impression of how the

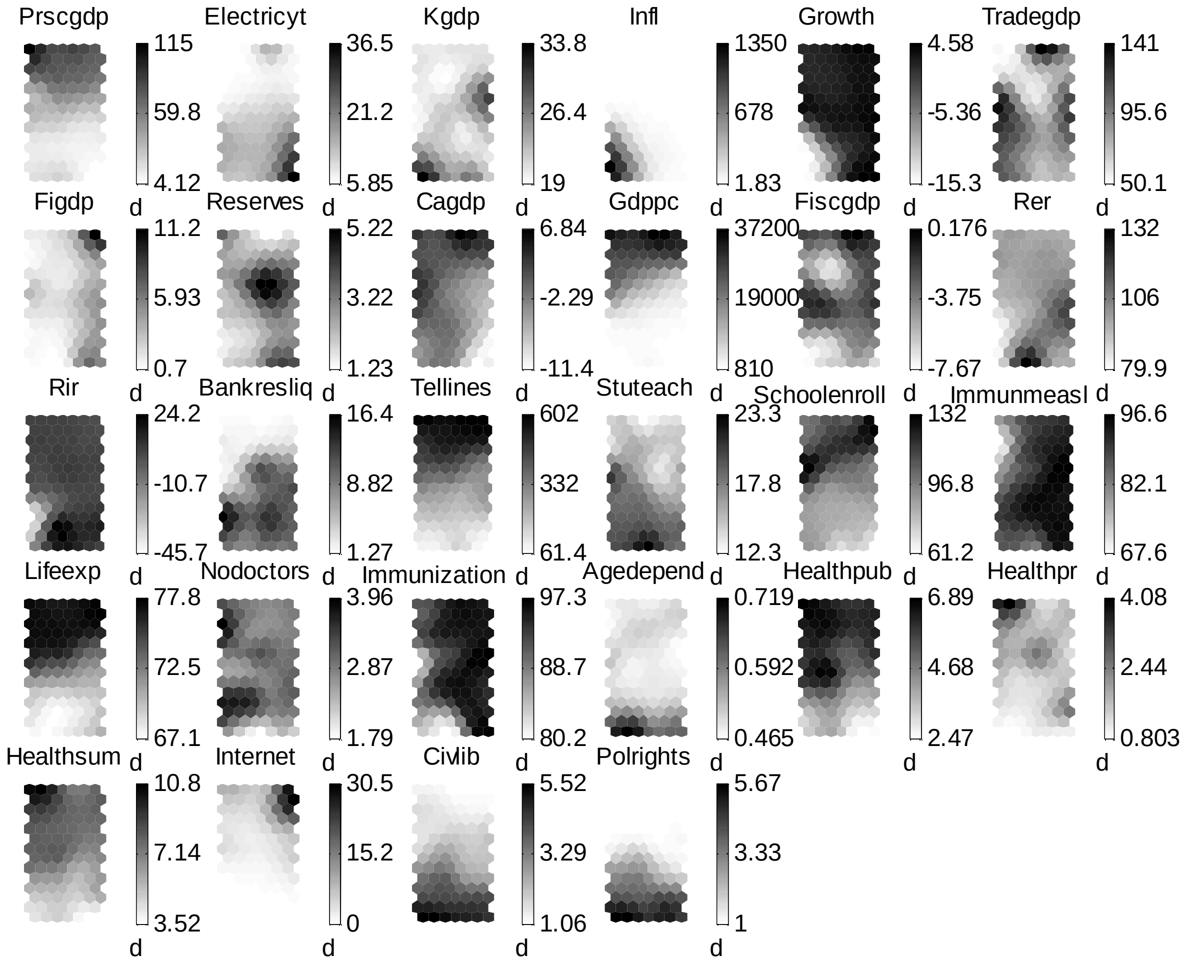


FIGURE 2. COMPONENT MAP OF ALL SOCIOECONOMIC VARIABLES
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

countries have fared since 1992 based upon the aggregate set of variables. Although these maps are useful for facilitating a holistic view of multifaceted convergence, they are cumbersome because they include a very complex set of social and economic variables. Accordingly, an analysis of the patterns in detail is not included at this point because they are more efficiently and effectively discussed in the next sections as distinct social and economic dimensions. Instead, Figure 3 is provided as an overarching summary of the main movements of clusters observed in the aggregate U-matrix. From this diagram, it may be asserted that there exists some evidence of positive change throughout Europe. Whether the transition states are indeed converging with the west or simply maintaining positions/falling behind is an issue that is best addressed with specific reference to the identified dimensions.

In an effort to glean a more explicit understanding of the dimensions/axes interpreted in Figure 3, the variables are now subdivided into (mainly) social and economic sub-sets. From these new maps, one may then glean clearer insights on the nature of the SOMs' axes, as well as the extent of convergence along these axes for all Eurasian states in the sample.

Analysis of Social Clusters and Dimensions

In order to evaluate social convergence, this method first identifies clusters of stable states, and then examines movement among clusters and individual states. Figure 4 provides a U-Matrix constructed on the basis of social variables only -infrastructure, health indicators, and political freedom measures, estimated values of which are shown individually in Figure 5. The U-matrix makes it possible to identify several groups, and ultimately combined with an inspection of Figure 5, to identify consistent dimensions and evaluate the degree of convergence over time.

Several groups are identified from Figure 4: a European Core including regionally cohesive sub-groups, the USA, and a former USSR-core state group including Russia, Belarus, and Ukraine. Outside of these groups, very little cohesion exists, and large distances separate each state, most of which tend to move quite substantially over time, with the exception of Turkey, Tajikistan, Armenia, and Albania.

The largest and most cohesive cluster in Figure 4 is the European Core. This includes most of the EU plus, at its edges, the Visegrád states (Czech Republic, Poland, Hungary, and Slovakia), and the Baltics (Estonia, Latvia, and Lithuania). This clustering of EU states with EU candidates is remarkable, underscoring longstanding social similarities that underlie recent economic differences. The clustering together of these states based upon several social variables lends credence to the argument that the Visegrád and Baltic states (formerly of the Warsaw Pact) are truly Western European on a social development level, while also supporting cultural assertions by Rosenbaum (2001).

Within the European Core, separated by sporadic yellow (grey in grey scale format) cells, three somewhat discrete clusters exist: first, a southern/central group (Italy, Greece, Austria, Germany). This group of welfare states is distinguished by a high number of doctors per 1000 population. On this specific measure the EU is similar to the group comprised of Russia, Ukraine, Belarus, and Kazakhstan (see Healthpub in Figure 5) where extensive public health services were extended to the population, a legacy of central planning in the Soviet core area. Second, a recent (late 1990s) Scandinavian group can be identified, distinguished by high levels of internet use, fewer doctors, more teachers, and higher school enrollment levels. Finally, Ireland stands alone throughout much of the decade, but is joined by Spain and Belgium in recent years. Separating these countries from the rest of Europe are larger school classes and much lower immunization rates. The USA is at the top of the social map, but clearly distinct from Europe. Again, by examining Healthpub in Figure 5, one may see how U-matrix positions can be attributed to an extreme estimated score on a specific variable, in this case the diminutive role of government in American healthcare. In association with US isolation on the left side of the U-matrix, this set of observations provides considerable insight for defining the overarching horizontal dimension as individual responsibility (left) versus government welfare (right).

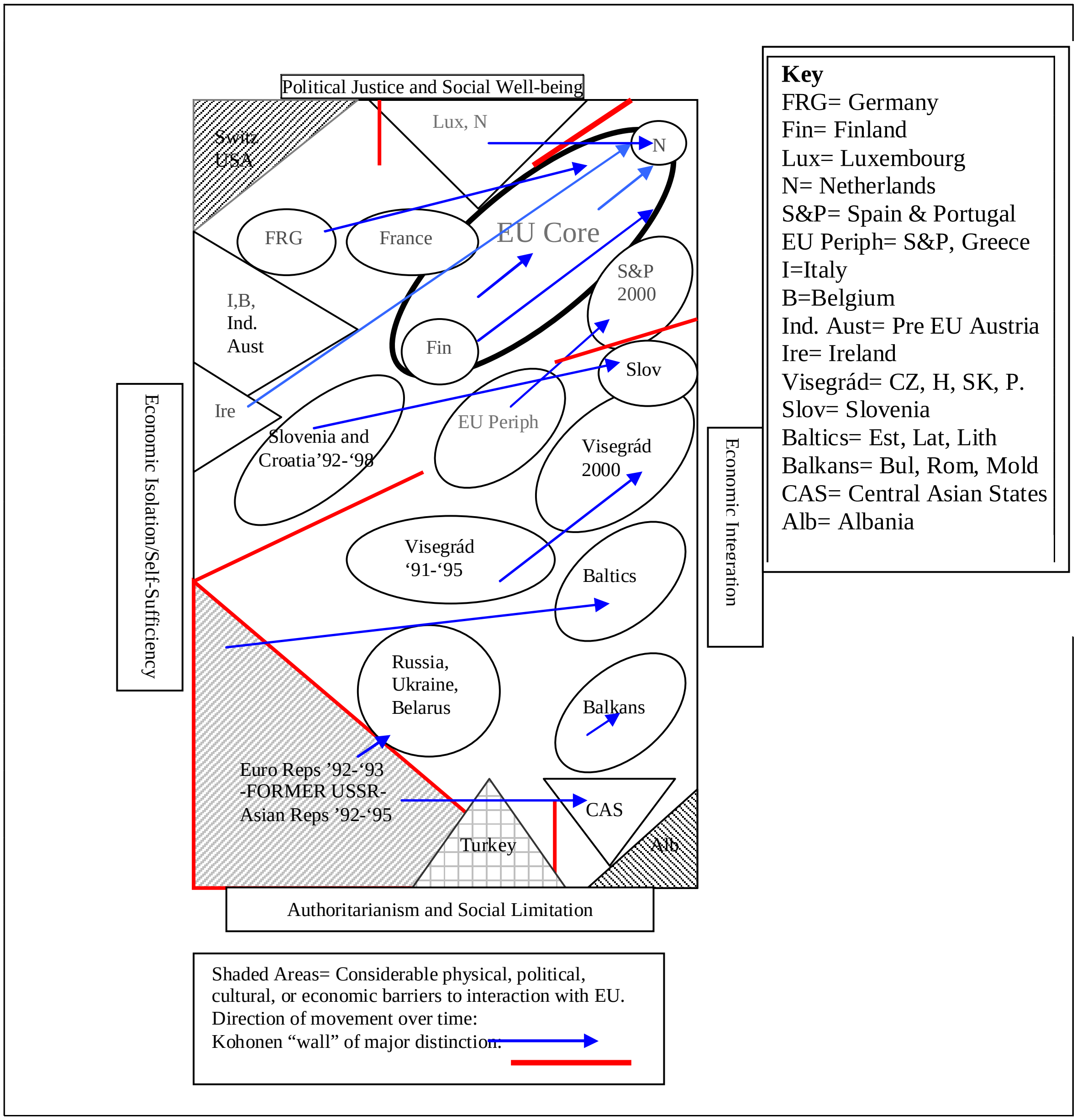


FIGURE 3. ABSTRACTED DIAGRAM OF GROUPS BASED ON FIGURE 1
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

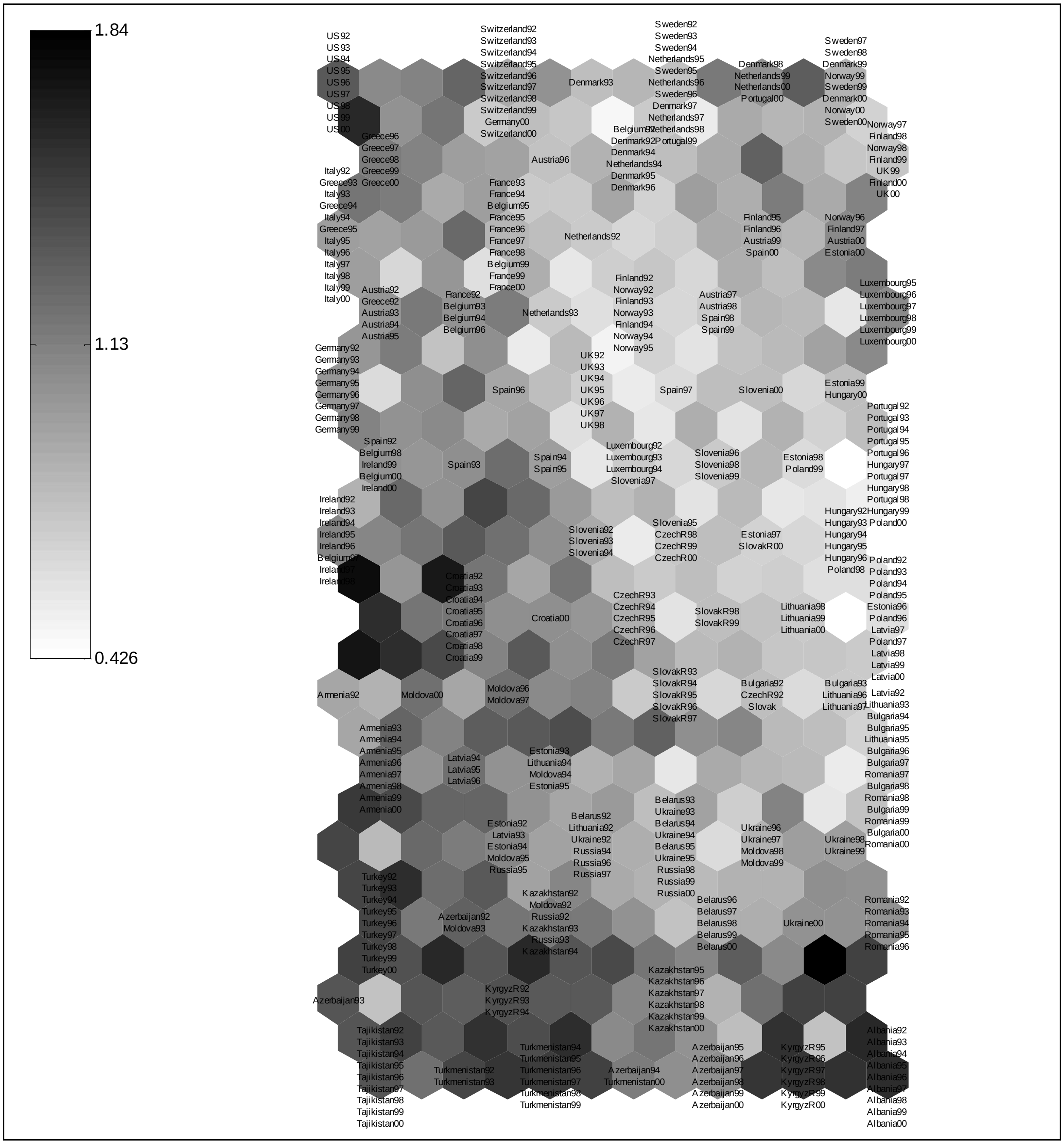


FIGURE 4. U-MATRIX BASED ON SOCIAL INDICATORS
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

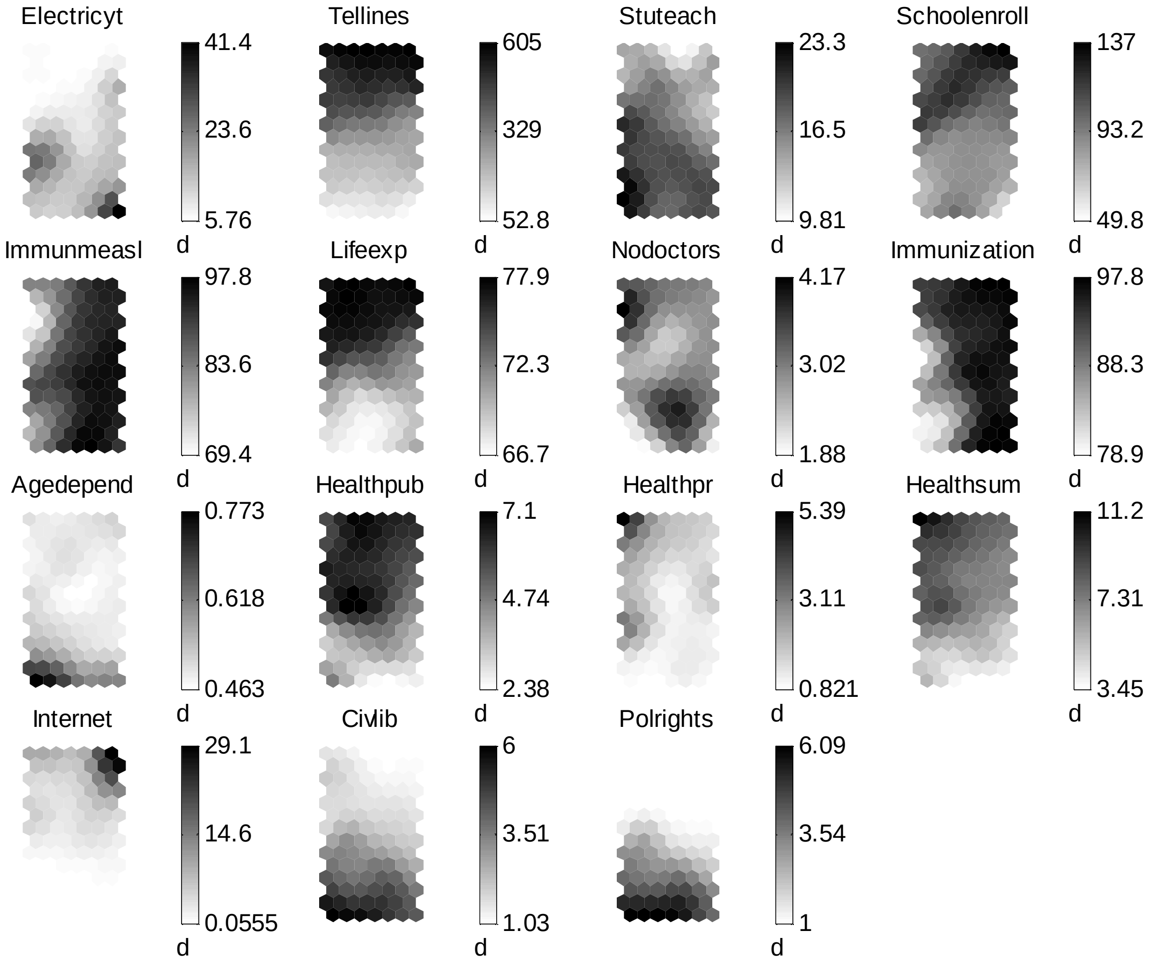


FIGURE 5. COMPONENT MAPS OF CONTRIBUTING SOCIAL VARIABLES
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

Several countries remain either completely or virtually stable between 1992 and 2000. Sharing few commonalities other than the fact that most of them are not EU-Core, these countries include Switzerland, the USA (our benchmark), Ireland, Turkey, Tajikistan, Albania, and Armenia. Several noteworthy differences were revealed by the estimated variables. First, the position of the US is clearly a result of high private versus public health expenditures, the only social variables in which the US varies notably from the European Core.

This means that although Americans on average enjoy a comparatively high standard of living, they are unique in how much they pay for healthcare. Second, Ireland has fewer teachers and doctors per thousand, and its infrastructure lags behind the European Core. Third, Turkey is isolated from the European Core by low scores on healthcare and education variables, as well as by civil liberties and political rights measures. Fourth, Albania’s stability is based upon high scores on democracy, which conflict with inadequate infrastructure (electricity losses) and

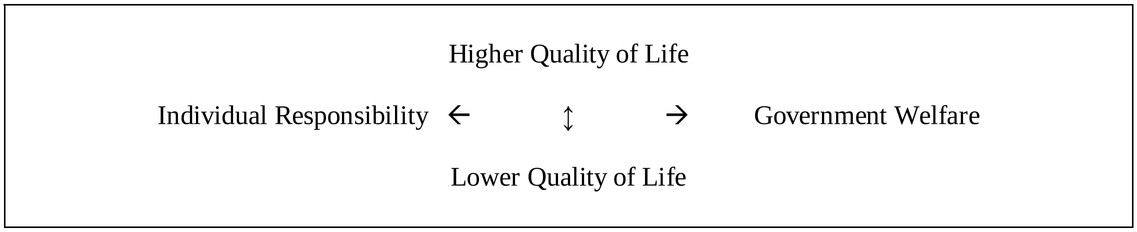


FIGURE 6. INTERPRETATION OF SOCIAL DIMENSIONS

poor estimated healthcare and education. Finally, Tajikistan seems to confirm the nature of the horizontal dimension with its high level of age dependency (Figure 5), which draws it to the left side of Figure 4. In drawing this conclusion, it is assumed that birth rates in Tajikistan are higher partially in response to an absence of state social security systems.

Given the aforementioned observations, the horizontal (left-right) dimension is interpreted as a continuum of social individualism vis-à-vis social welfare as exemplified by the relatively less individualistic European Core. Further, the quality of life variables (life expectancy, infrastructure, education, medical care, and political rights) along the vertical dimension lead us to conclude that the quality of life increases as one moves from the bottom to the top of the map (Figure 6).

Convergence on the Social Dimensions

Overall, the U-matrix of social variables (Figure 4) indicates relatively less movement than that which is found later on the economic map (Figure 7). This means that little evidence exists for convergence in the social dimension. In order to understand the movements in the map, both the component maps (Figure 5) and the original data file were consulted for dramatic changes in variable values. The largest jump and convergence to Western Europe occurs in Estonia. Although its starting point is similar to that of Latvia, it converges much faster to Europe and by the end of the decade groups together with the periphery European countries such as Portugal. Portugal in turn moves from center-right to top-center during the final two years of analysis due to a major improvement in school enrollment, internet use, and public health expenditures during these years.

Another major movement is that of Germany and Austria, which move from center-left to top-center in 2000. Austria's improved quality of live appears to be driven by an estimated increase in immunizations, internet use, and doctors, corresponding to the dates following its own EU accession in 1995. Similarly, Figure 5 hints that Germany's improvement is due to increased internet use, measles immunizations, and public health expenditures. This observation confirms a move toward a larger welfare state in Germany, which is in line with Germany's mid-1990s election of a Red-Green alliance government led by Social Democrats.

All of the Visegrád states witness an increase in the quality of life dimension during the final three to four years. Like Germany under the Social Democrats, these fledgling democracies appear to be moving toward the top right, more toward the model of a European welfare state than the individualistic model of the USA or Switzerland. As an example, the Czech Republic enjoyed improvements since 1997 in nearly all social indicators (except school enrollments and immunizations); these changes were faster than the average changes and suggest evidence of social convergence, especially in internet use, public and private health expenditures, availability of doctors, and quality of infrastructure.

Finland and Norway move from the top-center to the top-right, indicating again a recent improvement in the quality of life, as well as a modest increase in the role of government. This movement appears to be driven primarily by a large increase in estimated internet penetration, and an increase in the estimated number of doctors. Several states outside of the European Core show a very gradual horizontal move, but little or no vertical move. One such country is Turkmenistan, showing a gradual sign of improvement in political rights with no change in its quality of life dimension. Azerbaijan and the Kyrgyz Republic show movements similar to Turkmenistan in the political rights dimension; however, this improvement is accompanied by a worsening in the quality of life for both countries. Ukraine and Bulgaria show no evidence of social convergence to the European Core, but both show some modest positive changes on the political rights dimension.

Other countries show no change in the role of government dimension but indicate significant movements in the quality of life dimension. For example, Romania's sheer vertical move indicates improvements in quality of life since 1997. Similarly, Moldova shows no sign of change in its role of government, but it converges to the European criteria in the quality of life dimension, approaching Slovakia and Croatia. Kazakhstan contrasts with Moldova, showing deterioration in quality of life over time, while political rights have improved. Belarus shows signs of similar worsening in quality of life, with very slight improvements in political rights (similar to Kazakhstan).

Three countries — Russia, Latvia, and Lithuania — move along multiple dimensions. Russia has a very gradual increase in the quality of life, providing some evidence of convergence, accompanied by gradual improvements in the political and civil rights. Lithuania also shows similar positive movements, with signs of convergence to Western Europe. Latvia shows a more volatile pattern over the decade, but the end point is very similar to that of Lithuania. They retreat on both dimensions in 1993, but their recovery in 1996 results in net convergence to Western Europe over the decade.

To summarize the maps of social indicators, the European Core and its multiple fringes clearly corresponds to slight variations on Western Civilization (Rosenbaum, 2001; Jordan, 2002). For example, a Scandinavian cluster of welfare states seems to define the epicenter, surrounded by a Germanic cluster (Germany, Austria, Switzerland), a French cluster (France, Belgium), and a group that encompasses the Central European leading transition states (Visegrád and Baltics); this cluster is in slight contrast to the United States, which shares a high quality of life, but prescribes a smaller role for government. When considering these and other examples, considerable stability exists in the social dimensions, indicating that little convergence has occurred. It is likely that the convergence that has occurred in the region hearkens back to cultural linkages that preceded the superficial division of Europe by the Iron Curtain.

Analysis of Economic Clusters and Dimensions

On the basis of the earlier review of the literature (Rosenbaum, 2001), it is expected that economic change can be readily achieved if deeply-rooted cultural and societal values are in place. In an effort to confirm this expectation, Figure 7 was constructed using variables that represent only the economic measures of the states for comparison with Figure 4. The economic U-matrix includes a combination of absolute indicators and change indicators (such as inflation and growth), as well as domestic measures (GDP per capita, reserves) versus measures of international linkages (FDI, trade).

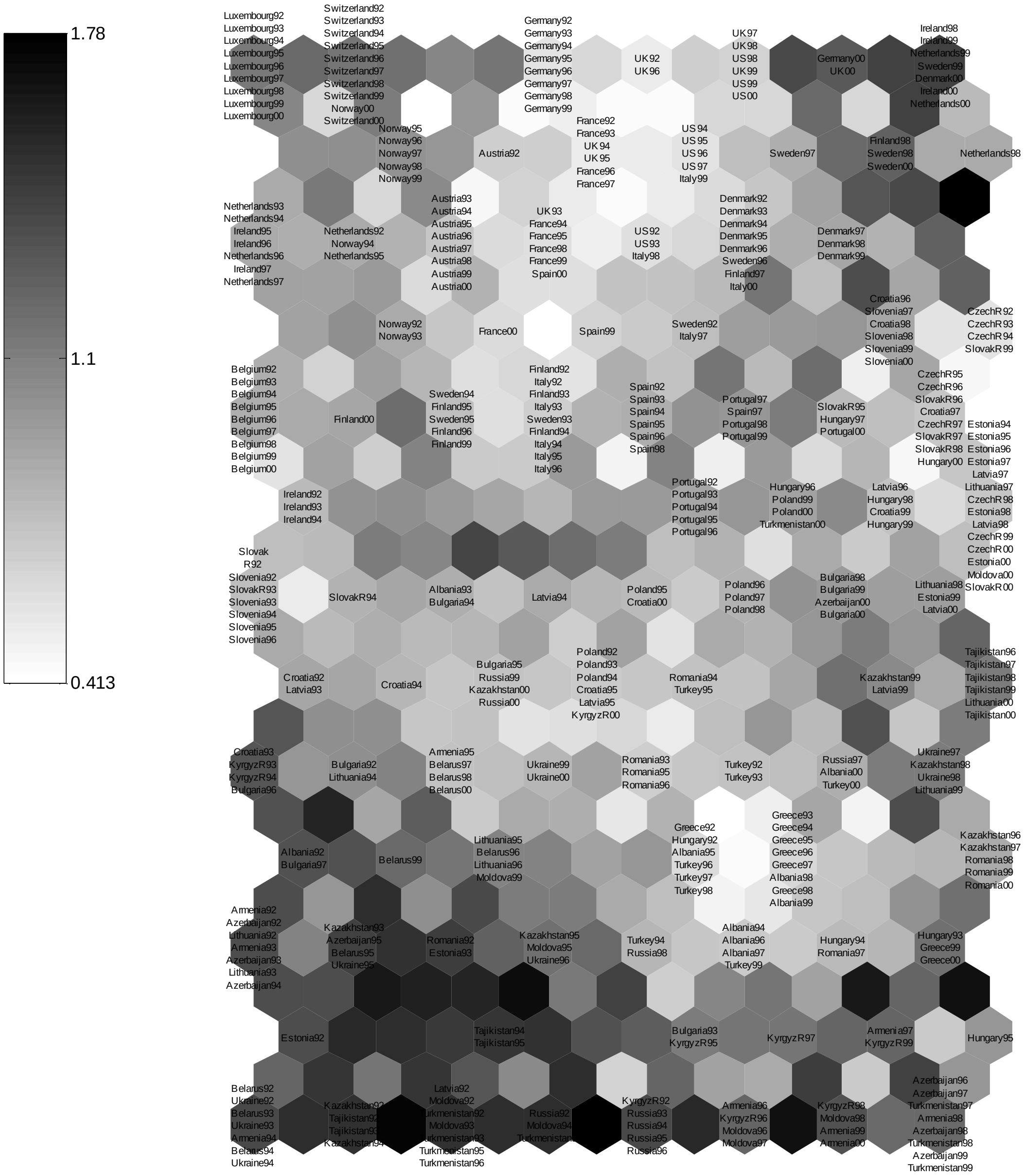


FIGURE 7. U-MATRIX BASED ON ECONOMIC INDICATORS
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

In many ways, this U-Matrix is more complex than the U-Matrix of social indicators. First, there is considerably more movement among the countries over time. Second, relatively more barriers exist that distinguish groups from one another, challenging the notion of a cohesive European Core group based upon economic characteristics. The more nebulous nature of this map is likely attributable to the fact that two of the economic variables are measures of rates of change (inflation, growth).

At a basic level, a European Core exists, made up of large EU countries and the US. These are all wealthy countries with plenty of capital—both public and private. Several wealthy, integrated countries with core locations (and geographically close to one another) remain stable throughout the entire period. These include Luxembourg, Belgium, Switzerland, Germany, and Austria. Within this core, a subgroup of countries can be identified in the upper-right that experienced exceptional growth at very specific time periods; examples include Ireland 98-2000 and Sweden 98-2000. Ireland's recent growth is widely attributed to relatively inexpensive, well-trained, English speaking labor force and targeted government policies, and the significant foreign direct investment these advantages attracted (see, e.g., Trauth, 2000; 2002).

The transition states are much more volatile on the economic map than on the map of social dimensions. This is unsurprising, as the lifestyles of Europeans, as established on the basis of the social map, are more homogeneous than their economic characteristics. Greece remains fairly stable in the bottom portion of this economic group of peripheral Southern Europe, which is periodically joined by shock-therapy Poland (1992-95), some unstable and rapidly changing former Soviet Republics including Ukraine 1999-2000 and Kazakhstan 2000, as well as Turkey (whose growth efforts are often derailed by economic crises, with correspondingly volatile economic growth) and Albania, which shows rapid change throughout the 1990s in response to far-reaching reforms.

Only a faint barrier distinguishes Europe's core (EU plus Switzerland and USA, minus Spain, Portugal, and Greece) from the transition states (which include the EU's periphery). This lack of clear distinction is attributed to the fact that there is a mixed bag of absolute and relative/change variables. In interpreting the patterns in Figure 7 on the basis of the specific variables in Figure 8, it may be seen that Europe's core has good initial conditions but has experienced less growth and fewer effects of reforms (in particular, growth, real GDP per capita growth).

In contrast to the Core, most transition states had worse initial conditions but have experienced more dramatic growth because of their reforms. As expected, it was found that slow starters converge faster (e.g., Albania, which features some of the worst initial conditions, but is propelled toward the top of Figure 7 by its growth rate throughout the nineties). The position of Turkey, which by comparison was much better off in 1992, remains closer to the bottom partially because of more modest changes since that time and constrained by the real effects of economic crises. These observations corroborate Barro and Sala-i Martin's (1992) assertion that high growth rates can be more easily achieved in economies with less advantageous initial conditions. The nature of economic variables therefore further complicates the position of each state, and in interpreting states' positions close attention should be paid to whether each measure is absolute (e.g., real exchange rate, real interest rate) or an indication of change (e.g., growth, inflation).

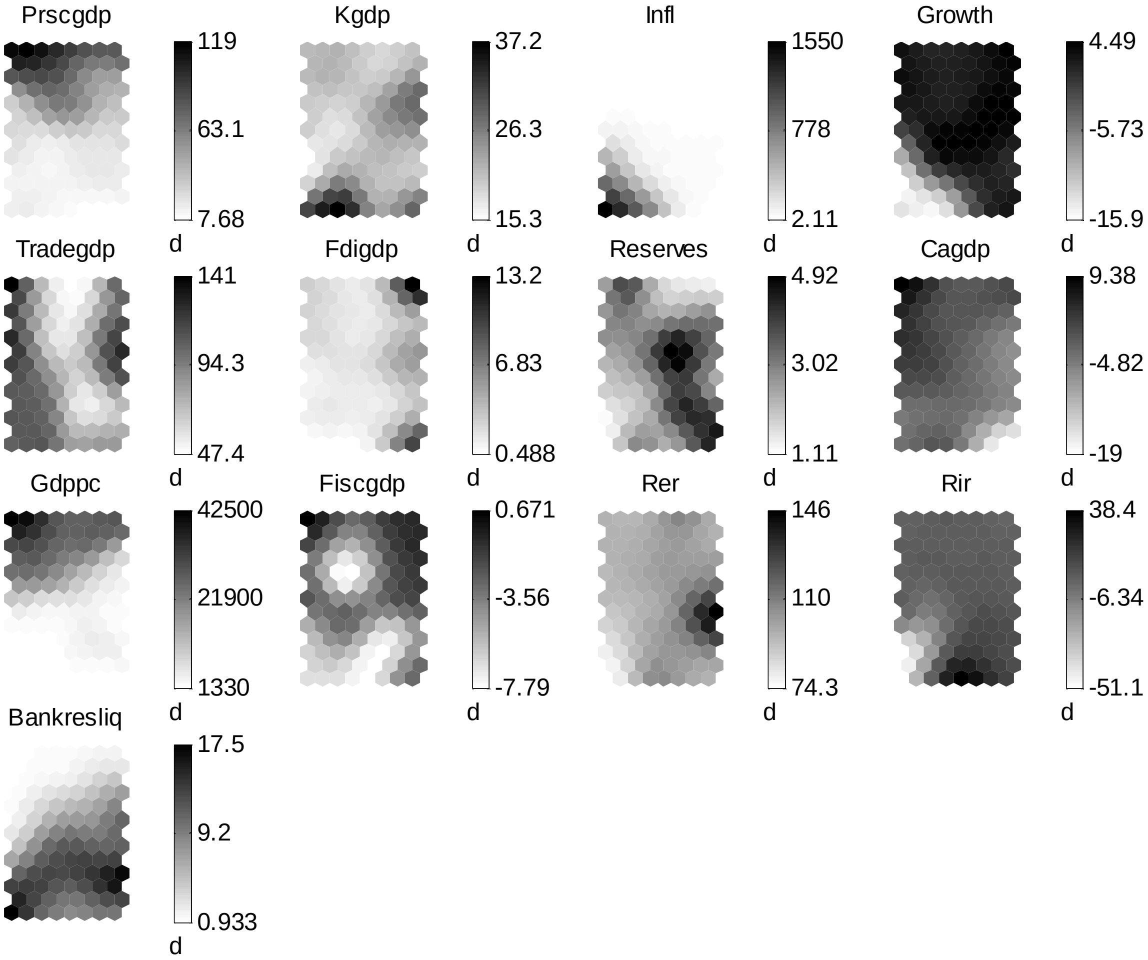


FIGURE 8. COMPONENT MAPS OF ECONOMIC INDICATORS
Note: See http://tbf.coe.wayne.edu/jmasm for map in color.

Convergence on the Economic Dimensions

Movement of countries—both from left to right and from bottom to top—provide evidence of migration to where the EU core rests, namely at the upper-right corner of the map. The EU core itself moves in this direction over time, as indicated in particular by the movements of the Scandinavian states plus Holland and Ireland. In this northeastwardly direction, the United States and United Kingdom move toward the top, with the same relative distance between them, and Italy moves toward the Core by itself. Note however the stability of Belgium and Luxembourg over the same period.

The Caucasus, Baltics, and Turkestan (see note 3 in Fig.9) all cover substantial space (even crossing red and yellow cells – very dark and dark cells in grey scale format) during the first few years of the 1990s. This migration represents a movement away from historical USSR-oriented trade and toward truly globally integrated trade and investment, which reinforces our interpretation of openness being both on the left and on the right side of the map. Among these states, the most profound moves are in the case of Tajikistan, which had been firmly lodged in the Soviet sphere of influence at the outset of our study (Figure 7), but approaches the Central European success stories by 2000.

France is a notable exception to the overarching upward trend, moving slightly toward the bottom and left. Similarly, the Czech Republic moves slightly toward the bottom, which is probably indicative of the difficult fiscal conditions during the late 1990s, corroborated by evidence in the component map that points to a shortage of capital. The Czechs responded by looking to FDI to treat their current account deficit since 2000.

Hungary, Estonia, Latvia and Lithuania demonstrate the greatest convergence among the sixteen transition economies in the data, both horizontally toward the right and vertically toward the top. Underscoring this substantial move across Figure 7 is the fact that these states cross a yellowish color (grey in grey scale format) barrier toward Europe, indicating a significant decline in economic distance. Among these four, Hungary is the lone state with a right-side starting point from which it moved vertically upwards. The three Baltic countries show a significant movement away from the other former European Soviet Republics that are concentrated in the lower-left corner.

The remaining countries show either extremely modest convergence or considerable volatility over time. Tajikistan and Kazakhstan show the largest vertical move in this group of eight. Specifically, they move in a mostly northeastward direction, incorporating the convergence features of both the horizontal and the vertical move. Belarus and Ukraine can be grouped together, moving mainly toward the top and right until the late 1990s when the direction seems to shift to the left. This is interpreted as a slowdown in their trajectories of convergence, but it could also be brought on by embracing foreign trade and investment. Armenia, Azerbaijian, Moldova, and Turkmenistan are propelled by increasing GDP per capita and growth rates, declining inflation, deeper financial markets, and improved fiscal balances. Moldova seems to have made a late move of convergence after 1998 when it separated from the others in this group. While some progress is evident, this cluster seems to have converged least among the Eurasian states.

The remaining states show considerable volatility. The Kyrgyz Republic demonstrates some of the most volatile movement among transition states, moving to the top and right in 1993-94 and then falling back in 1995, only to jump toward the top again in 2000. Given a lack of data to support this jump the sustainability and the evidence for a continuing convergence is not very clear at this point.

Distinct from the Kyrgyz Republic, but similar in volatility, Turkey and Bulgaria also show considerable circularity in their movements. Turkey is very unique in that it seems to complete a full circle in its move over the past decade. It finishes the decade at its starting point; the 1994 crisis pushes Turkey off the convergence path (toward the bottom of Figure 7) and the recovery brings Turkey back to its initial point with no further evidence for convergence through the end-point of the analysis. While it shows similar circularity during the 1990s, Bulgaria seems to have converged to Europe much more than Turkey.

Romania shows more of a horizontal move to the right, especially in the latter part of the decade. It also converges toward the EU Core significantly in 1994 before retreating again. This observation notwithstanding, Romania seems to be much more open to international goods and capital flows after this period. Finally, Russia's most dramatic period of convergence was 1997 toward the top and right, but following its 1998 economic crisis it returned to its approximate initial level. Russia's leftward movement can also be interpreted as a change toward integration, which lies in marked contrast to Russia's historical policy of autarky (self-sufficiency).

Taking into account the aforementioned movements and subsequent investigation of the component variables and data set, Figure 9 is

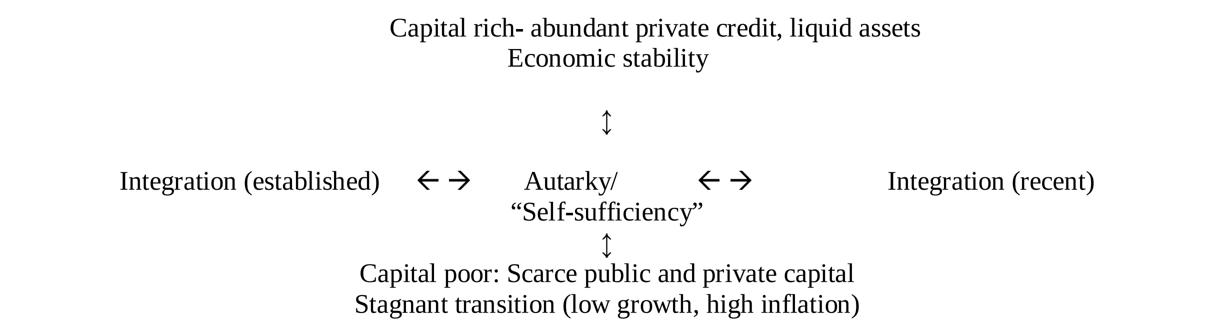

Capital rich- abundant private credit, liquid assets
Economic stability
↕
Integration (established)
←→
Autarky/
"Self-sufficiency"
←→
Integration (recent)
↕
Capital poor: Scarce public and private capital
Stagnant transition (low growth, high inflation)

FIGURE 9. INTERPRETATION OF ECONOMIC DIMENSIONS

presented as a conceptual simplification of the economic dimensions.

As in Figure 6, the labels are based upon the variables that have distinct top/bottom or left/right trends in Figure 8. For example, according to the measure of trade as a share of GDP in Figure 8, the countries on the right and on the left of the map clearly trade more than those in the middle, and the same holds true for FDI as a percentage of GDP; this signifies open (integrated) economies on both the left and the right side of Figure 7. Moreover, bank reserves, reserves, income levels, current account balances and private sector credit all tend to indicate that the top/bottom dimension represents a continuum of capital abundance/capital scarcity.

The meaning of the horizontal axis is less clear than the vertical axis. The right side seems to represent greater capital account openness because FDI as a share of GDP is higher from left to right. But the openness story is less clear when one considers trade as a share of GDP. In any case, the horizontal move seems to capture some of positive aspects of the vertical move as well, because lower inflation is evident when one moves toward the right (and top). Similarly, growth and GDP per capita, FDI as a share of GDP, all increase in that direction, and the fiscal balances improves toward the top and right. This interpretation could suggest that a move toward the right represents a stage in convergence; however, the full convergence occurs if the horizontal move is combined with the vertical move.

Along these lines, Wagner and Hlouskova (2001) also differentiate between convergence and loosely-speaking convergence, where the latter captures convergence in the economic structure of the countries involved on account of strengthened linkages via trade and foreign direct investment.

Conclusion

This article demonstrates the utility of Kohonen maps for visualizing Eurasian convergence over time (1992-2000) on the basis of 28 socioeconomic measures. It contributes to the literature by identifying and explaining the relative movements of states on a two-dimensional map, concurrently taking into account a large number of measures. In past work, measures had to be considered individually when discussing convergence, which explains why past work has to a certain extent led to sometimes conflicting conclusions. This analysis, thus, sheds some light on this debate.

[1] The USA is included in the analysis as a point of reference.
[2] To build our Kohonen maps, we used Matlab code (Laboratory of Computer and Information Science at the Helsinki University of Technology), available at www.cis.hut.fi/projects/somtoolbox/.
[3]Turkestan is the supranational physiographic and cultural region that includes Kazakhstan, Uzbekistan, Turkmenistan, Tajikistan, and Kyrgyzstan—all former Soviet Socialist Republics in Central Asia

In addition to the overall analysis that included the aggregate set of variables, the economic and social variables were analyzed separately. The resulting maps demonstrate several differences between the economic and social convergence processes. On one hand, the social variables seem to capture more stable traits of states than economic variables. Interestingly, the 2004 newcomers to the European Union are clearly clustered with most of the rest of the EU in the analysis of social variables, suggesting deeper cultural commonalities between these groups of states.

Nevertheless, there is clearer evidence to support economic convergence than social convergence. It is believed that this is because initial conditions vary, which is particularly evident in the Visegrád states and the Baltics, and to a lesser degree Russia, the other European former Soviet Republics, and the Balkans. This supports the argument regarding the effects of historical and cultural linkages presented by Rosenbaum (2001). This regional gradient of European-ness corresponds to the notion of Brussels distance introduced by Fisher, Sahay, and Végh (1998). Conversely, states that are culturally distant from Western Europe (such as Central Asian Republics or the region known as Turkestan) exhibit less economic convergence.

The main dimensions identified in the analysis suggest major higher-level constructs that can be used in interpretation of the results and potentially also in future research. In the overall analysis, the two major dimensions were the level of political justice and social well-being and the extent of economic integration. In the more detailed analyses, the dimensions of the social map were related to the quality of life and the respective roles of governments and individuals in providing social welfare. Finally, the analysis of the economic variables led to the identification of two dimensions related to the timing of economic integration and the availability of capital. Future research should use multiple methods to analyze the relevance of these constructs as well as specific policy reforms.

Methodologically this study demonstrates the usefulness of Kohonen maps to visualize large numbers of variables and complex sets of data in a two-dimensional space. It was found that the approach of using the state-year pairs as the basic unit of analysis, originally introduced by Costea et al. (2001), very useful in mapping the time-dependent changes in the relative positions of the states. Future research should pay special attention to the implications of analyzing absolute variables and measures of change, which may have impacted the results of this analysis.

In summary, this article provides an analysis of the socio-economic convergence of Eurasian states with the European Core. It demonstrates the usefulness of Kohonen maps as a tool for analyzing large sets of macroeconomic data over time. The study also distinguishes between economic and social factors, identifying much more proof of the former than the latter. This study identifies and reports indisputable evidence of economic convergence by European transition states that becomes less clear in countries farther to the east. It is argued that such convergence is either facilitated or constrained by preconditions that are either specific to each country or to a broader culture. This article lays the groundwork for further analysis of country-specific reforms and how they interact with initial conditions to impact convergence in the transition states.

References

Almström, L. (2000, February). What is Europe? *The Economist*, *12*.

Aslund, A., Boone, P., & Johnson, S. (1996). How to stabilize: Lessons from post-communist countries. *Brookings Papers Economic Activity*, *1*, 217-291.

Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economic*, *106*, 407-443.

Barro, R. J. & Sala-i Martin, X. (1992). Convergence. *Journal of Political Economy. 100*(2), 223-251.

Bartlett, D. L. (1997). *The political economy of dual transformations: Market reform and democratization in hungary*. Ann Arbor, MI: University of Michigan Press.

Jozef, M. B. (1998). Transformation, EU integration, and regional cooperation in eastern Europe. *Comparative Economic Studies*, *40*(4), 33-59.

Brada, J. & Kutan, A. (2001). The convergence of monetary policy between candidate countries and the European Union. *Economic System, 25*, 215-231.

Brada, J. Kutan, A., & Zhou, S. (2002). Real and monetary convergence within the european union and between the european union and candidate countries: A rolling cointegration approach. (William Davidson Institute Working Papers Series 458, William Davidson Institute, University of Michigan Business School, 2002).

Brezinski, H. & Fritsch, M. (1996). Bottom-up transformation: Prerequisites, scope and impediments. *International Journal of Social Economics, 23*(10/11).

Costea, A., Kloptchenko, A., Back, B. (2001). Analyzing economical performance of central-east-european countries using neural networks and cluster analysis. In Ivan, I. & Rosca, I (eds.), *Proceedings of the Fifth International Symposium on Economic Informatics*, 1006-1011.

Deboeck, G. (1998, January). Financial applications of self-organizing maps. *American Heuristics Electronic Newsletter*, 1-7.

Deichmann, J. I., Eshghi, A., Haughton, D., Sayek, S., Teebagy, N., & Topi, H. (2003). Geography matters: Kohonen classification of foreign direct investment in transition economies. *Journal of Business Strategies, 20*(1), 23-44.

DeMelo, M., Cevdet, D., & Gelb, A. (1996). *From plan to market*. World Bank Policy Research Working Paper No. 1564.

DeMelo, M., Cevdet, D., Gelb, A., & Stoyan, T. (2001). Circumstance and choice: The role of initial conditions and policies in transition economies. *World Bank Economic Review, 15*(1), 1-31.

Easterly, W. & Ross, L. (2003). Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics, 50*(1), 3-39.

Ellman, M. (1997). The political economy of transformation. *Oxford Review of Economic Policy, 13*(2), 23-32.

EBRD (European Bank for Reconstruction and Development, 1999). *Transition Report: Ten Years of Transition*. London: EBRD.

Estrin, S., Urga, G., & Lazarova, S. (2001). Testing for ongoing convergence in transition economies, 1970 to 1998. *Journal of Comparative Economics*, *29*, 677-691.

Falcetti, E., Raiser, M., & Sanfey, P. (2002). Defying the odds: Initial conditions, reforms, and growth in the first decade of transition. *Journal of Comparative Economics, 30*, 229-250.

Fisher, S., Sahay, R., & Végh, C. (1998). *How far is Eastern Europe from Brussels?* International Monetary Fund Working Paper, 98/53.

Freedom House. *Annual Freedom in the World Country Scores 1972-73 to 2001-2002*. www.freedomhouse.org/research/freeworld/FHSCORES.xls.

Gallup, J. L., Sachs, J. D., Mellinger, A. (1999). *Geography and Economic Development*. Center for International Development (CID) Working Paper. No. 1.

Genov, N. (1998). Transformation and anomie: Problems of quality of life in Bulgaria. *Social Indicators Research, 43*(1-2), 197-210.

Ghemawat, P. (2001). Distance still matters: The hard reality of global expansion. *Harvard Business Review*, September, 2001, 3-11.

Graham, B. & Hart, M. (1999). Cohesion and diversity in the european union: Irreconcilable forces? *Regional Studies, 33*(3), 259-68.

Jordan, T. (2002). *The european culture area: A systematic geography*. Lanham: Rowman & Littlefield.

Kaski, S., Kohonen, T. (1995, October). *Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world*. Proceedings of the Third International Conference on Neural Networks in the Capital Markets, London, England, 11-13.

Kloptchenko, A., Back, B., Vanharanta, H., Toivonen, J., & Visa, A. (2003). *Prototype-Matching System for Allocating Conference Papers*, Proceedings of the 36th Hawaii International Conference on System Sciences.

Kočenda, E. (2001). Macroeconomic convergence in transition countries. *Journal of Comparative Economics, 29*, 1-23.

Kohonen, T. (1982). Analysis of a simple self-organizing process. *Biological Cybernet. 44*, 135-140.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer Verlag.

Krueger, G. & Ciolko, M. (1998). A note on initial conditions and liberalization during transition. *Journal of Comparative Economics, 26,* 718-734.

Kutan, A. & Yigit, T. (2006, in press). Nominal and real stochastic convergence of transition economies. *Journal of Comparative Economics*.

Lang, I. (2003). Sustainable development: A new challenge for the countries in central and eastern europe. *Environment, Development and Sustainability, 5*(1-2), 167.

Oja, E. & Kaski, S. (1999). *Kohonen maps*. Amsterdam: Elsevier Science.

Popov, V. (2000). Shock therapy versus gradualism ten years down the road. *Comparative Economic Studies, 42*(3), 121-126.

Rosenbaum, E. F. (2001). Culture, cognitive models, and the performance of institutions in transformation countries. *Journal of Economic Issues, 35*(4), 889-909.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy. 94*(5), 1002-1037.

Sachs, J. D. & Andrew M. W. (1995). *Economic Convergence and Economic Policies*, NBER Working Paper Series, No. 5039, 1995.

Sachs, J. The transition at mid-decade. *American Economic Review Papers and Proceedings, 86*(2), 128-133.

Sachs, J. D. (2001). *Tropical underdevelopment*. NBER Working Paper Series, No. 8119, 2001.

Skupin, A. (2002, January/February). A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications*, 50-58.

Szalkowski, A. & Jankowicz, D. (1996). The ethical problems of personnel management in a transition economy. *International Journal of Social Economics, 26*(12), 14-18.

Trauth, E. M. *Using Public Policy to Influence Regional Futures: Lessons from Ireland's Information Economy*, www.greatvalley.org/nvc/events/pdf/Eilenn_trauth.pdf.

Trauth, E. M. (2000). *The culture of an information economy*. Norwell, MS: Kluwer Academic Publishers.

United Nations Economic Commission for Europe (2000, April). Catching up and Falling Behind: Economic Convergence in Europe, *UN/ECE News*.

Wagner, M. & Hlouskova, J. (2001). The CEEC10's Real Convergence Prospects. *Transition Economics Series, 20.*

Woo, W. T. (1998). Improving the performance of enterprises in transition economies. In Woo, W. T., Parker, S., & Sachs, J. (eds.), Economies in Transition. Cambridge, MS: MIT Press.

World Bank Development Indicators (2002). CD-ROM. Washington, D.C. World Bank.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 95-106

1538 – 9472/06/$95.00

Entropy Criterion In Logistic Regression And Shapley Value Of Predictors

Stan Lipovetsky
GfK Custom Research Inc.

Entropy criterion is used for constructing a binary response regression model with a logistic link. This approach yields a logistic model with coefficients proportional to the coefficients of linear regression. Based on this property, the Shapley value estimation of predictors' contribution is applied for obtaining robust coefficients of the linear aggregate adjusted to the logistic model. This procedure produces a logistic regression with interpretable coefficients robust to multicollinearity. Numerical results demonstrate theoretical and practical advantages of the entropy-logistic regression.

Keywords: entropy, logistic regression, multicollinearity, net effects, Shapley value.

Introduction

Logistic regression is a widely used tool in regression modeling for a data with a binary output (Pregibon, 1981; Arminger et al., 1995; Long, 1997; Hastie & Tibshirani, 1997; McCullagh & Nelder, 1997; Lloyd, 1999; Lipovetsky & Conklin, 2000). The logistic model is usually obtained by the maximum likelihood criterion applied to the binary output with the logistic link. In this article, the criterion of entropy is applied for constructing a logistic model. Various techniques based on the entropy criterion are well known in information theory, fuzzy data analysis, and other statistical applications (Lindley, 1956; Zeimer & Tranter, 1976; Dukhovny, 2002; Levene & Loizou, 2003; Maes & Netocny, 2003; Handscombe & Patterson, 2004; Bar-Yam, 1997, 2004). The entropy-logistic model yields the coefficients and forecasts very similar to multiple linear regression. It opens a possibility to apply some techniques developed in linear regression to binary modeling, particularly, for estimation of the predictor's contribution and construction of a model robust to the effects of multicollinearity.

Contribution of the predictors in a linear aggregate can be found by the net effects technique. In linear regression analysis the net effect of a predictor is a combination of the direct (as measured by its coefficient squared) and the indirect effects (measured by the combination of its correlations with other variables). The sum of the net effects equals the coefficient of multiple determination of the model. However, the net effect values themselves can be subjected to the multicollinearity in the data so that the estimated net effects can be negative, which is difficult to interpret.

Even in presence of multicollinearity, it is often desirable to keep all variables in the model if their comparative importance is evaluated. A regression model can be considered from the perspective of a coalition among players (predictors) to maximize the total value (quality of fitting). In the cooperative games a useful decision tool developed to evaluate the worth of participants is the Shapley Value imputation (Shapley, 1953; Roth, 1988; Straffin, 1993; Jones, 2000). The Shapley Value (SV) presents each player's input over all possible combinations of players. This technique proved to be very useful in various complicated estimation problems (Conklin et al., 2004; Conklin & Lipovetsky, 2005). In application to statistical modeling, this approach yields a model called

Stan Lipovetsky is an Analytical Services Manager for GfK Custom Research Inc. He serves as an internal and external consultant to GfK-CRI. His primary areas of research are multivariate statistics, multiple criteria decision making, econometrics, microeconomics, and marketing research.

Shapley Value regression (Lipovetsky & Conklin, 2001, 2004, 2005). In the current work, the SV approach to the logistic regression modeling is considered.

Entropy in Binary Response Modeling

Consider a data matrix with the elements x_{ij} of i-th observations ($i=1,\ ...,\ N$) by j-th variables ($j=0,\ 1,\ ...,\ n$), and a dependent variable y of the observed event's success or failure, presented by the binary output (y_i equals 1 if the event occurs, and 0 if it does not). The logistic probability function can be presented as:

$$p_i = \frac{1}{1+\exp(-z_i)}, \tag{1}$$

where z is a linear combination of the independent variables:

$$z_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_n x_{in}\,, \tag{2}$$

where the unknown parameters $a_0, a_1, a_2, ..., a_n$ correspond to the coefficients of the logistic regression model (1)-(2). Probability of the binary outcome is:

$$P_i = p_i^{y_i}(1-p_i)^{1-y_i}\,. \tag{3}$$

Maximum Likelihood objective is defined by the product of all probabilities (3):

$$ML = \prod_{i=1}^{N} P_i = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}\ , \tag{4}$$

or the logarithm of this ML is:

$$\begin{aligned} &\ln ML \\ &= \sum_{i=1}^{N} \ln P_i \qquad . \\ &= \sum_{i=1}^{N} \left(y_i \ln p_i + (1-y_i)\ln(1-p_i) \right) \end{aligned} \tag{5}$$

Maximizing (5) by the parameters in (1)-(2) yields the procedure for constructing a regular logistic regression, as it is known by the literature on categorical data modeling.

Instead of the ML (4) it is possible to consider an objective of a Gibbs distribution:

$$e^{-Entropy} = \prod_{i=1}^{N} P_i^{P_i}\ , \tag{6}$$

so its logarithm that defines the entropy of the data:

$$E \equiv Entropy = -\sum_{i=1}^{N} P_i \ln P_i\ , \tag{7}$$

where the binary probability outcome is defined in (3). The maximum entropy criterion (7) differs from the logarithm of maximum likelihood (5) by weighting the probabilities P_i by their logarithms. The first-order conditions for maximizing the objective (7) by the parameters of the aggregate (2) yields a gradient vector with the elements:

$$\begin{aligned} U_k &\equiv \frac{\partial(-E)}{\partial a_k} \\ &= \sum_i \left(\ln P_i \frac{\partial P_i}{\partial a_k} + P_i \frac{\partial \ln P_i}{\partial a_k} \right) \\ &= \sum_i P_i (1+\ln P_i) \frac{\partial \ln P_i}{\partial a_k} \\ &= \sum_i P_i (1+\ln P_i) \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \frac{\partial p_i}{\partial a_k} \\ &= \sum_i P_i (1+\ln P_i) \left(\frac{y_i - p_i}{p_i(1-p_i)} \right) p_i(1-p_i) \frac{\partial z_i}{\partial a_k} \\ &= \sum_i P_i (1+\ln P_i)(y_i - p_i) x_{ik} = 0, \end{aligned} \tag{8}$$

where the derivatives are sequentially taken from the functions (3), (1), and (2).

To solve a non-linear system of equations the Newton-Raphson algorithm can be applied. The vector with elements (8) is approximated as:

$$\begin{aligned} U &= U^{(0)} + \frac{\partial U}{\partial a}(a^{(t+1)} - a^{(t)}) \\ &= \frac{\partial(-E)}{\partial a'} + \frac{\partial^2(-E)}{\partial a \partial a'}(a^{(t+1)} - a^{(t)}), \\ &= 0 \end{aligned} \tag{9}$$

where a is a vector of the $(n+1)$-th order of all the coefficients a_k (2), and t denotes a step of iteration. The process of estimating the vector of parameters is:

$$a^{(t+1)} = a^{(t)} - \left(\frac{\partial^2(-E)}{\partial a \partial a'}\right)^{-1} \frac{\partial(-E)}{\partial a'} = a^{(t)} - H^{-1}U, \tag{10}$$

where H is a matrix of second derivatives, or Hessian, and H^{-1} is this matrix inversed.

Using (8), this matrix is constructed:

$$\begin{aligned} H_{jk} &= \frac{\partial^2(-E)}{\partial a_j \partial a_k} = \frac{\partial U_k}{\partial a_j} \\ &= \sum_i P_i \left((2 + \ln P_i)(y_i - p_i)\frac{\partial \ln P_i}{\partial a_j} - (1 + \ln P_i)\frac{\partial \ln p_i}{\partial a_j} \right) x_{ik} \\ &= \sum_i P_i \begin{Bmatrix} (2 + \ln P_i)\left[(y_i - p_i)^2 - p_i(1 - p_i)\right] \\ + p_i(1 - p_i) \end{Bmatrix} x_{ij} x_{ik}. \end{aligned} \tag{11}$$

In the brackets at the right-hand side (11), the difference of the items $(y_i - p_i)^2$ and $p_i(1 - p_i)$ of two forms of the variance estimations is always small. The total of these items is negligible (Becker & Le Cun, 1988; Bender, 2000), so (11) can be presented as:

$$\begin{aligned} H &= X' diag\left(P_i p_i (1 - p_i)\right) X \\ &= X' diag\left(p_i^{1+y_i}(1 - p_i)^{2-y_i}\right) X\ , \\ &\equiv X' W\ X \end{aligned} \tag{12}$$

where the diagonal matrix of weights W is defined using (1) and (3), and X is the data matrix in the aggregate (2) (with a uniform first column corresponded to the intercept). So (12) is a weighted matrix of the second moments of the predictors in the model (2).

The gradient vector (8) can be rewritten in a matrix form as:

$$U = X' diag\left(P(1 + \ln P)\right)(y - p), \tag{13}$$

where P, p, and y are the vectors with the elements P_i (3), p_i (1), and the binary output y_i, respectively. Then the iterative process (10) is:

$$\begin{aligned} a^{(t+1)} &= a^{(t)} - \left(X'WX\right)^{-1} X' diag\left(P(1 + \ln P)\right)(y - p) \\ &= \left(X'WX\right)^{-1} X'W \begin{Bmatrix} Xa^{(t)} \\ -diag\left(W^{-1}P(1 + \ln P)\right)(y - p) \end{Bmatrix} \\ &\equiv \left(X'WX\right)^{-1} X'W\xi^{(t)}, \end{aligned} \tag{14}$$

where $\xi^{(t)}$ is the so called working dependent variable that denotes the expression in figure parentheses (14). The right-hand side of the expression (14) presents the solution of the system (8) as a weighted linear regression with the adjusted response variable:

$$\begin{aligned} \xi_i^{(t)} &= (Xa^{(t)})_i - diag\left(W_i^{-1}P_i(1 + \ln P_i)\right)(y_i - p_i) \\ &= z_i^{(t)} - diag\left(\frac{1 + \ln P_i}{p_i(1 - p_i)}\right)\varepsilon_i^{(t)} \end{aligned}\ , \tag{15}$$

where $z^{(t)} = Xa^{(t)}$ is a vector of the linear aggregate (2), $\varepsilon^{(t)} = y - p$ is a vector of deviations between the empirical binary response and the theoretical probability (1). The solution (14) corresponds to the normal system of equations of the weighted least square problem $(X'WX)a = X'W\xi$ with the adjusted dependent variable (15), so the process (14)-(15)

is the Iteratively Reweighted Least Squares, or IRLS. Numerical simulations show that the weight matrix W in Hessian (12) quickly becomes approximately a scalar matrix, and the IRLS process converges already after several steps.

Consider numerical results from a real research project involving bank mortgages with the data elicited from 403 customers. The binary response defines the customers' "Satisfied or not" feeling on the bank performance with a mortgage, and the independent variables from x_1 to x_8 are shown in Table 1. The management of the bank is interested in estimating the predictors influence on increasing the client's satisfaction with the bank. Table 1 presents the pair correlations of the dependent with independent variables, and the coefficients (beginning from the intercept) with their t-statistics for the multiple linear, the regular logistic, and the entropy-logistic regressions. The entropy-logit model is constructed using the IRLS approach (14)-(15), and the *t*-statistics for the coefficients are estimated using bootstrapping.

Table 1 shows that the variables x_2, x_3, x_5, and also x_7 are the most significant predictors, while the other variables x_1, x_4, $x_{6,}$ and x_8 are unimportant in the models. In spite of all positive pair correlations with the binary dependent variable, the coefficients of the least significant variables change their sign in the models (negative sign for x_8 in the linear, for x_1 in the logit, and for both of them in the entropy-logit model). It is the effect of multicollinearity that distorts the estimation by the models.

Table 1. Binary models of customer satisfaction.

Variable		**Correlation**	**Linear regression**		**Regular Logistic**		**Entropy Logistic**	
			coeff	**t-stat**	**coeff**	**t-stat**	**coeff**	**t-stat**
Overall sat. w. mortgage loan	y	1	-.919	-6.73	-10.841	-7.73	-1.600	-6.68
Satisfaction with rate	x_1	.347	.0002	0.01	-.026	-0.34	-.0002	-0.01
Right type of loan	x_2	.402	.038	3.11	.233	2.89	.043	2.35
Feel like a valued customer	x_3	.498	.049	3.43	.340	3.76	.055	2.91
Bank knows customers needs	x_4	.438	.007	0.57	.060	0.79	.007	0.36
Communication	x_5	.423	.026	2.61	.120	1.98	.031	1.95
Handling mortgage payment	x_6	.359	.023	1.13	.127	0.92	.027	0.89
Posting payments accurately	x_7	.352	.039	1.76	.396	2.34	.044	1.29
Posting payments timely	x_8	.338	-.009	-0.40	.022	0.13	-.011	-0.32

Table 2 contains the ratios of the coefficients of the regular logit to the linear model, of the regular logit to the entropy-logit model, and of the entropy-logit to the linear model, respectively. The coefficients themselves vary differently in each model, and the ratios of the regular logit coefficients to the coefficients of the other models belong to a wide span of values. However, the ratio of the coefficients of the entropy-logit to the linear model is amazingly stable.

The last column in Table 2 shows that with exception of the intercept (that incorporates the influence of all the predictors), and slightly different ratios for the most insignificant variables x_1, x_4, and x_8, all absolute values of all the ratios are practically the same.

Denoting the theoretical, predicted values of the output as $\tilde{y}_{lin}$, $\tilde{y}_{\log}$, and $\tilde{y}_{ent}$ for the linear, logit, and entropy-logit models, respectively (where 0 and 1 values correspond to the rounded values of the probability below or above 0.5), and estimating the coefficient of pair correlation between the linear and entropy-logit predictions, it is possible to obtain a value of 0.9995, while the correlations between the predictions by the other models are about 0.94-0.95. Comparison of the models' predictive ability is presented in Table 3 by several cross-sections.

Section A of Table 3 presents the cross-tabulation of the empirical binary output y with the prediction $\tilde{y}_{lin}$ by the linear model, where 0 and 1 values are correctly identified 169 and 143 times, so the total of the correct forecasts is 312 within 403 observations, or 77.4%. The next section B in Table 3 shows the cross-tabulation of the empirical y with the prediction $\tilde{y}_{\log}$ by the regular logit model, where 0 and 1 outputs are correctly identified 173 and 138 times, with the total of correct forecasts equal 311, or 77.2%. Section C in this table presents the cross-tabulation of the empirical y with the prediction $\tilde{y}_{ent}$ by the entropy-logit model, that correctly identifies 0 and 1 outputs 167 and 143 times, so the total rate of correct forecasts is 310, or 76.9%. It is interesting to note that both linear and entropy-logit models better identify the level y=1 of the satisfied customers. The other sections D, E, and F of Table 3 compare predictions by each two of the three constructed models, where again the linear and entropy-logit models yield very close counts of 204 and 195 for 0 and 1 binary outputs, so the total rate of the coinciding results equals 99%.

The observed results are typical for various data sets. They show that all the considered models produce results of a similar quality. However, while a linear regression could yield an output beyond 0-1 interval in its prediction, both logistic regressions have the same link (1) with the linear aggregate of the predictors, so they always yield a probability in the 0-1 range. On the other hand, a close inspection of the results produced by the entropy-logit and linear models suggests a possibility to apply techniques developed for the linear models to a logistic model in its entropy-logit formulation. In the work (Lipovetsky and Conklin, 2001) the Shapley value regression was introduced for estimating the net effects of the predictors shares in the linear model. The proportionality between the coefficients of linear and entropy-logit models (see Table 2) suggests a possibility to extend the Shapley value net effects technique to the estimation of the contribution of the regressors into the linear aggregate (1) of the logistic link, and to adjust the coefficients of the logistic model using the obtained net effects.

Table 2. Ratios of the models' coefficients.

Variable	Logit to Linear	Logit to Entropy-Logit	Entropy-Logit to Linear
x_0	11.80	6.78	1.74
x_1	-168.62	116.64	-1.45
x_2	6.19	5.38	1.15
x_3	6.95	6.18	1.12
x_4	8.03	8.25	0.97
x_5	4.62	3.94	1.17
x_6	5.51	4.71	1.17
x_7	10.14	9.07	1.12
x_8	-2.45	-2.04	1.20

Table 3. Predictive ability of binary models.

A	$\tilde{y}_{lin}=0$	$\tilde{y}_{lin}=1$	B	$\tilde{y}_{\log}=0$	$\tilde{y}_{\log}=1$	C	$\tilde{y}_{ent}=0$	$\tilde{y}_{ent}=1$	D	$\tilde{y}_{lin}=0$	$\tilde{y}_{lin}=1$	E	$\tilde{y}_{ent}=0$	$\tilde{y}_{ent}=1$	F	$\tilde{y}_{ent}=0$	$\tilde{y}_{ent}=1$
$y=0$	169	53	$y=0$	173	49	$y=0$	167	55	$\tilde{y}_{\log}=0$	203	13	$\tilde{y}_{\log}=0$	201	15	$\tilde{y}_{lin}=0$	204	3
$y=1$	38	143	$y=1$	43	138	$y=1$	38	143	$\tilde{y}_{\log}=1$	4	183	$\tilde{y}_{\log}=1$	4	183	$\tilde{y}_{lin}=1$	1	195

Shapley Value Regression

A model of linear multiple regression can be presented as:

$$y = z + \varepsilon = X\,b + \varepsilon\,, \qquad (16)$$

where z is a linear aggregate (2) by the parameters b of the linear model , and ε denotes a vector of errors. The Least Squares (LS) objective for minimizing is:

$$\begin{aligned} \|\varepsilon\|^2 &= \varepsilon'\varepsilon \\ &= (y - Xb)'(y - Xb) \qquad . \\ &= y'y - 2b'X'\,y + b'X'Xb \end{aligned}$$

(17)

Minimization of (17) by its parameters yields a normal system of equations with the solution:

$$b = (X'X)^{-1}X'y\,. \qquad (18)$$

Substituting (18) into (17) gives a value of LS objective in minimum, or residual sum of squares $\varepsilon'\varepsilon$. The known LS relation $y'y = z'z + \varepsilon'\varepsilon$ says that the original sum of squares of the dependent variable equals the theoretical sum of squares around the regression plus residual sum of squares. The coefficient of multiple determination for the regression is:

$$R^2 = 1 - \frac{\varepsilon'\varepsilon}{y'y} = \frac{z'z}{y'y} = \frac{y'X(X'X)^{-1}X'y}{y'y} \qquad (19)$$

The minimum of the deviations (17) corresponds to the maximum regression quality estimated by R^2 (19). In the standardized variables the coefficient of multiple determination can be represented in a convenient form:

$$R^2 = y'X(X'X)^{-1}X'y = b'b_{pair} \equiv \beta'r\,, \qquad (20)$$

where b is the vector of multiple regression coefficients, and b_{pair} is a vector compounded from the coefficients of pairwise regressions of y by each x. The presentation $R^2 = \beta'r$ in (20) is given using a vector β of beta-coefficients of multiple regression (the coefficients of the standardized regression with all variables centered and normalized by their standard deviations), and vector r of pair correlations of y with each x (those correlations are equal to the coefficients in pair regressions by each predictor separately). Items of the scalar product at the right-hand side of total R^2 (20) define the so called Net Effects (*NEF*) of each j-th regressor:

$$NEF_j = \beta_j r_j\,. \qquad (21)$$

The multiple determination and net effects are widely used in practice for estimation of the regressors' contribution to the model.

Another measure of predictor comparative usefulness is utility U_j of each regressor that is estimated via the increment of multiple each determination of the models with and without particular x_j in the set of predictors (Darlington, 1968; Harris, 1975):

$$U_j = R^2 - R^2_{-j} \qquad (22)$$

Here R^2 denotes multiple determination in the model with all predictors including x_j, and R^2_{-j} denotes multiple determination in the model without x_j.

Consider the Shapley Value (SV) estimation of predictors' shares. SV assigns a value for each predictor calculated over all possible combinations of predictors in the linear model, so it includes the competitive influence of any subsets of predictors in the analysis. The SV is defined as each j-th participant's input to a coalition:

$$SV_j = \sum_{allM} \gamma_n(M)\left[\upsilon\left(M \bigcup \{j\}\right) - \upsilon(M)\right] \qquad (23)$$

with weights of proportions to enter into a coalition M defined as

$$\gamma_n(M) = m!(n-m-1)!/n!\,. \qquad (24)$$

In (23)-(24) n is the total number of participants, m is the number of participants in the M-th coalition, and $\upsilon(\;)$ is the characteristic function used for estimation of utility for each coalition. By $M\bigcup\{j\}$ a set of participants which includes the j-th participant is denoted, when M means a coalition without the j-th participant. In regression, the participants of the coalition game are predictors incorporated into the model.

As indicated above, the coefficient of multiple determination (20), net effects (21), and utility values (22) can be used as measures of quality in regression models. For ease of exposition, it is convenient to use notations A, B, C, etc., for variables x_1, x_2, x_3, etc., so R^2_{ABC}, for example, defines the multiple determination in the model with the corresponding predictors. The characteristic function υ (23) via these R^2 values are estimated by the results of linear modeling. For instance, if $n = 5$, the characteristic function for variable A is:

$$\begin{aligned} &\upsilon(0) = 0\,, \\ &\upsilon(A) = R_A^2\,, \\ &\upsilon(AB) = R_{AB}^2\ \ , ..., \qquad . \\ &\upsilon(ABCDE) = R_{ABCDE}^2 \end{aligned}$$

(25)

Substitution of characteristic function (25) into the SV (23) shows that each expression in brackets (23) coincides with the utility (22). So SV_A is a measure of the predictor A usefulness averaged by all the models that contain this predictor. The weights (24) are:

$$\gamma(0) = \gamma(4) = 0.20, \quad \gamma(1) = \gamma(3) = 0.05, \quad \gamma(2) = 0.033$$

(26)

Then the SV_A (23) for the variable A can be written explicitly as:

$$\begin{aligned} SV_A &= .2(U_A) + .05(U_{AB} + U_{AC} + U_{AD} + U_{AE}) \\ &+ .033(U_{ABC} + U_{ABD} + U_{ABE} + U_{ACD} + U_{ACE} + U_{ADE}) \\ &+ .05(U_{ABCD} + U_{ABCE} + U_{ACDE} + U_{ABDE}) + .2(U_{ABCDE}) \end{aligned}$$

(27)

with the values of utility (22):

$$\begin{aligned} &U_A = R^2{}_A\,, \\ &U_{AB} = R^2{}_{AB} - R^2{}_B\,, \dots\,, \\ &U_{ABC} = R^2{}_{ABC} - R^2{}_{BC}\,, \dots\,, \\ &U_{ABCD} = R^2{}_{ABCD} - R^2{}_{BCD}\,, \dots\,, \\ &U_{ABCDE} = R^2{}_{ABCDE} - R^2{}_{BCDE}\ . \end{aligned}$$

(28)

The items in sum (27) correspond to the utility margins from the variable A to all coalitions, and the SV_A is the mean margin over all coalitions. Similar formulas are used for each of the other variables B, C, D, and E, and their SV define margins from each of the predictors. The total of margins from all the variables equals the value of R^2 in the model with all the predictors together:

$$\sum_j^n SV_j = \upsilon(all) = R_{ABCDE}^2 \quad .$$

(29)

The SV are shares of total R^2 defining importance of each predictor in their aggregate.

Regrouping items in (27) with help of (28) represents the SV as following:

$$\begin{aligned} SV_A &= \left(R_A^2 - \bar{R}_1^2\right)/(n-1) \\ &+ \left(\bar{R}_{A*}^2 - \bar{R}_2^2\right)/(n-2) \\ &+ \left(\bar{R}_{A**}^2 - \bar{R}_3^2\right)/(n-3) + \dots \\ &+ \left(\bar{R}_{A*...*}^2 - \bar{R}_{n-1}^2\right)/(n-(n-1)) + R_{AB...Z}^2\,/\,n\,. \end{aligned}$$

(30)

The first item in sum (30) presents a difference of R_A^2 for the model with one predictor A and mean value $\bar{R}_1^2$ (marked by bar over R^2) for all the models with just one predictor (marked by sub-index 1). In the second item of this sum a difference between mean $\bar{R}_{A*}^2$ for all the models with two predictors one of which is A (marked by sub-index A* with asterisk denoting any other variable x) and mean $\bar{R}_2^2$ for all the models with any two predictors (marked by sub-index 2) is shown, etc.

The last item presents a share that the predictor A has in the total R^2 of the model with all predictors together. The important feature of the formula (30) is the presentation of sequential inputs of coalitions of the 1st, 2nd, etc. levels to the total SV. If the data is available only on the several initial stages of coalitions with one, two, and some other subsets of variables, it is possible to use (30) for approximation of the partial inputs to the total SV. Comparison of such cumulative values for each variable allows one to evaluate the stability of the SV imputation. This suggests an approach for reducing the computation time of the SV by limiting evaluation to the number of levels where stability is achieved. Each term in (30) is constructed via mean values of combinations with a predictor and without it, so these means can be estimated by sampling combinations.

The expression (29) presents the estimations of the net effects (20)-(21) obtained via the SV approach. So in place of the regular net effects one can use decomposition of the multiple determination by the SV net effects:

$$R^2 = \Sigma_j SV_j \quad . \tag{31}$$

Each item in (31) is a very robust estimate of the net effect because SV is an average across all possible models with different subsets of predictors. These values are not as volatile as the regular net effects, and they are not prone to multicollinearity. In difference to regular net effects (21), the SV net effects (31) are always positive, so they are interpretable and suggest an easy way for graphical (pie-charts) presentation of predictors' shares in their contribution to the linear aggregate of the model.

When the SV net effects are found, they can be used for adjusting the coefficients in the linear aggregate, that can be performed by the following procedure. The objective of multiple determination can be presented using (17) and (19) as:

$$\begin{aligned} R^2 &= 1-\varepsilon'\varepsilon \\ &= 1-(y-X\beta)'(y-X\beta) \\ &= 2\beta'X'y-\beta'X'X\beta \\ &= \beta'(2r-S\beta) \end{aligned} \quad , \tag{32}$$

where the standardized beta-coefficients are used, and *S* denotes a matrix of predictors' correlations. Equalizing items in sums (31) and (32) yields a system of quadratic equations that can be used for finding the coefficients of regression adjusted by the SV net effects:

$$\beta_j(2r-S\beta)_j = SV_j \ , \quad j=1,...,n. \tag{33}$$

Solution of the system (33) can be achieved by minimizing the objective:

$$F = \sum_{j=1}^{n}\left(\beta_j(2r-S\beta)_j - SV_j\right)^2 . \tag{34}$$

Initial value for the parameters in minimization (34) can be taken as $\beta_j = SV_j / r_j$ obtained from (21) where the SV net effects are used. Having the adjusted beta-coefficients of the standardized regression, one returns to the coefficients of the original regression (16) by the regular transformation $b_j = \beta_j\sigma_y / \sigma_j$, where σ_y and σ_j are the standard deviations of the dependent and the independent variables.

Using the obtained coefficients *b* of the adjusted SV regression (34) and the property of approximate proportion between the coefficients of the entropy-logit and linear models (see Table 2), it is possible to use a proportionality:

$$a_j = k\,b_j \ , \quad j=1,...,n\,, \tag{35}$$

with a constant *k* between the coefficients a_j of the logistic model and the SV regression coefficients b_j for all the predictors. Then, the logistic aggregate (2) can be presented as a linear transformation

$$z_i = q + k\,\tilde{y}_{lin}^{SV} \tag{36}$$

of the vector $\tilde{y}_{lin}^{SV}$ of theoretical estimation of the dependent variable by the adjusted SV model (34), with *q* and *k* as unknown parameters. The parameters of the transformation (36) can be found by a simple logistic model with only one variable $\tilde{y}_{lin}^{SV}$:

$$p = \frac{1}{1+\exp\left(-(q+k\,\tilde{y}_{lin}^{SV})\right)}, \tag{37}$$

using the original data on the binary output.

Table 4 in its left-hand side presents some additional estimates for the linear regression – there are columns of the net effects (21), their shares in the total coefficient of multiple determination (20), the SV net effects (31), and their shares in the same R^2. The last predictor in the linear regression has negative sign in the model (see Table 1), and its net effect is negative in Table 4. Estimated by SV, the net effects are all positive, so all the predictors contribute to the model, as it should be expected because any additional variable increases the quality of data fitting. Shares of the SV net effects are rather substantial even for the variables x_1, x_4, $x_{6,}$ and x_8 (considered as unimportant by the previous model – see the discussion by Table 1).

The right-hand section of Table 4 presents the results of the adjusted SV regressions. Procedure (34) yields the adjusted SV regression with all positive predictor coefficients, positive net effects, and R^2 =0.313 that is slightly less than R^2 =0.324 of the regular regression – this is a price of the trade-off for the adjusted model with interpretable coefficients and positive net effects. Although the coefficients of the regular and adjusted linear regressions are rather different, the SV net effect shares by the regular linear and the adjusted linear models are very similar. They can be used as the estimates of the variables role in increasing the clients' satisfaction with the bank's mortgage products.

The last column in Table 4 presents the logistic model constructed by the procedure (35)-(37). At first a vector $\tilde{y}_{lin}^{SV} = .015x_1 + .024x_2 + ... + .019x_8$ of the aggregate with the coefficients of the adjusted

Table 4. Net Effects, Shapley Value, Adjusted SV Linear and Logistic Models.

	Linear regression					**Adjusted SV regressions**		
Variable	**Net Effect**	**Share %**	**SV net effect**	**Share SV %**		**Linear model**	**Net Share %**	**Logistic model**
x_0						-0.943		-9.683
x_1	0.000	0.1	0.025	7.7		0.015	7.5	0.099
x_2	0.070	21.6	0.049	15.1		0.024	15.3	0.160
x_3	0.117	36.2	0.077	23.8		0.030	24.2	0.197
x_4	0.017	5.3	0.045	14.0		0.020	14.1	0.129
x_5	0.060	18.6	0.050	15.5		0.020	15.7	0.134
x_6	0.028	8.7	0.026	8.1		0.022	7.9	0.145
x_7	0.041	12.8	0.030	9.3		0.027	9.1	0.181
x_8	-0.010	-3.2	0.021	6.6		0.019	6.2	0.126
R^2	0.324	100.0	0.324	100.0		0.313	100	0.313

Table 5. Predictive ability of the SV logistic model.

	A		B		C		D	
	$y = 0$	$y = 1$	$\tilde{y}_{lin} = 0$	$\tilde{y}_{lin} = 1$	$\tilde{y}_{\log} = 0$	$\tilde{y}_{\log} = 1$	$\tilde{y}_{ent} = 0$	$\tilde{y}_{ent} = 1$
$\tilde{y}_{\log}^{SV} = 0$	169	46	201	14	204	11	200	15
$\tilde{y}_{\log}^{SV} = 1$	53	135	6	182	12	176	5	183

SV linear model is constructed. Then the parameters of the logistic model (37) are estimated as q = -9.683 and k = 6.617, and by (35) the coefficients of the adjusted SV logistic model are obtained (the last column in Table 4). In this model all the coefficients are positive, and the shares of the predictor contributions coincide with the net effect shares (Table 4, the column before the last one) because the proportionality of the coefficients (35) does not change the shares of the net effect (20)-(21).

The predictive ability of the SV logistic model in comparison with several others is presented in Table 5. There are cross-sections of the binary output $\tilde{y}_{\log}^{SV}$ of the SV logistic model with the empirical outcome y, and with the predictions $\tilde{y}_{lin}$, $\tilde{y}_{\log}$, and $\tilde{y}_{ent}$ by the linear, regular logit, and entropy-logit models, respectively.

Section A of Table 5 shows that the SV logistic correctly predicts (169+135)/403 or 75.4% of the original binary data. By Table 3, the rate of the correct identifications by the models with the coefficients non-adjusted to multicollinearity was about 77%. The next cross-sections in Table 5 show that the SV logit predictions coincides with the other models' predictions at the total rate of 95%. Thus, the adjusted SV logit model has both high predictive rate and interpretable coefficients of the model.

So the management of the bank can elaborate an appropriate program for improving the clients service based on the results of the adjusted SV logistic model.

Conclusion

The entropy criterion applied to the binary response data with the logistic link yields a logistic model with the coefficients proportional to the linear regression, and with the predictive ability similar to both linear and regular logistic models. Using the properties of the entropy-logistic regression, the Shapley value net effects are applied for estimating the contributions of the predictors in the logistic model, and for adjusting the coefficient of regression itself. The Shapley value logistic regression is robust, has interpretable coefficients, and demonstrates a high rate of predictive ability. The partnership of the entropy-logistic approach and the Shapley value binary response regressions can enrich theoretical possibilities and serve as a useful tool for categorical data modeling in practical applications.

References

Arminger G., Clogg C. C., & Sobel M. E. (Eds). (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York, London.

Bar-Yam Y. (1997). *Dynamics of complex systems*. Reading, M.A.: Addison-Wesley.

Bar-Yam Y. (2004). Multiscale complexity/entropy. *Advances in Complex Systems*, *7*, 47-63.

Becker, S. & Le Cun, Y. (1988). Improving the convergence of back-propagation learning with second order methods. In Touretzky D.S., Hinton G.E. and Sejnowski T.J. (Eds.), *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann, San Mateo, CA, 29-37.

Bender, E. A. (2000). *Mathematical methods in artificial intelligence*. Los Alamitos, C.A.: IEEE Computer Society Press.

Conklin, M., Powaga, K., & Lipovetsky, S. (2004). Customer satisfaction analysis: Identification of key drivers, *European Journal of Operational Research*, *154*, 819-827.

Conklin, M. & Lipovetsky, S. (2005). Marketing decision analysis by TURF and shapley value. *Information Technology and Decision Making*, *4*, 5-19.

Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *79*, 161-182.

Dukhovny, A. (2002). General entropy of general measures. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, *10*, 213-225.

Handscombe, R. D. & Patterson, E. A. (2004). *The entropy vector: Connecting science and business.* Sheffield, U.K.: World Scientific Publishing.

Harris, R. (1975). *A primer of multivariate statistics*. New York, N.Y.: Academic Press.

Hastie, T. J. & Tibshirani, R. J. (1997). *Generalized additive models*. London: Chapman and Hall.

Jones, A. J. (2000). *Game theory: Mathematical models of conflict*. Chichester: Horwood Publishing.

Levene, M. & Loizou, G. (2003). Computing the entropy of user navigation in the web. *International Journal of Information Technology and Decision Making*, *2*, 459-476.

Lindley, D. (1956). On a measure of the information provided by an experiment, *Annals of Mathematics and Statistics*, *27*, 986-1005.

Lipovetsky, S. & Conklin, M. (2000). Box-Cox generalization of logistic and algebraic binary response models. *International Journal of Operations and Quantitative Management*, *6*, 276-285.

Lipovetsky, S. & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, *17*, 319-330.

Lipovetsky, S. & Conklin, M. (2004). Decision making by variable contribution in discriminant, logit, and regression analyses. *Information Technology and Decision Making*, *3*, 265-279.

Lipovetsky, S. & Conklin, M. (2005). Incremental net effects in multiple regression. *International Journal of Mathematical Education in Science and Technology*, *36*, 361-373.

Lloyd, C. J. (1999). *Statistical analysis of categorical data.* New York, N.Y.: Wiley.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* London: Sage Publication.

Maes, C. & Netocny, K. (2003). Time reversal and entropy. *Journal of Statistical Physics*, *110*, 269-310.

McCullagh, P. & Nelder, J. A. (1997). *Generalized linear models.* London: Chapman and Hall.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*, 705-724.

Roth, A. E., (Ed.). (1988) *The Shapley Value - Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, Cambridge.

Shapley, L. S. (1953). A value for n-person games. In Kuhn H. W., Tucker A. W. (Eds.) *Contribution to the Theory of Games*, II, Princeton University Press, Princeton, NJ, 307-317.

Straffin, P. D. (1993). *Game theory and strategy*. The Mathematical Association of America.

Zeimer, R. E. & Tranter, W. H. (1976). Systems, modulation, and noise. Boston: Houghton-Mifflin.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 107-117

1538 – 9472/06/$95.00

The Efficiency Of OLS In The Presence Of Auto-Correlated Disturbances In Regression Models

Samir Safi
Department of Mathematics and Statistics
James Madison University*

Alexander White
Department of Mathematics and Statistics
Texas State University

The ordinary least squares (OLS) estimates in the regression model are efficient when the disturbances have mean zero, constant variance, and are uncorrelated. In problems concerning time series, it is often the case that the disturbances are correlated. Using computer simulations, the robustness of various estimators are considered, including estimated generalized least squares. It was found that if the disturbance structure is autoregressive and the dependent variable is nonstochastic and linear or quadratic, the OLS performs nearly as well as its competitors. For other forms of the dependent variable, rules of thumb are presented to guide practitioners in the choice of estimators.

Key words: Autocorrelation, autoregressive, ordinary least squares, generalized least squares, efficiency

Introduction

Let the relationship between an observable random variable y and k explanatory variables $X_1, X_2, \ldots, X_k$ in a T-finite system be specified in the following linear regression model:

$$y = X\beta + u \qquad (1)$$

where y is a $(T\times 1)$ vector of observations on a response variable, X is a $(T\times k)$ design matrix, β is a $(k\times 1)$ vector of unknown regression parameters, and u is a $(T\times 1)$ random vector of disturbances. For convenience, it is assumed that X is full column rank $k < T$ and its first column is 1's. *The ordinary least squares* (OLS) estimator of β in the regression model (1) is

$$\hat{\beta} = \left(X'X\right)^{-1} X'y \qquad (2)$$

In problems concerning time series, it is often the case that the disturbances are, in fact, correlated. Practitioners are then faced with a decision, use OLS anyway, or try to fit a more complicated disturbance structure. The problem is difficult because the properties of the estimators depend highly on the structure of the independent variables in the model. For more complicated disturbance structures, many of the properties are not well understood. If the disturbance term has mean zero, i.e. E(u) = 0, but is in fact, autocorrelated, i.e. $\mathrm{Cov}(u) = \sigma_u^2 \Sigma$, where Σ is a $T\times T$ positive definite matrix and the variance σ_u^2 is either known or unknown positive and finite scalar, then the OLS parameter estimates will continue to be unbiased, i.e. $E\left(\hat{\beta}\right) = \beta$. But it has a different covariance matrix;

$$\mathrm{Cov}_{\Sigma}\left(\hat{\beta}\right) = \sigma_u^2 \left(X'X\right)^{-1} X'\Sigma X \left(X'X\right)^{-1}. \quad (3)$$

*This article was accepted while Samir Safi was at James Madison University. He is now an Assistant Professor of Statistics at the Islamic University of Gaza. His research interests are in time series analysis, concerning the comparison of estimators in regression models with auto-correlated disturbances and efficiency of OLS in the presence of autocorrelated disturbances. Alexander White is Associate Professor in Mathematics Education. His research interests are in mathematics education, statistics and mathematical finance.

The most serious implication of autocorrelated disturbances is not the resulting inefficiency of OLS, but the misleading inference when standard tests are used. The autocorrelated nature of disturbances is accounted for in the generalized least squares (GLS) estimator given by:

$$\tilde{\beta} = \left(X'\Sigma^{-1}X\right)^{-1} X'\Sigma^{-1} y \qquad (4)$$

which is unbiased, i.e. $E\left(\tilde{\beta}\right) = \beta$, with covariance matrix

$$\mathrm{Cov}\left(\tilde{\beta}\right) = \sigma_u^2 \left(X'\Sigma^{-1}X\right)^{-1}. \qquad (5)$$

The superiority of GLS over OLS is due to the fact that GLS has a smaller variance. According to the Generalized Gauss Markov Theorem, the GLS estimator provides the *Best Linear Unbiased Estimator* (BLUE) of β. But the GLS estimator requires prior knowledge of the matrix correlation structure, Σ. The OLS estimator $\hat{\beta}$ is simpler from a computational point of view and does not require a prior knowledge of Σ.

A common approach for modeling univariate time series is the autoregressive model. The general finite order *autoregressive process of order p* or briefly, AR(p), is

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \varepsilon_t, \ \varepsilon_t \sim \text{i.i.d. } N\left(0, \sigma_\varepsilon^2\right) \qquad (6)$$

There are numerous articles describing the efficiency of the OLS coefficient estimator $\hat{\beta}$, which ignore the correlation of the error, relative to the GLS estimator $\tilde{\beta}$, which takes this correlation into account. One strand is concerned with conditions on regressors and error correlation structure, which guarantee that OLS is asymptotically as efficient as GLS (e.g. Chipman, 1979; Krämer, 1980). The efficiency of the OLS estimators in a linear regression containing an autocorrelated error term depends on the structure of the matrix of observations on the independent variables (e.g. Anderson, 1948; 1971; Grenander & Rosenblatt, 1957).

For a linear regression model with first order autocorrelated disturbances, several alternative estimators for the regression coefficients have been discussed in the literature, and their efficiency properties have been investigated with respect to the OLS and GLS estimators (e.g. Kadiyala, 1968; Maeshiro, 1976; 1979; Ullah et al., 1983).

The relative efficiency of GLS to OLS in the important cases of autoregressive disturbances of order one, AR(1), with autoregressive coefficient ρ and second order, AR(2), with autoregressive coefficients $\left(\phi_1, \phi_2\right)$ for specific choices of the design vector have been investigated.

Building on work on the economics and time series literature, the price one must pay for using OLS under suboptimal conditions required investigation. Different designs are being explored, under which relative efficiency of the OLS estimator to that of GLS estimator approaches to one or zero, determining ranges of first-order autoregressive coefficient, ρ, in AR(1) disturbance and second order of autoregressive coefficients, $\left(\phi_1, \phi_2\right)$ in AR(2) for which OLS is efficient and quantifying the effect of the design on the efficiency of the OLS estimator. Furthermore, a simulation study has been conducted to examine the sensitivity of estimators to model misspecification. In particular, how do estimators perform when an AR(2) process is appropriate and the process is incorrectly assumed to be an AR(1) or AR(4)?

Performance Comparisons

In this section, numerical results are presented using the formulas in (3) and (5). Focus will be placed on two issues; first, the relative efficiency of GLS estimator as compared with the OLS estimator when the structure of the design vector, X, is nonstochastic. For example, linear, quadratic, and exponential design vectors with an intercept term included in the design vector. Secondly, the relative efficiency of the GLS estimator as compared with the OLS for a stochastic design vector. In the example considered here, a standard Normal stochastic design vector of length 1000 was generated. The three finite sample sizes used are 50, 100, and 200 for selected values of the autoregressive coefficients. Both AR(1) and AR(2) error processes are considered to discuss the behavior of OLS as compared to GLS.

Performance Comparisons for AR (1) Process

The relative efficiencies of OLS to GLS are discussed when the disturbance term follows an AR(1) process, $u_t = \rho u_{t-1} + \varepsilon_t, t = 1, 2, \ldots, T$, assuming that the autoregressive coefficient, ρ, is known priori. The three finite sample sizes used are 50, 100, and 200 for the elected values of $|\rho| \leq .9$, evaluated in steps of .2.

Table (1) shows the relative efficiencies of the variances of GLS to OLS for a regression coefficient on linear trend with an intercept term included in the design. For estimating an intercept term, the relative efficiency of the OLS estimator as compared to the GLS estimator decreases with increasing values of $|\rho|$. For small and moderate sample sizes, the efficiency of the OLS estimator appears to be nearly as efficient as the GLS estimator for $|\rho| \leq .7$. In addition, for large size sample data, the OLS estimator performs nearly as efficiently as the GLS estimator for the additional values of $\rho = \pm.9$. Further, the efficiency for estimating the slope mimics the efficiency of the intercept, except for large sample size; the efficiency of the OLS estimator appears to be nearly as efficient as the GLS estimator for $\rho \neq \pm.9$.

The efficiency of GLS estimator to the OLS estimator for the quadratic design agrees with the behavior for the linear design vector. In contrast, the gain in efficiency of the GLS estimator for different design vectors such as exponential and 1000 standard Normal, N(0,1)

Table 1: Relative Efficiency of GLS to OLS for Linear Design

ρ	Intercept			Slope		
	T = 50	T =100	T = 200	T = 50	T =100	T = 200
-0.9	0.7097	0.8276	0.9047	0.6739	0.8012	0.8881
-0.7	0.9162	0.9552	0.9768	0.9024	0.9471	0.9724
-0.5	0.9694	0.9840	0.9918	0.9640	0.9810	0.9903
-0.3	0.9908	0.9952	0.9976	0.9891	0.9943	0.9971
-0.1	0.9991	0.9995	0.9998	0.9989	0.9994	0.9997
0.1	0.9991	0.9995	0.9998	0.9989	0.9994	0.9997
0.3	0.9911	0.9953	0.9976	0.9894	0.9944	0.9971
0.5	0.9717	0.9846	0.9920	0.9662	0.9816	0.9904
0.7	0.9288	0.9585	0.9777	0.9147	0.9503	0.9732
0.9	0.8359	0.8691	0.9164	0.8000	0.8418	0.8993

compared to the OLS estimator is substantial for moderate and large values of ρ. However, for small values of ρ the OLS appears to be nearly as efficient as GLS.

Performance Comparisons for AR (2) Process

The relative efficiencies of OLS to GLS are discussed for linear, quadratic, and exponential design vectors when the disturbance term follows an AR(2) process, $u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \varepsilon_t,\ t = 1, 2, .., T,$ assuming that the autoregressive coefficients ϕ_1 and ϕ_2 are known priori. The three finite sample sizes used are 50, 100, and 200 for the selected 45 pairs of the autoregressive coefficients. These coefficients were chosen according to stationary conditions $(\phi_1 + \phi_2 < 1,\ \phi_2 - \phi_1 < 1, \text{and} |\phi_2| < 1)$ and so that $\rho_1 = \phi_1(1 - \phi_2)^{-1}$ is positive. This second condition was chosen since this is the case in most econometric studies.

To demonstrate the efficiency of OLS, consider the linear design vector. When the disturbance term follows an AR(2) process for the linear design with small sample size, OLS performs nearly as efficiently as GLS for estimating the slope for all AR(2) parametrizations except when ϕ's are close to the stationary boundary. As the sample size increases, the difference between the performance of OLS and GLS decreases. Only when $\phi_2 = -.9$, does OLS perform badly regardless of the sample size. The efficiency of GLS to OLS for the quadratic design mimics the behavior for the linear design. Finally, for exponential and 1000 standard Normal design vectors, the efficiency of OLS appears to be nearly as efficient as GLS for $\phi_1 = .2$ and small values of $\phi_2\ {}^2$ for all sample sizes. Otherwise, OLS performs poorly.

Simulation Study

In this section, the robustness of various estimators are considered, including estimated generalized least squares (EGLS). These simulations examine the sensitivity of estimators to model misspecification. In particular, how do estimators perform when an AR(2) process is appropriate and it is incorrectly assumed that the process is an AR(1)? The finite sample efficiencies of the OLS estimator relative to four GLS estimators are compared: the GLS based on the correct disturbance model structures and known AR(2) coefficients denoted as GLS-AR(2); the GLS based on the correct disturbance model structures, but with estimated AR(2) coefficients denoted as EGLS-AR(2); the GLS based on AR(1) incorrect disturbance model structures with an estimated AR(1) coefficient denoted as EIGLS-AR(1); and the GLS based on AR(4) incorrect disturbance model structures with estimated AR(4) coefficients denoted as EIGLS-AR(4). This study focuses only on AR(p) GLS corrections disturbances which are widely used in econometric studies.

The Simulation Setup

Three finite sample sizes (50, 100, and 200) and three nonstochastic design vectors of the independent variable are used; linear, quadratic, and exponential. A standard Normal stochastic design vector of length 1000 is also generated (Assuming that the variance of the error term in AR(2) process $\sigma_\varepsilon^2 = 1$). Further, 1000 observations for each of the AR(2) error disturbances with four pairs of autoregressive coefficients; (.2,-.9), (.8,-.9), (.2,-.7), and (.2,-.1) were also generated. Table (2) shows the values of autocorrelation coefficients ρ_1, ρ_2, disturbance variances, σ_u^2, $\sigma_u^2 = [(1-\phi_2^2)(1-\rho_1^2)]^{-1}$ and the relative efficiencies for estimating an intercept β_0, and the slope, β_1 of GLS to OLS for linear design with T=50, denoted $\mathrm{RE}(\beta_0)$, and $\mathrm{RE}(\beta_1)$. Looking at the table, it may be seen that the choices (.2, -.9) and (.8, -.9) give the worst performance of OLS as compared to GLS for estimating (β_0, β_1) of the regression coefficients and the largest values of σ_u^2. However, the choices (.2, -.7) and (.2, -.1) give the moderate and best performance of OLS as compared to GLS and the smallest values of σ_u^2. Results for other sample sizes and designs demonstrate a similar pattern as in Table (2).

The regression coefficients β_0, and β_1 for an intercept and the slope were each chosen to be equal one. Breusch (1980) has shown that for a fixed design, the distribution of $\frac{\hat{\beta}_{EGLS} - \beta}{\sigma_u^2}$ does not depend on the choice for β and σ_u^2, and the result holds even if the covariance matrix Σ is misspecified. When the design vector is stochastic, the assumption of a fixed design can be constructed as conditioning upon a given realization of the design, provided that the design is independent of u_t, Koreisha et al. (2002).

Definition

The efficiency of the GLS estimates relative to that of OLS in terms of the mean squared error of the regression coefficient, $\hat{\zeta}_{\beta_j}$, is given by:

$$\hat{\zeta}_{\beta_j} = \frac{\sum_{i=1}^{k}\left(\tilde{\beta}_{ij,GLS} - \beta_j\right)^2}{\sum_{i=1}^{k}\left(\hat{\beta}_{ij,OLS} - \beta_j\right)^2} \qquad (7)$$

where j = 0,1, for four GLS estimates, and k is the number of simulations. A ratio less than one indicates that the GLS estimates is more efficient than OLS, and if $\hat{\zeta}_{\beta_j}$ is close to one, then the OLS estimate is nearly as efficient as GLS estimates.

The Simulation Results for $\hat{\zeta}_{\beta_j}$

Tables (3) through (6) show the complete simulation results of the ratios of the GLS estimators relative to the OLS estimator in terms of the mean squared error of the regression coefficients, $\hat{\zeta}_{\beta_0}$ and $\hat{\zeta}_{\beta_1}$ in (7), when the serially correlated disturbance follows an AR(2) process. Each table presents the results for the three sample sizes considered, as well as all four selected pairs of AR(2) parametrizations. Each of the different designs is presented in a separate table.

Note that regardless of the sample size, selected design vectors, and AR(2) parametrizations the efficiency in estimating an intercept, β_0, and the slope, β_1, of the regression coefficients is higher for the GLS-AR(2) estimator than OLS. This result emphasizes that GLS is the BLUE. However, OLS performs nearly as efficiently as GLS for all selected sample sizes and designs when Φ = (.2, -.1). This result is not surprising since the choice of Φ = (.2, -.1) gives the highest performance of OLS as compared to GLS, in addition, it gives the smallest values of ρ_1, ρ_2, and σ_u^2.

Table 2: Autocorrelation Coefficients, Disturbance Variances and the Relative Efficiencies of GLS to OLS for Standardized Linear Design with T = 50

(ϕ_1, ϕ_2)	ρ_1	ρ_2	σ_u^2	$RE(\beta_0)$	$RE(\beta_1)$
(.2, -.9)	.1053	-.8789	5.3221	.7656	.5645
(.8, -.9)	.4211	-.5632	6.3973	.8325	.6026
(.2, -.7)	.1176	-.6765	1.9883	.9414	.8531
(.2, -.1)	.1818	-.0636	1.0446	.9993	.9980

When the order of the disturbance term is under estimated, i.e. EIGLS-AR(1), the GLS estimate performs poorly. In fact, OLS is more efficient for nearly every situation considered here. For example, when Φ = (.8, -.9) for quadratic design with T = 50, $\left(\hat{\zeta}_{\beta_0}, \hat{\zeta}_{\beta_1}\right)$ = $\left(1.4179,\ 1.7296\right)$ as shown in Table (3).

This shows that EIGLS-AR(1) can be much less efficient than OLS. The poor performance of EIGLS-AR(1) relative to OLS is most marked when the sample size is relatively estimation is smaller than an appropriate estimated AR structure. This suggests the small (i.e. T = 50) and the order of the autoregressive process used in the GLS surprising result that OLS may often be better than assuming an AR(1) when the actual process is AR(2). However, for the choice of Φ = (.2, -.1) there is little difference between OLS and EIGLS-AR(1). For example, for linear design with T=200, $\left(\hat{\zeta}_{\beta_0}, \hat{\zeta}_{\beta_1}\right)$ = (.9998, .9984) as presented in Table (4).

Table 3: Efficiency for MSEs of the Regression Coefficients of the GLS Estimators Relative to OLS Estimator for Quadratic Design

		(Φ_1, Φ_2)							
		(.2, -.9)		(.8, -.9)		(.2, -.7)		(.2, -.1)	
Size	Estimator	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$
50	GLS-AR(2)	0.7929	0.5540	0.8321	0.6174	0.9435	0.8349	1.0002	0.9954
	EGLS-AR(2)	0.7934	0.5567	0.8342	0.6172	0.9453	0.8409	1.0079	1.0094
	EIGLS-AR(1)	1.0935	1.1973	1.4179	1.7296	1.0355	1.0861	1.0063	1.0050
	EIGLS-AR(4)	0.7968	0.5623	0.8399	0.6182	0.9564	0.8500	1.0385	1.0332
100	GLS-AR(2)	0.8660	0.6950	0.8849	0.7104	0.9638	0.9287	1.0003	0.9993
	EGLS-AR(2)	0.8661	0.6957	0.8844	0.7089	0.9676	0.9319	0.9993	0.9980
	EIGLS-AR(1)	1.0453	1.0963	1.2136	1.4127	1.0207	1.0348	0.9989	1.0001
	EIGLS-AR(4)	0.8651	0.6974	0.8861	0.7093	0.9723	0.9342	1.0078	1.0091
200	GLS-AR(2)	0.9410	0.8331	0.9700	0.8269	0.9628	0.9400	1.0004	0.9984
	EGLS-AR(2)	0.9409	0.8326	0.9702	0.8265	0.9637	0.9400	1.0016	0.9990
	EIGLS-AR(1)	1.0180	1.0417	1.0453	1.2683	1.0094	1.0187	1.0014	1.0023
	EIGLS-AR(4)	0.9418	0.8338	0.9707	0.8290	0.9627	0.9407	1.0018	1.0021

Table 4: Efficiency for MSEs of the Regression Coefficients of the GLS Estimators Relative to OLS Estimator for Linear Design

		(Φ_1, Φ_2)							
Size	Estimator	(.2, -.9)		(.8, -.9)		(.2, -.7)		(.2, -.1)	
		$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$
50	GLS-AR(2)	0.7472	0.5740	0.8214	0.6193	0.9740	0.8511	1.0012	0.9964
	EGLS-AR(2)	0.7485	0.5771	0.8219	0.6200	0.9773	0.8548	1.0091	1.0073
	EIGLS-AR(1)	1.1004	1.1893	1.3624	1.6995	1.0181	1.0895	1.0055	1.0079
	EIGLS-AR(4)	0.7490	0.5824	0.8255	0.6220	0.9868	0.8595	1.0122	1.0448
100	GLS-AR(2)	0.8756	0.6641	0.8992	0.7340	0.9724	0.9204	1.0005	0.9996
	EGLS-AR(2)	0.8766	0.6632	0.8992	0.7323	0.9718	0.9219	1.0025	1.0003
	EIGLS-AR(1)	1.0349	1.0995	1.1826	1.4783	1.0156	1.0266	1.0025	1.0023
	EIGLS-AR(4)	0.8782	0.6654	0.8992	0.7391	0.9758	0.9285	1.0133	1.0021
200	GLS-AR(2)	0.9127	0.8137	0.9584	0.8662	0.9623	0.9262	0.9990	0.9977
	EGLS-AR(2)	0.9127	0.8135	0.9586	0.8662	0.9621	0.9271	1.0000	0.9980
	EIGLS-AR(1)	1.0252	1.0464	1.0666	1.2104	1.0092	1.0175	0.9998	0.9984
	EIGLS-AR(4)	0.9117	0.8123	0.9584	0.8668	0.9618	0.9255	1.0022	1.0032

This result is expected because the choice of $\phi_2 = -.1$ indicates that the serially correlated disturbance very nearly AR(1) since ϕ_2 is close to zero.

To further demonstrate the efficiency of OLS, consider the quadratic and linear designs. OLS is nearly as efficient or more efficient in estimating (β_0, β_1) than the GLS estimators; EGLS-AR(2), and EIGLS-AR(4), for moderate and large sample sizes (i.e. T=100 and 200) with AR(2) parametrizations Φ = (.2, -.7) and (.2, -.1) Tables (3) and (4). However, there are examples where OLS performs poorly as well. For the exponential design, OLS is nearly as efficient as EGLS-AR(2), and EIGLS-AR(4) for all sample sizes only when Φ = (.2, -.1). Otherwise, OLS performs poorly as shown in Table (5). For example, when T = 50 with Φ = (.2, -.9), $\hat{\zeta}_{\beta_1}$ = (.2035, .2108). However, even in this case, the performance of the OLS estimator for estimating the intercept is not bad, $\hat{\zeta}_{\beta_0}$ = (.7561, .7606). In fact, the performance of OLS is always better for estimating the intercept than the slope.

For the standard Normal stochastic design model, OLS fares more poorly. Only for Φ = (.2, -.1) does the efficiency of OLS match GLS as shown in Table (6). However, regardless of the sample size, OLS performs as nearly as efficiently or better than EIGLS-AR(1) for all selected autoregressive coefficients for estimating β_0.

Table 5: Efficiency for MSEs of the Regression Coefficients of the GLS Estimators Relative to OLS Estimator for Exponential Design

		(Φ_1, Φ_2)							
		(.2, -.9)		(.8, -.9)		(.2, -.7)		(.2, -.1)	
Size	Estimator	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$	$\hat{\zeta}_{\beta_0}$	$\hat{\zeta}_{\beta_1}$
50	GLS-AR(2)	0.7529	0.1951	0.8208	0.2160	0.9394	0.5576	0.9986	0.9706
	EGLS-AR(2)	0.7561	0.2035	0.8256	0.2241	0.9464	0.5642	0.9987	1.0030
	EIGLS-AR(1)	1.0451	1.1030	1.1649	1.1683	1.0167	1.0656	0.9969	1.0019
	EIGLS-AR(4)	0.7606	0.2108	0.8293	0.2322	0.9473	0.5815	1.0042	1.0775
100	GLS-AR(2)	0.8922	0.1979	0.9139	0.2311	0.9668	0.5461	1.0009	0.9830
	EGLS-AR(2)	0.8895	0.2021	0.9163	0.2353	0.9682	0.5467	0.9993	0.9980
	EIGLS-AR(1)	1.0115	1.0803	1.0893	1.1383	1.0077	1.0575	0.9991	0.9965
	EIGLS-AR(4)	0.8904	0.2068	0.9149	0.2357	0.9692	0.5578	0.9997	1.0187
200	GLS-AR(2)	0.9168	0.2139	0.9771	0.2084	1.0162	0.5293	1.0022	0.9877
	EGLS-AR(2)	0.9164	0.2143	0.9782	0.2132	1.0150	0.5303	1.0008	0.9990
	EIGLS-AR(1)	1.0053	1.0645	1.0492	1.2352	0.9999	1.0390	1.0012	0.9900
	EIGLS-AR(4)	0.9171	0.2161	0.9802	0.2151	1.0149	0.5425	1.0006	1.0062

Table 6: Efficiency for MSEs of the Regression Coefficients of the GLS Estimators Relative to OLS Estimator Standard Normal Stochastic Design

		(Φ_1, Φ_2)							
		(.2, -.9)		(.8, -.9)		(.2, -.7)		(.2, -.1)	
Size	Estimator	$\hat{\xi}_{\beta_0}$	$\hat{\xi}_{\beta_1}$	$\hat{\xi}_{\beta_0}$	$\hat{\xi}_{\beta_1}$	$\hat{\xi}_{\beta_0}$	$\hat{\xi}_{\beta_1}$	$\hat{\xi}_{\beta_0}$	$\hat{\xi}_{\beta_1}$
50	GLS-AR(2)	0.6416	0.1127	0.7427	0.0667	0.8851	0.3334	0.9998	0.8838
	EGLS-AR(2)	0.6437	0.1149	0.7442	0.0676	0.8807	0.3428	1.0176	0.9513
	EIGLS-AR(1)	1.0536	0.9591	1.2168	0.5186	1.0129	0.9598	1.0091	0.9244
	EIGLS-AR(4)	0.6509	0.1211	0.7477	0.0737	0.8908	0.3652	1.0385	1.0221
100	GLS-AR(2)	0.7601	0.1055	0.8466	0.0640	0.9350	0.3230	0.9978	0.8902
	EGLS-AR(2)	0.7598	0.1060	0.8472	0.0639	0.9341	0.3274	0.9984	0.9076
	EIGLS-AR(1)	1.0241	0.9568	1.1141	0.5261	1.0121	0.9546	0.9994	0.9158
	EIGLS-AR(4)	0.7611	0.1109	0.8477	0.0668	0.9346	0.3386	1.0033	0.9359
200	GLS-AR(2)	0.8624	0.1002	0.9323	0.0720	0.9715	0.3226	1.0038	0.9194
	EGLS-AR(2)	0.8628	0.1006	0.9319	0.0725	0.9707	0.3245	1.0028	0.9331
	EIGLS-AR(1)	1.0141	0.9581	1.0400	0.5181	1.0021	0.9598	1.0033	0.9418
	EIGLS-AR(4)	0.8630	0.1033	0.9314	0.0748	0.9719	0.3307	1.0058	0.9512

Discussion

In investigating the simulation results in the previous section, the following significant results were observed. First and foremost, it was noticed that regardless of the sample size for all design structures and selected autoregressive coefficients, the efficiency in estimating an intercept, β_0, and the slope, β_1, of the regression model is higher for the GLS estimator based on the correct disturbance model structures and known AR(2) coefficients. This result is expected since GLS is BLUE, but because GLS requires a priori knowledge of Σ, this is not a viable option.

In addition, the relative efficiency of OLS is better than EIGLS-AR(1) in estimating (β_0, β_1) for all sample sizes and nonstochastic design vectors. The relative efficiency of OLS to be superior to that of EIGLS in estimating the slope when T=50 with AR(2) parametrization (.8, -.9) was also observed. This choice of (.8, -.9) gives the highest first-order autoregressive coefficient $\left(\rho_1 = .4211\right)$ and largest variance of the error process $\left(\sigma_u^2 = 6.3973\right)$ among the other choices of AR(2) parametrizations. This explains the poor relative performance of OLS to GLS for this choice of parameter.

However, from Table (3) through Table (6), it may be seen that the performance of EIGLS-AR(1) is even worse. This appears to occur because AR(2) parametrization (.8, -.9) produces large values of ρ_1, ρ_2 in absolute value ($\rho_2 = -.5632$) and disturbance variance comparing to the other parameter choices. This means using OLS is better than assuming another incorrect error process.

The third general conclusion from the simulation study is that regardless of the sample size, all of the estimators perform equally well with AR(2) parametrization (.2, -.1). This result is not surprising because the choice of (.2, -.1) gives the smallest variance of the process $\left(\sigma_u^2 = 1.0446\right)$, which is sufficiently close to the variance of standard OLS.

Fourth, for all stochastic and non-stochastic design vectors, the differences in the relative efficiency of OLS and all GLS estimators in estimating β_0 with a few expected exceptions are negligible. In fact, this is so even when the variance of the process is large, in other words, when AR(2) parametrizations are (.2, -.9) and (.8, -.9).

Similar to results for section 2, when the design vector is linear or quadratic, the relative efficiency of OLS is nearly as good as the EGLS-AR(2) and EIGLS-AR(4) estimators for moderate and large sample sizes for estimating β_1 with small variance of the disturbances.

It is observed that the differences in the relative efficiencies of GLS-AR(2), EGLS-AR(2), and EIGLS-AR(4) in estimating (β_0, β_1) are insignificant. Hence, when confronted with an error with unknown order p, it appears that using AR(4) is the best bet.

Finally, OLS may often be more preferable than assuming an AR(1) process when the actual process is AR(2). In other words, it is sometimes better to ignore the autocorrelation of the disturbance term and use the OLS estimation rather than to incorrectly assume the process is an AR(1).

Future Research

Perhaps, even more important than the efficiency of the different estimation methods in these models, is the effect on forecasting performance. Koreisha et al. (2004) investigated the impact that EIGLS correction may have on forecast performance. They developed a new procedure for generating forecasts for regression models with auto-correlated disturbances based on OLS and a finite AR process. They found that for predictive purposes there is not much gained in trying to identify the actual order and form of the auto-correlated disturbances or using more complicated estimation methods such as GLS or MLE procedures, which often require inversion of large matrices. It is necessary to extend Koreisha et al. (2004) results for different design vectors of the independent variables including both stochastic and nonstochastic designs instead of using one independent variable generated by an AR(1) process as in their investigation.

A second important consideration is the estimation of the standard errors of the estimators. In practice, if one were using a statistical package to compute the OLS estimators the variance estimate produced would be based on $\sigma_u^2 \left(X'X\right)^{-1}$, which may be biased for the true variance $\sigma_u^2 \left(X'X\right)^{-1} X'\Sigma X \left(X'X\right)^{-1}$. For GLS estimation ($\Sigma$ known), on the other hand, the variance estimate is unbiased for the true variance of the GLS estimator. It is unclear, however, how the variance estimators for EGLS estimation behave. The impact that the variance estimators may have on inference based on the OLS estimator is currently being investigated.

Finally, the long range goal is the creation of guidelines or rules of thumb which will aid the practitioner when deciding which regression estimation procedure to use.

Conclusion

This article has investigated an important statistical problem concerning estimation of the regression coefficients in the presence of autocorrelated disturbances. In particular, the comparison of efficiency of the ordinary least squares (OLS) estimation to alternative procedures such as generalized least squares (GLS) and estimated GLS (EGLS) estimators in the presence of autocorrelated disturbances was discussed. Both stochastic and non-stochastic design vectors were used with different sample sizes.

It was found that regardless of the sample size, design vector, and order of the auto-correlated disturbances, the relative efficiency of the OLS estimator generally increases with decreasing values of the disturbance variances. In particular, if the disturbance structure is a first or second order autoregressive and the dependent variable is nonstochastic and linear or quadratic, OLS performs nearly as well as its competitors for small values of the disturbance variances. The gain in efficiency of the GLS estimator for different design vectors such as exponential and standard Normal compared to the OLS estimator is substantial for moderate and large values of the autoregressive coefficient in the case of an AR(1) process and large values of the disturbance variance in the presence of an AR(2) process. However, for small values of the autoregressive coefficient and disturbance variance the OLS estimator appears to be nearly as efficient as the GLS estimator.

It was also found that if the error structure is autoregressive, and the dependent variable is nonstochastic and linear or quadratic, the OLS estimator performs nearly as well as its competitors. When faced with an unknown error structure, however, AR(4) may be the best choice.

References

Anderson, T. W. (1948). On the theory of testing serial correlation. *Skandinavisk Aktuarietid skrift, 31*, 88-116.

Anderson, T. W. (1971). The Statistical analysis of time series. New York:, N.Y: Wiley.

Breusch, T. (1980). Useful invariance results for generalized regression models. *Journal of Econometrics, 13*, 327-340.

Chipman, J. S. (1979). Efficiency of least squares estimation of linear trend when residuals are autocorrelated. *Econometrica*, *47*, 115-128.

Choudhury, A., Hubata, R. & Louis, R. (1999). Understanding time-series regression estimators. *The American Statistician*, *53*, 342-348.

Grenander, U. & Rosenblatt, M. (1957). *Statistical analysis of stationary time series*. New York, N.Y.: Wiley.

Judge, G. G., Griffiths, W. E., Hill, R. C.. Lutkepohl, H., & Lee, T. C. (1985). *The theory and practice of econometrics*. New York, N.Y.: Wiley & Sons Inc.

Kadiyala, K. R. (1968). A transformation used to circumvent the problem of autocorrelation. *Econometrica*, *36*, 93-96.

Koreisha, S. G. and Fang, Y. (2002). Generalized least squares with misspecified serial correlation structures. *Journal of the Royal Statistical Society*, *63*, Series B, 515-531.

Koreisha, S. G. and Fang, Y. (2004). Forecasting with serially correlated regression models. *Journal of Statistical Computations and Simulation*, *74*, 625-649.

Kramer, W. (1980). Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors. *Journal of the American Statistical Association*, *75*, 1005-1009.

Maeshiro, A. (1976). Autoregressive transformation, trended independent variables and autocorrelated disturbances terms. *Review of Economics and Statistics*, *58*, 497-500.

Maeshiro, A. (1979). On the retention of the first observations in serial correlation adjustment of regression models. *International Economic Review*, *20*, 259-65.

Ullah, A., Srivastava, V. K., Magee, L., & Srivastava, A. (1983). Estimation of linear regression model with autocorrelated disturbances. *Journal of Time Series Analysis*, *4*, 127-135.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 118-129

1538 – 9472/06/$95.00

Choosing Smoothing Parameters For Exponential Smoothing: Minimizing Sums Of Squared Versus Sums Of Absolute Errors

Terry Dielman
M.J. Neeley School of Business
Texas Christian University

When choosing smoothing parameters in exponential smoothing, the choice can be made by either minimizing the sum of squared one-step-ahead forecast errors or minimizing the sum of the absolute one-step-ahead forecast errors. In this article, the resulting forecast accuracy is used to compare these two options.

Key words: Exponential smoothing, forecasting accuracy, M-competition, outliers, parameter selection, Simulation

Introduction

In a number of comparisons of forecasting methods, exponential smoothing methods have been shown to be simple but relatively accurate techniques for generating forecasts (See Makridakis et al., 1982; Makridakis et al., 1993; Makridakis & Hibon, 2000). When using exponential smoothing methods to forecast a time series, a smoothing parameter (or parameters) must be chosen. One way this choice can be made is to choose the parameter or parameters that minimize some error criterion over the history of the data available. Typically, the choice made is to minimize the sum of squared one-step-ahead forecast errors (SSE). Another option would be to minimize the sum of the absolute one-step-ahead forecast errors (SAE). Minimizing SSE is the most often used criterion for choosing the smoothing parameter, but minimizing SAE could provide protection against outliers in the time series. This article examines the question of which of these choices might be best in practice.

In the context of regression models, forecasts generated from least squares (equivalent to SSE) coefficient estimates and least absolute value (equivalent to SAE) coefficient estimates were studied by Dielman (1986). When the disturbance distribution was long-tailed, presenting the opportunity for outliers, the least absolute value based forecasts were, on the whole, superior to the least squares based forecasts. These results were obtained from a simulation study assuming that an exogenous independent variable was available for use in the regressions. Whether the superiority of a least absolute value type criterion could exist for smoothing parameter choice and subsequent generation of forecasts in exponential smoothing methods is the issue considered in this article.

The analyses presented in this article support three main conclusions: First, while instances where outliers will degrade forecast performance may not be common, such instances do occur in practice. Second, minimizing SAE to determine exponential smoothing parameters can provide protection against such outliers. Finally, on average, minimizing SAE does not result in much, if any, deterioration in forecast accuracy over minimizing SSE when conditions are optimal for SSE.

Terry Dielman is Professor of Decision Sciences in the Information Systems and Supply Chain Management department, M.J. Neeley School of Business, Texas Christian University. Email: t.dielman@tcu.edu

Methodology

M1-Competition Data

Three exponential smoothing techniques are examined in this part of the study: single exponential smoothing, Brown's double exponential smoothing, and Holt's two-parameter exponential smoothing.

The one-period-ahead forecast for single exponential smoothing can be written as

$$\hat{y}_{T+1} = \alpha y_T + (1 - \alpha)\,\hat{y}_T \qquad (1)$$

All subsequent forecasts have the same value. The smoothing parameter, α, must be chosen to implement this forecasting technique. The choice is made by performing a grid search over the range 0.01, 0.02, …, 0.99 and choosing the value of α from this range that minimizes either the SSE or SAE.

Brown's double exponential smoothing is often suggested when data are trended. The m-period-ahead forecasts are generated from the following equations:

$$S_t^{'} = \alpha y_t + (1-\alpha)S_{t-1}^{'} \qquad (2)$$

$$S_t^{''} = \alpha S_t^{'} + (1-\alpha)S_{t-1}^{''} \qquad (3)$$

$$\hat{y}_{T+m} = a_T + m b_T \qquad (4)$$

where

$$a_t = 2S_t^{'} - S_t^{''} \qquad (5)$$

and

$$b_t = \frac{\alpha}{1-\alpha}(S_t^{'} - S_t^{''}) \qquad (6)$$

As with single exponential smoothing, the smoothing parameter, α, is chosen by performing a grid search over the range 0.01, 0.02, …, 0.99 and choosing the value of α from this range that minimizes either the SSE or SAE.

Holt's two-parameter exponential smoothing is also suggested when data are trended, but is somewhat more flexible than Brown's method because separate parameters are allowed for the two smoothing equations. The m-period-ahead forecasts are generated from the following equations:

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \qquad (7)$$

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \qquad (8)$$

$$\hat{y}_{T+m} = L_T + mT_T \qquad (9)$$

Values for two parameters, α and β, must be chosen in this case. Again, a grid search is used with values of 0.01, 0.02, …, 0.99 for each parameter. All possible parameter value combinations are examined and the pair of values that minimizes either the SSE or SAE is chosen.

The 1001 time series used in the M1 forecasting competition (See Makridakis et al., 1982) are used to evaluate the choice of criteria for choosing the smoothing parameter. The optimal values of the smoothing parameter(s) are chosen for each of the time series. The smoothing parameters for each method that minimize either the SSE or the SAE for each individual time series are chosen. One to six-period-ahead out-of-sample forecasts are then generated using the optimal values under the two criteria. The out-of-sample forecasts are compared to the actual values and accuracy measures are computed for the forecasts. The three accuracy measures reported in this article are the mean absolute percentage error (MAPE), the root mean square error (RMSE), and the mean absolute deviation (MAD). These accuracy measures will be presented to compare the forecasting accuracy for the parameter choices of each criterion.

A Brief Simulation

A small simulation was run to further compare forecast performance for the SAE and SSE criteria. Only single exponential smoothing was examined in this simulation. Single exponential smoothing provides optimal forecasts when the data generation process is ARIMA (0,1,1). This was the process used to generate the data for the simulation experiment. The procedures outlined in Dunne (1992) were used to generate data from an ARIMA (0, 1, 1) process. The following were factors considered in the experiment:

1. Sample sizes of T = 20, 30 and 50 were used.

2. The error distributions considered were:

a) Normal with mean zero and standard deviation one (Normal). The following distributions will be referred to as outlier-producing distributions:

b) Contaminated Normal with 0.75 probability of observations coming from a N(0,1) distribution and 0.25 probability from a N(0,5) distribution. The contamination was introduced in three different ways to assess potential situations where the minimum SAE criterion might outperform the minimum SSE criterion.
CNR5: The contamination was allowed to occur randomly throughout the time series.
CNB5: The first 25% of the observations were from the N (0, 5) distribution.
CNE5: The last 25% of the observations were from the N (0, 5) distribution.

c) Same as b but the contaminating distribution was N (0, 10) (CNR10, CNB10 and CNE10).

d) Cauchy with median zero and scale parameter one (Cauchy). These errors represent a pathological situation where extreme outliers are possible and should be the best-case scenario for minimizing SAE.

3. The true value of the exponential smoothing parameter was set at 0.2, 0.3, 0.5, 0.7 and 0.8.

For each experimental setting of the simulation, 10,000 time series were generated, the optimal value of the smoothing parameter was estimated using a grid search over the values 0.01, 0.02, …, 0.99, and one period ahead forecasts were computed using this parameter value. Out-of-sample forecasts were computed and were compared to the actual values (which were generated from the process used in the simulation) and the MAPE, RMSE, and MAD were computed for these 10,000 forecasts. All programs were written in FORTRAN and IMSL subroutines were used for random number generation.

Results

M1-Competition Results

Each of the three exponential smoothing methods was applied to each of the 1001 time series from the M1-competition. Optimal smoothing parameters to minimize both SSE and SAE were chosen and forecasts were generated. Table 1 shows the values of the accuracy measures for the one through six period ahead forecasts (combined). Table 2 shows the values for the one period ahead forecast. Cases where minimizing SAE results in greater accuracy are highlighted in bold. The choice of criterion is dependent to some extent on the accuracy measure. For example, in Table 1 the MAPE is smaller for the SAE criterion for single exponential smoothing, although the RMSE and MAD are both smaller for the SSE criterion. This experiment was conducted using seasonally adjusted data as well (where appropriate) with little difference in the results of the comparison. The forecast accuracy was improved regardless of criterion (because of the presence of seasonal series in the data set), but the difference in forecast accuracy between SAE and SSE did not change appreciably. The tables for the seasonally adjusted results have not been included in the article.

The results suggest that there are instances where the SAE forecasts provide improvement over the SSE forecasts according to some accuracy criterion. In other words, there are cases with outliers present that can affect forecast accuracy. The results from the simulation are intended to shed additional light on situations when the SAE forecasts might be most beneficial.

Table 1: Accuracy Measures One Through Six Period Ahead Forecasts

	MAPE		RMSE		MAD	
	SAE	SSE	SAE	SSE	SAE	SSE
Single	**17.5**	17.7	578348	572521	32884	32668
Brown	20.7	19.9	290890	272913	19475	18056
Holt	22.5	22.3	**290928**	389576	**19657**	25428

Table 2: Accuracy Measures One Period Ahead Forecasts

	MAPE		RMSE		MAD	
	SAE	SSE	SAE	SSE	SAE	SSE
Single	11.1	11.1	297704	291140	14001	13695
Brown	**13.3**	13.5	120026	119844	**9713**	10088
Holt	14.0	14.0	**123836**	170598	**10714**	13391

Simulation Results

Tables 3 through 17 summarize the simulation results. In all experimental settings when the disturbances were normal, there was little difference between accuracy measures for minimizing SAE versus SSE. In cases where there was a difference, the accuracy measures for minimizing SSE were smaller. In most of the outlier-producing distributions, the accuracy measures for minimizing SAE were smaller than those for minimizing SSE. The differences in the accuracy measures in favor of SAE are more pronounced in cases where the true smoothing constant is larger and where outliers are more likely. When the contaminated normal disturbances were used, the differences in the accuracy measures in favor of SAE occurred when the standard deviation was larger (10 rather than 5) and when the occurrence of the outliers was at the end or throughout the series rather than at the beginning.

Table 3: Accuracy Measures for Simulation using T = 50 and alpha = 0.2

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.03	1.02	0.82	0.81
CNR5	**1.64**	1.67	**2.69**	2.71	**1.63**	1.66
CNR10	**2.71**	2.84	**5.14**	5.21	**2.63**	2.77
CNB5	0.83	0.83	**1.03**	1.04	**0.82**	0.83
CNB10	**0.84**	0.85	**1.03**	1.05	**0.83**	0.84
CNE5	**4.16**	4.18	**5.19**	5.23	**4.14**	4.16
CNE10	**8.54**	8.58	**10.47**	10.55	**8.34**	8.38
Cauchy	**7.24**	7.90	**85.14**	86.00	**6.07**	6.72

Table 4: Accuracy Measures for Simulation using T = 30 and alpha = 0.2

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.03	1.02	0.82	0.81
CNR5	**1.65**	1.70	**2.69**	2.73	**1.64**	1.69
CNR10	**2.73**	2.91	**5.13**	5.25	**2.67**	2.85
CNB5	**0.82**	0.83	**1.03**	1.05	**0.82**	0.83
CNB10	**0.83**	0.86	**1.03**	1.07	**0.82**	0.85
CNE5	**4.15**	4.22	**5.21**	5.31	**4.13**	4.20
CNE10	**8.54**	8.67	**10.56**	10.76	**8.37**	8.50
Cauchy	**7.59**	9.26	**63.97**	98.24	**5.10**	7.21

Table 5: Accuracy Measures for Simulation using T = 20 and alpha = 0.2

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.83	0.83	1.04	1.04	0.83	0.83
CNR5	**1.68**	1.73	**2.73**	2.77	**1.68**	1.72
CNR10	**2.79**	2.96	**5.20**	5.34	**2.74**	2.91
CNB5	**0.84**	0.85	**1.05**	1.07	**0.83**	0.85
CNB10	**0.85**	0.89	**1.05**	1.11	**0.84**	0.88
CNE5	**4.12**	4.22	**5.15**	5.28	**4.11**	4.21
CNE10	**8.36**	8.75	**10.30**	10.80	**8.21**	8.59
Cauchy	**10.15**	11.86	1574.45	1570.05	23.62	23.60

Table 6: Accuracy Measures for Simulation using T = 50 and alpha = 0.3

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.82	1.03	1.02	0.82	0.82
CNR5	**1.64**	1.67	**2.69**	2.71	**1.62**	1.66
CNR10	**2.72**	2.85	**5.13**	5.21	**2.62**	2.76
CNB5	**0.83**	0.84	**1.03**	1.04	0.83	0.83
CNB10	**0.85**	0.86	**1.04**	1.06	**0.83**	0.85
CNE5	**4.15**	4.19	**5.18**	5.23	**4.13**	4.16
CNE10	**8.58**	8.63	**10.46**	10.54	**8.33**	8.39
Cauchy	**6.62**	7.17	**85.13**	86.17	**6.04**	6.68

Table 7: Accuracy Measures for Simulation using T = 30 and alpha = 0.3

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.03	1.02	0.82	0.81
CNR5	**1.64**	1.70	**2.68**	2.73	**1.64**	1.69
CNR10	**2.72**	2.91	**5.11**	5.23	**2.65**	2.84
CNB5	**0.83**	0.84	**1.03**	1.06	**0.82**	0.84
CNB10	**0.84**	0.87	**1.04**	1.09	**0.83**	0.86
CNE5	**4.15**	4.20	**5.20**	5.27	**4.13**	4.18
CNE10	**8.52**	8.60	**10.51**	10.62	**8.34**	8.42
Cauchy	**9.38**	10.39	**64.54**	103.00	**5.16**	7.36

Table 8: Accuracy Measures for Simulation using T = 20 and alpha = 0.3

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.83	0.83	1.05	1.04	0.83	0.83
CNR5	**1.68**	1.73	**2.72**	2.78	**1.67**	1.73
CNR10	**2.77**	2.98	**5.19**	5.33	**2.72**	2.92
CNB5	**0.83**	0.85	**1.04**	1.07	**0.83**	0.85
CNB10	**0.84**	0.89	**1.05**	1.10	**0.83**	0.88
CNE5	**4.13**	4.22	**5.16**	5.27	**4.11**	4.20
CNE10	**8.38**	8.69	**10.32**	10.71	**8.22**	8.52
Cauchy	**32.31**	32.89	1572.09	1569.45	**23.20**	23.50

Table 9: Accuracy Measures for Simulation using T = 50 and alpha = 0.5

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.82	1.03	1.02	0.82	0.81
CNR5	**1.65**	1.68	**2.69**	2.71	**1.62**	1.65
CNR10	**2.79**	2.91	**5.14**	5.19	**2.62**	2.73
CNB5	0.84	0.84	**1.04**	1.05	0.83	0.83
CNB10	**0.87**	0.89	**1.04**	1.07	**0.83**	0.85
CNE5	**4.17**	4.21	**5.18**	5.23	**4.12**	4.16
CNE10	**8.73**	8.79	**10.45**	10.55	**8.32**	8.39
Cauchy	**5.84**	6.25	**85.13**	87.23	**6.03**	6.68

Table 10: Accuracy Measures for Simulation using T = 30 and alpha = 0.5

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.03	1.02	0.82	0.81
CNR5	**1.64**	1.69	**2.67**	2.72	**1.63**	1.68
CNR10	**2.76**	2.92	**5.10**	5.21	**2.64**	2.80
CNB5	**0.83**	0.85	**1.04**	1.06	**0.82**	0.84
CNB10	**0.85**	0.89	**1.04**	1.10	**0.83**	0.86
CNE5	**4.14**	4.17	**5.17**	5.22	**4.11**	4.14
CNE10	**8.50**	8.58	**10.40**	10.50	**8.26**	8.34
Cauchy	**9.07**	9.84	**65.21**	124.25	**5.22**	7.77

Table 11: Accuracy Measures for Simulation using T = 20 and alpha = 0.5

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.84	0.83	1.05	1.04	0.84	0.83
CNR5	**1.67**	1.73	**2.71**	2.77	**1.66**	1.72
CNR10	**2.78**	2.97	**5.17**	5.31	**2.70**	2.89
CNB5	**0.85**	0.86	**1.06**	1.08	**0.84**	0.86
CNB10	**0.86**	0.90	**1.06**	1.12	**0.85**	0.88
CNE5	**4.15**	4.21	**5.19**	5.26	**4.13**	4.19
CNE10	**8.45**	8.69	**10.38**	10.69	**8.26**	8.50
Cauchy	**8.39**	10.76	**1568.82**	1569.82	**22.45**	23.53

Table 12: Accuracy Measures for Simulation using T = 50 and alpha = 0.7

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.82	1.03	1.02	0.82	0.81
CNR5	**1.66**	1.68	**2.69**	2.70	**1.62**	1.64
CNR10	**2.97**	3.07	**5.13**	5.18	**2.62**	2.71
CNB5	**0.84**	0.85	**1.03**	1.04	**0.82**	0.83
CNB10	**0.91**	0.93	**1.04**	1.07	**0.83**	0.85
CNE5	**4.19**	4.22	**5.16**	5.20	**4.11**	4.14
CNE10	**8.97**	9.05	**10.40**	10.52	**8.28**	8.37
Cauchy	**9.10**	9.42	**85.15**	89.06	**6.04**	6.74

Table 13: Accuracy Measures for Simulation using T = 30 and alpha = 0.7

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.02	1.02	0.82	0.81
CNR5	**1.65**	1.69	**2.67**	2.70	**1.63**	1.67
CNR10	**2.87**	3.00	**5.10**	5.19	**2.64**	2.77
CNB5	**0.83**	0.84	**1.03**	1.05	**0.82**	0.83
CNB10	**0.87**	0.90	**1.04**	1.08	**0.83**	0.85
CNE5	**4.11**	4.15	**5.13**	5.17	**4.07**	4.11
CNE10	**8.52**	8.63	**10.30**	10.43	**8.18**	8.30
Cauchy	**7.98**	8.52	**65.47**	161.36	**5.21**	8.51

Table 14: Accuracy Measures for Simulation using T = 20 and alpha = 0.7

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.83	0.83	1.05	1.04	0.83	0.83
CNR5	**1.67**	1.72	**2.71**	2.75	**1.66**	1.70
CNR10	**2.82**	2.97	**5.18**	5.28	**2.69**	2.84
CNB5	**0.85**	0.86	**1.06**	1.08	**0.84**	0.86
CNB10	**0.88**	0.91	**1.07**	1.11	**0.85**	0.88
CNE5	**4.16**	4.21	**5.20**	5.26	**4.14**	4.18
CNE10	**8.50**	8.78	**10.40**	10.75	**8.28**	8.55
Cauchy	**7.98**	8.92	**1567.46**	1571.71	**21.81**	23.71

Table 15: Accuracy Measures for Simulation using T = 50 and alpha = 0.8

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.82	0.81	1.02	1.02	0.82	0.81
CNR5	**1.67**	1.69	**2.69**	2.70	**1.62**	1.64
CNR10	**3.37**	3.45	**5.13**	5.17	**2.62**	2.70
CNB5	**0.84**	0.85	**1.03**	1.04	**0.82**	0.83
CNB10	**1.00**	1.01	**1.03**	1.06	**0.82**	0.84
CNE5	**4.19**	4.22	**5.14**	5.18	**4.10**	4.12
CNE10	**9.26**	9.34	**10.35**	10.48	**8.24**	8.33
Cauchy	**11.10**	11.77	**85.17**	89.92	**6.05**	6.73

Table 16: Accuracy Measures for Simulation using T = 30 and alpha = 0.8

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.81	0.81	1.02	1.02	0.81	0.81
CNR5	**1.66**	1.68	**2.67**	2.69	**1.63**	1.66
CNR10	**4.22**	5.07	**5.10**	5.17	**2.64**	2.75
CNB5	**0.83**	0.84	**1.03**	1.05	**0.82**	0.83
CNB10	**0.88**	0.91	**1.04**	1.08	**0.82**	0.85
CNE5	**4.10**	4.14	**5.11**	5.15	**4.06**	4.10
CNE10	**8.55**	8.69	**10.25**	10.43	**8.14**	8.29
Cauchy	**8.33**	8.70	**65.71**	181.03	**5.26**	8.91

Table 17: Accuracy Measures for Simulation using T = 20 and alpha = 0.8

	MAPE		RMSE		MAD	
Errors	SAE	SSE	SAE	SSE	SAE	SSE
Normal	0.83	0.82	1.04	1.03	0.83	0.82
CNR5	**1.67**	1.71	**2.71**	2.75	**1.65**	1.69
CNR10	**2.86**	2.98	**5.18**	5.26	**2.69**	2.81
CNB5	**0.85**	0.86	**1.06**	1.08	**0.84**	0.85
CNB10	**0.88**	0.91	**1.07**	1.11	**0.85**	0.88
CNE5	**4.16**	4.21	**5.18**	5.27	**4.13**	4.18
CNE10	**8.50**	8.85	**10.38**	10.83	**8.26**	8.59
Cauchy	**9.66**	10.72	**1567.28**	1572.77	**21.71**	23.72

Conclusion

The analyses presented suggest that minimizing SAE to determine exponential smoothing parameters can provide protection against outliers. Analysis of the M1-competition data suggests that cases where parameters selected by minimizing SAE result in superior forecasts do occur in practice. However, on average, minimizing SSE appears to provide forecasts that are reasonably robust to most outliers encountered. The simulation recommends that use of the SAE criterion would be most beneficial with the presence of outliers in conjunction with one or more of the following: larger values of the true smoothing parameter, outliers occurring near the end or throughout the series where forecasts are to be generated rather than at the beginning, and, obviously, cases where larger outliers are more likely. Further, even if outliers are not present, using the SAE criterion will not result in much deterioration in forecast accuracy.

References

Dielman, T. (1986). A comparison of forecasts from least absolute value and least squares regression. *Journal of Forecasting*, *5*, 189-195.

Dunne, A. (1992). Time series simulation. *The Statistician*, *41*, 3-8.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods. *Journal of Forecasting*, *1*, 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simpson, L. F. (1993). The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*, 5-23.

Makridakis, S. & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *15*, 451-476.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 130-139

Penalized Splines For Longitudinal Data With An Application In AIDS Studies

Hua Liang
Department of Biostatistics and Computational Biology
University of Rochester Medical Center

Yuanhui Xiao
Department of Mathematics and Statistics
Georgia State University

A penalized spline approximation is proposed in considering nonparametric regression for longitudinal data. Standard linear mixed-effects modeling can be applied for the estimation. It is relatively simple, efficiently computed, and robust to the smooth parameters selection, which are often encountered when local polynomial and smoothing spline techniques are used to analyze longitudinal data set. The method is extended to time-varying coefficient mixed-effects models. The proposed methods are applied to data from an AIDS clinical study. Biological interpretations and clinical implications are discussed. Simulation studies are done to illustrate the proposed methods.

Key words: Repeated measurements, varying-coefficient models, AIDS, ACTG315

Introduction

Recently, nonparametric regression has been used to analyze longitudinal data, which arise frequently in clinical trials and biological research and cannot be analyzed by traditional parametric approaches. The aims of nonparametric regression analysis include exploration of curves for a particular population and individual characteristic by introducing a mixed-effects framework. For parametric longitudinal data, for surveys, see Diggle, Liang and Zeger (1994), Davidian and Gilti-nan (1995), Vonesh and Chinchilli (1996) among others. Mixed-effects models provide a useful and flexible framework in which population characteristics are modeled as fixed effects, while individual variation is modeled as a random effect. Parametric mixed-effects models such as linear mixed-effects (LME) models (Laird & Ware 1982, Ware 1985, Diggle, et al. 1994) and nonlinear mixed-effects models (Davidian & Giltinan 1995, Vonesh & Chinchilli 1996) are widely used in longitudinal data analysis. Shi, Weiss, and Taylor (1996) and Rice and Wu (2001) proposed a nonparametric mixed-effects model for longitudinal data:

$$y_i(t) = \eta(t) + v_i(t\ ) + \varepsilon_i(t), \quad i = 1,2, \ldots, n, \tag{1}$$

where $\eta(t)$ models the population mean function, also called the fixed-effect or population curve; $v_i(t)$ models individual variations from $\eta(t)$ (these variation are called random-effect curves); $\varepsilon_i(t)$ are measurement errors; and $y_i(t)$ are response processes. The $v_i(t)$ and $\varepsilon_i(t)$ are assumed to be independent. $v_i(t)$'s can be considered as realizations of a zero mean process with a covariance function $\gamma(s,t) = E\{v_i(s)v_i(t)\}$, and $\varepsilon_i(t)$ can be regarded as realizations of an uncorrelated zero mean process with a variance function $\sigma^2(t)$. Let t_{ij}, $j = 1,2, \ldots, m_i$, be the design time points for the *i-th* individual, then model (1) becomes

Hua Liang is Associate Professor. His research interests are in methodologies for analyzing data in biomedical research, including longitudinal studies and clinical trials. Email: hliang@bst. rochester.edu. Yuanhui Xiao is Assistant Professor. His research interests are in time series analysis, statistical computation, statistical software development. E-mail: matyxx@langate.gsu.edu.

$$y_i(t_{ij}) = \eta(t_{ij}) + v_i(t_{ij}) + \varepsilon_i(t_{ij}), \quad j = 1, 2, \ldots, m_i; \quad i = 1, 2, \ldots, n, \tag{2}$$

where *n* is the number of subjects and mi is the number of measurements from subject *i*. For convenience, y_{ij} is denoted as being equal to $y_i(t_{ij})$ and ε_{ij} as being equal to $\varepsilon_i(t_{ij})$.

The primary goal is to estimate the fixed-effect (population) curve $\eta(t)$ and random-effect curves $v_i(t)$ or individual curves $s_i(t) = \eta(t) + v_i(t)$, for $i = 1, 2, \ldots, n$. The mean function $\eta(t)$ is important because it reflects the overall trend or progress of an underlying population process and can be used as an important index for the population response to a drug or a treatment in a clinical or biomedical study. The estimation of $v_i(t)$ or $s_i(t)$ is also important. The estimates of $v_i(t)$ are crucial for the estimation of the covariance of $y_i(t)$, which, in turn, can be used to better the estimate of the population curve $\eta(t)$ (see later sections). Because an individual curve $s_i(t)$ may represent an individual response to a treatment in a study, a good estimate of $s_i(t)$ may help investigators to make a better decision about individual treatment. The estimates of individual curves $s_i(t)$ are also useful if the investigators wish to group or classify the subjects on the basis of individual response curves.

Several methods have been proposed for the nonparametric modeling of longitudinal data. Diggle and Hutchison (1989), Altman (1990), Hart (1991), Rice and Silverman (1991) and others proposed modifications to criteria for selection of smoothing parameters. These modifications include leave-one-subject-out cross-validation (CV) or generalized cross-validation (GCV) to indirectly account for the correlations among data. Zhang et al. (1998) considered the correlation structure of longitudinal data in their smoothing spline semi-parametric mixed-effects models, but only the population curve (mean function) is modeled non-parametrically.

Wang (1998a, b) included the correlation in a mixed-effects smoothing spline models, but the special correlation structure of longitudinal data was not emphasized. Hart and Wehrly (1986) and Fan and Zhang (2000) suggested a two-step approach (local averaging or local regression first, then smoothing) to indirectly account for the data correlation. Hoover et al. (1998) and Wu, Chiang and Hoover (1998) proposed a standard local polynomial kernel method for varying-coefficient model with longitudinal data. Lin and Carroll (2000) propose a local polynomial generalized estimating equation (GEE) method for clustered data that may also be used to estimate the population curve $\eta(t)$ in our model. More recently, Wu and Zhang (2002) suggested that $\eta(t)$ and $v_i(t)$ be estimated simultaneously by combining LME models and local polynomial techniques, and they propose new bandwidth selection methods that are hybrid approaches of leave-one-subject-out and leave-one-point-out CV.

Although all of these approaches have demonstrated promise, several potential weaknesses exist.

(a) All these existing methods, except that of Wu and Zhang (2002), did not consider estimating the random-effect curves $v_i(t)$ or individual curves $s_i(t)$, which are very important in the application of the models to data from clinical and biological studies.

(b) The approach of Wu and Zhang (2002) has been shown to be more efficient than the other approaches, and the authors considered individual curves $s_i(t)$, but the computation of their methods is very expensive and sometimes unstable for bandwidth variation.

(c) Even if these weaknesses are ignored, the selection of smoothing parameters depends heavily upon selection criterion such as AIC, BIC or GCV.

Here, a new method is proposed to simultaneously estimate $\eta(t)$ and $v_i(t)$ by combining LME models (Laird & Ware 1982) and penalized techniques (Carroll & Ruppert 1999). The resulting estimators are called penalized spline LME (PSLME) estimators. This approach overcomes the above weakness, and is simple, easily and quickly implemented and robust to smoothing parameters.

An approach similar to the one proposed here has been used for common nonparametric regression. Parise, Ruppert, Ryan and Wand (2001) proposed penalized spline model to study the relationship between animal body weight and tumor onset by incorporating variation from one experiment to another. A similar mixed model was used to analyze the data from a study of the Utah Valley respiratory health/air pollution study by Coull, Schwartz and Wand (2001), and from a study of ragweed pollen data by Coull, Ruppert and Wand (2001).

The rest of the paper is organized as follows. Section 2 shows the derivation of the PSLME estimators and an extension to time varying coefficient mixed-effects model. As an illustration, an application of the model to a data set from an AIDS study is shown in section 3.1. A simulation study is presented in section 3.2. Some discussions are given in section 4.

Estimation Framework

Before the estimation framework is established, the principle of penalized spline for classic non-parametric regression is briefly introduced. More details were described by Ruppert and Carroll (1999).

The penalized least-squares estimator

The data (X_i, Y_i) follow $Y_i = m(X_i) + e_i$ for i = 1, 2, …, n, where X_i is univariate. To estimate m, β is equal to $(\beta_0, \beta_1, \ldots, \beta_p)^T$ and a regression spline model

$$m(x;\beta) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} b_k (x - \zeta_k)_+^p$$

is used to approximate $m(x)$, where $p \geq 1$ is an integer and $\zeta_1 < \cdots < \zeta_K$ are fixed knots, $a_+ = \max(a, 0)$. The traditional method of "smoothing" the estimate is knot selection. The estimator $\hat{\beta}(\alpha)$ of β is defined as the minimizer of

$$\sum_{i=1}^{n} \{Y_i - m(X_i;\beta)\}^2 + \alpha \sum_{k=1}^{K} b_k^2 \,, \tag{3}$$

where α is a smoothing parameter.

As shown by Brumback, Ruppert and Wand (1999), the estimator $\hat{\beta}(\alpha)$ based on equation (3) is equivalent to the estimator of β based on an LME model

$$y = X\beta + Zb + \varepsilon \,,$$

where

$$X = \begin{pmatrix} 1 & X_1 & \cdots & X_1^p \\ 1 & X_2 & \cdots & X_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n & \cdots & X_n^p \end{pmatrix},$$

$$Z = \begin{pmatrix} (X_1 - \zeta_1)_+^p & (X_1 - \zeta_2)_+^p & \cdots & (X_1 - \zeta_K)_+^p \\ (X_2 - \zeta_1)_+^p & (X_2 - \zeta_2)_+^p & \cdots & (X_2 - \zeta_K)_+^p \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \zeta_1)_+^p & (X_n - \zeta_2)_+^p & \cdots & (X_n - \zeta_K)_+^p \end{pmatrix}$$

$$b = (b_1, \ldots, b_K)^T \sim N(0, \sigma_b^2) \,,$$

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma_\varepsilon^2) \,,$$

and

$$\alpha = \alpha_\varepsilon^2 / \sigma_b^2 \,.$$

This fact implies that penalized spline smoother under the framework in equation (3) is equivalent to a standard LME. The solution can be obtained through the use of an LME macro available for S-PLUS software. The penalized parameter α is automatically estimated as $\hat{\alpha} = \hat{\alpha}_\varepsilon^2 / \hat{\alpha}_b^2$ by a restricted maximum likelihood (RML) approach.

Estimation Procedures for Model (3)

Motivated by the idea stated in Section 2.1, an estimation approach is proposed as follows. First, $\{(t_{ij}, Y_{ij})\}$ ($j = 1,2, ..., m_i$ and $i = 1,2, ..., n$) are the data drawn from the model in (2). The fixed effects functions $\eta(t)$ are approximated by

$$\tilde{\eta}(t, \beta, u) = \sum_{k=0}^{p} \beta_k t^k + \sum_{k=1}^{K} u_k (t - \zeta_k)_+^p$$

and those of $v_i(t)$ are approximated by

$$\tilde{v}_i(t, b_i, w_i) = \sum_{k=0}^{p} b_{ik} t^k + \sum_{k=1}^{K} w_{ik} (t - \zeta_k)_+^p$$

Here

$$\beta = (\beta_0, \ldots, \beta_p)^T, \quad u = (u_1, \ldots, u_K)^T,$$
$$b_i = (b_{i0}, \ldots, b_{ip})^T, \quad w_i = (w_{i1}, \ldots, w_{iK})^T.$$

Assume that $\{u_k\} \sim N(0, \sigma_u^2), \{b_{ik}\} \sim N(0, \sigma_{b,k}^2)$ and $\{w_{ik}\} \sim N(0, \sigma_w^2)$ for $k = 1, \ldots, K$ and $i = 1, ..., n$. Then $\tilde{\eta}(t, \beta, u) + \tilde{v}_i(t, b_i, w_i)$ is the individual curve of the i^{th} subject. Define the following matrix notation.

$$X_i = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^p \\ 1 & t_{i2} & \cdots & t_{i2}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \cdots & t_{im_i}^p \end{pmatrix},$$

$$Z_i = \begin{pmatrix} (t_{i1} - \zeta_1)_+^p & (t_{i1} - \zeta_2)_+^p & \cdots & (t_{i1} - \zeta_K)_+^p \\ (t_{i2} - \zeta_1)_+^p & (t_{i2} - \zeta_2)_+^p & \cdots & (t_{i2} - \zeta_K)_+^p \\ \vdots & \vdots & \ddots & \vdots \\ (t_{im_i} - \zeta_1)_+^p & (t_{im_i} - \zeta_2)_+^p & \cdots & (t_{im_i} - \zeta_K)_+^p \end{pmatrix},$$

$$\eta_i(t_i) = \{\eta_i(t_{i1}), \ldots, \eta_i(t_{im_i})\}^T,$$
$$y_i = (y_{i1}, \ldots, y_{im_i})^T, \quad X = (X_1^T, \ldots, X_n^T)^T,$$
$$y = (y_1^T, \ldots, y_n^T)^T,$$
$$\Lambda = diag(X_1, \ldots, X_n), \quad Z = (Z_1^T, \ldots, Z_n^T)^T,$$
$$\Gamma = diag(Z_1, \ldots, Z_n),$$
$$b = (b_1^T, \ldots, b_n^T)^T, \quad w = (w_1^T, \ldots, w_n^T)^T.$$

The approximation of the model in (2) can be rewritten as

$$y = X\beta + \Lambda b + Zu + \Gamma w + \varepsilon$$

This standard LME has unknown population parameters β and unknown individual effects b, u and w. The estimates $\hat{\beta}$, $\hat{b}$, $\hat{u}$ and $\hat{w}$ of the parameter vector can be easily given closed forms, and the well-developed SAS and S-plus macros can be directly applied for computation. As a consequence, population and individual curves can be obtained from the estimates $\tilde{\eta}(t, \hat{\beta}, \hat{u})$ and $\tilde{v}_i(t, \hat{b}_i, \hat{w}_i)$.

For a common penalized spline, the penalty parameter α and the number of knots K must be selected. Relatively speaking smoothing is controlled by the penalty parameter α, and the number of knots K is not a crucial parameter. See also Ruppert (2002) for a detailed discussion. As indicated in section 2.2, the formulation of mixed-effects model automatically derives an estimated of α. Only K needs to be specified. Computation experience indicates $\max(10, n/4)$ is a good choice as a value of K and that the results are very insensitive to different values of K. The knots are then at equally spaced sample quantiles of $\{t_{ij}\}$.

Extension to Time Varying-coefficient Models

As an effective approach to reduce curse of dimensionality suffered in high-dimension non-parametric regression, time varying-coefficient models were first proposed in longitudinal data structure by Hoover, Rice, Wu and Yang (1998) and Wu, Chiang and Hoover (1998). The standard time-varying coefficient models (Hoover et al. 1998, Wu et al. 1998) can be written as

$$y_i(t) = c^T(t) + \eta_i(t) + \varepsilon_i(t), \quad i = 1, \dots, n, \tag{4}$$

Where $c(t) = \{1, c_1(t), \dots, c_L(t)\}^T$ and $\eta_i(t) = \eta(t) + v_i(t)$ with $\eta(t) = \{\eta_0(t), \dots, \eta_L(t)\}^T$ and $v_i(t) = \{v_{0i}(t), \dots, v_{Li}(t)\}^T$. The functions $\eta_l(t)$ and $\eta_l(t) + v_{li}(t)$ indicate the population and individual effects of $c_l(t)$ for subject i.

Both smoothing spline and local polynomial kernel regression methods are proposed by Hoover et al. (1998). Alternatively, Fan and Zhang (2000) proposed a two-step method for the same model. However, none of these methods efficiently considered the important features of longitudinal data such as between-subject and within-subject variation, and the special correlation structure of longitudinal data. Lin and Carroll (2000), however, showed that specifying the correlation structure when using kernel methods to estimate the nonparametric function results an asymptotically less efficient estimator than the one obtained assuming independence among repeated measures. Welsh, Lin and Carroll (2000) showed regression and smoothing splines do not suffer from this difficulty.

Local polynomial estimates of Hoover et al. (1998) rely upon one bandwidth to smooth all coefficient curves, but these estimates may not be enough to capture smoothness of all coefficient curves simultaneously. The smoothing spline method of Hoover et al. (1998) permits the use of multiple smoothing parameters, but the computation is very intensive even only a single smoothing parameter is included when the number of distinct observation time is large.

More recently, Liang, Wu and Carroll (2003) proposed a global fitting method for a varying-coefficient model based on basis spline approximation. The purpose of their method is to approximate the coefficient functions by the basis spline. Their approach is shown to be simple to estimation and inference for the timing varying coefficient models.

Approximate $\eta_l(t)$ and $v_{li}(t)$ by using the following:

$$\tilde{\eta}_l(t, \beta_l, u_l) = \sum_{k=0}^{p_l} \beta_{lk} t^k + \sum_{k=1}^{K_{l1}} u_{lk} (t - \zeta_{lk})_+^{p_l}$$

and

$$\tilde{v}_{li}(t, b_{li}, w_{li}) = \sum_{k=0}^{q_l} b_{lik} t^k + \sum_{k=1}^{K_{l2}} w_{lik} (t - \zeta_{lk})_+^{q_l}$$

for $l = 1, \dots, L$. After notation similar to that in section 2.2 is introduced, model (4) can be approximated by

$$y = \sum_{l=1}^{L} X_l \beta_l + \sum_{l=1}^{L} (\Lambda_l b_l + Z_l u_l + \Gamma_l w_l) + \varepsilon.$$

Again, the estimates for all parameters and subsequent population and individual curves can be derived.

Numerical Examples

Analyses of Data from the ACTG 315 Study

ACTG 315 was a single-arm clinical trial in which 53 enrolled subjects with moderately advanced HIV-1 infection received combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir for 48 weeks. The primary objective of the study was to assess whether the treatment was associated with evidence of immunologic restoration. Of the 53 subjects (49 men, 4 women, age range 6-63 years), 44 remained on treatment for at least 9 of the first 12 weeks. Lederman et al. (1998) reported the results of the study after 12 weeks of follow-up. The lower limit of quantification of HIV-1 RNA viral-load is 100 copies/ml. The HIV-1 RNA measures below this limit are not considered reliable; therefore, we censor values that are below 100 copies/ml. HIV-1 RNA measurements were observed on days 0, 2, 7, 10, and weeks 2, 3, 4, 8, 12, 24, and 48 of follow-up.

One aim of the ACTG 315 study is to characterize the viral load trajectory in the population and the individual patients during antiviral treatment. The population estimate of the viral-load trajectory was obtained as a function of treatment time by using the PSLME method. The estimated curves are presented in Figure 1 in dotted lines. The PSLME method was used to estimate the viral-load trajectories for individual patients. The ability to estimate values for population and individual characteristics is another important advantage of the PSLME method. The individual estimates of viral-load traj ectory for four selected patients are shown in Figure 1, which indicates that individual viral-load trajectories may differ from that estimate for the population. The viral-load trajectory of subject 18 is identical to the viral-load trajectory of the population and the pattern of viral-load trajectory in subject 1 is similar to that of the population, but the difference in magnitude is obvious. Other large differences between individual viral-load trajectory in subjects 23 and 3 5 and that in the population are observed. The estimated trajectories of viral-load in individual patients can provide more accurate information for physicians with which to individualize treatment management for individual patients with AIDS.

To study the relationship between virologic and immunologic responses, repeatedly measured by HIV RNA levels (viral load) and CD4+ cell counts respectively in an AIDS clinical trial ACTG 315, observe that the viral load and CD4+ cell counts are negatively and approximately linearly related in most of the treatment times, but the regression coefficients may not be constant during the whole treatment period. Motivated by this feature of the data, Liang, Wu and Carroll (2003) proposed a mixed-effects varying-coefficient model. The model captures population and individual relationships for the two longitudinal variables. The method proposed above is used to analyze this data set again. In the implementation, set $p = q = 2$, and $K_{l1} = 6$ and $K_{l2} = 10$. Other values were tried, and the results are very stable. The discoveries are similar to what Liang, Wu, and Carroll (2003) obtained. The viral load and CD4+ cell counts are inversely related in the study population during the treatment. However, the strength of the association varies smoothly, where the association is very strong at the beginning of the treatment to the weakest about 4 weeks of treatment. The association gradually recovered and is strongest from week 4 to week 24. See the dotted line in Figure 2 for the population curve.

Figure 2 also shows the individual estimates of $\beta_1(t)$ from four arbitrarily selected patients and the corresponding population estimate of $\beta_1(t)$. Not only the magnitude but also the patterns differ between the population and individual estimates of $\beta_1(t)$ (Figure 2). The pattern for subject 18 is almost identical to that of the population pattern. The patterns for subjects 1 and 47 are similar to the population pattern. However, the viral load and CD4+ cell counts of subject 1 was positive correlated with those of subject 47 during the early treatment stage. For subject 47, there is a negative correlation between viral load and CD4+ cell counts in the later stage. Interestingly we also observe discordance between patterns of the population estimate and individual estimates of $\beta_1(t)$. See pattern for subject 2 shown in Figure 2. Because of the large between-subject variation, the individual estimates become very important in individualizing treatment and care for patients with AIDS.

A Simulation Study

A simulation model is designed as $y_i(t) = \eta(t) + \gamma_i(t) + \varepsilon_i(t)$, where $\eta(t) = 1 + \cos(2\pi t) + \sin(2\pi t)$ and $\gamma_i(t) = a_{i0} + a_{i1}\cos(2\pi t) + a_{i2}\sin(2\pi t)$ with $(a_{i0}, a_{i1}, a_{i2})^T \sim N((0, 0, 0)^T, I_{3\times 3})$, and $\varepsilon_i(t) \sim N(0,1)$, for $i = 1, \ldots, n = 20$. The design time points are $t_{ij} = j/(1+m)$ for $j = 1, \ldots, m = 35$. To mimic the unbalanced data feature in longitudinal studies, randomly remove y_{ij} with a probability of r_m = 0.35 (i.e., r_m is the missing rate of the data). Thus, there are an average of 23 observations for each subject and 460 observations in total. Note that the data from different subjects are independent, but the within-subject data are correlated. The within-subject correlation coefficient can be calculated as:

$$\rho_y = corr\{y_i(t), y_i(s)\} = \{1 + \cos 2\pi(t-s)\}/2$$

for $s \neq t$. In this simulation experiment and in later examples, let p = 3 and K = 8 set $\sigma_w^2 = 0$. When a Dell PC machine (2GHz CPU) was used, the computation for the simulation experiment require only 8 seconds. The estimated value of the penalized parameter α is $\hat{\alpha} = 0.034$.

Figure 3 shows the profiles of data for 6 arbitrarily selected subjects. The generated data, the real population curve, the estimated population and individual curves are depicted for comparison. Although the population estimate is similar to the true characteristic of the population, the estimated individual curves more precisely describe individual trends than the estimated population curves. For comparison, this simulation data was set for p = 2 and K = 10,15, 20. The corresponding results are not distinguishable from those in Figure 3.

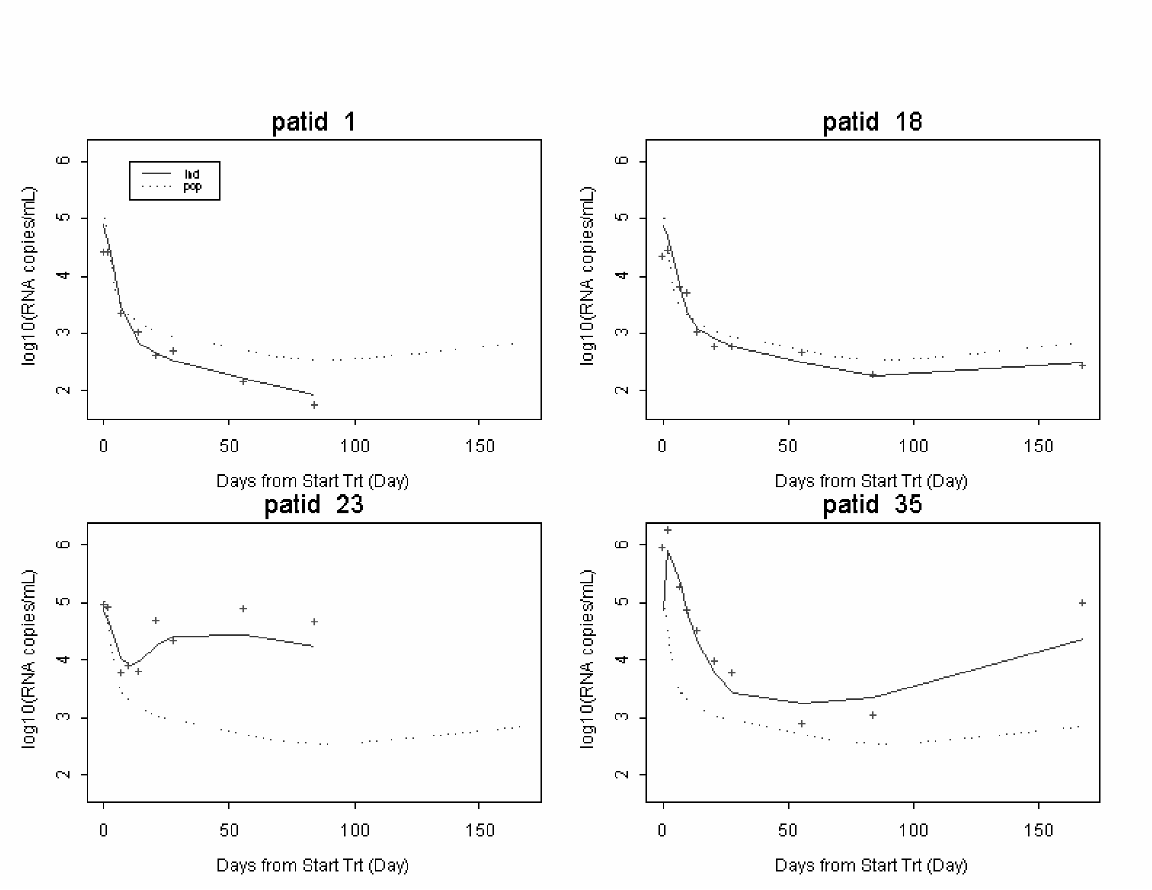


Figure 1.

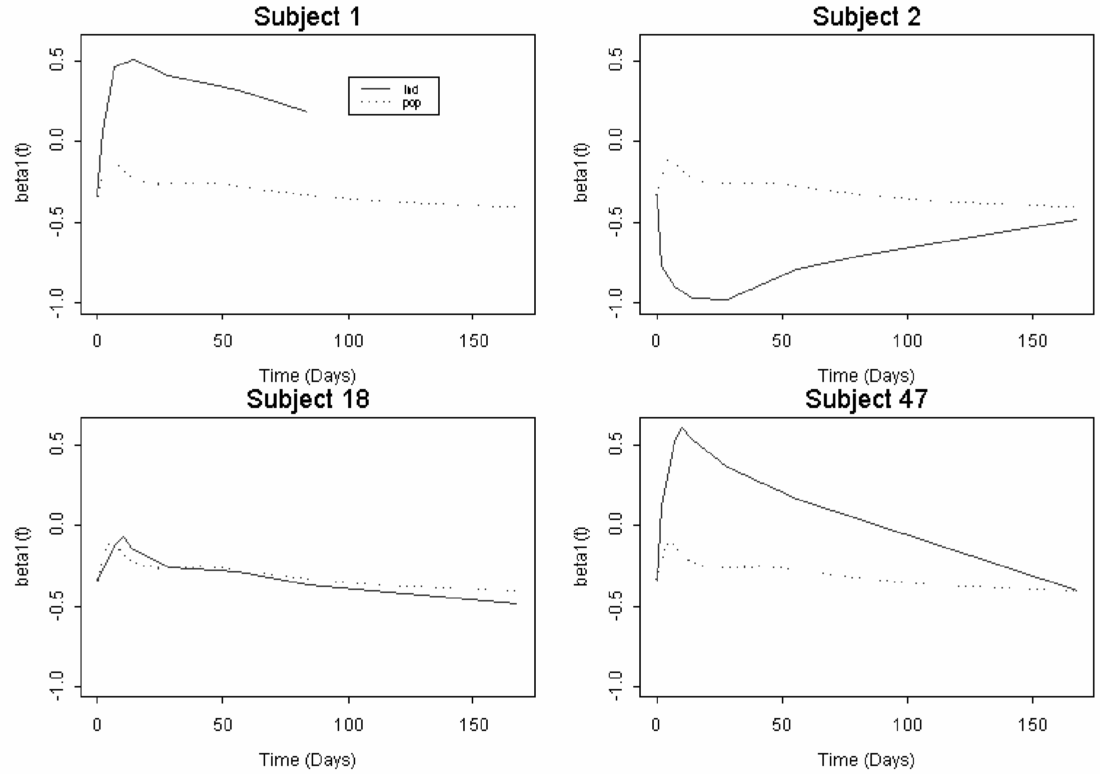

Subject 1
Subject 2
Subject 18
Subject 47
Ind
pop
beta1(t)
Time (Days)
0.5
0.0
-0.5
-1.0
0
50
100
150

Figure 2.

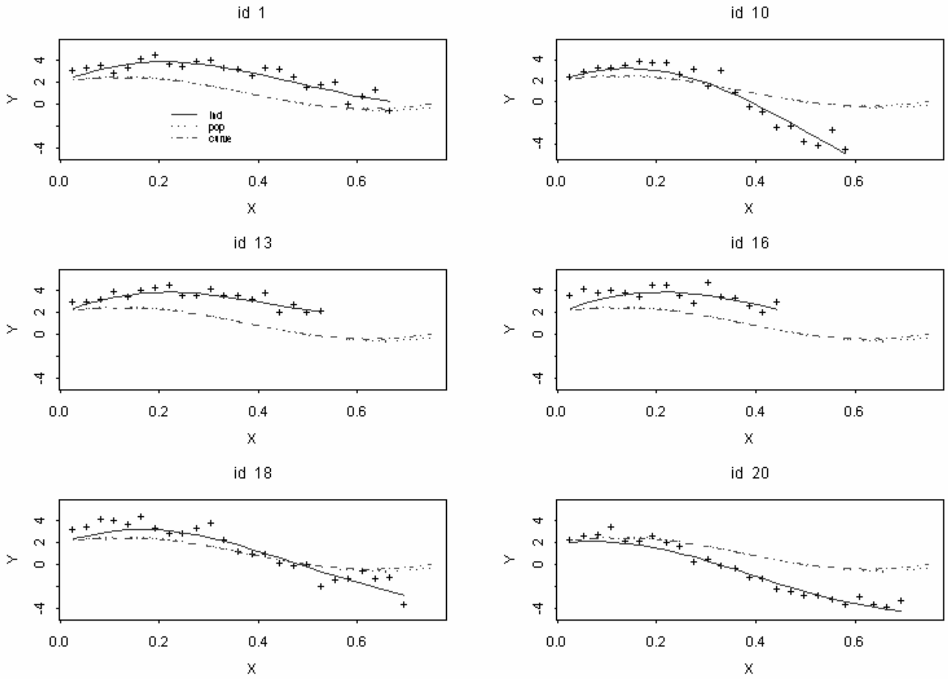


Figure 3.

Conclusion

Considering nonparametric regression modeling for longitudinal data, a very effective routine is proposed by combining a penalized spline technique and LME models. The principal advantage of this approach is that it avoids computational challenges that occur when local kernel smoothing or smoothing spline techniques in which bandwidths or smoothing penalty parameters have to be selected are used. This approach avoids these challenges by using a concern of LME. Penalty parameters were automatically calculated out. Curves for population and individual characteristics are easily derived. The approach is also effective to time varying coefficient mixed-effects models. The method has been shown to be useful in analyzing AIDS data set. It is believed that the approach can be used to other clinical trail or biological data.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association, 85*, 749-759.

Brumback, B. A. & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association, 93*, 961-994.

Brumback, B. A., Ruppert, D., & Wand, M. (1999). Comments on “Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior” by Shively, Kohn, and Wood. *Journal of the American Statistical Association, 94*, 794-797.

Coull, B. A., Ruppert, D., & Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics, 57*, 539-545.

Coull, B.A., Schwartz, J., & Wand, M. P. (2001). Respiratory health and air pollution: Additive mixed model analyses. *Biostatistics, 2*, 337-349.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* Chapman & Hall, New York.

Diggle, P. J., & Hutchison, M. F. (1989). On spline smoothing with autocorrelated errors. *Australian Journal of Statistics, 31*, 166-168.

Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Sciences, 11*, 89-121.

Fan, J. & Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B, 62*, 303-322.

Hoover, D. R., Rice, J. A., Wu, C. O., & Yang, L. P. (1998). Nonparametric smoothing estimates of timing-varying coefficient models with longitudinal data. *Biometrika, 85*, 809-822.

Laird, N. M. & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics, 38*, 963-974.

Lederman, M. M., Connick, E., Landay, A, et al. (1998). Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of Zidovudine, Lamivudine and Ri-tonavir: Results of AIDS clinical trials group protocol 315. *The Journal of Infectious Diseases, 178*, 70-79.

Liang, H., Wu, H. L., & Carroll, R. J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics, 4*, 297-312.

Lin, X. & Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association, 95*, 520-534.

Parise, H., Ruppert, D., Ryan, L., & Wand, W. P. (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics, 50*, 31-42

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*, New York: Springer.

Rice, J. A. & Silverman B. W. (1991). Estimating the mean and covariance structure nonpara-metrically when the data are curves. *Journal of the Royal Statistical Society, Series B, 53*, 233-243.

Rice, J. A. & Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics, 57*, 253-259.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistic (in press).*

Ruppert, D. & Carroll, R. (1999). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics, 42*, 205-253.

Shi, M., Weiss, R. E., & Taylor, J. M. G. (1996). An analysis of pediatrics CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics, 45*, 151-163.

Vonesh, E. F. & Chinchilli, V. M. (1996), *Linear and nonlinear models for the analysis of repeated measurements,* New York: Marcel Dekker, Inc.

Wang, Y. D. (1998a). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association, 93*, 341-348.

Wang, Y. D., (1998b). Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B, 60*, 159-174.

Welsh, A., Lin, X., & Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association, 97*, 482-493.

Wu, C. O., Chiang, C. T., & Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association, 93*, 1388-1402.

Wu, H. & Zhang, J. T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association, 97*, 883-897.

Zeger, S. L. & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics, 50*, 689-699.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 140-151

Properties Of The GAR(1) Model For Time Series Of Counts

Vasiliki Karioti Chrys Caroni
Department of Mathematics
National Technical University of Athens

Models for time series count data include several proposed by Zeger and Qaqish (1988), subsequently generalized into the *GARMA* family. The *GAR(1)* model is examined in detail. The maximum likelihood estimation of the parameters will be discussed and the properties of Pearson and randomized residuals will be examined.

Key words: time series, count data, GARMA, GAR(1), residuals

Introduction

Many of the time series recorded in practice consist of count data, in which each observation represents the number of events occurring at a point in time or in a given time interval. Examples include the number of cases of a particular disease reported each month. Especially when the counts are low, standard Gaussian time series models may need to be replaced by other models more suitable for count data, based on the Poisson distribution or another discrete distribution on the non-negative integers.

A number of models of this type have been developed. In this article, regression models for time series count data will be examined. These models, originally proposed by Zeger and Qaqish (1988), have been considered subsequently by several other authors (see, in particular, Kedem and Fokianos, 2002) and extended by Benjamin, Rigby, and Stasinopoulos (2003). In these models, each observation y_t in the series is represented as a Poisson variate which is conditionally independent of previous observations, given its mean, but whose mean depends on the previous observations $y_{t\text{-}1},...,y_1$ and possibly on covariates. These are examples of observation-driven models for time-dependent data in the terminology introduced by Cox (1981). In the simplest case, with first-order dependence and no covariates:

$$y_t \mid y_{t\text{-}1} \sim Poisson\ (\mu_t)$$

where

$$\mu_t = \mu_t(y_{t\text{-}1})\ .$$

In this article, the basic model is examined from several points of view relevant to its practical application to data. Principally, the performance of maximum likelihood estimation of the parameters and the properties of the residuals from the models are examined.

Models

Following Zeger and Qaqish (1988), let y_t be an outcome random variable and x_t an mx1 vector of covariates at time t. Define $\mu_t = E(y_t \mid D_t)$ where $D_t = \{x_t\ , x_{t\text{-}1},\ldots,\ y_{t\text{-}1}, \ldots, y_1\}$ includes past outcomes and the past and present covariates. It is assumed that

$$g(\mu_t) = x_t^{'}\beta + \sum_{i=1}^{p}\theta_i f_i(D_t)$$

Vasiliki Karioti's main interest is in time series analysis. Email: vaskar@otenet.gr. Chrys Caroni is Associate Professor in the Department of Mathematics. Her research work is mainly on multivariate outliers and reliability. E-mail: ccar@math.ntua.gr

where g is a link function, the f_i are functions of the past data and the parameters β and $\theta = (\theta_1,\ldots,\theta_p)'$ are to be estimated. Because the link function is applied to the lagged observations y_{t-j}, this model goes beyond standard generalized linear models (GLM) with independent data (McCullagh and Nelder, 1989). A general model for μ_t is:

$$\begin{aligned} g(\mu_t) &= \eta_t \\ &= x_t'\beta + \sum_{j=1}^{p} \theta_j \left\{ g(y_{t-j}) - x_{t-j}'\beta \right\} \\ &\quad + \sum_{j=1}^{q} \phi_j \left\{ g(y_{t-j}) - \eta_{t-j} \right\} \end{aligned} \tag{1}$$

This defines a class of models called generalized autoregressive moving average models (GARMA: Benjamin, Rigby, and Stasinopoulos, 2003). A special case of GARMA arises when the conditional distribution for y_t (given D_t) is Poisson and g the canonical link function as in standard GLM, that is, the logarithm. Equation (1) becomes:

$$\begin{aligned} g(\mu_t) &= \log(\mu_t) \\ &= x_t'\beta + \sum_{j=1}^{p} \theta_j \left\{ \log(y_{t-j}^*) - x_{t-j}'\beta \right\} \\ &\quad + \sum_{j=1}^{q} \phi_j \left\{ \log(y_{t-j}^*) - \eta_{t-j} \right\} \end{aligned} \tag{2}$$

where $y_{t-1}^* = \max(y_{t-1}, c)$, $0 < c < 1$ (Zeger and Qaqish, 1988; Benjamin, Rigby, and Stasinopoulos, 2003). The effect of using y_t^* in place of y_t is that zero values of y_t are replaced by c. This device is adopted in order to avoid an absorbing state at $y = 0$. If $\phi_j = 0$, for $j = 1,\ldots,q$, the model is autoregressive order p, *GAR(p)*. If $\theta_j = 0$, for $j = 1,\ldots,p$, it is a moving average model of order q, *GMA(q)* (Li, 1994). In the special case of $\phi_j = 0$, and $p = 1$, the model (2) is *GAR(1)* with the form:

$$\log(g(\mu_t)) = x_t'\beta + \theta_1 \left\{ \log(y^*_{t-1}) - x_{t-1}'\beta \right\} \tag{3}$$

If there are no covariates x, then writing $x_t'\beta = \mu =$ constant, equation (3) becomes:

$$\mu_t = \exp(\mu) \left[\frac{y_{t-1}^*}{\exp(\mu)} \right]^{\theta_1} \tag{4}$$

Positive values of θ_1 represent positive autocorrelation within the series and negative values represent negative autocorrelation. Zeger and Qaqish (1988) also proposed another way of solving the problem of the absorbing state. Instead of introducing y_t^*, this model defines:

$$\mu_t = \exp(\mu) \left[\frac{y_{t-1} + c}{\exp(\mu) + c} \right]^{\theta_1} \tag{5}$$

where c is a constant added to each observation rather than only to zero outcomes. In some situations it might be interpreted as an immigration rate. This model is not part of the GARMA family.

Maximum Likelihood Estimation

The likelihood function conditional on the first term of the series is given by

$$L(y_2, y_3, \ldots, y_m \mid y_1) = \prod_{t=2}^{m} P[Y_t = y_t \mid y_{t-1}] = \prod_{t=2}^{m} \left(\frac{e^{-\mu_t} \mu_t^{y_t}}{y_t!} \right)$$

with log-likelihood

$$\ell = \ln L = \sum_{t \geq 2} \{ -\mu_t + y_t \ln \mu_t \} - \sum_{t \geq 2} \ln y_t!$$

Let the vector of model parameters to be estimated be denoted by η. Then

$$\frac{\partial \ell}{\partial \eta_i} = \sum_{t \geq 2} \left(\frac{y_t}{\mu_t} - 1 \right) \frac{\partial \mu_t}{\partial \eta_i}$$

$$\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j} = \sum_{t \geq 2} \left(\frac{y_t}{\mu_t} - 1 \right) \frac{\partial^2 \mu_t}{\partial \eta_i \partial \eta_j} - \sum_{t \geq 2} \left(\frac{y_t}{\mu_t^2} \right) \left(\frac{\partial \mu_t}{\partial \eta_i} \right) \left(\frac{\partial \mu_t}{\partial \eta_j} \right)$$

Closed-form expressions are not available for the estimation of η. Consequently, the likelihood must be maximised numerically. The BCOAH subroutine was used from the IMSL library to minimize the negative of the log-likelihood. This employs a modified Newton method and a user-supplied Hessian. Zeger and Qaqish (1988) fitted their models by quasi-likelihood estimation. Benjamin, Rigby, and Stasinopoulos (2003) fitted GARMA models by maximum likelihood using iteratively weighted least squares.

Simulation study

To examine the *GAR(1)* model from several points of view relevant to its practical application to data, a numerical study of simulated data was carried out. The limitation to first-order autoregression is common throughout the time series literature, chiefly for practical reasons (Greene, 2000). Because there is only one autoregressive parameter θ_1, its subscript will be dropped from this point on. To generate a realization of a time series of length *m* for selected values of μ, θ and *c*, the *GAR(1)* model (4) was used to generate a sequence of *m* + 50 counts, starting from a Poisson deviate. The pseudorandom number generator *RNPOI* from the IMSL library was used to generate Poisson deviates. The first 50 counts were discarded and the remaining *m* values were retained for analysis. A relatively short series of *m* = 50 observations and longer series of *m* = 150 observations were examined.

From (4), the parameter *c* appears in the likelihood only in the terms, if any, that immediately follow a zero. If there are few zeros in the series, then there is very little information available for the estimation of *c*. If desired, its estimation can be avoided in order to simplify the likelihood equations. As well, a very flat likelihood surface (with respect to *c*) can be avoided by dividing the series into blocks. A block ends when a zero occurs, and the following block starts with the next non-zero outcome. The overall likelihood is the product of the likelihoods of the separate blocks, each of which is conditional on the first member of the block, and it is a function of θ and μ only. The minor drawback of this procedure is that some information is lost, because the overall likelihood consists not of *m* - 1 but *m - 1* - m_0 terms, where m_0 is the number of zeros occurring within the series.

Results

Table 1 shows summary statistics for the estimates of θ in the *GAR(1)* model. Difficulties with the numerical fitting procedure prevented the use of the larger values of $|\theta|$ when μ was small. It appears that the maximum likelihood estimate of μ is effectively unbiased, although a minor downward bias appears as θ increases to large positive values. The precision of the estimate of μ increases as μ increases, and appears to be a decreasing function of θ being lowest when θ takes large positive values. Table 2 shows results for the estimation of θ. There is some downwards bias in $|\theta|$, larger when $\theta > 0$ than when $\theta \le 0$, and larger for series of length 50 than ones of length 150. The precision of estimation of θ is also a decreasing function of θ but depends less heavily on the value of μ. Comparison of mean squared errors between Tables 1 and 2 shows that μ is estimated relatively much more precisely than θ.

Table 3 shows the correlation between estimates of θ and μ. Correlations appear to be a decreasing function of θ and also of μ, but do not depend heavily on the length of the series. For the larger values of μ (= 4, 6) and for θ positive or moderately negative, the estimates of the two parameters are virtually uncorrelated.

Table 1. Average and mean squared error of maximum likelihood estimate of μ in the GAR(1) model. Each entry is based on 2,000 simulated sets of data.

		Length 50		Length 150	
μ	θ	mean	m.s.e.	mean	m.s.e.
2	-0.6	2.000	0.0011	2.000	0.0004
	-0.3	1.998	0.0017	2.001	0.0057
	0	1.996	0.0028	1.999	0.0010
	0.3	1.996	0.0057	1.998	0.0021
	0.6	1.976	0.0215	1.990	0.0068
4	-0.8	4.000	0.0001	4.000	0.00004
	-0.6	4.000	0.0001	4.000	0.00005
	-0.3	3.999	0.0002	4.000	0.0001
	0	3.999	0.0004	4.000	0.0001
	0.3	3.999	0.0007	4.000	0.0003
	0.6	3.997	0.0023	3.999	0.0008
	0.8	3.988	0.0102	3.995	0.0032
6	-0.8	6.000	0.00002	6.000	0.00001
	-0.6	6.000	0.00002	6.000	0.00001
	-0.3	6.000	0.00003	6.000	0.00001
	0	6.000	0.00005	6.000	0.00002
	0.3	6.000	0.0001	6.000	0.00004
	0.6	5.999	0.0003	6.000	0.00011
	0.8	5.998	0.0013	5.999	0.0018

Table 2. Average and mean squared error of maximum likelihood estimate of θ in the GAR(1) model. Each entry is based on 2,000 simulated sets of data.

		Length 50		Length 150	
μ	θ	mean	m.s.e.	mean	m.s.e.
2	-0.6	-0.586	0.0087	-0.596	0.0026
	-0.3	-0.300	0.0129	-0.298	0.0044
	0	-0.014	0.0162	-0.007	0.0053
	0.3	0.267	0.0173	0.288	0.0054
	0.6	0.549	0.0165	0.584	0.0046
4	-0.8	-0.776	0.0086	-0.790	0.0025
	-0.6	-0.586	0.0127	-0.596	0.0036
	-0.3	-0.303	0.0168	-0.301	0.0056
	0	-0.021	0.0186	-0.007	0.0065
	0.3	0.258	0.0206	0.285	0.0064
	0.6	0.539	0.0186	0.580	0.0049
	0.8	0.727	0.0168	0.777	0.0033
6	-0.8	-0.777	0.0094	-0.791	0.0027
	-0.6	-0.587	0.0130	-0.597	0.0041
	-0.3	-0.307	0.0177	-0.302	0.0059
	0	-0.021	0.0202	-0.009	0.0069
	0.3	0.258	0.0202	0.285	0.0064
	0.6	0.542	0.0184	0.578	0.0053
	0.8	0.729	0.0162	0.775	0.0034

Table 3. Correlations between maximum likelihood estimates of μ and θ.

	Length 50			Length 150		
θ	μ=2	4	6	μ=2	4	6
-0.8		0.179	0.082		0.201	0.081
-0.6	0.309	0.145	0.040	0.351	0.125	0.049
-0.3	0.246	0.084	0.033	0.246	0.043	0.035
0	0.168	0.060	0.054	0.168	0.069	0.001
0.3	0.108	0.065	0.025	0.150	0.055	0.059
0.6	0.021	0.070	0.006	0.055	-0.015	-0.027
0.8		0.063	0.015		0.008	0.028

Residuals

In any regression model, it is important to examine residuals in order to assess the model's adequacy. Our ability to do this depends quite heavily on whether or not the residuals follow the normal distribution; otherwise it may be difficult to draw conclusions from their behavior. Benjamin, Rigby, and Stasinopoulos (2003) advocated using Dunn and Smyth's (1996) randomized quantile residuals for this purpose, because they expected Pearson or deviance residuals to be highly non-normally distributed for count data, at least when the mean count is low. Randomized quantile residuals are defined by

$$r_t = \Phi^{-1}\left(u_t\right) \qquad (6)$$

where Φ^{-1} is the inverse standard normal cumulative distribution function, u_t is a random variable uniformly distributed on the interval $\left[F\left(y_t\text{-}1;\, \hat{\mu}_t\right),\, F\left(y_t;\, \hat{\mu}_t\right)\right]$ and $F\left(y_t;\, \hat{\mu}_t\right)$ is the fitted Poisson cumulative distribution function.

Figures 1-4 show examples of the behavior of ordinary Pearson residuals $\left(y_t - \hat{\mu}_t\right)/\hat{\mu}_t^{1/2}$ and randomized quantile residuals in series of length 50, first within a series (all residuals from one simulated series) and then across series (the residual for t=20 examined across all 2000 simulations of the same set of parameter values). Figure 1 shows that, even though the counts are quite low (μ=2), the Pearson residuals within a series do not depart from normality as much as might be expected, so although the randomized quantile residuals (Figure 2) give an improvement, this does not seem to be important. However, across series the Pearson residuals depart markedly from a normal distribution (Figure 3) in the extreme tails whereas the randomized quantile residuals have much better behavior (Figure 4).

In the corresponding Figures 5-8 for series of length 150, it can be seen that the Pearson residuals are quite satisfactory; therefore there is little scope for the randomized quantile residuals to offer any improvement.

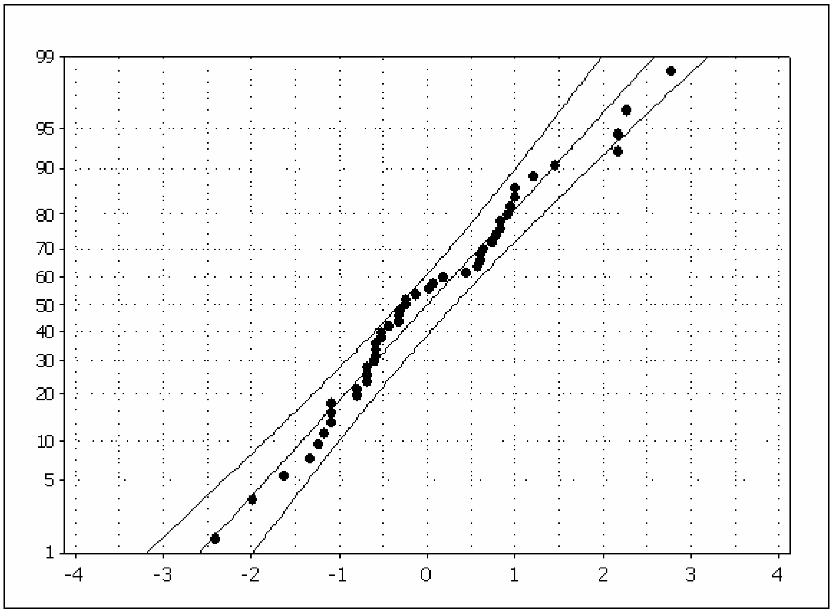


Figure 1. Normal probability plot of Pearson residuals from one realization of *GAR(1)* with $m = 50$, $\mu = 2$, $\theta = 0.3$.

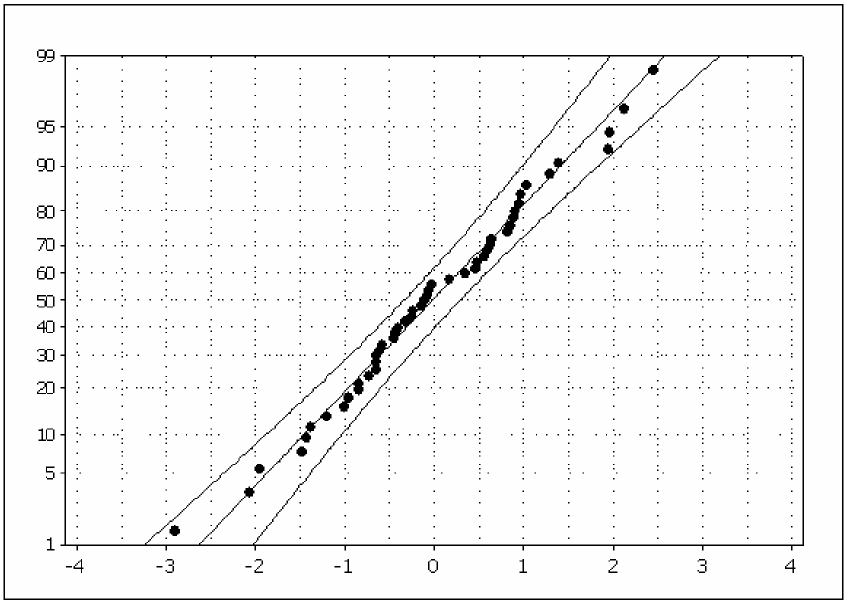


Figure 2. Normal probability plot of randomized residuals from one realization of *GAR(1)* with $m = 50$, $\mu = 2$, $\theta = 0.3$.

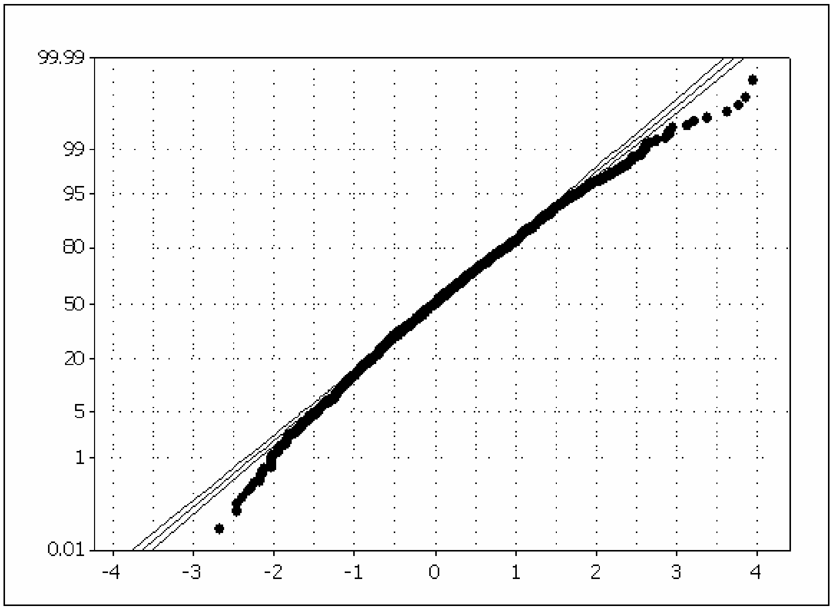


Figure 3. Normal probability plot of Pearson residuals at t = 20 from 2000 realizations of *GAR(1)* with m = 50, μ = 2, θ = 0.3.

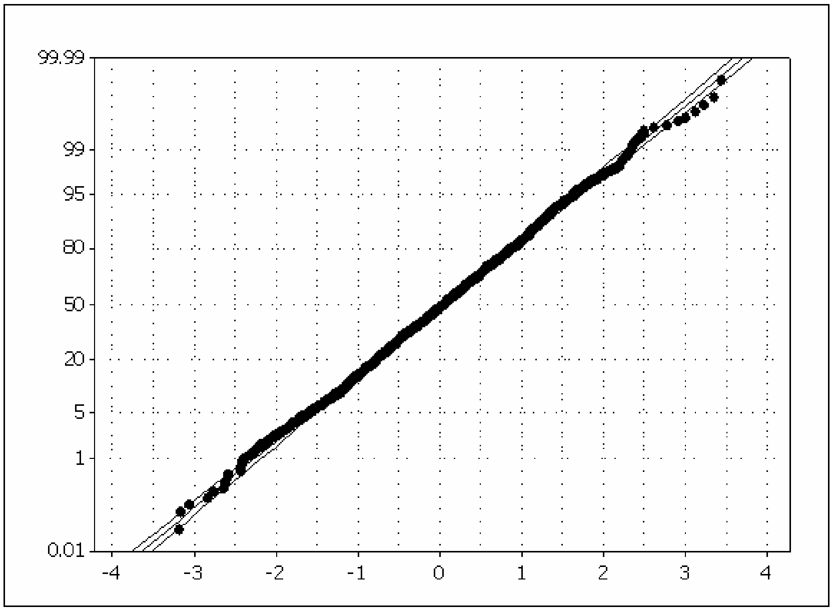


Figure 4. Normal probability plot of randomized residuals at t = 20 from 2000 realizations of *GAR(1)* with m = 50, μ = 2, θ = 0.3.

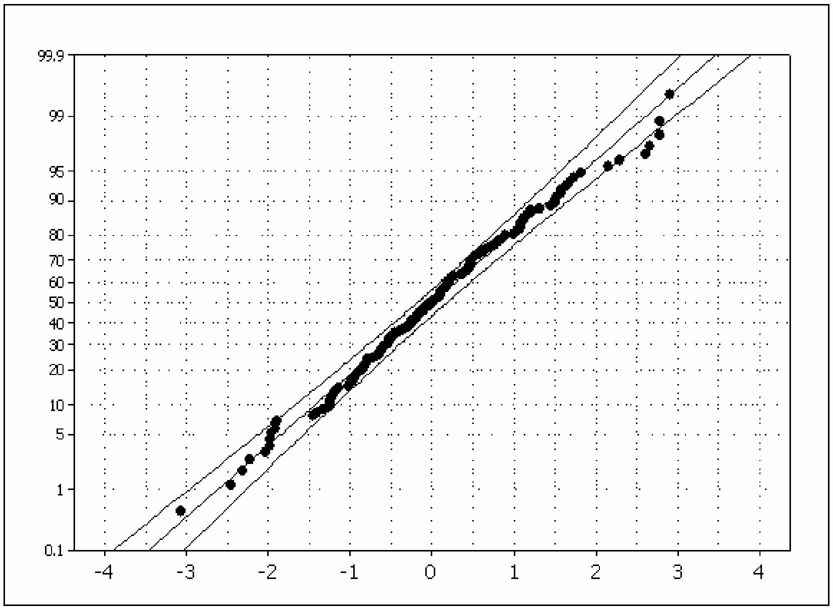


Figure 5. Normal probability plot of Pearson residuals from one realization of *GAR(1)* with m = 150, μ = 4, θ = -0.6.

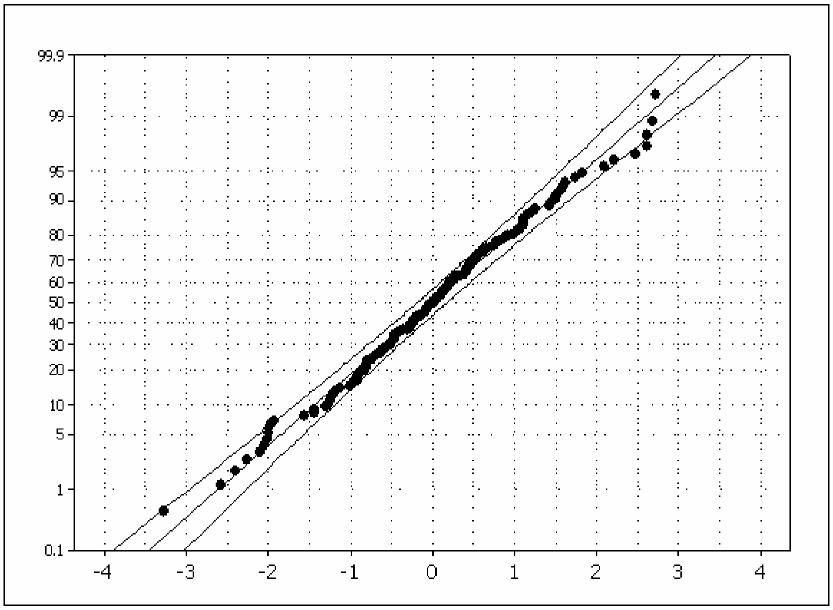


Figure 6. Normal probability plot of randomized residuals from one realization of *GAR(1)* with m = 150, μ = 4, θ = -0.6.

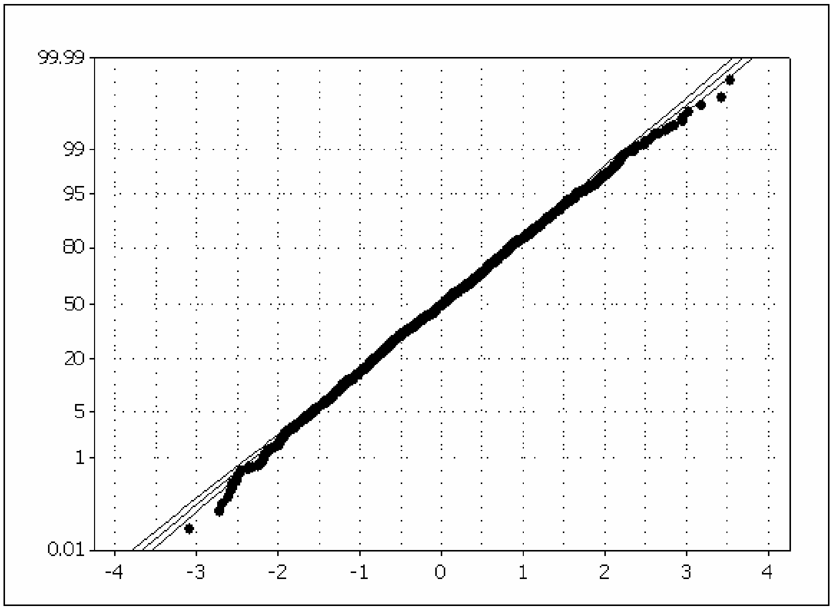


Figure 7. Normal probability plot of Pearson residuals at $t = 20$ from 2000 realizations of *GAR(1)* with $m = 150$, $\mu = 4$, $\theta = -0.6$.

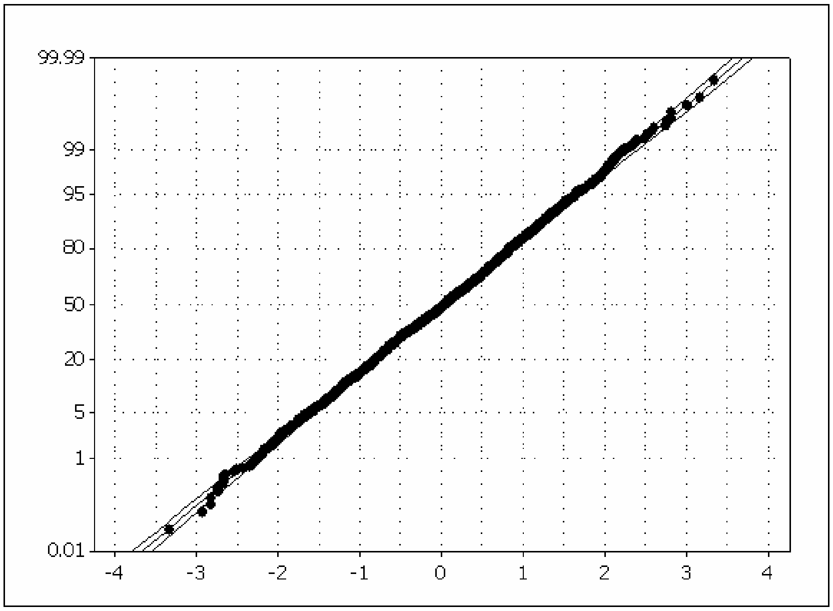


Figure 8. Normal probability plot of randomized residuals at $t = 20$ from 2000 realizations of *GAR(1)* with $m = 150$, $\mu = 4$, $\theta = -0.6$.

Table 4 presents results on the distribution of the residuals in relation to the 5% and 1% critical values of the standard normal distribution. Binomial standard errors of these simulated exceedance probabilities with *n* = 2000 are about 0.5% for the 5% point and about 0.2% for the 1% point. There is a moderate tendency for the exceedance probabilities to be lower than the nominal values, which would lead to conservative tests based on the normal distribution. Fitting logistic regression models with factors μ, θ and the type of residual (Pearson or randomized) confirmed a difference between the exceedance probabilities of the two residuals for *m* = 50 at the 5% point (logistic regression coefficient for randomized versus Pearson = 0.154 with standard error 0.029) but not at the 1% point (-0.067, s.e. 0.067).

Table 4. Simulated exceedance probabilities (x1000) of normal 5% and 1% critical values of a randomly selected Pearson residual (P) and randomized residual (R). Each entry is based on 2,000 simulations of the GAR(1) model.

		Length 50				Length 150			
		5%		1%		5%		1%	
μ	θ	P	R	P	R	P	R	P	R
2	-0.6	430	445	90	90	475	485	135	80
	-0.3	430	440	60	80	425	470	100	85
	0	500	550	130	130	510	470	65	95
	0.3	460	510	105	95	430	500	90	80
	0.6	370	435	65	65	435	465	80	75
4	-0.8	435	420	110	105	550	565	110	110
	-0.6	415	410	80	75	480	510	75	95
	-0.3	490	495	85	105	445	455	85	100
	0	450	485	115	85	445	435	90	65
	0.3	440	460	110	100	570	560	70	90
	0.6	465	465	130	125	525	505	95	90
	0.8	440	420	85	85	460	485	85	75
6	-0.8	490	485	120	100	505	510	80	90
	-0.6	435	420	80	70	490	495	100	100
	-0.3	415	410	35	40	485	490	95	105
	0	500	505	125	120	515	525	105	115
	0.3	520	545	140	140	545	550	110	130
	0.6	460	470	55	55	505	515	105	115
	0.8	500	490	110	110	535	525	130	135

Conclusion

These results suggest that the *GAR(1)* model without covariates is numerically well behaved, except in the case of the combination of small μ and large $|\theta|$. Restricting the study to *GAR(1)* is not unreasonable, because this is likely to be the most important practical case. According to Greene (2000), "The first-order autoregression has withstood the test of time and experimentation as a reasonable model for underlying processes that probably, in truth, are impenetrably complex" (p.531).

The results also show that the Pearson residuals do not depart from normality as much as might have been expected. However, the randomized residuals are available for use, if preferred, and their distribution seems to be very close to normal. Sometimes there are objections to using randomization within statistical analysis but, as Dunn and Smyth (1996) pointed out, these do not apply when the aim is to look at the overall pattern of residuals, which is what happens when all the residuals within one run are being considered. On the other hand, the random element does become an issue when specific residuals are being examined. This is the case when, for instance, extreme values are under consideration as potential outliers. Although the simulation results show that the normal distribution applies quite well even at the 1% points, outlier detection may be based on much more extreme values than this (for example, when Bonferroni adjustments are used). Figure 4 compared to Figure 3 and to a lesser extent, Figure 8 compared to Figure 7, show that the randomized residuals would work far better than the Pearson residuals for this purpose. One way of obtaining the advantage of adjusting the residuals, but avoiding randomization, is as follows. Instead of definition (6), define adjusted residuals by

$$r_t^* = \Phi^{-1}\left(u_t^*\right)$$

where u_t^* is the mid-point of the interval $\left[F\left(y_t - 1;\, \hat{\mu}_t\right),\, F\left(y_t;\, \hat{\mu}_t\right)\right]$. In other words, the random variable u_t in (6) is replaced by its expected value. The distribution of these adjusted residuals across series in the simulations was very close to the distribution of the randomized residuals shown in Figures 4 and 8.

One unsatisfactory feature of the model (2) or (4) is the necessity for introducing y_t^*. This is an artificial device to enable the series to restart from zero, which otherwise would be an absorbing state. As remarked above, the amount of information available on the parameter *c* is very small and it is preferred to ignore it entirely by dividing the series up into blocks. This is only an issue when μ is small, because otherwise the chances of reaching zero are negligible. On the other hand, this case may be the most interesting for the application of these models. It is noted that Benjamin, Rigby, and Stasinopoulos (2003) did not discuss this problem and in their example (which includes many zeroes) they appear simply to have used *c* = 0.1 without estimation. Kedem and Fokianos (2002) used examples without zeroes.

During the course of the investigations, the alternative model (5) was also examined. It was found that the likelihood surface tends to be very flat with respect to *c*. Because of this practical problem, but especially because of the dislike of the unrealistic device of adding a constant to every observation, this work has not been pursued and was not reported in this article. Another model, replacing both (4) and (5), could allow a random quantity (independent of other parts of the model and other time periods) to be added to each observation. This could be a much more satisfactory physical model of immigration from elsewhere than is offered by the existing proposals.

References

Benjamin, M. A., Rigby, R. A. & Stasinopoulos, M. D. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association, 98,* 214-223.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics, 8,* 93-115.

Dunn, P. K. & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics, 5*, 236-244.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, N.J.: Prentice Hall.

Kedem, B. & Fokianos, K. (2002). *Regression models for time series analysis*. New Jersey: John Wiley.

Li, W. K. (1994). Time series models based on generalized linear models: Some further results. *Biometrics, 50*, 506-511.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika, 75*, 621-629.

Zeger, S. L. & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics, 44*, 1019-1031.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 152-170

1538 – 9472/06/$95.00

Analysis Of Type-II Progressively Hybrid Censored Competing Risks Data

Debasis Kundu
Department of Mathematics
Indian Institute of Technology

Avijit Joarder
Reserve Bank of India

A Type-II progressively hybrid censoring scheme for competing risks data is introduced, where the experiment terminates at a pre-specified time. The likelihood inference of the unknown parameters is derived under the assumptions that the lifetime distributions of the different causes are independent and exponentially distributed. The maximum likelihood estimators of the unknown parameters are obtained in exact forms. Asymptotic confidence intervals and two bootstrap confidence intervals are also proposed. Bayes estimates and credible intervals of the unknown parameters are obtained under the assumption of gamma priors on the unknown parameters. Different methods have been compared using Monte Carlo simulations. One real data set has been analyzed for illustrative purposes.

Key words: Competing risk; maximum likelihood estimator; Type-I and Type-II censoring; Fisher information matrix; asymptotic distribution; bayesian inference; exponential distribution; gamma distribution; Type-II progressive censoring scheme.

Introduction

In medical studies or in reliability analysis, it is quite common that more than one cause or risk factor may be present at the same time. In analyzing the competing risks model, it is assumed that data consists of a failure time and an indicator denoting the cause of failure. Several studies have been carried out under this assumption for both the parametric and the non-parametric set up. For the parametric set up it is assumed that different lifetime distributions follow some special parametric distribution, namely exponential, Weibull or gamma. Several authors, for example Berkson and Elveback (1960), Cox (1959), David and Moeschberger (1978) considered the problem from the parametric point of view. In the non-parametric set up, no specific lifetime distribution is assumed. Kaplan and Meier (1958), Efron (1967) and Peterson (1991) analyzed the non-parametric version of this model.

The two most common censoring schemes, namely Type-I and Type-II censoring schemes, are widely used in practice. Briefly, they can be described as follows. Consider n items are under observations in a particular experiment. In the conventional Type-I censoring scheme, the experiment continues up to a pre-specified time T. On the other hand, the conventional Type-II censoring scheme requires the experiment to continue until a pre-specified number of failures m ≤ n occurs. In this scenario, only the smallest lifetimes are observed. The mixture of Type-I and Type-II censoring schemes is known as the hybrid censoring scheme. This hybrid censoring scheme was first introduced by Epstein (1954; 1960). But, recently it becomes quite popular in the reliability and life-testing experiments. See for example the work of Chen and Bhattacharya (1988), Childs, Chandrasekhar, Balakrishnan, and Kundu (2003), Draper and Guttman (1987),

Debasis Kundu is Professor of Statistics. His research interests include Statistical Signal Processing, Reliability Analysis, Statistical Computing and Competing Risks Models. Email him at kundu@iitk.ac.in. Avijit Joarder is Research Officer in Reserve Bank of India. His areas of interest are reliability, survival analysis and numerical analysis. The views in this article are his personal views and not those of the Reserve Bank of India.

Fairbanks, Madasan and Dykstra (1982), Gupta and Kundu (1998), and Jeong, Park and Yum (1996).

One of the drawbacks of the conventional Type-I, Type-II, or hybrid censoring schemes is that they do not allow for removal of units at points other than the terminal point of the experiment. When the items are highly reliable it might be necessary to know the causes for which the items are failed and also necessary to remove items in between the experiment (at the time of each failure) for efficient estimation of the parameters. Because of this, one censoring scheme known as progressive censoring scheme under competing risks becomes very popular for the last few years. It can be described as follows: Consider n items in a study and assume that there is K causes of failure, which are known. Suppose m < n is fixed before the experiment. Moreover, m other integers, $R_1, \ldots, R_m$ are also fixed before so that $R_1 + \ldots + R_m + m = n$. At the time of the first failure $X_{1:m:n}$, R_1 of the remaining units are randomly removed. Similarly, at the time of the second failure $X_{2:m:n}$, R_2 of the remaining units are randomly removed and so on. Finally, at the time of the m^th^ failure $X_{m:m:n}$, the rest of the R_m units are removed. It is also known that the first failure takes place due to cause δ_1, similarly the second failure takes place due to cause δ_2 and so on, finally the m^th^ failure takes place due to cause δ_m. For an exhaustive list of references and further details on Type-II progressive censoring, the readers may refer to the book by Balakrishnan and Aggarwala (2000).

In this article, a Type-II progressively hybrid censoring scheme under competing risk is introduced. As the name suggests, it is a mixture of Type-II progressive and hybrid censoring schemes under the competing risk data. In this new censoring scheme, the likelihood inference of the unknown parameters is obtained, under the assumptions that the lifetime distributions of the different causes are independent identically distributed (i.i.d.) exponential random variables. It is observed that the maximum likelihood estimators of the unknown parameters always exist and one obtains the explicit form of the maximum likelihood estimators (MLEs) of the unknown parameters. One also obtains the asymptotic confidence intervals and proposed two bootstrap confidence intervals. Bayes estimates and credible intervals are also obtained under the assumption of the gamma priors on the unknown parameters. Different methods are compared using Monte Carlo simulations and for illustrative purposes, one real data set is analyzed.

Model Description and Notation

Suppose n identical items are put on a test and the lifetime distributions of the n items are denoted by $X_1, \ldots, X_n$. The integer $m < n$ is pre-fixed and also $R_1, \ldots, R_m$ are m pre-fixed integers satisfying $R_1 + \ldots + R_m + m = n$. T is a pre-fixed time point. At the time of first failure R_1 of the remaining units are randomly removed. Similarly at the time of the second failure R_2 of the remaining units are removed and so on. If the m^th^ failure occurs before the time point T, the experiment stops at the time point $X_{m:m:n}$. On the other hand, suppose the m^th^ failure does not occur before time point T and only J failures occur before the time point T, where $0 \le J < m$, then at the time point T all the remaining R_J^* units are removed and the experiment terminates at the time point T. Note that $R_J^* = n - (R_1 + \ldots + R_J) - J$. The two cases are denoted as Case I and Case II respectively and this censoring scheme is referred to as the Type-II progressively hybrid censoring scheme under competing risk data. In the presence of Type-II progressively hybrid censoring scheme under competing risks data, the following is a type of observation:

Case I: $\{(X_{1:m:n}, \delta_1, R_1), \ldots, (X_{m:m:n}, \delta_m, R_m)\}$; if $X_{m:m:n} < T$, or Case II: $\{(X_{1:m:n}, \delta_1, R_1), \ldots, (X_{J:m:n}, \delta_J, R_J), (T, R_J^*)\}$; if $X_{J:m:n} < T < X_{J+1:m:n}$.

Note that for Case II, $X_{J:m:n} < T < X_{J+1:m:n} < \ldots < X_{m:m:n}$ and $X_{J+1:m:n} < \ldots < X_{m:m:n}$ are not observed.

The conventional Type-I progressive censoring scheme needs the pre-specification of $R_1, \ldots, R_m$ and also $T_1, \ldots, T_m$, see Cohen (1963; 1966) for details. The choices of $T_1, \ldots, T_m$ are not trivial. For the conventional Type-II progressive censoring scheme the experimental time is unbounded. In the proposed censoring scheme, the choice of T depends upon how much maximum experimental time the experimenter can afford to spend. Moreover, the experimental time is bounded.

Without loss of generality, it is assumed that there are only two independent causes of failure i.e. K = 2. It may be extended to the case of K > 2. Before progressing further, the following notations are introduced/ reviewed:

X_{ji} : lifetime of the ith individual under cause j; for j = 1, 2 and i = 1, . . . , n

$X_{i:m:n}$: ith observed failure time; i = 1, . . . ,m

f(.) : probability density function (PDF) of X_i

F(.) : cumulative distribution function (CDF) of X_i

$F_j(.)$: cumulative distribution function (CDF) of X_{ji}

m_1 : the number of failures observed before termination due to cause 1 for Case I

m_2 : the number of failures observed before termination due to cause 2 for Case I

m : total number of failures observed before termination for Case I; *i.e.* $m = m_1 + m_2$

J_1 : the number of failures observed before termination due to cause 1 for Case II

J_2 : the number of failures observed before termination due to cause 2 for Case II

J : total number of failures observed before termination for Case II; *i.e.* $J = J_1 + J_2$

D_1 : the number of failures due to cause 1, *i.e.* $D_1 = m_1$ for Case I and $D_1 = J_1$ for Case II

D_2 : the number of failures due to cause 2, *i.e.* $D_2 = m_2$ for Case I and $D_2 = J_2$ for Case II

D : total number of failures, *i.e.* $D = m = m_1 + m_2$ for Case I and $D = J = J_1 + J_2$ for Case II

R_i : the number of units removed at the time of ith failure; $R_i \geq 0$

R_J^* : the number of remaining units left at the time point T for Case II

δ_i : indicator variable denoting the cause of failure of the ith individual

e(λ) : exponential random variable with PDF $\lambda e^{-\lambda x}$

gamma(α, λ) : gamma random variable with PDF $\frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$

It is assumed that $(X_{1i,} X_{2i})$; i = 1, . . ., n are n *i.i.d.* exponential random variables. Further, X_{1i} and X_{2i} are independent for all i = 1, . . ., n and $X_i = \min(X_{1i}, X_{2i})$. Now, the MLEs of the unknown parameters are provided when X_{ji}'s (for I = 1, . . ., n) are *i.i.d.* exp(λ_j), for j= 1, 2.

Maximum Likelihood Estimator

Based on the observations as discussed in the previous subsection, the log-likelihood function (without the constant term) can be written as;

$$L(\lambda_1, \lambda_2) = D_1 \ln \lambda_1 + D_2 \ln \lambda_2 - (\lambda_1 + \lambda_2)W, \quad (1)$$

where

$$D_1 = m_1,\ D_2 = m_2,\ W = \sum_{i=1}^{m} (1 + R_i) x_{i:m:n}$$

for Case I and

$$D_1 = J_1,\ D_2 = J_2,\ W = \sum_{i=1}^{J} (1 + R_i) x_{i:m:n} + TR_J^*$$

for Case II. From (1), it is clear that the MLEs of λ_1 and λ_2 always exists and they are

$$\hat{\lambda}_1 = \frac{D_1}{W} \qquad \text{and} \qquad \hat{\lambda}_2 = \frac{D_2}{W}. \qquad (2)$$

It is not possible to obtain the exact distribution of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ because of the complicated nature of the conditional distributions of $X_{1:m:n}$, . . ., $X_{m:m:n}$ given $X_{m:m:n} < T$. Interestingly, the distribution of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the mixture of discrete and continuous distributions. They have positive masses at the point 0 and have the bounded supports. Since, the exact distributions of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are not known, the exact confidence intervals also cannot be obtained.

Confidence Intervals

In this section, three different confidence intervals are proposed. One is based on the asymptotic distribution of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ and two different bootstrap confidence intervals.

Asymptotic Confidence Interval

In this section, we present the Fisher Information matrix of λ_1 and λ_2. Let I(λ_1, λ_2) = (I$_{ij}$(λ_1, λ_2)); i, j =1, 2, denote the Fisher Information matrix of the parameters λ_1 and λ_2, where

$$I_{ij}(\lambda_1,\lambda_2) = -E\left[\frac{\partial^2 L(\lambda_1,\lambda_2)}{\partial\lambda_i\partial\lambda_j}\right] \qquad (3)$$

From (1) it follows that

$$I_{11}(\lambda_1,\lambda_2) = \frac{E(D_1)}{\lambda_1^{\ 2}},$$

$$I_{12}(\lambda_1,\lambda_2) = I_{21}(\lambda_1,\lambda_2) = 0$$

and

$$I_{22}(\lambda_1,\lambda_2) = \frac{E(D_2)}{\lambda_2^{\ 2}}.$$

Simple calculation shows that

$$E(D_1) = \sum_{i=1}^{m_1} P(X_{i:m:n} < T)$$

and

$$E(D_2) = \sum_{i=1}^{m_2} P(X_{i:m:n} < T).$$

It is not easy to compute P($X_{i:m:n}$ < T) for general i, because $X_{i:m:n}$ is a sum of i independent, but not identically distributed exponential random variables. Therefore, for $D_1 > 0$ and $D_2 > 0$, the following approximate 100(1-α)% confidence interval for λ_1 and λ_2 are proposed,

$$\hat{\lambda}_1 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\lambda}_1^{\ 2}}{D_1}}$$

and

$$\hat{\lambda}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\lambda}_2^{\ 2}}{D_2}} \qquad (4)$$

respectively.

Bootstrap Confidence Intervals

In this subsection, two confidence intervals based on the bootstrapping are proposed. The two bootstrap methods that are widely used in practice are:

(1) The percentile bootstrap (Boot-p) proposed by Efron (1982), and

(2) The bootstrap-t method (Boot-t) proposed by Hall (1988).

It is observed that in this type of situations (Kundu, Kannan, & Balakrishnan, 2004), the non-parametric bootstrap method does not work well. Hence, the following two parametric bootstrap confidence intervals for λ_1 and λ_2 are proposed. The procedure is illustrated for the parameter λ_1. For the other parameter (λ_2), a confidence interval may be constructed in an analogous manner.

Boot-p Method

1. Estimate $\hat{\lambda}_1$ and $\hat{\lambda}_2$ from the sample using (2).
2. Generate a bootstrap sample $\{X^*_{1:m:n},\ldots,X^*_{D^*:m:n}\}$, using $\hat{\lambda}_1$ and $\hat{\lambda}_2$, $R_1, \ldots, R_m$ and T. Obtain the bootstrap estimate of λ_1 say, $\hat{\lambda}_1^*$ using the bootstrap sample.
3. Repeat Step 2 NBOOT times.
4. Let $\hat{CDF}(x) = P(\hat{\lambda}_1^* \le x)$, be the cumulative distribution function of $\hat{\lambda}_1^*$. Define $\hat{\lambda}_{1Boot-p}(x) = \hat{CDF}^{-1}(x)$ for a given x. The approximate 100(1-α)% confidence interval for λ_1 is given by:

$$\left(\hat{\lambda}_{1Boot-p}\left(\frac{\alpha}{2}\right), \hat{\lambda}_{1Boot-p}\left(1-\frac{\alpha}{2}\right)\right).$$

Boot-t Method

1. Estimate $\hat{\lambda}_1$ and $\hat{\lambda}_2$ from the sample using (2) as before.
2. Generate a bootstrap sample $\{X^*_{1:m:n},\ldots,X^*_{D^*:m:n}\}$, using $\hat{\lambda}_1$ and $\hat{\lambda}_2$, $R_1; \ldots; R_m$ and T. Also compute $\hat{V}(\hat{\lambda}_1^*) = \frac{\hat{\lambda}_1^{*2}}{D_1^*}$ for $D_1^* > 0$.
3. Determine the T_1^* statistic

$$T_1^* = \frac{(\hat{\lambda}_1^* - \hat{\lambda}_1)}{\sqrt{\hat{V}(\hat{\lambda}_1^*)}}$$

4. Repeat Steps 2 - 3 NBOOT times.
5. Let $\hat{CDF}(x) = P(T_1^* \le x)$, be the cumulative distribution function of T_1^*. For a given x, define $\hat{\lambda}_{1Boot-t}(x) = \hat{\lambda}_1 + \sqrt{\hat{V}(\hat{\lambda}_1^*)}\,\hat{CDF}^{-1}(x)$. The approximate 100(1-$\alpha$)% confidence interval for λ_1 is given by

$$\left(\hat{\lambda}_{1Boot-t}\left(\frac{\alpha}{2}\right), \hat{\lambda}_{1Boot-t}\left(1-\frac{\alpha}{2}\right)\right).$$

Bayesian Analysis

In this section, the problem is approached from the Bayesian point of view. In the context of exponential lifetimes, λ_1 and λ_2 may be reasonably modelled by the gamma priors. It is assumed that λ_1 and λ_2 are independently distributed as *gamma* (a_1, b_1) and *gamma* (a_2, b_2) priors, respectively. The gamma parameters a_1, b_1, a_2 and b_2 are all assumed to be positive. When $a_1 = b_1 = 0$ ($a_2 = b_2 = 0$), one obtains the non-informative priors of λ_1 (λ_2). The posterior density of λ_1 and λ_2 based on the gamma priors is given by

$$l(\lambda_1, \lambda_2 | data) \propto \lambda_1^{D_1+a_1-1} \lambda_2^{D_2+a_2-1} e^{-\lambda_1(W+b_1)} e^{-\lambda_2(W+b_2)} \tag{5}$$

From (5), it is clear that the posterior density functions of λ_1 and λ_2, say l(λ_1|*data*) and l(λ_2|*data*), respectively, are independent. Further, l(λ_1|*data*) is the density function of a *gamma*(D_1 + a_1, W + b_1) random variable, and l(λ_2|*data*) is the density function of a *gamma*(D_2 + a_2, W + b_2) random variable.

Therefore, the Bayes estimates of λ_1 and λ_2 under squared error loss functions are

$$\hat{\lambda}_{1Bayes} = \frac{D_1 + a_1}{W + b_1}$$

and

$$\hat{\lambda}_{2Bayes} = \frac{D_2 + a_2}{W + b_2} \tag{6}$$

respectively. Interestingly, when the non-informative priors $a_1 = b_1 = a_2 = b_2 = 0$, the Bayes estimators coincide with the corresponding MLEs.

The credible intervals for λ_1 and λ_2 can be obtained using the posterior distributions of λ_1 and λ_2. Note that *a posteriori* $Z_1 = 2\lambda_1$ (W + b$_1$) and $Z_2 = 2\lambda_2$ (W + b$_2$) follow χ^2 distributions with $2(D_1 + a_1)$ and $2(D_2 + a_2)$ degrees of freedom respectively, provided both $2(D_1 + a_1)$ and $2(D_2 + a_2)$ are positive integers. Therefore, 100(1-α)% credible intervals for λ_1 and λ_2 are

$$\left[\frac{\chi^2_{2(D_1+a_1),1-\frac{\alpha}{2}}}{2(W+b_1)}, \frac{\chi^2_{2(D_1+a_1),\frac{\alpha}{2}}}{2(W+b_1)}\right]$$

and

$$\left[\frac{\chi^2_{2(D_2+a_2),1-\frac{\alpha}{2}}}{2(W+b_2)}, \frac{\chi^2_{2(D_2+a_2),\frac{\alpha}{2}}}{2(W+b_2)}\right] \tag{7}$$

respectively for $(D_1 + a_1) > 0$ and $(D_2 + a_2) > 0$. Here $\chi^2_{k,\frac{\alpha}{2}}$ and $\chi^2_{k,1-\frac{\alpha}{2}}$ denote the lower and upper $\frac{\alpha}{2}$-th percentile points of a χ^2 distribution with k degrees of freedom. Note that if $2(D_1 + a_1)$ and $2(D_2 + a_2)$ are not integer values, then gamma distribution can be used to construct the credible intervals. If no prior information is available, then non-informative priors can be used to compute the credible intervals for λ_1 and λ_2. Alternatively, using the suggestion of Congdon (2001), very small positive values of a_1, b_1, a_2 and b_2 can be used to construct the Bayes estimates or the corresponding credible intervals.

Numerical Results and Discussions

Since the performance of the different methods cannot be compared theoretically, Monte Carlo simulations are used to compare different methods for different parameter values and for different sampling schemes. The term different sampling schemes means for different sets of R_i's and for different T values. All the computations are performed using Pentium IV processor and using the random number generation algorithm RAN2 of Press, Flannery, Teukolsky, & Vetterling.(1991). All the programs are written in FORTRAN and they can be obtained from the authors on request.

Before progressing further, first a description of how the Type-II progressively hybrid censored competing risk data was generated for a given set n, m, $R_1, \ldots, R_m$ and T. The following transformation as suggested in Balakrishnan and Aggarwala (2000) is used.

$$\begin{aligned} Z_1 &= nX_{1:m:n} \\ Z_2 &= (n - R_1 - 1)(X_{2:m:n} - X_{1:m:n}) \\ &\vdots \\ Z_m &= (n - R_1 - \ldots - R_{m-1} - m + 1)(X_{m:m:n} - X_{m-1:m:n}). \end{aligned} \tag{8}$$

It is known that if X_i's are *i.i.d. exp*($\lambda_1 + \lambda_2$), then the spacings Z_i's are also *i.i.d. exp*($\lambda_1 + \lambda_2$) random variables. From (8) it follows that

$$\begin{aligned} X_{1:m:n} &= \frac{1}{n} Z_1 \\ X_{2:m:n} &= \frac{1}{n - R_1 - 1} Z_2 + \frac{1}{n} Z_1 \\ &\vdots \\ X_{m:m:n} &= \frac{1}{n - R_1 - \ldots - R_{m-1} - m + 1} Z_m + \ldots + \frac{1}{n} Z_1. \end{aligned} \tag{9}$$

Using (9), Type-II progressively hybrid censored competing risk data can be easily generated as follows. For a given n, m, $R_1,\ldots,R_m$, $X_{1:m:n},\ldots,X_{m:m:n}$ is generated using (9). Again using the random number generation algorithm RAN2 of Press *et al.* (1991), a new random variable U(i), for i = 1…m is generated. Now if U(i) < $\frac{\lambda_1}{\lambda_1+\lambda_2}$, then assign δ_i = 1 otherwise, δ_i = 2. If $X_{m:m:n}$ < T. Then, one has Case I and the corresponding sample is $\{(X_{1:m:n},\delta_1,R_1),\ldots,(X_{m:m:n},\delta_m,R_m)\}$otherwise, one has Case II and J, such that $X_{J:m:n} < T < X_{J+1:m:n}$ is found. The corresponding sample is$\{(X_{1:m:n},\delta_1,R_1),\ldots,(X_{m:m:n},\delta_m,R_m),(T,R^*{}_J)\}$, where R^*_J is same as defined before.

Different n, m, T, λ_1, λ_2 and R_i's are considered. In all of the simulation experiments, λ_1 = 1.0 and λ_2 = 0.8 is taken. The following are taken n = 15, 25, 50, 100, m = 5, 10, 15, T = 0.25, 0.50, 1.00, 2.00 and three different sampling schemes. Scheme 1: $R_1 = \ldots = R_{m-1} = 0$ and R_m = n - m. Scheme 2: R_1 = n - m and $R_1 = \ldots = R_m = 0$. Scheme 3: $R_1 = \ldots = R_{m-1} = 1$ and R_m = n -2m + 1. For each case, the MLEs and the 95% confidence intervals of $\lambda_1$and λ_2 are computed using all three of the proposed methods. For comparison purposes, the 95% credible intervals are computed using non-informative prior. The process is replicated 1000 times in each case and the average bias, mean squared errors, and the coverage percentages are reported. The results are reported in Tables 1 - 9.

Some of the important observations are as follows. For fixed n as m increases the biases and MSEs of both λ_1 and λ_2 decrease for all cases as expected. But, interestingly for fixed m as n increases the biases increase and the MSEs decrease for both λ_1 andλ_2. This phenomenon is quite counter intuitive and a proper explanation cannot be found for this. Now, comparing different confidence intervals in terms of their average lengths and coverage percentages, it is observed that the MLEs, BOOT-T confidence intervals and Bayes credible intervals behave quite satisfactory unless the T is very small.

Otherwise, most of the cases of these three confidence intervals maintain the nominal coverage probabilities. Since BOOT-T method is involved numerically and the confidence intervals based on the asymptotic distributions are slightly larger than the Bayes credible intervals, it is recommended to use the Bayes credible intervals for all cases. Among the different schemes, it is observed that scheme 1 produces the smallest confidence intervals, followed by scheme 3 and scheme 2.

Data Analysis

In this section, one real-life dataset originally analyzed by Hoel (1972) is considered. The data arose from a laboratory experiment in which male mice received a radiation dose of 300 roentgens at 5 to 6 weeks of age. The cause of death for each mouse was determined by autopsy to be thymic lymphoma, reticulum cell sarcoma, or other causes. For the purpose of analysis, reticulum cell sarcoma is considered as cause 1 and the other causes of death are combined as cause 2. There were n = 77 observations in the data. A progressively type-II censored sample was generated from the original measurements.

Table 1: n = 15, m = 5[*].

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.2406 (1.2953)	0.2834 (1.2330)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1422 (0.6589)	0.1754 (0.6266)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.8876 (86.4)	2.9185 (93.3)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4473 (90.5)	2.4790 (89.6)	2.4801 (89.6)	2.4801 (89.6)
1	Boot-P	λ_1	4.0095 (88.3)	4.0829 (91.1)	4.0721 (91.6)	4.0717 (91.6)
		λ_2	3.2510 (87.0)	3.3224 (89.1)	3.3175 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.6389 (87.7)	2.8758 (90.7)	2.9050 (90.6)	2.9055 (90.6)
		λ_2	2.1035 (89.8)	2.3166 (88.7)	2.3436 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.7977 (93.1)	2.8322 (93.8)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3545 (88.9)	2.3885 (91.6)	2.3895 (91.6)	2.3895 (91.6)
		λ_1	0.2280 (1.7153)	0.2247 (1.3883)	0.2417 (1.2802)	0.2759 (1.2423)
		λ_2	0.1689 (1.0298)	0.1461 (0.7663)	0.1475 (0.6577)	0.1706 (0.6320)
	MLE	λ_1	3.6133 (79.0)	3.1929 (88.3)	2.9571 (90.7)	2.9142 (92.8)
		λ_2	3.0330 (69.5)	2.6902 (81.5)	2.5017 (87.5)	2.4762 (89.2)
2	Boot-P	λ_1	4.1914 (77.3)	4.0090 (85.5)	4.0136 (90.7)	4.0654 (89.9)
		λ_2	3.3645 (67.7)	3.2375 (79.9)	3.2395 (86.2)	3.3093 (88.9)
	Boot-T	λ_1	3.3581 (78.7)	2.9655 (87.4)	2.8422 (91.3)	2.8636 (90.8)
		λ_2	2.6215 (69.4)	2.3683 (80.9)	2.2597 (88.1)	2.3070 (89.0)
	Bayes	λ_1	3.4450 (77.3)	3.0707 (87.1)	2.8612 (92.9)	2.8273 (93.6)
		λ_2	2.8805 (67.8)	2.5721 (80.6)	2.4046 (88.0)	2.3851 (91.0)
		λ_1	0.2199 (1.3079)	0.2804 (1.2382)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1269 (0.6734)	0.1725 (0.6300)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.9090 (89.5)	2.9144 (92.6)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4540 (87.9)	2.4755 (89.3)	2.4801 (89.6)	2.4801 (89.6)
3	Boot-P	λ_1	3.9577 (89.2)	4.0778 (90.5)	4.0734 (91.6)	4.0717 (91.6)
		λ_2	3.2041 (85.2)	3.3183 (88.9)	3.3180 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.6347 (91.1)	2.8461 (90.7)	2.9038 (90.6)	2.9055 (90.6)
		λ_2	2.0913 (88.2)	2.2907 (88.6)	2.3413 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8142 (92.0)	2.8282 (93.7)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3580 (86.2)	2.3848 (91.1)	2.3895 (91.6)	2.3895 (91.6)

[*] In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 2: n = 25, m = 5*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.2825 (1.2347)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1741 (0.6284)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.9170 (93.1)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4770 (89.6)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
1	Boot-P	λ_1	4.0845 (90.8)	4.0726 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3214 (89.3)	3.3178 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.8529 (90.8)	2.9056 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.2954 (88.9)	2.3428 (88.7)	2.3437 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8308 (93.6)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3864 (91.2)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)
		λ_1	0.2370 (1.6967)	0.2279 (1.3813)	0.2414 (1.2803)	0.2759 (1.2423)
		λ_2	0.1712 (1.0103)	0.1482 (0.7633)	0.1483 (0.6561)	0.1715 (0.6314)
	MLE	λ_1	3.6058 (80.1)	3.1899 (88.8)	2.9538 (90.9)	2.9139 (92.8)
		λ_2	3.0232 (70.7)	2.6895 (81.9)	2.5017 (87.7)	2.4777 (89.3)
2	Boot-P	λ_1	4.2070 (78.3)	4.0052 (85.3)	4.0114 (90.8)	4.0654 (90.0)
		λ_2	3.3690 (68.8)	3.2410 (79.5)	3.2438 (86.4)	3.3097 (88.9)
	Boot-T	λ_1	3.4596 (79.9)	2.9826 (87.5)	2.8495 (90.8)	2.8646 (90.7)
		λ_2	2.6999 (69.9)	2.3953 (81.5)	2.2670 (88.0)	2.3073 (89.0)
	Bayes	λ_1	3.4403 (78.2)	3.0685 (87.7)	2.8583 (93.0)	2.8271 (93.6)
		λ_2	2.8724 (69.2)	2.5718 (81.3)	2.4047 (88.2)	2.3866 (91.1)
		λ_1	0.2812 (1.2368)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1718 (0.6308)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.9159 (92.4)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4744 (89.3)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
3	Boot-P	λ_1	4.0860 (90.7)	4.0736 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3216 (89.1)	3.3181 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.8364 (90.4)	2.9047 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.2802 (88.8)	2.3412 (88.7)	2.3437 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8297 (94.2)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3838 (90.8)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 3: n = 25, m = 10*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.0812 (0.3105)	0.1225 (0.2790)	0.1225 (0.2789)	0.1225 (0.2789)
		λ_2	0.0560 (0.2404)	0.0882 (0.2188)	0.0891 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.8802 (90.8)	1.8411 (94.0)	1.8406 (93.9)	1.8406 (93.9)
		λ_2	1.6573 (92.5)	1.6259 (92.7)	1.6261 (92.7)	1.6261 (92.7)
1	Boot-P	λ_1	2.1524 (91.4)	2.1440 (94.0)	2.1319 (94.1)	2.1317 (94.1)
		λ_2	1.8623 (88.6)	1.8597 (91.8)	1.8537 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.7514 (92.6)	1.8218 (93.7)	1.8341 (93.7)	1.8340 (93.7)
		λ_2	1.5029 (89.7)	1.5810 (90.8)	1.5951 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.8460 (92.8)	1.8120 (94.3)	1.8116 (94.1)	1.8116 (94.1)
		λ_2	1.6194 (91.1)	1.5932 (93.6)	1.5935 (93.6)	1.5935 (93.6)
		λ_1	0.0753 (0.5199)	0.0778 (0.3620)	0.0984 (0.3136)	0.1181 (0.2821)
		λ_2	0.0400 (0.4258)	0.0497 (0.2902)	0.0733 (0.2355)	0.0828 (0.2208)
	MLE	λ_1	2.5991 (90.3)	2.1705 (91.5)	1.9260 (92.9)	1.8488 (93.7)
		λ_2	2.2059 (85.2)	1.8888 (87.7)	1.7022 (91.6)	1.6304 (92.7)
2	Boot-P	λ_1	2.7334 (91.7)	2.3661 (92.2)	2.1893 (93.5)	2.1398 (93.9)
		λ_2	2.2943 (85.3)	2.0360 (92.0)	1.8917 (89.8)	1.8541 (91.3)
	Boot-T	λ_1	2.4446 (91.5)	2.0895 (91.9)	1.8889 (93.4)	1.8255 (93.8)
		λ_2	2.0044 (85.7)	1.7540 (91.0)	1.6192 (89.9)	1.5852 (91.1)
	Bayes	λ_1	2.5100 (90.7)	2.1177 (92.9)	1.8908 (93.4)	1.8191 (94.4)
		λ_2	2.1189 (83.9)	1.8330 (92.0)	1.6633 (92.9)	1.5971 (93.4)
		λ_1	0.0752 (0.3272)	0.1142 (0.2855)	0.1226 (0.2788)	0.1225 (0.2789)
		λ_2	0.0445 (0.2500)	0.0823 (0.2222)	0.0890 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.9918 (90.5)	1.8449 (94.0)	1.8407 (93.9)	1.8406 (93.9)
		λ_2	1.7386 (88.3)	1.6301 (92.3)	1.6261 (92.7)	1.6261 (92.7)
3	Boot-P	λ_1	2.2036 (92.2)	2.1502 (93.5)	2.1335 (94.1)	2.1317 (94.1)
		λ_2	1.9051 (89.8)	1.8606 (91.3)	1.8547 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.8715 (92.3)	1.8015 (93.6)	1.8326 (93.7)	1.8340 (93.7)
		λ_2	1.5931 (89.6)	1.5596 (91.0)	1.5940 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.9504 (92.7)	1.8152 (94.0)	1.8117 (94.1)	1.8116 (94.1)
		λ_2	1.6939 (90.7)	1.5968 (93.7)	1.5935 (93.6)	1.5935 (93.6)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 4: n = 50, m = 5*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
1	MLE	λ_1	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
		λ_1	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
	Boot-P	λ_1	4.0723 (91.6)	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3176 (89.4)	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.9049 (90.6)	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.3430 (88.7)	2.3437 (88.7)	2.3438 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)
2	MLE	λ_1	0.2378 (1.6791)	0.2302 (1.3733)	0.2427 (1.2795)	0.2757 (1.2485)
		λ_2	0.1761 (1.0055)	0.1494 (0.7596)	0.1493 (0.6548)	0.1716 (0.6312)
		λ_1	3.5945 (80.7)	3.1875 (89.5)	2.9530 (90.8)	2.9136 (92.8)
		λ_2	3.0208 (71.5)	2.6866 (82.2)	2.5029 (87.8)	2.4777 (89.3)
	Boot-P	λ_1	4.2231 (78.9)	4.0181 (85.7)	4.0113 (90.4)	4.0653 (90.1)
		λ_2	3.3637 (69.2)	3.2376 (79.8)	3.2436 (86.2)	3.3096 (88.9)
	Boot-T	λ_1	3.4955 (80.4)	2.9977 (87.6)	2.8515 (90.9)	2.8656 (90.7)
		λ_2	2.7151 (70.4)	2.3951 (81.7)	2.2697 (87.8)	2.3087 (89.0)
	Bayes	λ_1	3.4304 (78.9)	3.0669 (88.0)	2.8577 (92.8)	2.8267 (93.6)
		λ_2	2.8714 (70.1)	2.5696 (81.4)	2.4060 (88.5)	2.3866 (91.0)
3	MLE	λ_1	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
		λ_1	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
	Boot-P	λ_1	4.0726 (91.6)	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3178 (89.4)	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.9056 (90.6)	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.3428 (88.7)	2.3437 (88.7)	2.3438 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 5: n = 50, m = 10[*].

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.1226 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)
		λ_2	0.0890 (0.2183)	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.8408 (93.9)	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)
		λ_2	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)
1	Boot-P	λ_1	2.1406 (94.0)	2.1318 (94.1)	2.1317 (94.1)	2.1317 (94.1)
		λ_2	1.8576 (91.7)	1.8536 (91.8)	1.8536 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.8280 (93.7)	1.8340 (93.7)	1.8340 (93.7)	1.8340 (93.7)
		λ_2	1.5886 (91.1)	1.5950 (91.2)	1.5950 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.8118 (94.1)	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)
		λ_2	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)
		λ_1	0.0812 (0.5127)	0.0794 (0.3626)	0.1002 (0.3127)	0.1183 (0.2816)
		λ_2	0.0405 (0.4190)	0.0510 (0.2876)	0.0733 (0.2343)	0.0831 (0.2204)
	MLE	λ_1	2.5875 (90.1)	2.1628 (91.3)	1.9254 (93.4)	1.8488 (93.6)
		λ_2	2.1918 (85.7)	1.8825 (87.8)	1.7004 (91.7)	1.6306 (92.9)
2	Boot-P	λ_1	2.7158 (92.1)	2.3613 (92.3)	2.1873 (93.3)	2.1396 (93.8)
		λ_2	2.3004 (86.0)	2.0385 (91.6)	1.8924 (90.2)	1.8550 (91.3)
	Boot-T	λ_1	2.4721 (91.7)	2.0908 (91.5)	1.8900 (93.3)	1.8256 (93.8)
		λ_2	2.0481 (86.1)	1.7653 (90.9)	1.6233 (90.3)	1.5857 (91.1)
	Bayes	λ_1	2.5003 (91.0)	2.1106 (92.4)	1.8904 (93.5)	1.8191 (94.5)
		λ_2	2.1061 (84.8)	1.8274 (91.9)	1.6616 (93.0)	1.5972 (93.6)
		λ_1	0.1225 (0.2790)	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)
		λ_2	0.0882 (0.2188)	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.8411 (94.0)	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)
		λ_2	1.6259 (92.7)	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)
3	Boot-P	λ_1	2.1440 (94.0)	2.1319 (94.1)	2.1317 (94.1)	2.1317 (94.1)
		λ_2	1.8597 (91.8)	1.8537 (91.8)	1.8536 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.8218 (93.7)	1.8341 (93.7)	1.8340 (93.7)	1.8340 (93.7)
		λ_2	1.5810 (90.8)	1.5951 (91.2)	1.5950 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.8120 (94.3)	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)
		λ_2	1.5932 (93.6)	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)

[*] In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 6: n = 50, m = 15*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.0800 (0.1570)	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)
		λ_2	0.0336 (0.1174)	0.0366 (0.1150)	0.0366 (0.1150)	0.0366 (0.1150)
	MLE	λ_1	1.4553 (93.5)	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)
		λ_2	1.2720 (93.1)	1.2687 (93.7)	1.2687 (93.7)	1.2687 (93.7)
1	Boot-P	λ_1	1.6128 (93.6)	1.5828 (94.3)	1.5826 (94.3)	1.5826 (94.3)
		λ_2	1.4223 (93.1)	1.4045 (93.5)	1.4043 (93.5)	1.4043 (93.5)
	Boot-T	λ_1	1.4274 (94.0)	1.4515 (93.9)	1.4516 (93.9)	1.4516 (93.9)
		λ_2	1.2578 (93.0)	1.2819 (93.5)	1.2817 (93.5)	1.2817 (93.5)
	Bayes	λ_1	1.4400 (94.0)	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)
		λ_2	1.2545 (95.9)	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)
		λ_1	0.0746 (0.3559)	0.0651 (0.2411)	0.0682 (0.1739)	0.0819 (0.1545)
		λ_2	0.0313 (0.2689)	0.0270 (0.1677)	0.0275 (0.1314)	0.0332 (0.1180)
	MLE	λ_1	2.1969 (87.6)	1.7837 (90.7)	1.5448 (93.3)	1.4626 (94.1)
		λ_2	1.8902 (90.7)	1.5599 (92.3)	1.3513 (92.6)	1.2771 (92.9)
2	Boot-P	λ_1	2.2113 (91.7)	1.8593 (94.5)	1.6663 (94.0)	1.5974 (94.7)
		λ_2	1.8917 (91.8)	1.6091 (92.0)	1.4683 (94.4)	1.4134 (93.4)
	Boot-T	λ_1	2.0680 (91.0)	1.7434 (94.6)	1.5346 (93.4)	1.4580 (93.9)
		λ_2	1.7138 (91.4)	1.4864 (91.5)	1.3445 (93.0)	1.2842 (93.3)
	Bayes	λ_1	2.1411 (93.0)	1.7534 (92.2)	1.5258 (93.6)	1.4471 (94.3)
		λ_2	1.8314 (92.3)	1.5262 (93.1)	1.3298 (94.4)	1.2594 (95.2)
		λ_1	0.0686 (0.1630)	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)
		λ_2	0.0241 (0.1216)	0.0365 (0.1151)	0.0366 (0.1150)	0.0366 (0.1150)
	MLE	λ_1	1.4702 (93.2)	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)
		λ_2	1.2846 (93.1)	1.2687 (93.6)	1.2687 (93.7)	1.2687 (93.7)
3	Boot-P	λ_1	1.6215 (93.1)	1.5844 (94.3)	1.5826 (94.3)	1.5826 (94.3)
		λ_2	1.4262 (93.3)	1.4056 (93.4)	1.4043 (93.5)	1.4043 (93.5)
	Boot-T	λ_1	1.4336 (94.1)	1.4499 (93.9)	1.4516 (93.9)	1.4516 (93.9)
		λ_2	1.2587 (93.3)	1.2813 (93.5)	1.2817 (93.5)	1.2817 (93.5)
	Bayes	λ_1	1.4539 (93.7)	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)
		λ_2	1.2660 (94.9)	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 7: n = 100, m = 5[*].

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
1		λ_1	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
	Boot-P	λ_1	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.3438 (88.7)	2.3438 (88.7)	2.3438 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)
2		λ_1	0.2398 (1.6732)	0.2317 (1.3679)	0.2428 (1.2792)	0.2759 (1.2422)
		λ_2	0.1783 (1.0011)	0.1500 (0.7576)	0.1512 (0.6542)	0.1715 (0.6313)
	MLE	λ_1	3.5902 (80.8)	3.1872 (89.8)	2.9520 (90.7)	2.9141 (92.7)
		λ_2	3.0201 (71.6)	2.6851 (82.3)	2.5047 (87.9)	2.4775 (89.3)
	Boot-P	λ_1	4.2216 (78.9)	4.0150 (85.8)	4.0098 (90.5)	4.0650 (90.1)
		λ_2	3.3769 (69.5)	3.2425 (79.8)	3.2461 (86.2)	3.3100 (88.9)
	Boot-T	λ_1	3.4957 (80.4)	2.9995 (87.4)	2.8521 (90.9)	2.8666 (90.7)
		λ_2	2.7357 (71.0)	2.4007 (81.6)	2.2715 (87.9)	2.3092 (89.0)
	Bayes	λ_1	3.4270 (78.9)	3.0669 (88.4)	2.8568 (92.8)	2.8272 (93.6)
		λ_2	2.8711 (70.6)	2.5683 (81.5)	2.4079 (88.5)	2.3865 (91.0)
3		λ_1	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)	0.2842 (1.2314)
		λ_2	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)	0.1759 (0.6258)
	MLE	λ_1	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)	2.9192 (93.4)
		λ_2	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)	2.4801 (89.6)
	Boot-P	λ_1	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)	4.0717 (91.6)
		λ_2	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)	3.3172 (89.4)
	Boot-T	λ_1	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)	2.9055 (90.6)
		λ_2	2.3437 (88.7)	2.3438 (88.7)	2.3438 (88.7)	2.3438 (88.7)
	Bayes	λ_1	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)	2.8331 (93.9)
		λ_2	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)	2.3895 (91.6)

[*] In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 8: n = 100, m = 10*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
		λ_1	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)
		λ_2	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)
		λ_2	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)
1	Boot-P	λ_1	2.1318 (94.1)	2.1317 (94.1)	2.1317 (94.1)	2.1317 (94.1)
		λ_2	1.8536 (91.8)	1.8536 (91.8)	1.8536 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.8340 (93.7)	1.8340 (93.7)	1.8340 (93.7)	1.8340 (93.7)
		λ_2	1.5950 (91.2)	1.5950 (91.2)	1.5950 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)
		λ_2	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)
		λ_1	0.0833 (0.5097)	0.0795 (0.3643)	0.1005 (0.3126)	0.1182 (0.2817)
		λ_2	0.0418 (0.4155)	0.0512 (0.2890)	0.0729 (0.2342)	0.0830 (0.2204)
	MLE	λ_1	2.5789 (90.0)	2.1578 (91.4)	1.9246 (93.5)	1.8485 (93.6)
		λ_2	2.1851 (86.0)	1.8791 (87.9)	1.6989 (91.7)	1.6303 (92.9)
2	Boot-P	λ_1	2.7055 (91.9)	2.3619 (92.4)	2.1864 (93.3)	2.1397 (93.9)
		λ_2	2.3012 (86.6)	2.0384 (91.4)	1.8924 (90.3)	1.8552 (91.3)
	Boot-T	λ_1	2.4757 (91.7)	2.0947 (91.7)	1.8898 (93.3)	1.8258 (93.9)
		λ_2	2.0653 (86.3)	1.7689 (90.7)	1.6233 (90.5)	1.5857 (91.1)
	Bayes	λ_1	2.4928 (91.4)	2.1060 (92.5)	1.8896 (93.7)	1.8189 (94.5)
		λ_2	2.1004 (85.2)	1.8243 (91.8)	1.6603 (93.0)	1.5970 (93.6)
		λ_1	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)	0.1225 (0.2789)
		λ_2	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)	0.0891 (0.2182)
	MLE	λ_1	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)	1.8406 (93.9)
		λ_2	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)	1.6261 (92.7)
3	Boot-P	λ_1	2.1318 (94.1)	2.1317 (94.1)	2.1317 (94.1)	2.1317 (94.1)
		λ_2	1.8536 (91.8)	1.8536 (91.8)	1.8536 (91.8)	1.8536 (91.8)
	Boot-T	λ_1	1.8340 (93.7)	1.8340 (93.7)	1.8340 (93.7)	1.8340 (93.7)
		λ_2	1.5950 (91.2)	1.5950 (91.2)	1.5950 (91.2)	1.5950 (91.2)
	Bayes	λ_1	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)	1.8116 (94.1)
		λ_2	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)	1.5935 (93.6)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Table 9: n = 100, m = 15*.

Scheme	Methods		T = 0.25	T = 0.50	T = 1.00	T = 2.00
1		λ_1	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)
		λ_2	0.0366 (0.1150)	0.0366 (0.1150)	0.0366 (0.1150)	0.0366 (0.1150)
	MLE	λ_1	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)
		λ_2	1.2687 (93.7)	1.2687 (93.7)	1.2687 (93.7)	1.2687 (93.7)
	Boot-P	λ_1	1.5826 (94.3)	1.5826 (94.3)	1.5826 (94.3)	1.5826 (94.3)
		λ_2	1.4044 (93.5)	1.4043 (93.5)	1.4043 (93.5)	1.4043 (93.5)
	Boot-T	λ_1	1.4516 (93.9)	1.4516 (93.9)	1.4516 (93.9)	1.4516 (93.9)
		λ_2	1.2818 (93.5)	1.2817 (93.5)	1.2817 (93.5)	1.2817 (93.5)
	Bayes	λ_1	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)
		λ_2	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)
2		λ_1	0.0739 (0.3503)	0.0675 (0.2395)	0.0678 (0.1735)	0.0819 (0.1545)
		λ_2	0.0343 (0.2643)	0.0264 (0.1671)	0.0275 (0.1315)	0.0332 (0.1180)
	MLE	λ_1	2.1841 (87.9)	1.7816 (90.9)	1.5434 (93.3)	1.4625 (94.2)
		λ_2	1.8860 (90.7)	1.5555 (92.0)	1.3503 (92.4)	1.2770 (92.9)
	Boot-P	λ_1	2.2098 (92.0)	1.8572 (94.6)	1.6646 (94.0)	1.5972 (94.7)
		λ_2	1.8977 (91.8)	1.6063 (92.6)	1.4677 (94.4)	1.4136 (93.4)
	Boot-T	λ_1	2.0764 (91.3)	1.7421 (94.2)	1.5339 (93.3)	1.4576 (93.9)
		λ_2	1.7271 (91.6)	1.4871 (91.7)	1.3446 (93.1)	1.2843 (93.3)
	Bayes	λ_1	2.1292 (92.6)	1.7515 (91.8)	1.5245 (93.7)	1.4469 (94.3)
		λ_2	1.8280 (92.5)	1.5221 (93.0)	1.3289 (94.4)	1.2593 (95.2)
3		λ_1	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)	0.0862 (0.1520)
		λ_2	0.0366 (0.1150)	0.0366 (0.1150)	0.0366 (0.1150)	0.0366 (0.1150)
	MLE	λ_1	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)	1.4530 (94.0)
		λ_2	1.2687 (93.7)	1.2687 (93.7)	1.2687 (93.7)	1.2687 (93.7)
	Boot-P	λ_1	1.5828 (94.3)	1.5826 (94.3)	1.5826 (94.3)	1.5826 (94.3)
		λ_2	1.4045 (93.5)	1.4043 (93.5)	1.4043 (93.5)	1.4043 (93.5)
	Boot-T	λ_1	1.4515 (93.9)	1.4516 (93.9)	1.4516 (93.9)	1.4516 (93.9)
		λ_2	1.2819 (93.5)	1.2817 (93.5)	1.2817 (93.5)	1.2817 (93.5)
	Bayes	λ_1	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)	1.4379 (94.4)
		λ_2	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)	1.2515 (94.8)

* In each cell, the first row of λ_1 and λ_2 represents the average biases and the corresponding mean squared errors are reported within brackets for the MLEs. The second, third, fourth and fifth rows of λ_1 and λ_2 represent the average 95% confidence lengths of asymptotic confidence intervals, Boot-p confidence intervals, Boot-t confidence intervals and the credible intervals with respect to the non-informative priors respectively. The corresponding coverage percentages are reported within brackets.

Example 1: In this case, n = 77 and m = 25, T = 700, $R_1 = R_2 = \ldots = R_{24} = 2$ and $R_{25} = 4$ are taken. Thus, the Type II progressively hybrid censored sample is:

(40, 2), (42, 2), (62, 2), (163, 2), (179,2), (206, 2), (222, 2), (228, 2), (252, 2), (259, 2), (318, 1), (385, 2), (407, 2), (420, 2), (462, 2), (507, 2), (517, 2), (524, 2), (525, 1), (528, 1), (536, 1), (605, 1), (612, 1), (620, 2), (621, 1).

In this case, $D_1 = 7$, $D_2 = 18$ and W = $\sum_{i=1}^{25} (1 + R_i) x_{i:m:n}$ = 28962. Therefore,

$$\hat{\lambda}_1 = \frac{7}{28962} = 2.41696 \times 10^{-4}$$

and

$$\hat{\lambda}_2 = \frac{18}{28962} = 6.21504 \times 10^{-4}.$$

The 95% asymptotic, Boot-P, Boot-t confidence intervals and also the 95% credible intervals of λ_1 and λ_2 are reported in Table 10.

It is clear that although all of them provided almost similar confidence/credible intervals, but Bayes credible intervals have the smallest lengths. Now, the data using T = 600 instead of T = 700 is generated, while m and R(i)'s are the same as before.

Example 2: In this case the progressively hybrid censored sample obtained as:

(40, 2), (42, 2), (62, 2), (163, 2), (179,2), (206, 2), (222, 2), (228, 2), (252, 2), (259, 2), (318, 1), (385, 2), (407, 2), (420, 2), (462, 2), (507, 2), (517, 2), (524, 2), (525, 1), (528, 1), (536, 1).

Here $D_1 = 4$, $D_2 = 17$ and W = $\sum_{i=1}^{21}(1+R_i)x_{i:m:n}$ = 20346. Therefore, the following is obtained:

$$\hat{\lambda}_1 = \frac{4}{28746} = 1.39150 \times 10^{-4}$$

and

$$\hat{\lambda}_2 = \frac{17}{28746} = 20.23809 \times 10^{-4}.$$

In this case, the 95% asymptotic, Boot-P, Boot-t confidence intervals and also the 95% credible intervals of λ_1 and λ_2 are reported in Table 11.

From Table 11, it is observed that T plays a major role for the estimation of λ's and for the construction of the corresponding confidence intervals. As T decreases, the lengths of the confidence/credible intervals for both the parameters are as expected. It is also important to note that Boot-p and Boot-t are the most affected due to T and the Bayes confidence intervals are the least affected. Therefore, Bayes confidence intervals are quite robust also with respect to T.

Conclusion

In this article, a new censoring scheme is discussed, namely the Type II progressively hybrid censoring scheme under competing risks data. Assuming that the lifetime distributions are exponentially distributed, one may obtain the maximum likelihood estimators of the unknown parameter and propose different confidence intervals using asymptotic distributions as well as using bootstrap methods. Bayesian estimates of the unknown parameters are also proposed and it is observed that the Bayes credible intervals with respect to non-informative prior work quite well in this case and it has several desirable properties. Although it is assumed that the lifetime distributions are exponential, most of the methods may be extended for other distributions also, such as the Weibull or gamma distribution.

Table 10.

Methods	λ_1	λ_2
Asymptotic	$\left(0.62645\times10^{-4},4.20747\times10^{-4}\right)$	$\left(3.34384\times10^{-4},9.08624\times10^{-4}\right)$
Boot-p	$\left(0.76099\times10^{-4},4.52108\times10^{-4}\right)$	$\left(3.47439\times10^{-4},10.52984\times10^{-4}\right)$
Boot-t	$\left(0.58039\times10^{-4},4.26943\times10^{-4}\right)$	$\left(2.71588\times10^{-4},9.46895\times10^{-4}\right)$
Credible	$\left(0.97174\times10^{-4},4.50918\times10^{-4}\right)$	$\left(3.60913\times10^{-4},9.31153\times10^{-4}\right)$

Table 11.

Methods	λ_1	λ_2
Asymptotic	$\left(0.02783\times10^{-4},2.75517\times10^{-4}\right)$	$\left(10.61752\times10^{-4},29.85867\times10^{-4}\right)$
Boot-p	$\left(0.00000\times10^{-4},3.02527\times10^{-4}\right)$	$\left(14.13159\times10^{-4},32.89348\times10^{-4}\right)$
Boot-t	$\left(0.00000\times10^{-4},3.63490\times10^{-4}\right)$	$\left(11.92432\times10^{-4},27.94359\times10^{-4}\right)$
Credible	$\left(0.37913\times10^{-4},3.04992\times10^{-4}\right)$	$\left(3.37047\times10^{-4},8.95152\times10^{-4}\right)$

References

Balakrishnan, N. & Aggarwala, R. (2000). *Progressive censoring: Theory, methods, and applications*. Boston: BirkhÄauser.

Berkson, J. & Elveback, L. (1960). Competing exponential risks with particular inference to the study of smoking lung cancer. *Journal of the American Statistical Association, 87*, 84-89.

Chen, S. M. & Bhattacharya, G. K. (1988). Exact confidence bounds for an exponential parameter hybrid censoring. *Communications in Statistics – Theory and Methods, 17*, 1858 - 1870.

Childs, A., Chandrasekhar, B., Balakrishnan, N., & Kundu, D. (2003). Exact likelihood inference based on type-I and type-II hybrid censored samples from the exponential distribution. *Annals of the Institute of Statistical Mathematics, 55*, 319-330.

Cohen, A. C. (1963). Progressively censored samples in life testing. *Technometrics, 5*, 327 - 329.

Cohen, A. C. (1966). Life-testing and early failure. *Technometrics, 8*, 539 - 549.

Congdon, P. (2001). *Bayesian statistical modeling*. New York: John Wiley.

Cox, D. R. (1959). The analysis of exponentially distributed lifetimes with two types of failures. *Journal of Royal Statistics, 21*(B), 411-421.

Draper, N. & Guttman, T. (1987). Bayesian analysis of hybrid life-test with exponential failure times. *Annals of the Institute of Statistical Mathematics*, *39*, 219-255.

David, H. A. & Moeschberger, M. L. (1978). *The theory of competing risks*. London: Griffin.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium Math. Stat. Prob*., 831 - 853.

Efron, B. (1982). *The jackknife, the Bootstrap and Other Re-sampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, *38*, SIAM, Philadelphia, PA.

Epstein B. (1954). Truncated life-tests in the exponential case. *Annals of Mathematical Statistics*, *25*, 555-564.

Epstein, B. (1960). Estimation from life-test data. *Technometrics*, *2*, 447 - 454.

Fairbanks, K., Madasan, R., & Dykstra, R. (1982). A confidence interval for an exponential parameter from hybrid life-test. *Journal of the American Statistical Association*, *77*, 137 - 140.

Gupta, R. D. & Kundu, D. (1998). Hybrid censoring schemes with exponential failure distribution, *Communications in Statistics - Theory and Methods*, *27*, 3065-3083.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, *16*, 927-953.

Hoel, D. G. (1972). A representation of mortality data by competing risks. *Biometrics*, *28*, 475-488.

Jeong, H. S., Park, J. I., & Yum, B. J. (1996). Development of (r;T) hybrid sampling plans for exponential lifetime distributions. *Journal of Applied Statistics*, *23*, 601-607.

Kundu, D., Kannan, N., & Balakrishnan, N. (2004). Analysis of progressively censored competing risks data. In N. Balakrishnan, and C. R. Rao (Eds.), *Handbook of Statistics*, *23*. NY: Elsevier.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observation. *Journal of the American Statistical Association*, *53*, 457-481.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1991). *Numerical recipes: The art of scientific computing*. Cambridge, U.K.: Cambridge University Press.

Peterson, A. P. (1977). Expressing the Kaplan-Meier estimator as a function of empirical survival functions. *Journal of the American Statistical Association*, *72*, 854-858.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 171-180

1538 – 9472/06/$95.00

Comparison Of Some Simple Estimators Of The Lognormal Parameters Based On Censored Samples

Ayman Baklizi Mohammed Al-Haj Ebrahem
Department of Statistics
Yarmouk University

Point estimation of the parameters of the lognormal distribution with censored data is considered. The often employed maximum likelihood estimator does not exist in closed form and iterative methods that require very good starting points are needed. In this article, some techniques of finding closed form estimators to this situation are presented and extended. An extensive simulation study is carried out to investigate and compare the performance of these techniques. The results show that some of them are highly efficient as compared with the maximum likelihood estimator.

Keywords: Modified maximum likelihood estimator, least squares estimators, lognormal distribution, mean squared error, Persson Rootzen estimators

Introduction

Let the random variable Y be normally distributed with mean μ and variance σ^2. Let $T = e^Y$, then T is said to have a lognormal distribution. The probability density function of T is given by (Lawless, 1982);

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right), \ 0 < t < \infty . \tag{1}$$

The many special features of the lognormal distribution together with its relation with the normal distribution have allowed it to be used as a model in various real life applications. It is used in analyzing biological data (Koch, 1966), and for analyzing data in workplace exposure to contaminants (Lyles & Kupper, 1996). It is also of importance in modeling lifetimes of products and individuals (Lawless, 1982). Various other motivations and applications of the lognormal distribution can be found in Johnson et al. (1994) and Schneider (1986).

In most life testing experiments, one is faced with censored data (Lawless, 1982) arising from either terminating the experiment at a certain prespecified time (Type 1 censoring) or when a predetermined number of failures occur (Type 2 censoring). Censoring is often employed because of time and cost considerations. However, complications do often arise in inference from censored data and usually likelihood based inference procedures are used. Assume that the data is Type 2 censored, whereby the following is observed: $t_{(1)}, \ldots, t_{(r)}$, $r \le n$. The likelihood function is given by

$$L(\mu,\sigma) = \left(\prod_{i=1}^{r} \frac{1}{\sigma t_{(i)}} \phi\left(\frac{\ln t_{(i)} - \mu}{\sigma}\right)\right)\left(Q\left(\frac{\ln t_{(r)} - \mu}{\sigma}\right)\right)^{n-r} \tag{2}$$

where $\phi(\ )$ and $Q(\ )$ are the probability density and the survival functions of the standard normal

Ayman Baklizi is an Associate Professor of Applied Statistics. His research interests are in accelerated life tests, censored data, nonparametric statistics, and simulation and resampling methods. Email him at baklizi1@yahoo.com. Mohammed Al-Haj is an Assistant Professor. His research interest is in reliability, accelerated life test and non-parametric regression models. Email: m_hassanb@hotmail.com

distribution. The likelihood function corresponding to Type 1 censoring is obtained by replacing $\ln t_{(r)}$ by $\ln t_0$, the censoring time under Type 1 censoring. The maximum likelihood estimator is obtained by finding $\hat{\mu}$ and $\hat{\sigma}$ that maximize the likelihood function. This is often done by equating the first partial derivatives of the log-likelihood function to zero and solving for μ and σ simultaneously by applying an iterative numerical procedure for root finding like the Newton-Raphson method. However, this is problematic unless very good starting values are available (Lawless, 1982); the problem becomes serious when the proportion of censored observations is large, especially when the total sample size is relatively small to moderate. In such cases, alternatives to the maximum likelihood estimator are needed, either on their own or as initial approximations to the maximum likelihood estimators. The books of Lawless (1982), Schneider (1986) and Balakrishnan and Cohen (1991) survey much of the work in this area.

In this article, the performances of three techniques for point estimation of parameters in the case of censored data from a lognormal distribution will be extended, investigated, and compared. The first technique is based on finding the least squares estimator by regressing certain estimators of the linearized distribution function on a function of the observations themselves. This approach is used in Hossain and Howlader (1996) and Hossain and Zimmer (2003) for the parameters of the Weibull distribution. Their results showed that the estimators are a reasonable substitute for the maximum likelihood estimator in most situations.

The second technique is due to Perrson and Rootzen (1977) where they presented some modified likelihood function with Type 1 censored data whose maximizing point does not require iterative techniques. The last technique is based on expanding certain terms in the first derivatives of the log-likelihood function in an appropriate Taylor series to get a new system of likelihood equations whose solution exists in closed form. This last approach was studied for Type 2 censored data. An account of this work can be found in Balakrishnan and Cohen (1991). Recently Al-Haj Ebarahem and Baklizi (2005) used the first and the last techniques to estimate the parameters of the Log-Logistic distribution based on complete and censored samples

Least Squares Estimators

The distribution function of the lognormal random variable is given by

$$F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right).$$

Linearization of this distribution function gives $\Phi^{-1}(F(t)) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}\ln t$.which is a linear regression model between $\Phi^{-1}(F(t))$ and $\ln t$. Let $T_{(1)},\ldots,T_{(r)}$ be the observed censored sample and let S_i be an estimate of $\Phi^{-1}(F(T_{(i)}))$, then the least squares estimators of $b = \frac{1}{\sigma}$ and $a = -\frac{\mu}{\sigma}$ are given respectively by

$$\hat{b} = \frac{\sum_{i=1}^{r} S_i \ln T_i - r\overline{\ln T}\,\overline{S}}{\sum_{i=1}^{r} (\ln T_i)^2 - r\left(\overline{\ln T}\right)^2}$$

and

$$\hat{a} = \overline{S} - \hat{b}\overline{\ln T},$$

where

$$\overline{\ln T} = \sum_{i=1}^{r} \ln T_i \Big/ r$$

and

$$\overline{S} = \sum_{i=1}^{r} S_i \Big/ r.$$

An estimate of S_i, $i = 1, \ldots, r$ is now required. Two methods of estimation of $F\left(T_{(i)}\right)$ and hence S_i will be considered:

a) Let $\hat{F}\left(T_{(i)}\right) = 1 - R_{(i)}, i = 1, \ldots, r$

where

$$R_{(i)} = \frac{r_i}{r_i + 1} R_{(i-1)},\ R_{(0)} = 1$$

and

$$r_i = n - r_i' + 1$$

where r_i' is the rank of the i-th failure in the original sample. Hence, $S_i = \Phi^{-1}\left(1 - R_{(i)}\right)$. Substituting these values in $\hat{b}$ and $\hat{a}$, one obtains the estimators $\hat{\mu}_1$ and $\hat{\sigma}_1$.

b) Use $R_{(i)} = \dfrac{r_i - 0.5}{r_{i-1} - 0.5} R_{(i-1)}$. In this case the new least squares based estimators are based on $\hat{\mu}_2$ and $\hat{\sigma}_2$.

Approximate Maximum Likelihood Estimators

Let $T_{(1)} \le T_{(2)} \le \ldots \le T_{(r)}$ be a Type 2 censored sample consisted of the smallest r ordered observations obtained from the lognormal population with probability distribution function given by (1), the remaining $(n - r)$ observations being censored at $T_{(r)}$. Let $Y_i = \ln T_{(i)}, i = 1, \ldots, r$ be the corresponding order statistics from the normal distribution. The likelihood function of (μ, σ) is given by equation (2). The maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ are given as the solution to the following simultaneous system of nonlinear equations (Lawless, 1982);

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{r} (y_i - \mu) + \frac{(n-r)}{\sigma} \left(\frac{\phi\left(\frac{y_r - \mu}{\sigma}\right)}{Q\left(\frac{y_r - \mu}{\sigma}\right)} \right)$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{r}{\sigma}$$

$$+ \frac{1}{\sigma^3} \sum_{i=1}^{r} (y_i - \mu)^2 + \frac{(n-r)}{\sigma} \left(\frac{\frac{y_r - \mu}{\sigma} \phi\left(\frac{y_r - \mu}{\sigma}\right)}{Q\left(\frac{y_r - \mu}{\sigma}\right)} \right) \tag{3}$$

The likelihood equations corresponding to Type 1 censoring are obtained by replacing $y_r = \ln t_{(r)}$ by $y_0 = \ln t_0$, the censoring time under Type 1 censoring. As stated in the introduction, the system of equations (3) does not admit a closed form solution and a numerical method is needed to find the solution (the MLE). In the following two subsections, some modifications of these likelihood equations will be presented to obtain a closed form solution.

The Persson-Rootzen Approach

Consider the likelihood function (2) given by

$$L(\mu, \sigma) = \left(\prod_{i=1}^{r} \frac{1}{\sigma} \phi\left(\frac{y_{(i)} - \mu}{\sigma} \right) \right) \left(Q\left(\frac{y_{(r)} - \mu}{\sigma} \right) \right)^{n-r}$$

Putting

$$x_i = y_i - y_L$$

and

$$\theta = \frac{y_L - \mu}{\sigma} \tag{4}$$

where

$$y_L = \begin{cases} \ln t_0, & \text{for type 1 censoring} \\ y_{(r)}, & \text{for type 2 censoring} \end{cases}$$

where t_0 is the censoring time, write:

$$L(\mu,\sigma)=\frac{1}{\sigma^r}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{r}(x_i+\theta\sigma)^2\right)Q(\theta)^{n-r} \tag{5}$$

Persson and Rootzen (1977) suggested replacing the survival function $Q(\theta)$ in (4) by its nonparameteric estimator $\dfrac{n-r}{n}$ and therefore replacing θ by $\theta^* = Q^{-1}\left(\dfrac{n-r}{n}\right)$, the $(r/n)th$ quantile of the standard normal distribution. Substituting these quantities in (4), one obtains a function of σ alone which is maximized by

$$\hat{\sigma}_3=\frac{1}{2}\left(\frac{\theta^*}{r}\sum_{i=1}^{r}x_i+\left(\left(\frac{\theta^*}{r}\sum_{i=1}^{r}x_i\right)^2+\frac{4}{r}\sum_{i=1}^{r}x_i^2\right)^{1/2}\right) \tag{6}$$

Substituting $\hat{\sigma}_3$ in (4) yields

$$\hat{\mu}_3 = y_L - \theta^*\hat{\sigma}_3 \tag{7}$$

Approximate MLE Based on Taylor Series Expansion

Consider the likelihood equations given by (3)

$$\frac{\partial \log L}{\partial \mu}=\frac{1}{\sigma^2}\sum_{i=1}^{r}(y_i-\mu)+\frac{(n-r)}{\sigma}\left(\frac{\phi\left(\frac{y_r-\mu}{\sigma}\right)}{Q\left(\frac{y_r-\mu}{\sigma}\right)}\right)$$

$$\frac{\partial \log L}{\partial \sigma}=-\frac{r}{\sigma}$$

$$+\frac{1}{\sigma^3}\sum_{i=1}^{r}(y_i-\mu)^2+\frac{(n-r)}{\sigma}\left(\frac{\frac{y_r-\mu}{\sigma}\phi\left(\frac{y_r-\mu}{\sigma}\right)}{Q\left(\frac{y_r-\mu}{\sigma}\right)}\right)$$

Let $Z_i = \dfrac{y_i-\mu}{\sigma}$, $i=1,\ldots,r$ and noting that $\phi'(z)=-z\phi(z)$ obtains

$$\frac{\partial \log L}{\partial \mu}=\frac{1}{\sigma}\left(\sum_{i=1}^{r}z_i+(n-r)\frac{\phi(z_r)}{Q(z_r)}\right)=0$$

$$\frac{\partial \log L}{\partial \sigma}=-\frac{1}{\sigma}\left(r-(n-r)z_r\frac{\phi(z_r)}{Q(z_r)}-\sum_{i=1}^{r}z_i^2\right)=0 \tag{8}$$

Expanding the function $\dfrac{\phi(z_r)}{Q(z_r)}$ in a Taylor series about the point $\xi_r = \Phi^{-1}(p_r)$, where $\Phi^{-1}(.)$ is the inverse of the distribution function of the standard normal distribution and $p_r = \dfrac{r}{n+1}$. Setting $q_r = 1-p_r$ obtains $\dfrac{\phi(z_r)}{Q(z_r)} \cong \gamma + \delta z_r$, where

$$\gamma = \phi(\zeta_r)\left(1+\xi_r^2-\xi_r\phi(\zeta_r)/q_r\right)/q_r$$

and

$$\delta = \phi(\zeta_r)(\phi(\zeta_r)-q_r\xi_r)/q_r^2 .$$

Substituting these quantities in the likelihood equations obtains

$$\frac{\partial \log L}{\partial \mu}\cong\frac{1}{\sigma}\left(\sum_{i=1}^{r}z_i+(n-r)\gamma+(n-r)\delta z_r\right)=0$$

$$\frac{\partial \log L}{\partial \sigma}\cong-\frac{1}{\sigma}\left(r-(n-r)\gamma z_r-(n-r)\delta z_r^2-\sum_{i=1}^{r}z_i^2\right)=0 \tag{9}$$

Solving these equations yields the following:

$$\hat{\mu}_4 = B - \hat{\sigma}_4 C$$

$$\hat{\sigma}_4 = \left(-D+\left(D^2+4rE\right)^{1/2}\right)/2r , \tag{10}$$

where

$$B=\left(\sum_{i=1}^{r}y_i+(n-r)\delta y_r\right)\Big/m ,$$

$$C = -(n-r)\gamma/m,$$

$$D = -(n-r)\gamma(y_r - B),$$

$$E = \sum_{i=1}^{r} y_i^2 + (n-r)\delta y_r^2 - mB^2$$

and

$$m = r + (n-r)\delta .$$

Performance of the Estimators

A simulation study is conducted to investigate the performance of the estimators. The simulation indices are the sample size $n = 10,15,20,30,40,50,60,80,100,150$. The censoring proportion cp: 0.1, 0.3, 0.5, $a = 1 - cp$. For each combination of the simulation indices, 2,000 pairs of samples are generated and the maximum likelihood estimator $(\hat{\mu}, \hat{\sigma})$ and the closed form estimators $(\hat{\mu}_i, \hat{\sigma}_i), i = 1\ldots,4$ are calculated. Their biases $B\hat{\mu}, B\hat{\sigma}$ and $B\hat{\mu}_i, B\hat{\sigma}_i, i = 1,\ldots,4$ and their mean squared errors and the relative efficiencies $ef\hat{\mu}_i = \frac{MSE(\hat{\mu})}{MSE(\hat{\mu}_i)}$ and $ef\hat{\sigma}_i = \frac{MSE(\hat{\sigma})}{MSE(\hat{\sigma}_i)}$ $, i = 1\ldots,4$ are obtained.

Results

The results are given in Tables 1 – 4. The biases of the estimators are given in Tables 1 – 2 and the efficiencies of the estimators are given in tables 3 – 4. Inspection of the simulation numerical results lead to the following observations and conclusions. It appears that, under Type 1 censoring, $\hat{\mu}_1$ and $\hat{\mu}_2$ are positively biased when the censoring proportion is moderate to heavy. This is true for all sample sizes. In all other cases, all estimators tend to be negatively biased, regardless of the sample size. It appears that $\hat{\mu}_3$ has the highest bias, and the least bias is achieved by $\hat{\mu}_3$ for light censoring and $\hat{\mu}_2$ and $\hat{\mu}_5$ for moderate to heavy censoring.

For estimators of the scale parameter σ under Type 1 censoring, it appears that $\hat{\sigma}$ has the least bias followed by $\hat{\sigma}_3$ and $\hat{\sigma}_4$. The performances of $\hat{\sigma}_3$ and $\hat{\sigma}_4$ in terms of bias is about similar. However, $\hat{\sigma}_1$ tends to have the largest bias among the estimators considered.

The relative performance of estimators under Type 2 censoring is similar to that of Type 1 censoring. In all cases, the bias decreases as the sample size increases. It is also smaller for lighter censoring.

Concerning the relative efficiencies of the estimators under Type 1 censoring, it appears that the following schemes hold, $\hat{\mu}_4 > \hat{\mu}_3 > \hat{\mu}_2 > \hat{\mu}_1$ under heavy censoring regardless of the sample size and $\hat{\mu}_4 > \hat{\mu}_1 > \hat{\mu}_2 > \hat{\mu}_3$ for moderate to light censoring, where (>) means more efficient. It also appears that the relative efficiencies of $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\mu}_3$ do not depend on the sample size. However, the relative efficiency of $\hat{\mu}_4$ increases as sample size increases. The relative efficiencies of $\hat{\mu}_2$ and $\hat{\mu}_3$ increase as the censoring proportion becomes smaller, while it decreases for $\hat{\mu}_4$.

The results show that, under Type 1 censoring $\hat{\mu}_4$ are more efficient than the MLE. With regard to scale estimators under Type 1 censoring, it appears that $\hat{\sigma}_4 > \hat{\sigma}_3 > \hat{\sigma}_2 > \hat{\sigma}_1$, whereas before (>) indicated more efficient. It appears that the relative efficiencies of the scale estimators do not depend on n; however, they depend on the censoring proportion. As the censoring proportion becomes smaller, the relative efficiencies of $\hat{\sigma}_1, \hat{\sigma}_2$ and $\hat{\sigma}_4$ increases and it decreases for $\hat{\sigma}_3$. Surprisingly, in all cases considered, the approximate estimators $\hat{\sigma}_4$ are more efficient than the corresponding MLE.

Table 1. Bias of the Estimators Under Type 1 Censoring

n	a	$B\hat{\mu}_1$	$B\hat{\mu}_2$	$B\hat{\mu}_3$	$B\hat{\mu}_4$	$B\hat{\mu}$	$B\hat{\sigma}_1$	$B\hat{\sigma}_2$	$B\hat{\sigma}_3$	$B\hat{\sigma}_4$	$B\hat{\sigma}$
10	0.5	0.106	0.039	-0.121	-0.114	-0.099	0.268	0.239	-0.194	-0.195	-0.188
10	0.7	0.040	-0.033	-0.083	-0.055	-0.041	0.231	0.193	-0.141	-0.138	-0.125
10	0.9	-0.000	-0.086	-0.103	-0.016	-0.015	0.191	0.131	-0.149	-0.108	-0.099
15	0.5	0.088	0.043	-0.075	-0.066	-0.056	0.221	0.203	-0.118	-0.118	-0.112
15	0.7	0.030	-0.019	-0.047	-0.032	-0.018	0.175	0.149	-0.085	-0.086	-0.073
15	0.9	-0.007	-0.069	-0.103	-0.010	-0.014	0.153	0.107	-0.115	-0.062	-0.057
20	0.5	0.079	0.047	-0.062	-0.059	-0.051	0.184	0.171	-0.094	-0.096	-0.091
20	0.7	0.027	-0.008	-0.041	-0.028	-0.021	0.139	0.122	-0.074	-0.074	-0.066
20	0.9	0.005	-0.038	-0.050	-0.009	-0.001	0.126	0.096	-0.069	-0.055	-0.041
30	0.5	0.078	0.057	-0.036	-0.033	-0.026	0.147	0.139	-0.063	-0.064	-0.061
30	0.7	0.025	0.001	-0.024	-0.018	-0.010	0.108	0.096	-0.046	-0.049	-0.041
30	0.9	0.007	-0.021	-0.031	-0.007	0.003	0.098	0.079	-0.041	-0.035	-0.021
40	0.5	0.051	0.036	-0.039	-0.033	-0.033	0.117	0.111	-0.050	-0.050	-0.049
40	0.7	0.013	-0.004	-0.026	-0.019	-0.016	0.089	0.081	-0.033	-0.033	-0.029
40	0.9	-0.000	-0.022	-0.030	-0.008	-0.003	0.071	0.057	-0.038	-0.030	-0.022
50	0.5	0.050	0.038	-0.030	-0.025	-0.024	0.102	0.097	-0.041	-0.041	-0.040
50	0.7	0.015	0.001	-0.020	-0.013	-0.010	0.079	0.072	-0.025	-0.025	-0.022
50	0.9	0.002	-0.014	-0.022	-0.006	0.000	0.066	0.054	-0.026	-0.021	-0.012
60	0.5	0.051	0.041	-0.022	-0.019	-0.016	0.103	0.099	-0.024	-0.024	-0.022
60	0.7	0.013	0.002	-0.014	-0.011	-0.007	0.065	0.060	-0.023	-0.024	-0.020
60	0.9	0.001	-0.012	-0.019	-0.005	-0.001	0.053	0.044	-0.025	-0.020	-0.014
80	0.5	0.035	0.027	-0.019	-0.016	-0.016	0.076	0.074	-0.020	-0.020	-0.019
80	0.7	0.014	0.006	-0.008	-0.005	-0.003	0.050	0.047	-0.019	-0.020	-0.017
80	0.9	-0.002	-0.012	-0.016	-0.004	-0.003	0.036	0.029	-0.022	-0.019	-0.015
100	0.5	0.034	0.028	-0.014	-0.012	-0.011	0.069	0.067	-0.014	-0.015	-0.014
100	0.7	0.009	0.003	-0.010	-0.007	-0.005	0.048	0.045	-0.011	-0.011	-0.009
100	0.9	-0.001	-0.010	-0.014	-0.006	-0.002	0.036	0.030	-0.013	-0.012	-0.007
150	0.5	0.026	0.022	-0.008	-0.007	-0.006	0.048	0.046	-0.011	-0.012	-0.011
150	0.7	0.005	0.001	-0.009	-0.005	-0.005	0.035	0.033	-0.008	-0.007	-0.006
150	0.9	-0.001	-0.006	-0.008	-0.004	-0.001	0.025	0.022	-0.008	-0.008	-0.004

Table 2. Bias of the Estimators Under Type 2 Censoring

n	a	$B\hat{\mu}_1$	$B\hat{\mu}_2$	$B\hat{\mu}_3$	$B\hat{\mu}_4$	$B\hat{\mu}$	$B\hat{\sigma}_1$	$B\hat{\sigma}_2$	$B\hat{\sigma}_3$	$B\hat{\sigma}_4$	$B\hat{\sigma}$
10	0.5	0.117	0.049	-0.114	-0.092	-0.091	0.285	0.256	-0.185	-0.178	-0.178
10	0.7	0.050	-0.023	-0.064	-0.028	-0.027	0.221	0.184	-0.141	-0.128	-0.127
10	0.9	-0.000	-0.086	-0.108	-0.016	-0.015	0.201	0.140	-0.145	-0.094	-0.093
15	0.5	0.100	0.057	-0.109	-0.093	-0.092	0.231	0.214	-0.143	-0.138	-0.138
15	0.7	0.059	0.011	-0.040	-0.014	-0.014	0.193	0.170	-0.089	-0.081	-0.080
15	0.9	0.007	-0.048	-0.064	-0.009	-0.009	0.160	0.124	-0.087	-0.059	-0.059
20	0.5	0.088	0.056	-0.060	-0.048	-0.047	0.192	0.180	-0.092	-0.089	-0.089
20	0.7	0.034	-0.001	-0.032	-0.013	-0.013	0.140	0.122	-0.072	-0.065	-0.065
20	0.9	0.007	-0.036	-0.048	0.000	0.001	0.123	0.094	-0.071	-0.043	-0.043
30	0.5	0.078	0.057	-0.039	-0.029	-0.029	0.149	0.141	-0.065	-0.062	-0.062
30	0.7	0.027	0.003	-0.025	-0.010	-0.010	0.115	0.104	-0.043	-0.037	-0.037
30	0.9	0.002	-0.026	-0.034	-0.001	-0.001	0.084	0.065	-0.052	-0.033	-0.033
40	0.5	0.063	0.047	-0.034	-0.026	-0.026	0.123	0.117	-0.049	-0.047	-0.047
40	0.7	0.022	0.005	-0.018	-0.007	-0.007	0.089	0.081	-0.035	-0.030	-0.030
40	0.9	0.004	-0.017	-0.025	0.001	0.001	0.069	0.055	-0.039	-0.024	-0.024
50	0.5	0.047	0.035	-0.035	-0.028	-0.028	0.101	0.097	-0.043	-0.042	-0.041
50	0.7	0.020	0.007	-0.013	-0.004	-0.004	0.076	0.069	-0.027	-0.024	-0.024
50	0.9	-0.000	-0.017	-0.024	-0.002	-0.002	0.061	0.050	-0.029	-0.016	-0.016
60	0.5	0.041	0.031	-0.027	-0.023	-0.023	0.090	0.086	-0.033	-0.032	-0.032
60	0.7	0.019	0.007	-0.014	-0.004	-0.004	0.067	0.061	-0.025	-0.022	-0.022
60	0.9	-0.001	-0.015	-0.020	-0.003	-0.003	0.053	0.043	-0.023	-0.013	-0.013
80	0.5	0.040	0.033	-0.014	-0.011	-0.011	0.076	0.073	-0.022	-0.021	-0.021
80	0.7	0.011	0.002	-0.012	-0.006	-0.006	0.054	0.050	-0.016	-0.014	-0.014
80	0.9	0.001	-0.009	-0.015	-0.001	-0.001	0.039	0.032	-0.022	-0.013	-0.013
100	0.5	0.034	0.028	-0.012	-0.009	-0.009	0.060	0.058	-0.022	-0.021	-0.021
100	0.7	0.016	0.009	-0.005	0.000	0.000	0.048	0.045	-0.012	-0.010	-0.010
100	0.9	0.002	-0.005	-0.009	0.001	0.001	0.035	0.030	-0.014	-0.008	-0.008
150	0.5	0.028	0.024	-0.007	-0.005	-0.005	0.050	0.049	-0.010	-0.009	-0.009
150	0.7	0.010	0.005	-0.004	-0.001	-0.001	0.031	0.029	-0.010	-0.009	-0.009
150	0.9	0.001	-0.004	-0.006	0.001	0.001	0.026	0.022	-0.008	-0.004	-0.004

Table 3. Efficiencies of the Estimators Under Type 1 Censoring

n	a	$ef\hat{\mu}_1$	$ef\hat{\mu}_2$	$ef\hat{\mu}_3$	$ef\hat{\mu}_4$	$ef\hat{\sigma}_1$	$ef\hat{\sigma}_2$	$ef\hat{\sigma}_3$	$ef\hat{\sigma}_4$
10	0.5	0.741	0.835	0.977	1.718	0.388	0.419	0.994	1.054
10	0.7	0.917	0.957	0.932	2.095	0.421	0.477	0.975	1.143
10	0.9	1.000	0.937	0.822	1.563	0.472	0.600	0.863	1.200
15	0.5	0.745	0.811	0.981	2.109	0.399	0.425	0.991	1.097
15	0.7	0.930	0.952	0.927	2.208	0.459	0.510	0.964	1.180
15	0.9	0.999	0.935	0.755	1.438	0.495	0.623	0.792	1.278
20	0.5	0.732	0.787	0.966	2.320	0.446	0.467	0.989	1.098
20	0.7	0.891	0.915	0.936	2.373	0.496	0.537	0.959	1.179
20	0.9	0.997	0.971	0.810	1.485	0.535	0.626	0.856	1.293
30	0.5	0.674	0.714	0.989	2.521	0.439	0.454	1.000	1.100
30	0.7	0.878	0.902	0.939	2.520	0.534	0.565	0.971	1.243
30	0.9	0.983	0.973	0.832	1.438	0.551	0.625	0.855	1.335
40	0.5	0.736	0.767	0.966	2.727	0.489	0.503	0.993	1.126
40	0.7	0.897	0.910	0.925	2.753	0.548	0.575	0.968	1.291
40	0.9	0.989	0.973	0.814	1.494	0.635	0.701	0.837	1.377
50	0.5	0.725	0.752	0.973	2.847	0.512	0.524	0.994	1.132
50	0.7	0.890	0.905	0.930	2.827	0.571	0.594	0.972	1.291
50	0.9	0.986	0.978	0.813	1.505	0.613	0.670	0.852	1.358
60	0.5	0.707	0.734	0.970	3.018	0.490	0.501	0.992	1.145
60	0.7	0.884	0.898	0.935	2.867	0.601	0.624	0.963	1.306
60	0.9	0.991	0.982	0.804	1.528	0.663	0.715	0.859	1.354
80	0.5	0.712	0.730	0.977	3.119	0.518	0.528	0.993	1.145
80	0.7	0.910	0.924	0.911	3.171	0.625	0.643	0.969	1.277
80	0.9	0.991	0.980	0.801	1.571	0.754	0.798	0.836	1.447
100	0.5	0.702	0.718	0.975	3.224	0.532	0.541	0.993	1.145
100	0.7	0.901	0.911	0.919	3.152	0.616	0.632	0.975	1.307
100	0.9	0.988	0.978	0.801	1.482	0.725	0.764	0.821	1.437
150	0.5	0.719	0.733	0.972	3.309	0.588	0.595	0.998	1.158
150	0.7	0.913	0.918	0.923	3.307	0.677	0.691	0.956	1.351
150	0.9	0.988	0.983	0.806	1.528	0.758	0.789	0.833	1.436

Table 4. Efficiencies of the Estimators Under Type 2 Censoring

n	a	$ef\hat{\mu}_1$	$ef\hat{\mu}_2$	$ef\hat{\mu}_3$	$ef\hat{\mu}_4$	$ef\hat{\sigma}_1$	$ef\hat{\sigma}_2$	$ef\hat{\sigma}_3$	$ef\hat{\sigma}_4$
10	0.5	0.723	0.821	0.978	0.999	0.370	0.400	0.992	0.999
10	0.7	0.921	0.972	0.929	1.000	0.445	0.505	0.961	0.999
10	0.9	0.999	0.934	0.807	0.999	0.452	0.577	0.869	0.999
15	0.5	0.688	0.753	0.980	0.999	0.395	0.416	0.996	0.999
15	0.7	0.853	0.910	0.954	0.999	0.425	0.463	0.980	0.999
15	0.9	0.978	0.950	0.866	0.999	0.487	0.575	0.900	0.999
20	0.5	0.709	0.764	0.975	0.999	0.429	0.449	0.992	0.999
20	0.7	0.908	0.939	0.917	1.000	0.507	0.547	0.963	0.999
20	0.9	0.982	0.961	0.856	1.000	0.531	0.622	0.842	1.000
30	0.5	0.693	0.733	0.974	0.999	0.439	0.454	0.996	0.999
30	0.7	0.880	0.907	0.919	1.000	0.499	0.529	0.975	0.999
30	0.9	0.990	0.970	0.814	1.000	0.621	0.698	0.851	0.999
40	0.5	0.687	0.720	0.982	0.999	0.455	0.468	1.001	0.999
40	0.7	0.896	0.919	0.919	1.000	0.549	0.576	0.971	0.999
40	0.9	0.986	0.976	0.825	1.000	0.639	0.703	0.864	0.999
50	0.5	0.700	0.725	0.978	0.999	0.503	0.515	0.991	1.000
50	0.7	0.890	0.909	0.936	1.000	0.572	0.595	0.974	1.000
50	0.9	0.992	0.977	0.796	1.000	0.652	0.710	0.846	0.999
60	0.5	0.716	0.738	0.977	0.999	0.492	0.502	1.001	0.999
60	0.7	0.882	0.900	0.926	0.999	0.590	0.611	0.972	0.999
60	0.9	0.994	0.981	0.795	1.000	0.670	0.722	0.847	1.000
80	0.5	0.709	0.729	0.970	1.000	0.525	0.534	0.995	1.000
80	0.7	0.903	0.915	0.912	1.000	0.610	0.629	0.968	1.000
80	0.9	0.986	0.981	0.833	1.000	0.722	0.765	0.844	0.999
100	0.5	0.728	0.745	0.974	1.000	0.572	0.581	0.994	1.000
100	0.7	0.902	0.917	0.917	1.000	0.616	0.633	0.973	1.000
100	0.9	0.987	0.985	0.815	1.000	0.725	0.765	0.830	1.000
150	0.5	0.713	0.726	0.981	1.000	0.573	0.580	0.998	1.000
150	0.7	0.912	0.921	0.918	1.000	0.689	0.701	0.967	1.000
150	0.9	0.988	0.986	0.823	1.000	0.751	0.780	0.865	1.000

The performance of the estimators under Type 2 censoring is similar to their performance under Type 1 censoring. However it appears that $\hat{\sigma}_3$ and $\hat{\sigma}_4$ are about as efficient as the MLE for all sample sizes and censoring proportions, except for $\hat{\sigma}_3$ when the censoring proportion is small, in which case $\hat{\sigma}_3$ is less efficient.

Conclusion

It appears that good substitutes to the MLE in closed form do exist. The performance of some of them is highly competent with that of the MLE and sometimes they are better, as is the case with the approximation based on the Taylor series expansion $\hat{\mu}_4$ and $\hat{\sigma}_4$.

References

Balakrishnan & Cohen (1991). *Order statistics and inference*. Chapman and Hall.

Hossain, A. & Howlader H. A. (1996). Unweighted least squares estimation of Weibull parameters. *Journal of Statistical Computation and Simulation, 54*, 265-271.

Hossain, M., Anwar, & Zimmer, J. W. (2003). Comparison of estimation methods for Weibull parameters: Complete and censored samples. *Journal of Statistical Computation and Simulation, 73*(2), 145-153

Johnson, N. L, Kotz, S., & Balakrishnan. (1994). *Continuous univariate distributions* (vol. 1). New York: Wiley.

Koch, A. L. (1966). The logarithm in biology. *Journal of Theoretical Biology, 12,* 276–290.

Lawless, J. F. (1982). *Statistical models and methods for lifetime data.* NewYork: Wiley.

Lyles, R. H., Kupper, L. L., & Rappaport, S. M. (1997). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects ANOVA model. *Journal of Agricultural, Biological, and Environmental Statistics, 2*, 64–86.

Ebrahem, M. A. & Baklizi, A. (2005). Comparison of some simple estimators of the log-logistic parameters based on complete and censored samples. *Journal of Statistical Theory and Applications, 4*(2): 151 – 160.

Persson, T. & Rootzen, H. (1977). Simple and highly efficient estimators for a type 1 censored normal sample. *Biometrika, 64,* 123–128.

Schneider, H. (1986). *Truncated and censored samples from normal populations.* New York: Marcel Dekker.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 181-193

1538 – 9472/06/$95.00

Properties Of Bound Estimators On Treatment Effect Heterogeneity For Binary Outcomes

Edward J. Mascha
Department of Quantitative Health Sciences
The Cleveland Clinic Foundation

Jeffrey M. Albert
Department of Epidemiology and Biostatistics
Case Western Reserve University

Variability in individual causal effects, treatment effect heterogeneity (TEH), is important to the interpretation of clinical trial results, regardless of the marginal treatment effect. Unfortunately, it is usually ignored. In the setting of two-arm randomized studies with binary outcomes, there are estimators for bounds on the probability of control success and treatment failure for an individual, or the treatment risk. Here, those bounds were refined and the sampling properties were assessed using simulations of correlated multinomial data via the Dirichlet multinomial. Results indicated low bias and mean squared error. Moderate to high intraclass correlation (ICC) and large numbers of clusters allow narrower confidence interval widths for the treatment risk.

Key words: Blocked or clustered data, bounds, causal effects, Dirichlet multinomial, intraclass correlation, marginal treatment effect, randomized trial, potential outcomes, treatment effect heterogeneity, unit-treatment interaction.

Introduction

In randomized clinical trials comparing an experimental treatment (T) to a control (C), the focus is usually on the marginal treatment effect, (i.e., mean causal effect) estimated by the difference in means or the difference in the proportion having a successful outcome. Unfortunately, the amount of variability of the individual causal effects is usually ignored. Recent work has seen the development of bounds on a treatment effect heterogeneity parameter for binary outcomes (Gadbury, Iyer, & Albert, 2004; Albert, Gadbury, & Mascha, 2005). The latter provided bound estimates and confidence intervals in the case of blocked binary outcomes. However, no study has been yet conducted to evaluate the properties and practicality of these methods.

Treatment effect heterogeneity (TEH), also called unit-treatment interaction (Gadbury & Iyer, 2000) or subject-treatment interaction (Gadbury, Iyer, & Allison, 2001), is the amount of variability in the causal effect of T versus C on some outcome Y. The causal effect for an individual is defined as the difference in the individual's potential outcomes (Neyman, 1923; Rubin, 1974; 2000) on T and C, respectively. This is an unobservable latent variable since only one of the two potential outcomes may be observed for an individual. For example, consider a binary outcome scenario with success proportions of 0.50 and 0.30 for treatments T and C, respectively, giving a marginal treatment effect of 0.20. With these marginals, the minimum possible TEH would be that no patients who succeed on C would fail on T, implying that 0.20 of the patients would fail on C and succeed on T. With the same marginals, the maximum possible TEH would be that 0.30 of patients would succeed on C but fail on T, and that 0.50 would fail on C but succeed on T.

Thus, in the case of a binary outcome for two treatments, individuals fall into a category based on their potential outcomes: (1) failure on both T and C, (2) success on T and C, or (3) success on one but not the other. The

Edward J. Mascha, Ph. D., is an Assistant Staff Biostatistician. His interests include causal effects and correlated data methods. Email him at maschae@ccf.org. Jeffrey M. Albert, is an Assistant Professor of Biostatistics. His research interests include causal inference. Email him at jma13@case.edu.

probabilities of membership into each of these categories are denoted as π_{00}, π_{01}, π_{10}, π_{11}, where indices indicate response (1=success, 0=failure) to T and C, respectively. The probability of doing worse on a new treatment (T) than on standard treatment (C), π_{01}, may be understood as the treatment risk because patients would not expect to do worse on the new treatment. Although this quantity is typically overlooked in analyses of clinical trials, it would be of potential interest for both individual treatment decisions and the understanding of the population impact of treatment.

Albert, Gadbury and Mascha (2005, AGM) provided bounds and bound estimators for the treatment risk π_{01} (referred to by AGM as π_2). However, the AGM bounds cannot be reliably used in practice until their sampling properties have been assessed. Such is the purpose of this article.

Background

Gadbury and Iyer (2000) derived bounds for the probability of an unfavorable individual treatment effect where the outcome is continuous; for example, an individual doing better (higher value) on control than on treatment. They assumed a trivariate normal distribution between the potential outcomes on treatment X and control Y, and a covariate Z which is measured on all patients. Such methods are not easily applicable to binary outcomes because of the difficulty in specifying a meaningful multivariate distribution for the binary setting.

New methods are available, however, to estimate bounds on treatment effect heterogeneity for binary outcomes. These include simple bounds and bounds which make use of clustering. Based on the fact that π_{11}, π_{00}, π_{10}, π_{01} sum to 1.0 and that $\pi_{10} - \pi_{01} = \pi_T - \pi_C$; Gadbury, Iyer, and Albert (2004), which is referred to as GIA, derived simple bounds for π_{01} such that

$$\max(0, \pi_C - \pi_T) \equiv L_s \leq \pi_{01} \leq U_s \equiv \min(1-\pi_T, \pi_C) \tag{1}$$

For example, with true marginal successes π_T = .80 and π_C =.70, simple bounds for π_{01} are (0, .20), and by substituting π_T for π_C and visa versa, the simple bounds for π_{10} are (.10, .30). The marginal proportions π_T and π_C have a large effect on the possible range of unit-treatment interaction in the binary outcome case. A proportion close to 0 or 1 greatly limits the range of TEH, and so allows tighter bounds on the parameters of interest. When neither of the marginals is close to 0 or 1, there is a wider range of possible heterogeneity, and therefore greater opportunity for narrowing through more refined methods.

GIA also give more refined bounds on π_{01}, first using a matched-pairs design in which one member of a pair is randomly assigned to receive treatment and the other member receives control. They construct bounds which narrow as the quality of the matching improves. Further, they consider an extended matched-pairs design, in which some pairs are randomized to either both treatment or both control, which allows the refined bounds to be estimated.

Gadbury, Iyer, and Albert (2004) defined the probability that a treatment unit fails ($Y_T(u_1)$=0) and the matched control unit has success ($Y_C(u_2)$=1), i.e., control beats treatment, or,

$$g_2 = P(Y_T(u_1) = 0, Y_C(u_2) = 1)$$

where u_1 and u_2 are two members of a matched pair. GIA also define h_T and h_C as probabilities of success for both members of a pair of randomly chosen matched treated or control units, respectively, such that

$$h_T = P(Y_T(u_1) = 1, Y_T(u_2) = 1)$$

and

$$h_C = P(Y_C(u_1) = 1, Y_C(u_2) = 1)$$

Higher h_T and h_C indicate better matching and will lead to tighter bounds. Lower and upper bounds for π_{01}, with the "*B*" subscript referring to the blocked (in the present case, the extended matched pairs) design, are as follows:

$$L_{\mathrm{C}} \equiv \mathrm{Max}\ (0, g_2 - \min(\pi_{\mathrm{T}} - h_{\mathrm{T}}, \pi_{\mathrm{C}} - h_{\mathrm{C}}))$$
$$U_{\mathrm{C}} \equiv \mathrm{Min}\ (1, g_2 + \min(\pi_{\mathrm{T}} - h_{\mathrm{T}}, \pi_{\mathrm{C}} - h_{\mathrm{C}})) \tag{2}$$

The bounds for π_{01} (equation 2) were derived by first expressing g_2 , h_T and h_C as functions of the underlying parameters of interest, and then adding terms to the expression for g_2 so that the resulting form consisted of quantities for which one has estimators.

In the latest development, Albert, Gadbury, and Mascha (2005, AGM) used bounds with the same form as (2) for π_{01}, but extend definitions to the more general blocked or clustered design. That is, the pair of individuals u_1 and u_2 in the definitions of g_2, h_T and h_C, is now considered as belonging to the same cluster. In many cases this is more realistic than the matched or extended pairs design. Blocks can be created post-hoc. Good blocking or matching gives narrower bounds.

AGM provide non-parametric estimators of the bounds in (2). Each represents a proportion with the given outcome combination, and is estimated as the ratio of the sum across clusters of the number of pairs observed with the given outcome combination to the number of pairs with the given treatment assignments. For example,

$$\hat{g}_2 = \frac{\sum_j n_{C1j} n_{T0j}}{\sum_j n_{Cj} n_{Tj}},$$

is the estimator for g_2, and is the proportion of observed pairs with treatment failure and control success out of the total number of possible treatment-control pairs. Substitution into (2) yields estimated cluster bounds $\hat{L}_B$ and $\hat{U}_B$. AGM (equations 6 through 11) give variances and covariances for estimators of the lower and upper bounds on π_{01} and for their components. Refer to their article for details on the formulae, which are quite extensive.

In this study, the AGM estimators for bounds on π_{01} are first refined. Then, through simulations their statistical properties, including bias, variance, MSE, and coverage are evaluated. Because the AGM bound estimators depend on clustering in the data, a simulation method that allows specification of the intraclass correlation (ICC) as well as the underlying probabilities has been devised. Simultaneous confidence intervals for the lower and upper bounds are shown to provide at least 1-α coverage of π_{01}, the real parameter of interest. Properties are shown to depend on degree of ICC, TEH, marginal success, number of clusters, and sample size.

Methodology

First, a refinement to the AGM bounds is proposed, and then the Dirichlet-multinomial (DMN) is introduced as the model for the potential outcomes. Finally, the treatment effect heterogeneity scenarios and simulation methods used to assess statistical properties of the estimators for bounds on π_{01} and their components are outlined.

Refinement to AGM Bounds

With good blocking, the AGM cluster bounds in (2) are narrower than the simple bounds (1) on π_{01}. However, it can be shown that the cluster bounds are the same or wider than the simple bounds when subjects are independent from each other (and thus, $h_T = \pi_T^2$ and $g_2 = (1-\pi_T)\ \pi_C$), which would occur if the matching or clustering were at random or non-existent. Therefore, a modification of the AGM cluster bounds to be the narrower of the simple and AGM cluster bounds is proposed, such that:

$$L_{\mathrm{MC}} \equiv Max(L_{\mathrm{S}}, \mathrm{g}_2 - \min(\pi_{\mathrm{T}}\text{-}\mathrm{h}_{\mathrm{T}}, \pi_{\mathrm{C}}\text{-}\mathrm{h}_{\mathrm{C}}))$$
$$U_{\mathrm{MC}} \equiv Min(U_{\mathrm{S}}, \mathrm{g}_2 + \min(\pi_{\mathrm{T}}\text{-}\mathrm{h}_{\mathrm{T}}, \pi_{\mathrm{C}}\text{-}\mathrm{h}_{\mathrm{C}})) \tag{3}$$

With random matching, the modified AGM cluster bounds (MAGM) and simple bounds are identical, and the cluster bound width will always be at least as narrow as the simple bound width, sometimes significantly narrower, depending on the TEH scenario, the marginals, and the amount of clustering.

Property assessment

In order to assess the statistical properties of the bound estimators for $\pi_{01,}$ a model of the underlying (i.e., latent) correlated multinomial data was needed, where each unit or subject belongs to one of the four potential outcome categories (C_{00}, C_{01}, C_{10}, C_{11}), indexed by the latent response to treatment and control, respectively, with probabilities $\pi_{00,}$ $\pi_{01,}$ $\pi_{10,}$ π_{11}, and where units are correlated within clusters. Various approaches to modeling correlated multinomial data have been used (Gange, 1995, Morel & Nagaraj, 1993, Banergee & Paul, 1999). Mosimann (1962) and Brier (1980) extol the Dirichlet multinomial (DMN) distribution, also called the multivariate beta-binomial distribution, as a natural way to model over-dispersed multinomial data. The DMN is used because it also allows direct specification of the intra-class correlation and there is no need to assume an underlying continuous distribution of the data. It is less computationally intensive than some of the other methods and can therefore be used with large numbers of clusters and units per cluster, r, where the method of Gange (1995), for example, cannot.

It is assumed that each unit latently falls into one of the four population categories with the corresponding probabilities $\pi_{00,}$ $\pi_{01,}$ $\pi_{10,}$ π_{11}, denoted as the vector π. Each cluster's set of probabilities deviates randomly from the underlying vector according to the Dirichlet distribution and the counts within each cluster are independent multinomial data conditional on the realized cluster probabilities. The unconditional counts in the 4 categories are distributed as DMN, or $DMN_4(n,\pi,k)$, where k is a structural parameter related to the ICC, the correlation among units within the same cluster and category, such that k= (1-ICC)/ICC, and so ICC=1/ (1+k). This relationship between k and the ICC is used to induce varying levels of correlation among subjects within clusters in the simulations.

The statistical properties of the MAGM and AGM estimators for bounds on π_{01} and estimators for their components (g_2, H_T,$H_{C,}$ π_T and π_C) were evaluated under five treatment effect heterogeneity (TEH) scenarios (Table 1). Scenarios are distinguished by the level of TEH (low, medium or high value of π_{01} for the given marginals) and the marginal success proportions π_T and π_C: one marginal close to zero (π_T =.20, π_C =.10) or both close to .50 (π_T =.45, π_C =.55). Each scenario is also described by the amount of correlation among the potential outcomes on T and C, or ρ_{PO}. This correlation is a function of π_{01} and the marginal success proportions, so that zero ρ_{PO} indicates independence of the potential outcomes, in which case π_{01} and π_{10} are the product of the corresponding marginals, and which may be the most natural case. Negative ρ_{PO} indicates high TEH (π_{01} and π_{10} are higher than under independence) and positive ρ_{PO} indicates low TEH (π_{01} and π_{10} are lower than expected under independence). Within each scenario, the ICC (.15, .50, and .85), the total sample size N (600, 3000), and the number of clusters C (20, 40, and 100) are varied to assess the effect of each factor on the estimator properties.

A set of simulations was conducted for each TEH scenario from Table 1, for each variation of ICC, total sample size, and number of clusters. For each cluster i , Dirichlet random deviates $p_1^{(i)},\ldots,p_4^{(i)}$ were formed of success probabilities from the underlying vector π as the ratio of random gamma deviates over the sum of the associated four gamma deviates (Jensen, 1998), where subscripts 1, …, 4 indicate the four population categories C_{00}, C_{01}, C_{10}, C_{11}, respectively. The parameter for each of the four gamma deviates is the clustering parameter k times the probability of the associated underlying population category. Next, n units (where n=N /C) were randomly sampled from the four population categories according to a multinomial distribution with probabilities $p_1^{(i)},\ldots,p_4^{(i)}$ for the i^{th} cluster. Each unit within each cluster was randomly assigned to have either the response to Y_T or Y_C observed. Finally, the estimated bounds (and estimated bound components) for π_{01}, plus individual and simultaneous (lower, upper bound) confidence intervals for the bounds were calculated. This was repeated 1,000 times for each scenario combination (each particular scenario, sample size, ICC and number of clusters combination) and summarized across simulations.

For the AGM and MAGM bound estimators and their components within each scenario, the expected value (mean over 1,000 simulations), bias, true variance (variance of the estimated values over the simulations), mean estimated variance and mean squared error (MSE) were assessed. Formula-based 95% confidence intervals (CI) and their widths for lower and upper bounds were then obtained. Approximate confidence intervals were calculated using a normal approximation for the distribution of the bound estimators. For example, a 100(1-α) % confidence interval (CI) for the AGM upper bound, U_B, is $\hat{U}_B \pm z_{1-\alpha/2}(\hat{V}(\hat{U}_B))^{1/2}$, where $z_{1-\alpha/2}$ is the (1-α/2) percentile of the standard normal distribution. A CI for the lower bound, L_B , was obtained similarly. Finally, coverage of the true bounds for both the lower and upper bound estimators was obtained.

Simultaneous (i.e., joint) asymptotic (1-α)% confidence intervals intended to have at least a 1- α probability of containing the true population values of both the lower and upper bounds were also obtained. These were formed by the estimated lower 95% CL of the lower bound and the estimated upper 95% CL of the upper bound from the AGM formulae. Because the formed intervals are designed to have the given nominal probability of containing the true bounds on π_{01}, by definition they should have at least as great a probability of containing the true π_{01}, the parameter of interest. Using these intervals, the mean estimated width, the simultaneous estimated coverage of the true bounds, and the estimated coverage of the true parameter $\tilde{\pi}_{01}$ are reported.

For comparison purposes, and because the joint distribution of the lower and upper bounds is not readily available (assumed to be independent in forming the confidence intervals above), joint confidence intervals were also estimated using a bootstrap method which naturally accounts for dependency between the bounds and also allows non-symmetric intervals around the estimators. Bickel and Friedman (1981) proved that the bootstrap can be used to construct confidence intervals for two unknown parameters simultaneously. Horowitz and Manski (2000) use the bootstrap to put bounds on the treatment effect for missing-value data, where either baseline covariates and/or outcomes are missing for some subjects. The same method was used to provide a joint confidence interval for a pair of lower and upper cluster bounds on the parameter $\pi_{01.}$ The goal was to create an interval of the form [$\hat{L}$ – d_α, $\hat{U}$ + d_α], where $\hat{L}$ and $\hat{U}$. An appropriate value of a constant d_α was chosen such that the interval contains the true parameters L and U with probability 1- α asymptotically. The delta was applied non-symmetrically in hopes of achieving even better coverage with equivalent or smaller confidence interval widths as with the formula method.

Results

Tables 2 and 3 report bias, variance and MSE of the MAGM lower and upper bound estimators for two representative scenarios: scenario 1, the combination of low treatment heterogeneity (π_{01}= .01) and marginals close to zero and scenario 5, the combination of high treatment heterogeneity (π_{01}= .40) and marginals close to .50. Bias of the lower and upper bound estimators and their components is consistently low, typically much less than 5% of the expected value of the estimator for low, medium, or high ICC for each scenario assessed. Bias decreases with increasing ICC. Higher ICC increases the mean estimated variance of the lower and upper bound estimators and components and therefore the MSE, given the consistently low bias. As expected, the mean estimated variances and covariances of the bound estimators across simulations using the AGM formulas are also very close to the true variances and covariances for each estimator. Having a larger number of clusters for a fixed ICC and sample size steadily decreases the variance of all estimators and their associated MSEs. Similar properties and relationships were observed for scenarios 2, 3, and 4 (results not shown).

Confidence interval width and coverage results of both the individual and the simultaneous lower and upper bound estimators on π_{01} are given in Tables 4 and 5 for scenarios

1 and 5, respectively, and in Figures 1 (all scenarios, 20 clusters) and 2 (scenarios 1, 3 and 5 for 20, 40 and 100 clusters). As expected from results on the variance of the bound estimators, CI widths for the individual lower and upper bounds were in general much narrower for scenario 1 (Table 4) and scenario 2 (data not shown), where at least one of the marginal success proportions is close to 0 or 1. Mean CI widths for the lower and upper bounds increase substantially as the ICC increases from 0.15 to 0.85, and this is a function of the variance increasing with ICC. Widths decrease substantially with increasing number of clusters (but C=100 also has a larger total N). The MAGM and AGM methods produce very similar or identical simultaneous (lower, upper) bound widths in cases where the ICC is at least 0.50 (Tables 4, 5) or where neither marginal is close to 0 or 1 (Table 5). The MAGM method has widths that are a 0-20% narrower than the AGM for low ICC and marginals close to 0 or 1 (Table 4, ICC=.15).

Joint CI width of the lower and upper bounds is much narrower when either marginal is close to zero, especially with low to moderate ICC (Figures 1 and 2). The average width of the simultaneous intervals narrows by as much as 50% as the ICC increases from 0.15 to 0.85, and this is more pronounced with larger total sample size. The average joint CI width also decreases substantially as the number of clusters is increased within a fixed sample size, particularly when the ICC is 0.50 or 0.85 (Figure 2). Across all of the scenarios assessed, the average width of the joint intervals is only 3-15 percentage points wider than the width of the true bounds. Higher values of π_{01} (and thus higher TEH) for fixed marginals increase the joint CI width (Figure 1).

Coverage of the individual true bounds was between 90% and 100% for both the AGM and MAGM methods in most situations (Tables 4 and 5, columns H and M). Coverage was above 90% under all scenarios when the ICC was 0.15 or when it was 0.50 and with 30 or more clusters (data shown for 40 and 100 clusters). However, it dropped below 90% with the combined scenario of smaller number of clusters (20), marginals closer to zero, and moderate to high ICC. In a few situations with only 10 clusters (not shown), the coverage was as low as 65-70%. With the unlikely ICC of 0.85 and marginals close to zero or one, forty or more clusters were sometimes needed to obtain coverage of at least 90%.

Simultaneous coverage of the true bounds (column O in Tables 4 and 5) is at least 90% in most cases, and often above 95%. It follows a pattern similar to coverage of the individual bounds, being best when the ICC is moderate or low and with a non-trivial number of clusters (20 or more). In most situations, the coverage was close to or slightly better than the worst of the individual lower and upper bound coverages for that scenario. The width of the simultaneous interval was sometimes narrower for the bootstrap method, but the slightly narrower width was usually accompanied by lower coverage of the true bounds. In general, coverage of the true MAGM bounds was better with the variance formula method than for the bootstrap method (as much as 0.15 better) for similar CI width.

Finally, coverage of the unobservable quantity π_{01} using the simultaneous confidence intervals (column P in Tables 4 and 5) is often 100% and nearly always above 95%. It is affected by the ICC, number of clusters, TEH scenario and total sample size with the same pattern as for the simultaneous bounds coverage.

Conclusion

AGM and refined AGM estimators have good statistical properties (low bias, MSE) and can thus be used in practice to estimate bounds for treatment effect heterogeneity with a binary outcome. Moderately or highly clustered data result in narrower confidence intervals for the measure of treatment heterogeneity π_{01}, the probability of treatment failure and control success, which is termed the treatment risk. Higher ICC is preferable because the bounds themselves move considerably closer to the parameter they are bounding, π_{01}, for larger ICC, and this phenomenon leads to narrower confidence interval widths for the simultaneous bounds as well as for $\pi_{01.}$. A moderate or large number of clusters (at least 20) and larger sample size allow more narrow confidence

Table 1. Simulation scenarios used to assess π_{01} bound estimators and components.

Scenario	Marginal Success		Heterogeneity Descriptions		Prob (Y_T=i, Y_C=j)			
	π_T	π_C	TEH	ρ_{PO}^{1}	π_{00}	π_{01}	π_{10}	π_{11}
1	0.20	0.10	Low	.58	.79	.01	.11	.09
2	“	“	Med	.00	.72	.08	.18	.02
3	0.55	0.45	Low	.78	.44	.01	.11	.44
4	“	“	Med	.00	.25	.20	.30	.25
5	“	“	High	-.80	.05	.40	.50	.05

Note: [1] = correlation among potential outcomes on T, C

Table 2. Bias, variance and MSE for Scenario #1 (low TEH + marginals near 0).

			PROPERTY					
θ	ICC	# Clusters	E(θ)	E($\hat{\theta}$)	E($\theta - \hat{\theta}$)	E($\hat{V}(\hat{\theta})$)	$V\left(\hat{\theta}\right)$	MSE
LB	0.15	20	0.0000	0.0012	0.0012	0.0001	0.0000	0.0000
		40	.	0.0009	0.0009	0.0001	0.0000	0.0000
		100	.	0.0001	0.0001	0.0000	0.0000	0.0000
	0.5	20	0.0000	0.0063	0.0063	0.0003	0.0001	0.0002
		40	.	0.0061	0.0061	0.0002	0.0001	0.0001
		100	.	0.0030	0.0030	0.0001	0.0000	0.0000
	0.85	20	0.0070	0.0149	0.0079	0.0006	0.0005	0.0005
		40	.	0.0162	0.0092	0.0005	0.0003	0.0004
		100	.	0.0107	0.0037	0.0001	0.0001	0.0001
UB	0.15	20	0.1000	0.0998	-.0002	0.0014	0.0010	0.0010
		40	.	0.1008	0.0008	0.0010	0.0006	0.0006
		100	.	0.1004	0.0004	0.0003	0.0002	0.0002
	0.5	20	0.0900	0.0804	-.0096	0.0015	0.0016	0.0017
		40	.	0.0831	-.0069	0.0009	0.0009	0.0009
		100	.	0.0874	-.0026	0.0003	0.0003	0.0003
	0.85	20	0.0340	0.0268	-.0072	0.0010	0.0009	0.0009
		40	.	0.0304	-.0036	0.0007	0.0006	0.0006
		100	.	0.0344	0.0004	0.0002	0.0002	0.0002

Notes:Marginals: π_T = .20 , π_C= .10; P(Y_T=i,Y_C=j): π_{00}= .79 , π_{01}= .01, π_{10}= .11 , π_{11}= .09; Total N=600 (for C=20, 40), N=300 (for C=100); 1,000 simulations per scenario.

Table 3. Bias, variance and MSE for Scenario #5 (high TEH + marginals near 0.5).

					PROPERTY			
θ	ICC	# Clusters	E(θ)	E($\hat{\theta}$)	E($\theta - \hat{\theta}$)	E($\hat{V}(\hat{\theta})$)	$V\left(\hat{\theta}\right)$	MSE
LB	0.15	20	0.0218	0.0377	0.0159	0.0021	0.0014	0.0017
		40	.	0.0334	0.0116	0.0014	0.0010	0.0011
		100	.	0.0266	0.0049	0.0004	0.0004	0.0004
	0.5	20	0.1775	0.1891	0.0116	0.0068	0.0066	0.0068
		40	.	0.1863	0.0088	0.0039	0.0039	0.0040
		100	.	0.1815	0.0040	0.0014	0.0015	0.0015
	0.85	20	0.3333	0.3447	0.0114	0.0106	0.0112	0.0114
		40	.	0.3434	0.0102	0.0058	0.0060	0.0061
		100	.	0.3382	0.0050	0.0022	0.0021	0.0022
UB	0.15	20	0.4425	0.4284	-.0141	0.0022	0.0021	0.0023
		40	.	0.4248	-.0177	0.0015	0.0013	0.0016
		100	.	0.4365	-.0060	0.0005	0.0004	0.0005
	0.5	20	0.4250	0.4084	-.0166	0.0063	0.0062	0.0065
		40	.	0.4082	-.0168	0.0035	0.0035	0.0038
		100	.	0.4192	-.0058	0.0013	0.0013	0.0014
	0.85	20	0.4075	0.3954	-.0121	0.0103	0.0105	0.0106
		40	.	0.3930	-.0145	0.0056	0.0054	0.0056
		100	.	0.4047	-.0028	0.0021	0.0019	0.0019

Notes: Marginals: π_T =.55, π_C=.45; P(Y_T=i, Y_C=j): π_{00}= .05 , π_{01}= .40, π_{10}= .50, π_{11}= .05
Total N=600 (for C=20, 40), N=300 (for C=100); 1000 simulations per scenario.

Table 4. CI width and coverage of bounds on π_{01} for scenario 1: Low heterogeneity and marginals near zero.

													Simultaneous		
			Lower Bound(LB)					Upper Bound(UB)					Lower, Upper		
ICC	#C/#U	Meth	True	L95	U95	W	Cov	True	L95	U95	W	Cov	W	Cov	$C\pi_{01}$
.15	20/30	AGM	.000	.00	.02	.02	1.0	.15	.07	.22	.15	.90	.22	.92	1.0
		MAGM	.000	.00	.02	.02	1.0	.10	.03	.17	.15	.97	.17	.97	1.0
	40/15	AGM	.000	.00	.02	.02	1.0	.15	.09	.21	.12	.93	.21	.95	1.0
		MAGM	.000	.00	.02	.02	1.0	.10	.04	.16	.12	.97	.16	.98	1.0
	100/30	AGM	.000	.00	.01	.01	1.0	.15	.11	.18	.07	.94	.18	.96	1.0
		MAGM	.000	.00	.01	.01	1.0	.10	.07	.13	.07	.98	.13	.99	1.0
.50	20/30	AGM	.000	.00	.03	.03	1.0	.09	.02	.16	.14	.88	.16	.89	1.0
		MAGM	.000	.00	.03	.03	.99	.09	.01	.15	.14	.86	.15	.87	1.0
	40/15	AGM	.000	.00	.03	.03	1.0	.09	.03	.15	.12	.91	.15	.91	1.0
		MAGM	.000	.00	.03	.03	1.0	.09	.03	.14	.11	.89	.14	.89	1.0
	100/30	AGM	.000	.00	.02	.02	1.0	.09	.06	.12	.07	.94	.12	.95	1.0
		MAGM	.000	.00	.02	.02	1.0	.09	.05	.12	.07	.93	.12	.94	1.0
.85	20/30	AGM	.007	.00	.05	.05	.87	.03	.00	.08	.08	.76	.08	.76	.92
		MAGM	.007	.00	.05	.05	.87	.03	.00	.08	.08	.75	.08	.75	.92
	40/15	AGM	.007	.00	.05	.05	.96	.03	.00	.08	.08	.89	.08	.89	.98
		MAGM	.007	.00	.05	.05	.96	.03	.00	.08	.08	.89	.08	.89	.98
	100/30	AGM	.007	.00	.03	.03	.96	.03	.01	.06	.06	.92	.06	.93	1.0
		MAGM	.007	.00	.03	.03	.96	.03	.01	.06	.06	.92	.06	.93	1.0

Legend: Table values are means over 1000 simulations, except for columns labeled 'True' values ICC= Dirichlet multinomial correlation; #C= number of clusters, #U= number of units per cluster AGM=Equation 2.2; MAGM=Equation 2.6; W=width of 95% CI= U95-L95; Cov=coverage; Simultaneous: coverage of both Lb and UB using L95 of LB, U95 of UB; $C\pi_{01}$: coverage of π_{01} using L95 of LB, U95 of UB

Table 5. CI width and coverage of bounds on π_{01} for scenario 5: High heterogeneity and marginals near 0.50.

													Simultaneous		
			Lower Bound(LB)					Upper Bound(UB)					Lower, Upper		
ICC	#C/#U	Meth	True	L95	U95	W	Cov	True	L95	U95	W	Cov	W	Cov	Cπ_{01}
.15	20/30	AGM	.022	.00	.12	.12	.99	.44	.35	.53	.18	.93	.53	.94	1.0
		MAGM	.022	.00	.12	.12	.99	.44	.34	.52	.18	.92	.52	.92	1.0
	40/15	AGM	.022	.00	.11	.11	.98	.44	.36	.51	.15	.95	.51	.94	1.0
		MAGM	.022	.00	.11	.11	.98	.44	.35	.50	.15	.94	.50	.92	1.0
	100/30	AGM	.022	.00	.07	.06	.98	.44	.40	.48	.08	.94	.48	.94	1.0
		MAGM	.022	.00	.07	.06	.98	.44	.39	.48	.08	.94	.48	.93	1.0
.50	20/30	AGM	.178	.04	.35	.31	.93	.43	.26	.57	.31	.93	.53	.93	.97
		MAGM	.178	.04	.35	.31	.93	.43	.25	.56	.31	.92	.52	.92	.97
	40/15	AGM	.178	.07	.31	.24	.95	.43	.30	.53	.23	.93	.46	.93	.98
		MAGM	.178	.07	.31	.24	.95	.43	.29	.52	.23	.93	.46	.92	.98
	100/30	AGM	.178	.11	.25	.15	.94	.43	.35	.49	.14	.95	.38	.94	1.0
		MAGM	.178	.11	.25	.15	.94	.43	.35	.49	.14	.95	.38	.94	1.0
.85	20/30	AGM	.333	.14	.55	.40	.93	.41	.20	.60	.40	.92	.45	.92	.95
		MAGM	.333	.14	.55	.40	.93	.41	.20	.59	.40	.92	.45	.92	.94
	40/15	AGM	.333	.19	.49	.30	.95	.41	.25	.54	.29	.93	.35	.93	.96
		MAGM	.333	.19	.49	.30	.95	.41	.25	.54	.29	.93	.35	.92	.95
	100/30	AGM	.333	.25	.43	.18	.95	.41	.31	.50	.18	.96	.25	.95	.98
		MAGM	.333	.25	.43	.18	.95	.41	.31	.50	.18	.96	.25	.95	.98

Legend: Table values are means over 1000 simulations, except for columns labeled 'True' values ICC= Dirichlet multinomial correlation; #C= number of clusters, #U= number of units per cluster AGM=Equation 2.2; MAGM=Equation 2.6; W=width of 95% CI= U95-L95; Cov=coverage; Simultaneous: coverage of both Lb and UB using L95 of LB, U95 of UB; Cπ_{01} : coverage of π_{01} using L95 of LB, U95 of UB

Fig 1. Width of 95% CI for $\pi01$ by ICC and heterogeneity scenario

N=20 clusters, 30 units per cluster

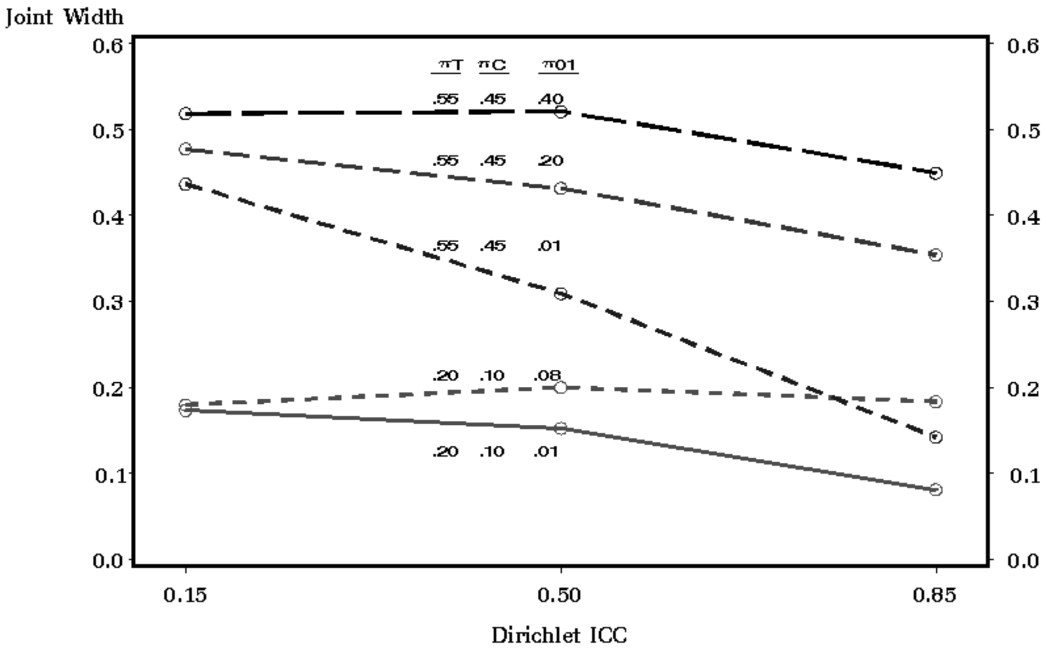
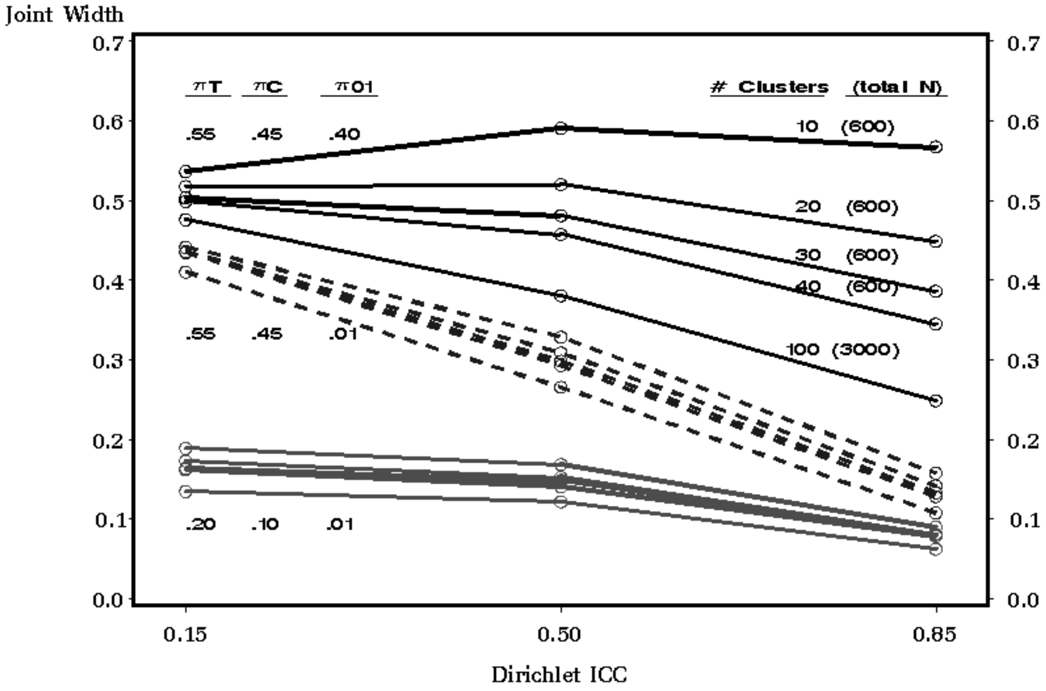


Fig 2. Width of 95% CI for $\pi01$ by ICC, heterogeneity scenario and # clusters



intervals for the individual bounds, the simultaneous bounds and for π_{01}.

The effect of ICC on confidence interval widths is more dramatic in the case where the marginal success probabilities are closer to 0.5. In this case, when there is high heterogeneity (π_{01}=0.4), 95% CI widths for π_{01} are reduced from around 0.5 (at ICC=0.15) to as low as 0.3 (at ICC = 0.8), and a similar reduction in width (from roughly 0.4 to 0.2) is seen in the low heterogeneity (π_{01}=0.01). This is important because CI widths of more than .20 or so are unlikely to be very useful.

Although nominal or near-nominal coverage of the true bounds was attained for most of the scenarios considered, the estimators did not give sufficient coverage of either the individual bounds or the simultaneous bounds with the combination of very high ICC and small number of clusters (20 or less) when using the fixed total sample size of 600. In results not presented, it was found that using less than 20 clusters (specifically, 10) gave very poor coverage in most scenarios. Creating a confidence interval estimator which directly takes into account the number of clusters and the ICC might greatly improve the coverage in these outlying situations.

These methods assume that the observed data consist of clusters (or blocks) that are either natural or can be created post-hoc. Post-hoc clusters can be created by first predicting the observed outcome on either T or C using all available baseline covariables, excluding treatment group, and then grouping patients by percentiles of their predicted probability of success. In order to be able to apply these methods and obtain appropriately narrow confidence intervals on bound estimators, studies would best collect data on as many baseline covariables as feasible. SAS macros will soon be available to calculate the bound estimators and confidence intervals.

Confidence intervals for the treatment risk could be used in several ways in practice. First is the case where the lower confidence limit on treatment risk is zero, and the interval width is small. Being able to conclude that the new intervention is expected to be successful for a certain proportion of the existing treatment failures, but not likely to change any of the existing treatment successes, seems ideal. But a non-zero upper bound estimate would imply that the treatment risk may be non-zero, and this may provoke interest, concern and perhaps more research. Second, if the lower estimated confidence limit was above zero, non-zero treatment risk would be concluded, and researchers would best search for patient subsets that would be better off with the standard treatment. Researchers for a new drug or treatment would likely be more satisfied with an intervention that had very low probability of failing in patients already expected or known to have success on the standard treatment.

For individual decision-making, the confidence intervals on treatment risk might be useful in some situations. An individual with no experience with either intervention might well choose the one with the largest observed marginal success, regardless of the estimated bounds on the treatment risk. On the other hand, if it was believed that the treatment risk was high, an individual with known or supposed success on the control might be hesitant to switch to an intervention with greater marginal success, even with fewer expected side effects. The gamble would be more likely if the treatment risk was thought to be low. In future work, study of the methods of using covariate information to help predict an individual’s underlying category is planned.

The Dirichlet multinomial (DMN) was found to be a useful model for assessing the statistical properties of estimators for bounds on treatment effect heterogeneity because the ICC can be directly specified and because of the natural clumping of the data with higher ICC. One potential limitation of the DMN for this work is that the covariance structure is based on the underlying proportion of individuals in each category, and the corresponding structure of the intraclass between-category correlations may not be intuitive for some real situations. However, there is no reason to believe that an underlying model, allowing full specification of the covariance between the four categories of interest, would yield substantially different property assessment results. Because the parameters of interest are non-estimable (only one of two potential outcomes is observed for each unit or individual), without distributional assumptions, at best bounds may be put on the parameters of interest.

References

Albert, J. M., Gadbury, G. L, & Mascha, E. J. (2005) Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biometrical Journal, 47,* 662-673.

Bickel, P. & Friedman, D. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics, 9*, 1196-1217.

Brier, S. (1980). Analysis of contingency tables under cluster sampling. *Biometrika, 67*(3), 591-596.

Banerjee, T. & Paul, S. (1999). An extension of Morel-Nagaraj's finite mixture distribution for modeling multinomial clustered data. *Biometrika, 86*(3), 723-727.

Gadbury, G. L., & Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics, 56,* 882-885.

Gadbury, G. L, Iyer, H. K., & Albert, J. M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference, 121,* 163-174.

Gadbury, G. L., Iyer, H. K., & Allison, D. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics, 11*(4) , 313-333.

Gange, S. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician, 95*(49), 134-138.

Horowitz, J. & Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association, 95*, 77-88.

Jensen, D. R. (1998). Multivariate distributions. In *Encyclopedia of Biostatistics*. Chichester, England: John Wiley & Sons, 2857.

Mascha, E. J. & Albert, J. M. (2006). Estimating treatment effect heterogeneity for binary outcomes via Dirichlet multinomial constraints. *Biometrical Journal* (in press).

Morel, J. & Nagaraj, N. (1993). A finite mixture distribution for modeling multinomial extra variation. *Biometrika, 80* (2), 363-71.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate B-distribution and correlation among proportions. *Biometrika, 49,* 65-82.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Essay on Principles*, Section 9. Translated in *Statistical Science* (1990) 5, 465-480.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. B. (2000). Comment on 'Causal inference without counterfactuals', by A. P. Dawid, *Journal of the American Statistical Association*, 95, 435-437.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 194-216

Two New Unbiased Point Estimates Of A Population Variance

Matthew E. Elam
Department of Industrial Engineering
The University of Alabama

Two new unbiased point estimates of an unknown population variance are introduced. They are compared to three known estimates using the mean-square error (MSE). A computer program, which is available for download at http://program.20m.com, is developed for performing calculations for the estimates.

Key words: Unbiased, point estimate, variance, range, standard deviation, moving range.

Introduction

The statistical analysis of sample data often involves determining point estimates of unknown population parameters. A desirable property for these point estimates is that they be unbiased. An unbiased point estimate has an expected value (or mean) equal to the unknown population parameter it is being used to estimate. For example, consider the mean $\overline{x}$ and variance v calculated from a random sample of size n (x_1, x_2, …, x_n) obtained from a population with unknown mean μ and variance σ^2. The equations for these two statistics are equations (1) and (2):

$$\overline{x} = \sum_{i=1}^{n} x_i / n \tag{1}$$

$$v = \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 / (n-1) \tag{2}$$

It is well known that $\overline{x}$ and v are unbiased point estimates of μ and σ^2, respectively (e.g., see Theorems 8.2.1 and 8.2.2, respectively, in Bain & Engelhardt, 1992). This means the expected value of the sampling distribution of $\overline{x}$ is equal to μ (i.e., $E(\overline{x})=\mu$) and the expected value of the sampling distribution of v is equal to σ^2 (i.e., $E(v)=\sigma^2$).

It is important to have a sample that is random when calculating unbiased point estimates of unknown population parameters. In a random sample, each value comes from the same population distribution. If the values come from different population distributions (i.e., populations with different distributions, means, and/or variances), then the point estimates they are used to calculate will be inaccurate. For example, if the values come from population distributions with different means, then v calculated from this sample using equation (2) will be inflated.

Many situations exist in which it is difficult to obtain a random sample. One of these is when the population is not well-defined, as is the case when studying on-going processes. On-going processes are often encountered in manufacturing situations. An approach to obtain unbiased point estimates of unknown population parameters from these types of processes is to collect data as some number m of subgroups, each having size n. This is the procedure that is used when constructing control charts to monitor the centering and/or spread of a process. The idea is for the data within a subgroup to come from the same process distribution. If any changes are to occur in the process distribution, it is desirable for them to show up between subgroups. An additional procedure in control chart construction, which may be called a delete-and-revise (D&R) procedure, is performed as an additional safeguard to ensure data within subgroups has the same distribution.

Matthew E. Elam is Assistant Professor of Industrial Engineering at The University of Alabama. He is a member of the ASQ and IIE, and is an ASQ Certified Quality Engineer. Email him at melam@bama.ua.edu.

Two new unbiased point estimates of an unknown population variance are introduced. They are derived assuming the sample data is drawn from an on-going process as m subgroups, each of size n. The Methodology section has an example showing how the control charting procedure works. Also, it presents the three known unbiased point estimates used in the situation considered in this article, it derives the two new unbiased point estimates, and it explains a Mathcad (1999) computer program that performs calculations for the unbiased point estimates. The Results section has mean-square error (MSE) results for the unbiased point estimates. These are useful for the purpose of comparing the unbiased point estimates. The Conclusion section summarizes the interpretations of the analyses in the Results section.

Methodology

Control Charting Procedure. Consider the data in Table 1 obtained from a normally distributed process with μ=100.0 and σ=7.0 (the data was generated in Minitab (2003) and a few changes were made to simulate a process with a nonconstant mean). The true unknown variability for the process is estimated using within subgroup variability. A control chart for spread may be used to determine if data within a subgroup comes from the same process distribution. The control chart for spread used here is the range (R) chart. It is constructed using equations (3a)-(3c):

$$\mathrm{UCL} = \mathrm{D}_4 \times \overline{\mathrm{R}} \tag{3a}$$

$$\mathrm{CL} = \overline{\mathrm{R}} \tag{3b}$$

$$\mathrm{LCL} = \mathrm{D}_3 \times \overline{\mathrm{R}} \tag{3c}$$

UCL, CL, and LCL are the upper control limit, center line, and lower control limit, respectively, for the R chart. Values for the control chart factors D_4 and D_3 for various n are widely available in control chart factor tables (e.g., see Table M in the appendix of Duncan, 1974). The value $\overline{\mathrm{R}}$ (Rbar) is the mean of the m subgroup ranges. The subgroup ranges are calculated for each subgroup as the maximum value in the subgroup minus the minimum value in the subgroup (these calculations are in the "R" column of Table 1). Equations (4a)-(4c) are the R chart control limit calculations for the data in Table 1:

$$\mathrm{UCL} = \mathrm{D}_4 \times \overline{\mathrm{R}} = 2.282 \times 13.584 = 30.999 \tag{4a}$$

$$\mathrm{CL} = \overline{\mathrm{R}} = 13.584 \tag{4b}$$

$$\mathrm{LCL} = \mathrm{D}_3 \times \overline{\mathrm{R}} = 0.0 \times 13.584 = 0.0 \tag{4c}$$

Figure 1 is the R control chart generated in Minitab (2003).

The delete-and-revise (D&R) procedure involves identifying any subgroup ranges that are greater than the UCL or less than the LCL. The identified subgroups are then removed from the analysis as long as, in this case, each identified subgroup was an indication of a shift in the process mean. The R chart control limits are recalculated using the remaining subgroups. For the Table 1 data, the range (R) for subgroup seven is above the UCL (see the point marked with a "1" in Figure 1). The new value for $\overline{\mathrm{R}}$ calculated using the remaining m=19 subgroups after subgroup seven is removed is shown as the Revised $\overline{\mathrm{R}}$ in Table 1. The revised control limits are calculated in equations (5a)-(5c):

$$\mathrm{UCL} = \mathrm{D}_4 \times \overline{\mathrm{R}} = 2.282 \times 12.604 = 28.762 \tag{5a}$$

$$\mathrm{CL} = \overline{\mathrm{R}} = 12.604 \tag{5b}$$

$$\mathrm{LCL} = \mathrm{D}_3 \times \overline{\mathrm{R}} = 0.0 \times 12.604 = 0.0 \tag{5c}$$

Because all of the remaining subgroup ranges are between the revised control limits, the conclusion is that the data within each subgroup comes from the same process distribution.

Table 1. Data Collected as m=20 Subgroups, Each of Size n=4

Subgroup	X_1	X_2	X_3	X_4	R
1	89.558	99.593	99.069	91.211	10.035
2	98.263	98.745	96.959	102.132	5.173
3	93.246	108.054	98.811	102.767	14.808
4	95.493	94.852	109.277	98.418	14.425
5	109.667	108.467	88.994	105.678	20.673
6	94.636	105.764	93.755	88.376	17.388
7	88.000	108.000	113.203	81.000	32.203
8	112.215	104.877	97.752	104.484	14.463
9	87.578	90.221	108.198	99.202	20.620
10	100.029	92.639	96.211	94.332	7.390
11	97.998	101.717	98.704	92.989	8.728
12	107.147	102.370	103.020	95.581	11.566
13	94.597	105.221	103.527	94.565	10.656
14	110.381	93.632	103.740	102.841	16.749
15	96.551	104.145	102.043	102.206	7.594
16	108.505	100.040	99.048	110.904	11.856
17	107.918	104.065	94.514	93.943	13.975
18	114.000	116.000	121.000	123.000	9.000
19	109.304	99.160	97.338	114.353	17.015
20	96.920	104.280	100.290	101.984	7.360
				$\overline{R}$	13.584
				Revised $\overline{R}$	12.604

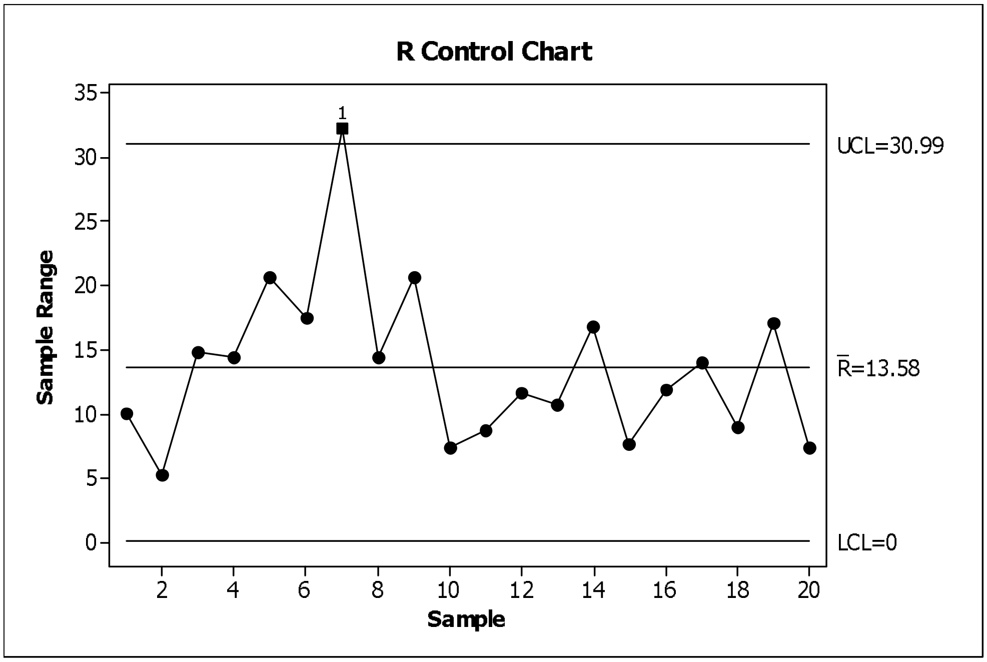


Figure 1. R Control Chart for the Data in Table 1

The next two subsections, Known Unbiased Point Estimates of σ^2 and Two New Unbiased Point Estimates of σ^2, explain how data collected and cleaned in this manner is used to obtain an unbiased point estimate of an unknown process variance using the following statistics:

- $\bar{v}$, the mean of the subgroup variances, where each subgroup variance is calculated using equation (2).
- v_c, the variance of the m×n data values grouped together as one sample. It is calculated using equation (2) with n replaced by m×n and with $\bar{x}$ calculated using equation (1), also with n replaced by m×n. It should be noted that v_c cannot be used when cleaning subgrouped data using a delete-and-revise (D&R) procedure as explained in this subsection. The reason is it would include between subgroup variability, which would inflate its value if the process from which the data is collected is operating under multiple distributions.
- $\bar{R}$, as previously demonstrated.
- $\bar{s}$, the mean of the subgroup standard deviations, where each subgroup standard deviation is calculated using the square root of equation (2).
- $\overline{MR}$, the mean of the moving ranges. When data is collected as m individual values, m-1 moving ranges may be calculated as the absolute value of the difference between consecutive individual values. In this case, the subgroup size n is taken to be two. For example, if the first three individual values are 5.1, 5.3, and 4.8, the first two moving ranges are |5.1-5.3|=0.2 and |5.3-4.8|=0.5.

Known Unbiased Point Estimates of σ^2

The three known unbiased point estimates of σ^2 calculated from data collected as m subgroups, each of size n, considered in this article are $\bar{v}$, v_c, and $\left(\bar{R}/d_2^*\right)^2$. The unbiasedness of $\bar{v}$ is shown in the Appendix of Elam and Case (2003). Wheeler (1995), in his Tables 3.6, 3.7, and 4.2, indicated the unbiasedness for $\bar{v}$ (listed as the pooled variance) as well as for $\left(\bar{R}/d_2^*\right)^2$. The value d_2^* may be called an unbiasing factor, as $\left(\bar{R}\right)^2$ by itself is a biased point estimate of σ^2. The value d_2^* is tabled for various m and n (e.g., see Table D3 in the appendix of Duncan, 1974).

David (1951) gave the equation for d_2^* (i.e., d2star) as equation (6):

$$\mathrm{d2star} = \sqrt{\mathrm{d2}^2 + \mathrm{d3}^2/\mathrm{m}} \qquad (6)$$

The value d2 (i.e., d_2) is the mean of the distribution of the range W. Its values for various n are widely available in control chart factor tables. Assuming a normal population with mean μ and variance equal to one, Harter (1960) gave the equation for d2 as equation (7) (with some modifications in notation):

$$\mathrm{d2} = n\times(n-1)\times\int_{-\infty}^{\infty}\left[\int_0^{\infty} W\times(F(x+W) - F(x))^{n-2}\times f(x+W)dW\right]\times f(x)dx \qquad (7)$$

The function F(x) is the cumulative distribution function (cdf) of the standard normal probability density function (pdf) f(x). The value d3 (i.e., d_3) is the standard deviation of the distribution of the range W. Its values for various n are widely available in control chart factor tables. It is calculated using equation (8):

$$\mathrm{d3} = \sqrt{\mathrm{EW2} - \mathrm{d2}^2} \qquad (8)$$

Harter (1960) gave the equation for EW2, the expected value of the second moment of the distribution of the range W for subgroups of size n sampled from a normal population with mean μ and variance equal to one, as equation (9) (with some modifications in notation):

$$\mathrm{EW2} = n\times(n-1)\times\int_{-\infty}^{\infty}\left[\int_0^{\infty} W^2\times(F(x+W) - F(x))^{n-2}\times f(x+W)dW\right]\times f(x)dx \qquad (9)$$

Equations (6)-(9) are the forms used in the Mathcad (1999) computer program explained in the Mathcad (1999) Computer Program subsection.

Two New Unbiased Point Estimates of σ^2

Elam and Case (2005a), in their Appendix 7, derived the equation for the factor that allows for an unbiased point estimate of σ^2 to be calculated using $\bar{s}$. Elam and Case (2005a) denoted this factor as c_4^* (i.e., c4star) and gave its equation as equation (10):

$$\text{c4star} = \sqrt{\text{c4}^2 + \text{c5}^2/\text{m}} \qquad (10)$$

The fact that $\left(\bar{s}/c_4^*\right)^2$ is an unbiased point estimate of σ^2 is shown in the Appendix. In equation (10), the value c4 (i.e., c_4) is the mean of the distribution of the standard deviation. Its values for various n are widely available in control chart factor tables. Mead (1966) gave the equation for c4 as equation (11) when σ=1.0 (with some modifications in notation):

$$\text{c4} = \sqrt{2/(\text{n}-1)} \times \exp(\text{gammln}(\text{n}/2) - \text{gammln}((\text{n}-1)/2)) \qquad (11)$$

The equivalency of this form to that given by Mead (1966) is shown in Appendix 3 of Elam and Case (2005a). The function gammln represents the natural logarithm of the gamma (Γ) function. The value c5 (i.e., c_5) is the standard deviation of the distribution of the standard deviation. Mead (1966) also gave the equation for c5 as equation (12) when σ=1.0 (with some modifications in notation):

$$\text{c5} = [(2/(\text{n}-1)) \times [\exp(\text{gammln}((\text{n}+1)/2) - \text{gammln}((\text{n}-1)/2)) - \exp(2 \times (\text{gammln}(\text{n}/2) - -\text{gammln}((\text{n}-1)/2)))]]^{0.5} \qquad (12)$$

The equivalency of this form to that given by Mead (1966) is shown in Appendix 4 of Elam and Case (2005a). The value c5 is also equal to $\sqrt{1-c_4^2}$. Equations (10)-(12) are the forms used in the Mathcad (1999) computer program explained in the Mathcad (1999) Computer Program subsection.

Elam and Case (2006a), in Appendix 1, derived the equation for the factor that allows for an unbiased point estimate of σ^2 to be calculated using $\overline{\text{MR}}$. Elam and Case (2006a) denoted this factor as $d_2^*(\text{MR})$ (i.e., d2starMR) and gave its equation as equation (13):

$$\text{d2starMR} = \sqrt{\text{d2n2}^2 + \text{d2n2}^2 \times \text{r}} \qquad (13)$$

The fact that $\left(\overline{\text{MR}}/d_2^*(\text{MR})\right)^2$ is an unbiased point estimate of σ^2 is shown in the Appendix. In equation (13), the value d2n2 is d2 when n is equal to two. Harter (1960) gave the equation for d2n2 as equation (14) (with some modifications in notation):

$$\text{d2n2} = 2/\sqrt{\pi} \qquad (14)$$

The value r is the ratio of the variance to the squared mean, both of the distribution of the mean moving range $\overline{\text{MR}}/\sigma$, an approximation to which is derived in Elam and Case (2006a). Palm and Wheeler (1990) gave the equation for r as equation (15):

$$\text{r} = ((4 \times \pi - 18 + 2 \times 3^{1.5}) \times (\text{m}-1) - \pi + 12 - 2 \times 3^{1.5})/(6 \times (\text{m}-1)^2) \qquad (15)$$

Equations (13)-(15) are the forms used in the Mathcad (1999) computer program explained in the Mathcad (1999) Computer Program subsection.

Mathcad (1999) Computer Program

A computer program was coded in Mathcad (1999) with the Numerical Recipes Extension Pack (1997) in order to calculate the unbiasing factors d_2^*, c_4^*, and $d_2^*(\text{MR})$ in equations (6), (10), and (13), respectively, regardless of the number of subgroups m and the subgroup size n. The program is in the Appendix and is named UEFactors.mcd. It is on one page which is divided into seven sections. Download instructions for the program are available at http://program.20m.com.

The first section of the program is the data entry section. The program requires the user to enter m (number of subgroups) and n (subgroup size). Before a value can be entered, the cursor must be moved to the right side of the appropriate equal sign. This may be done using the arrow keys on the keyboard or by moving the mouse arrow to the right side of the equal sign and clicking once with the left mouse button. The program is activated by paging down once the last entry is made. The user is allowed to immediately page down to the output section of the program (explained later) after the last entry is made.

In section 1.1 of the program, the value TOL is the tolerance. The calculations that use this value will be accurate to ten places to the right of the decimal. The functions dnorm(x, 0, 1) and pnorm(x, 0, 1) in Mathcad (1999) are the pdf and cdf, respectively, of the standard normal distribution.

Section 1.2 of the program has the equations for d2, d3, and EW2 given earlier as equations (7), (8), and (9), respectively. Section 1.3 of the program has the equations for c4 and c5 given earlier as equations (11) and (12), respectively. The function gammln is a numerical recipe in the Numerical Recipes Extension Pack (1997). Using it in equations (11) and (12) allows for c4 and c5, respectively, to be calculated for large values of n. Section 1.4 of the program has the equations for d2n2 and r, given earlier as equations (14) and (15), respectively. Section 1.5 of the program has the equations for d2star, c4star, and d2starMR, given earlier as equations (6), (10), and (13), respectively.

The last section of the program has the output. The two values entered at the beginning of the program are given. Accurate values for the unbiasing factors d_2^*, c_4^*, and $d_2^*(\mathrm{MR})$ are also given. The value for $d_2^*(\mathrm{MR})$ is always calculated for n=2, regardless of the value for n entered at the beginning of the program. To copy results into another software package (like Excel), follow the directions from Mathcad's (1999) help menu or highlight a value and copy and paste it into the other software package. When highlighting a value with the mouse arrow, place the arrow in the middle of the value, depress the left mouse button, and drag the arrow to the right. This will ensure just the numerical value of the result is copied and pasted.

Results

The two new unbiased point estimates of σ^2 are compared to the three known unbiased point estimates of σ^2 using the mean-square error (MSE) calculation in equation (16), which is based on Luko's (1996) equation (A3):

$$\mathrm{MSE}(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) + [\mathrm{E}(\hat{\sigma}^2) - \sigma^2]^2 \qquad (16)$$

The value $\hat{\sigma}^2$ represents $\bar{v}$, v_c, $\left(\bar{R}/d_2^*\right)^2$, $\left(\bar{s}/c_4^*\right)^2$, or $\left(\overline{\mathrm{MR}}/d_2^*(\mathrm{MR})\right)^2$, and Var represents the variance as calculated in equation (2). Because these five point estimates of σ^2 are all unbiased, $\mathrm{E}(\hat{\sigma}^2) - \sigma^2 = 0$. Therefore, calculating their MSEs is identical to calculating their variances. Better point estimates are those with smaller MSEs.

MSEs for $\bar{v}$, v_c, $\left(\bar{R}/d_2^*\right)^2$, and $\left(\bar{s}/c_4^*\right)^2$ are calculated using the FORTRAN (1994) computer program named "simulate" in the Appendix. The program simulates the random sampling of m subgroups (m: 1-20, 25, 30, 50, 75, 100, 150, 200, 250, 300), each of size n (n: 2-8, 10, 25, 50), from a standard normal distribution (uniform (0, 1) random variates are generated using the Marse-Roberts code (1983)). This process is repeated 5000 times for each combination of m and n in order to generate 5000 values each of $\bar{v}$, v_c, $\left(\bar{R}/d_2^*\right)^2$, and $\left(\bar{s}/c_4^*\right)^2$ so that their variances can be determined. The necessary values for d_2^* and c_4^* are taken from Table A1 in Appendix III: Tables of Elam and Case (2001) and Table A.1 in Appendix II of Elam and Case (2005b), respectively.

MSEs for $\left(\overline{\mathrm{MR}}/d_2^*(\mathrm{MR})\right)^2$ are calculated using the FORTRAN (1994) computer program named "simulate_MR" in the Appendix. The program simulates the random sampling of m subgroups (m: 2-20, 25, 30, 50, 75, 100, 150, 200, 250, 300) from a standard normal distribution (uniform (0, 1) random variates are generated using the Marse-Roberts code (1983)). This process is repeated 5000 times for each m in order to generate 5000 $\left(\overline{MR}/d_2^*(MR)\right)^2$ values so that the variance can be determined. The necessary values for $d_2^*(MR)$ are taken from Table A.1 in Appendix 2 of Elam and Case (2006b).

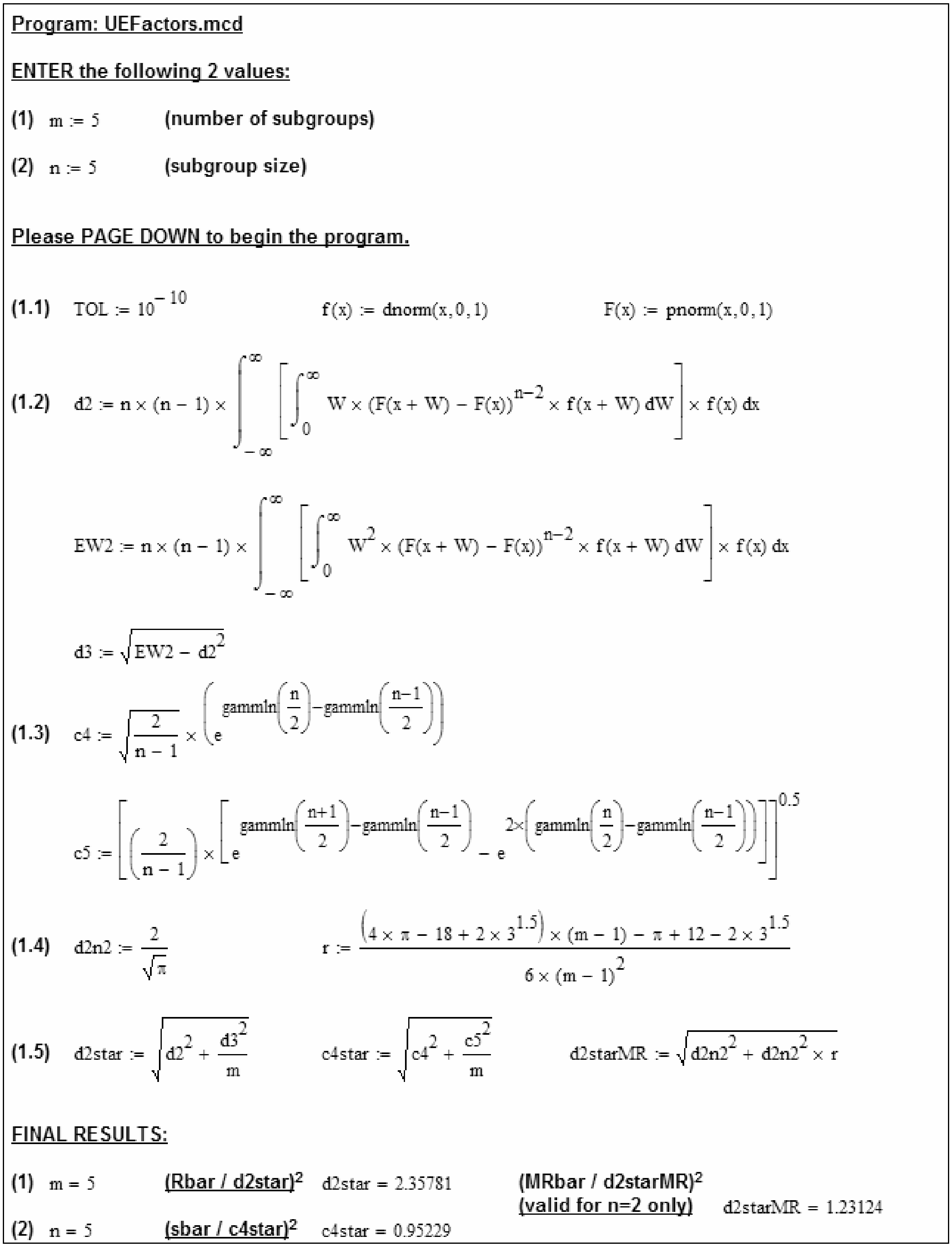
The Appendix has the MSE results for $\overline{v}$, v_c, $\left(\overline{R}/d_2^*\right)^2$, $\left(\overline{s}/c_4^*\right)^2$, and $\left(\overline{MR}/d_2^*(MR)\right)^2$ in its Tables A.1-A.5, respectively. As m increases for any n, or as n increases for any m, the MSEs in Tables A.1-A.4 decrease. As m increases, the MSEs decrease in Table A.5. This is not surprising because as more information about the process is at hand, the unbiased estimates should perform better. Only the MSEs for $\overline{v}$, v_c, $\left(\overline{R}/d_2^*\right)^2$, and $\left(\overline{s}/c_4^*\right)^2$ when n=2 and m=1 can be compared to the MSE for $\left(\overline{MR}/d_2^*(MR)\right)^2$ when m=2. In this case, the moving range is interpreted to be the same as the range. These results are the same.

Tables A.6-A.8 in the Appendix have the percent change in MSE ($\overline{v}$) over MSE (v_c), $MSE\left[\left(\overline{s}/c_4^*\right)^2\right]$ over MSE ($\overline{v}$), and $MSE\left[\left(\overline{R}/d_2^*\right)^2\right]$ over $MSE\left[\left(\overline{s}/c_4^*\right)^2\right]$, respectively. The calculations in Tables A.6-A.8 were performed using Excel's full accuracy. Because most of the percentages in these tables are zero or positive, it can be stated that, in general, MSE (v_c) ≤ MSE ($\overline{v}$) ≤ $MSE\left[\left(\overline{s}/c_4^*\right)^2\right]$ $\leq MSE\left[\left(\overline{R}/d_2^*\right)^2\right]$. The following additional conclusions can be drawn from Tables A.6-A.8:

- In Tables A.6 and A.7, the percent changes decrease as n increases for any m. This means the MSEs for $\overline{v}$, v_c, and $\left(\overline{s}/c_4^*\right)^2$ converge to each other as n increases for any m.
- The MSEs for $\overline{v}$, v_c, and $\left(\overline{s}/c_4^*\right)^2$ are the same when m=1.
- The MSE for $\left(\overline{R}/d_2^*\right)^2$ when n=2 and m=1 is almost identical to that for $\overline{v}$, v_c, and $\left(\overline{s}/c_4^*\right)^2$; however, as n gets larger for m=1 (or any m), the MSEs for $\left(\overline{R}/d_2^*\right)^2$ grow larger than those for $\overline{v}$, v_c, and $\left(\overline{s}/c_4^*\right)^2$. This is because of the well known fact that the range calculation loses efficiency as the size of the sample from which it is calculated increases.
- The MSEs for $\left(\overline{R}/d_2^*\right)^2$ and $\left(\overline{s}/c_4^*\right)^2$ when n=2 are almost identical. This is because the range and standard deviation calculations differ by only a constant when n=2.

Conclusion

From the analyses in the Results section, it may be concluded that $\left(\overline{s}/c_4^*\right)^2$ is at least as good of an unbiased point estimate of σ^2 as $\left(\overline{R}/d_2^*\right)^2$. In fact, as n increases for any m, $\left(\overline{s}/c_4^*\right)^2$ becomes a much better unbiased point estimate of σ^2 than $\left(\overline{R}/d_2^*\right)^2$. Also, the performance of $\left(\overline{s}/c_4^*\right)^2$ approaches that of $\overline{v}$ and v_c as n increases for any m. Additionally, $\left(\overline{MR}/d_2^*(MR)\right)^2$ appears to be an adequate unbiased point estimate of σ^2, as indicated by its reasonably small MSE values. This means that, for the first time, there is an alternative to equation (2) for obtaining an unbiased point estimate of σ^2 from individual values.



Program: UEFactors.mcd

ENTER the following 2 values:

(1) $m := 5$ **(number of subgroups)**

(2) $n := 5$ **(subgroup size)**

Please PAGE DOWN to begin the program.

(1.1) $TOL := 10^{-10}$ $\quad f(x) := dnorm(x, 0, 1)$ $\quad F(x) := pnorm(x, 0, 1)$

(1.2)

$$d2 := n \times (n-1) \times \int_{-\infty}^{\infty} \left[\int_{0}^{\infty} W \times (F(x+W) - F(x))^{n-2} \times f(x+W)\, dW \right] \times f(x)\, dx$$

$$EW2 := n \times (n-1) \times \int_{-\infty}^{\infty} \left[\int_{0}^{\infty} W^2 \times (F(x+W) - F(x))^{n-2} \times f(x+W)\, dW \right] \times f(x)\, dx$$

$$d3 := \sqrt{EW2 - d2^2}$$

(1.3)

$$c4 := \sqrt{\frac{2}{n-1}} \times \left(e^{gammln\left(\frac{n}{2}\right) - gammln\left(\frac{n-1}{2}\right)} \right)$$

$$c5 := \left[\left(\frac{2}{n-1} \right) \times \left[e^{gammln\left(\frac{n+1}{2}\right) - gammln\left(\frac{n-1}{2}\right)} - e^{2 \times \left(gammln\left(\frac{n}{2}\right) - gammln\left(\frac{n-1}{2}\right) \right)} \right] \right]^{0.5}$$

(1.4) $d2n2 := \frac{2}{\sqrt{\pi}}$ $\quad r := \frac{\left(4 \times \pi - 18 + 2 \times 3^{1.5}\right) \times (m-1) - \pi + 12 - 2 \times 3^{1.5}}{6 \times (m-1)^2}$

(1.5) $d2star := \sqrt{d2^2 + \frac{d3^2}{m}}$ $\quad c4star := \sqrt{c4^2 + \frac{c5^2}{m}}$ $\quad d2starMR := \sqrt{d2n2^2 + d2n2^2 \times r}$

FINAL RESULTS:

(1) $m = 5$ **(Rbar / d2star)^2** $d2star = 2.35781$ **(MRbar / d2starMR)^2** **(valid for n=2 only)** $d2starMR = 1.23124$

(2) $n = 5$ **(sbar / c4star)^2** $c4star = 0.95229$

```
program simulate
implicit none
INTEGER, parameter :: DOUBLE=SELECTED_REAL_KIND(p=15)
real(kind=double) :: mean, sd, pi, d2star, c4star, r1, r2, X, large, small, v, s, R, vc
real(kind=double) :: sumvc, sumvc2, sumvbar, sumvbar2, sumsbar2, sumsbar22, sumRbar2, sumRbar22
real(kind=double) :: sumX, sumX2, sumv, sums, sumR, sumXsv, sumX2sv
real(kind=double) :: vbar, sbar2, Rbar2, varvc, varvbar, varsbar2, varRbar2
INTEGER :: c, b, a, rep, i, j, seed = 1973272912
integer, dimension(1:29) :: m
integer, dimension(1:10) :: n
open(unit=1, file="simulate.txt")
open(unit=2, file="d2star.txt")
open(unit=3, file="c4star.txt")

mean = 0.0
sd = 1.0
pi = ACOS(-1.0)
m = (/ (c, c = 1, 20), 25, 30, 50, 75, 100, 150, 200, 250, 300 /)
n = (/ 2, 3, 4, 5, 6, 7, 8, 10, 25, 50 /)

write(1, 5) "n", "m", "c4star", "d2star", "varvc", "varvbar", "varsbar2", "varRbar2"
5 format(2X, A, 3X, A, 2X, A, 2X, A, 5X, A, 8X, A, 5X, A, 5X, A)

do b = 1, 10
! n loop

  do a = 1, 29
!   m loop

    sumvc = 0.0
    sumvc2 = 0.0
    sumvbar = 0.0
    sumvbar2 = 0.0
    sumsbar2 = 0.0
    sumsbar22 = 0.0
    sumRbar2 = 0.0
    sumRbar22 = 0.0

    read(2, *) d2star
    read(3, *) c4star

    do rep = 1, 5000
!     replication loop

      sumX = 0.0
      sumX2 = 0.0
      sumv = 0.0
      sums = 0.0
      sumR = 0.0

      do i = 1, m(a)

        sumXsv = 0.0
        sumX2sv = 0.0

!       new subgroup

        do j = 1, n(b)

          call random(r1, seed)
          call random(r2, seed)

          X = mean + sd * ((SQRT(-2. * LOG(r1))) * (COS(2. * pi * r2)))

          sumX = sumX + X
          sumX2 = sumX2 + X**(2.0)
          sumXsv = sumXsv + X
          sumX2sv = sumX2sv + X**(2.0)

          if (j == 1) then
            large = X
            small = X
          else
            if (X > large) large = X
            if (X < small) small = X
          end if

        end do
        v = (sumX2sv - ((sumXsv)**(2.0)) / n(b)) / (n(b)-1)
        s = v**(0.5)
        R = large - small

        sumv = sumv + v
        sums = sums + s
        sumR = sumR + R

      end do

      vc = (sumX2 - ((sumX)**(2.0)) / (m(a)*n(b))) / (m(a)*n(b)-1.0)
      vbar = sumv / m(a)
      sbar2 = ((sums / m(a))/c4star)**2
      Rbar2 = ((sumR / m(a))/d2star)**2

      sumvc = sumvc + vc
      sumvc2 = sumvc2 + vc**(2.0)
      sumvbar = sumvbar + vbar
      sumvbar2 = sumvbar2 + vbar**(2.0)
      sumsbar2 = sumsbar2 + sbar2
      sumsbar22 = sumsbar22 + sbar2**(2.0)
      sumRbar2 = sumRbar2 + Rbar2
      sumRbar22 = sumRbar22 + Rbar2**(2.0)

!     replication loop
    end do

    varvc = (sumvc2 - ((sumvc)**(2.0)) / (rep - 1.0)) / (rep - 2.0)
    varvbar = (sumvbar2 - ((sumvbar)**(2.0)) / (rep - 1.0)) / (rep - 2.0)
    varsbar2 = (sumsbar22 - ((sumsbar2)**(2.0)) / (rep - 1.0)) / (rep - 2.0)
    varRbar2 = (sumRbar22 - ((sumRbar2)**(2.0)) / (rep - 1.0)) / (rep - 2.0)

    write(1, 10) n(b), m(a), c4star, d2star, varvc, varvbar, varsbar2, varRbar2
10 format(1X, I2, 1X, I3, 1X, F7.5, 1X, F7.5, 1X, F12.10, 1X, F12.10, 1X, F12.10, 1X, F12.10)

!   m loop
  end do

! n loop
end do

stop

contains

subroutine random(uniran, seed)
!
! ********************************************************
! ***** This subroutine generates Uniform (0, 1)     *****
! ***** random variates using the Marse-Roberts code *****
! ********************************************************
!
    implicit none
    INTEGER, parameter :: DOUBLE=SELECTED_REAL_KIND(p=15)
    REAL(KIND=DOUBLE), INTENT(OUT) :: uniran
    INTEGER, INTENT(IN OUT) :: seed
    INTEGER :: hi15, hi31, low15, lowprd, ovflow
    INTEGER, PARAMETER :: mult1 = 24112, mult2 = 26143, &
                          b2e15 = 32768, b2e16 = 65536, &
                          modlus = 2147483647
!
    hi15 = seed / b2e16
    lowprd = (seed - hi15 * b2e16) * mult1
    low15 = lowprd / b2e16
    hi31 = hi15 * mult1 + low15
    ovflow = hi31 / b2e15
    seed = (((lowprd - low15 * b2e16) - modlus) + &
            (hi31 - ovflow * b2e15) * b2e16) + ovflow
!
    if (seed < 0) seed = seed + modlus
```

```
!
    hi15 = seed / b2e16
    lowprd = (seed - hi15 * b2e16) * mult2
    low15 = lowprd / b2e16
    hi31 = hi15 * mult2 + low15
    ovflow = hi31 / b2e15
    seed = (((lowprd - low15 * b2e16) - modlus) + &
            (hi31 - ovflow * b2e15) * b2e16) + ovflow
!
    if (seed < 0) seed = seed + modlus
!
    uniran = (2 * (seed / 256) + 1) / 16777216.0
!
    return
  end subroutine random
!
end program simulate



program simulate_MR
implicit none
INTEGER, parameter :: DOUBLE=SELECTED_REAL_KIND(p=15)
real(kind=double) :: mean, sd, pi, d2starMR, r1, r2, X, first, second, MR
real(kind=double) :: sumMRbar2, sumMRbar22, sumMR, MRbar2, varMRbar2
INTEGER :: c, a, rep, i, seed = 1973272912
integer, dimension(1:28) :: m
open(unit=1, file="simulate_MR.txt")
open(unit=2, file="d2starMR.txt")

mean = 0.0
sd = 1.0
pi = ACOS(-1.0)
m = (/ (c, c = 2, 20), 25, 30, 50, 75, 100, 150, 200, 250, 300 /)

write(1, 5) "m", "d2starMR", "varMRbar2"
5 format(3X, A, 2X, A, 3X, A)

do a = 1, 28
! m loop

  sumMRbar2 = 0.0
  sumMRbar22 = 0.0

  read(2, *) d2starMR

  do rep = 1, 5000
!   replication loop

    sumMR = 0.0

    do i = 1, m(a)

      call random(r1, seed)
      call random(r2, seed)
      X = mean + sd * ((SQRT(-2. * LOG(r1))) * (COS(2. * pi * r2)))

      if (i == 1) then
        first = X
      else
        second = X
        MR = abs(first - second)
        sumMR = sumMR + MR
        first = second
      end if

    end do
    MRbar2 = ((sumMR / (m(a) - 1))/d2starMR)**2
    sumMRbar2 = sumMRbar2 + MRbar2
    sumMRbar22 = sumMRbar22 + MRbar2**(2.0)

!   replication loop
  end do

  varMRbar2 = (sumMRbar22 - ((sumMRbar2)**(2.0)) / (rep - 1.0)) / (rep - 2.0)

  write(1, 10) m(a), d2starMR, varMRbar2
10 format(1X, I3, 2X, F7.5, 2X, F12.10)

! m loop
end do

stop

contains

subroutine random(uniran, seed)
!
! ********************************************************
! ***** This subroutine generates Uniform (0, 1)     *****
! ***** random variates using the Marse-Roberts code *****
! ********************************************************
!
    implicit none
    INTEGER, parameter :: DOUBLE=SELECTED_REAL_KIND(p=15)
    REAL(KIND=DOUBLE), INTENT(OUT) :: uniran
    INTEGER, INTENT(IN OUT) :: seed
    INTEGER :: hi15, hi31, low15, lowprd, ovflow
    INTEGER, PARAMETER :: mult1 = 24112, mult2 = 26143, &
                          b2e15 = 32768, b2e16 = 65536, &
                          modlus = 2147483647
!
    hi15 = seed / b2e16
    lowprd = (seed - hi15 * b2e16) * mult1
    low15 = lowprd / b2e16
    hi31 = hi15 * mult1 + low15
    ovflow = hi31 / b2e15
    seed = (((lowprd - low15 * b2e16) - modlus) + &
            (hi31 - ovflow * b2e15) * b2e16) + ovflow
!
    if (seed < 0) seed = seed + modlus
!
    hi15 = seed / b2e16
    lowprd = (seed - hi15 * b2e16) * mult2
    low15 = lowprd / b2e16
    hi31 = hi15 * mult2 + low15
    ovflow = hi31 / b2e15
    seed = (((lowprd - low15 * b2e16) - modlus) + &
            (hi31 - ovflow * b2e15) * b2e16) + ovflow
!
    if (seed < 0) seed = seed + modlus
!
    uniran = (2 * (seed / 256) + 1) / 16777216.0
!
    return
  end subroutine random
!
end program simulate_MR
```

Table A.1. MSE of $\bar{v}$

m	n									
	2	3	4	5	6	7	8	10	25	50
1	1.752	0.988	0.670	0.476	0.410	0.331	0.288	0.222	0.084	0.040
2	1.039	0.504	0.313	0.244	0.184	0.169	0.145	0.111	0.042	0.021
3	0.667	0.334	0.223	0.165	0.135	0.107	0.099	0.071	0.027	0.013
4	0.527	0.245	0.167	0.127	0.096	0.086	0.074	0.056	0.021	0.011
5	0.395	0.202	0.132	0.102	0.080	0.067	0.057	0.045	0.016	0.008
6	0.338	0.163	0.112	0.084	0.067	0.056	0.048	0.037	0.014	0.007
7	0.294	0.145	0.094	0.071	0.055	0.047	0.040	0.031	0.012	0.006
8	0.245	0.127	0.085	0.063	0.050	0.042	0.036	0.027	0.011	0.005
9	0.224	0.109	0.074	0.054	0.044	0.039	0.031	0.025	0.009	0.005
10	0.200	0.098	0.067	0.050	0.039	0.034	0.028	0.022	0.008	0.004
11	0.181	0.094	0.062	0.046	0.037	0.031	0.025	0.020	0.008	0.004
12	0.163	0.086	0.056	0.043	0.035	0.027	0.023	0.018	0.007	0.003
13	0.151	0.077	0.050	0.038	0.031	0.025	0.022	0.017	0.006	0.003
14	0.142	0.072	0.047	0.036	0.028	0.023	0.021	0.015	0.006	0.003
15	0.134	0.068	0.045	0.033	0.026	0.022	0.020	0.015	0.006	0.003
16	0.127	0.062	0.042	0.032	0.025	0.020	0.017	0.014	0.005	0.003
17	0.118	0.059	0.039	0.030	0.023	0.020	0.017	0.013	0.005	0.002
18	0.110	0.057	0.038	0.027	0.022	0.018	0.016	0.012	0.004	0.002
19	0.101	0.053	0.035	0.026	0.021	0.018	0.015	0.012	0.004	0.002
20	0.100	0.051	0.033	0.025	0.019	0.017	0.014	0.011	0.004	0.002
25	0.079	0.041	0.027	0.020	0.016	0.013	0.012	0.009	0.003	0.002
30	0.066	0.034	0.022	0.017	0.014	0.012	0.010	0.007	0.003	0.001
50	0.041	0.020	0.014	0.010	0.008	0.006	0.006	0.004	0.002	0.001
75	0.028	0.013	0.009	0.007	0.005	0.004	0.004	0.003	0.001	0.001
100	0.021	0.010	0.007	0.005	0.004	0.003	0.003	0.002	0.001	0.000
150	0.013	0.007	0.004	0.003	0.003	0.002	0.002	0.001	0.001	0.000
200	0.010	0.005	0.003	0.002	0.002	0.002	0.001	0.001	0.000	0.000
250	0.008	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000
300	0.007	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000

Table A.2. MSE of v_c

m	n									
	2	3	4	5	6	7	8	10	25	50
1	1.752	0.988	0.670	0.476	0.410	0.331	0.288	0.222	0.084	0.040
2	0.683	0.400	0.268	0.216	0.168	0.156	0.136	0.104	0.041	0.021
3	0.396	0.247	0.184	0.143	0.120	0.097	0.092	0.066	0.026	0.013
4	0.300	0.181	0.134	0.107	0.086	0.077	0.068	0.051	0.021	0.011
5	0.214	0.142	0.104	0.085	0.070	0.059	0.052	0.041	0.016	0.008
6	0.178	0.116	0.087	0.071	0.056	0.049	0.043	0.034	0.014	0.007
7	0.159	0.100	0.073	0.056	0.047	0.042	0.036	0.028	0.011	0.006
8	0.132	0.085	0.065	0.051	0.043	0.037	0.032	0.025	0.010	0.005
9	0.118	0.076	0.057	0.045	0.037	0.034	0.028	0.023	0.009	0.005
10	0.106	0.067	0.052	0.041	0.033	0.029	0.025	0.020	0.008	0.004
11	0.095	0.064	0.047	0.038	0.031	0.027	0.023	0.018	0.007	0.004
12	0.087	0.060	0.042	0.035	0.029	0.024	0.021	0.017	0.007	0.003
13	0.078	0.054	0.040	0.030	0.026	0.022	0.019	0.015	0.006	0.003
14	0.077	0.050	0.035	0.029	0.024	0.021	0.018	0.014	0.006	0.003
15	0.070	0.045	0.033	0.027	0.022	0.019	0.017	0.013	0.005	0.003
16	0.067	0.042	0.032	0.025	0.022	0.018	0.015	0.013	0.005	0.002
17	0.061	0.039	0.030	0.024	0.019	0.017	0.015	0.011	0.005	0.002
18	0.058	0.038	0.028	0.022	0.018	0.016	0.014	0.011	0.004	0.002
19	0.052	0.036	0.026	0.021	0.018	0.016	0.013	0.011	0.004	0.002
20	0.050	0.034	0.025	0.020	0.016	0.015	0.012	0.010	0.004	0.002
25	0.041	0.028	0.020	0.016	0.013	0.011	0.010	0.008	0.003	0.002
30	0.033	0.023	0.017	0.014	0.012	0.010	0.008	0.007	0.003	0.001
50	0.020	0.014	0.010	0.008	0.007	0.005	0.005	0.004	0.002	0.001
75	0.014	0.009	0.007	0.005	0.004	0.004	0.003	0.003	0.001	0.001
100	0.010	0.007	0.005	0.004	0.003	0.003	0.002	0.002	0.001	0.000
150	0.007	0.004	0.003	0.003	0.002	0.002	0.002	0.001	0.001	0.000
200	0.005	0.003	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000
250	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000
300	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.000	0.000

Table A.3. MSE of $\left(\overline{R}/d_2^*\right)^2$

m	n									
	2	3	4	5	6	7	8	10	25	50
1	1.752	1.000	0.693	0.517	0.452	0.378	0.331	0.273	0.131	0.085
2	1.089	0.531	0.331	0.262	0.204	0.190	0.165	0.135	0.064	0.044
3	0.709	0.356	0.240	0.181	0.150	0.125	0.113	0.088	0.042	0.028
4	0.586	0.264	0.181	0.136	0.108	0.098	0.086	0.068	0.032	0.022
5	0.441	0.222	0.142	0.113	0.089	0.076	0.067	0.055	0.025	0.017
6	0.366	0.180	0.122	0.093	0.075	0.064	0.055	0.046	0.022	0.014
7	0.333	0.158	0.103	0.079	0.061	0.052	0.047	0.038	0.019	0.012
8	0.289	0.137	0.092	0.068	0.055	0.048	0.042	0.033	0.016	0.011
9	0.250	0.121	0.081	0.060	0.048	0.044	0.037	0.030	0.014	0.009
10	0.222	0.107	0.073	0.055	0.044	0.039	0.032	0.026	0.013	0.008
11	0.205	0.104	0.067	0.051	0.041	0.035	0.030	0.025	0.012	0.008
12	0.182	0.093	0.060	0.048	0.038	0.031	0.027	0.022	0.010	0.007
13	0.178	0.084	0.055	0.041	0.035	0.029	0.026	0.021	0.010	0.006
14	0.163	0.078	0.051	0.040	0.032	0.026	0.024	0.019	0.009	0.006
15	0.154	0.074	0.049	0.036	0.029	0.025	0.023	0.018	0.009	0.006
16	0.144	0.067	0.046	0.035	0.029	0.023	0.020	0.017	0.008	0.005
17	0.132	0.064	0.043	0.032	0.025	0.022	0.020	0.015	0.008	0.005
18	0.124	0.062	0.042	0.031	0.025	0.021	0.018	0.015	0.007	0.005
19	0.113	0.058	0.039	0.028	0.023	0.021	0.017	0.014	0.007	0.005
20	0.111	0.056	0.036	0.028	0.022	0.019	0.016	0.013	0.007	0.004
25	0.090	0.045	0.030	0.022	0.018	0.015	0.013	0.011	0.005	0.003
30	0.076	0.037	0.024	0.019	0.016	0.013	0.011	0.009	0.004	0.003
50	0.046	0.022	0.015	0.011	0.009	0.007	0.006	0.005	0.003	0.002
75	0.032	0.014	0.010	0.008	0.006	0.005	0.004	0.004	0.002	0.001
100	0.024	0.011	0.007	0.006	0.005	0.004	0.003	0.003	0.001	0.001
150	0.015	0.007	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.001
200	0.011	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.001	0.000
250	0.009	0.004	0.003	0.002	0.002	0.002	0.001	0.001	0.001	0.000
300	0.008	0.004	0.002	0.002	0.002	0.001	0.001	0.001	0.000	0.000

Table A.4. MSE of $\left(\bar{s}/c_4^*\right)^2$

m	n									
	2	3	4	5	6	7	8	10	25	50
1	1.752	0.988	0.670	0.476	0.410	0.331	0.288	0.222	0.084	0.040
2	1.089	0.517	0.322	0.250	0.187	0.171	0.147	0.113	0.042	0.021
3	0.709	0.354	0.234	0.171	0.139	0.111	0.102	0.072	0.027	0.013
4	0.586	0.259	0.174	0.131	0.100	0.089	0.076	0.058	0.021	0.011
5	0.441	0.220	0.138	0.107	0.083	0.069	0.058	0.046	0.016	0.008
6	0.366	0.177	0.118	0.087	0.069	0.058	0.049	0.038	0.014	0.007
7	0.333	0.156	0.100	0.075	0.056	0.048	0.042	0.032	0.012	0.006
8	0.289	0.135	0.090	0.066	0.052	0.043	0.037	0.028	0.011	0.005
9	0.250	0.120	0.079	0.057	0.045	0.040	0.033	0.026	0.010	0.005
10	0.222	0.106	0.071	0.052	0.041	0.035	0.028	0.022	0.008	0.004
11	0.205	0.103	0.065	0.049	0.039	0.032	0.026	0.021	0.008	0.004
12	0.182	0.093	0.059	0.046	0.036	0.029	0.024	0.019	0.007	0.003
13	0.178	0.083	0.053	0.040	0.033	0.026	0.023	0.017	0.006	0.003
14	0.163	0.078	0.050	0.038	0.030	0.024	0.021	0.016	0.006	0.003
15	0.154	0.073	0.047	0.035	0.027	0.023	0.020	0.015	0.006	0.003
16	0.144	0.066	0.044	0.033	0.027	0.021	0.018	0.015	0.005	0.003
17	0.132	0.064	0.042	0.031	0.023	0.020	0.017	0.013	0.005	0.002
18	0.124	0.062	0.041	0.029	0.023	0.019	0.016	0.012	0.005	0.002
19	0.113	0.057	0.038	0.027	0.021	0.019	0.016	0.012	0.004	0.002
20	0.111	0.056	0.035	0.027	0.020	0.018	0.014	0.011	0.004	0.002
25	0.090	0.045	0.029	0.021	0.016	0.014	0.012	0.009	0.003	0.002
30	0.076	0.036	0.024	0.018	0.015	0.012	0.010	0.008	0.003	0.001
50	0.046	0.022	0.014	0.011	0.008	0.007	0.006	0.004	0.002	0.001
75	0.032	0.014	0.009	0.007	0.006	0.005	0.004	0.003	0.001	0.001
100	0.024	0.011	0.007	0.005	0.004	0.003	0.003	0.002	0.001	0.000
150	0.015	0.007	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.000
200	0.011	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.000	0.000
250	0.009	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.000	0.000
300	0.008	0.004	0.002	0.002	0.001	0.001	0.001	0.001	0.000	0.000

Table A.5. MSE of $\left(\overline{MR}/d_2^*(MR)\right)^2$

m	MSE
2	1.752
3	1.498
4	1.015
5	0.790
6	0.677
7	0.519
8	0.440
9	0.397
10	0.366
11	0.340
12	0.297
13	0.270
14	0.248
15	0.242
16	0.213
17	0.199
18	0.199
19	0.183
20	0.178
25	0.135
30	0.118
50	0.068
75	0.043
100	0.032
150	0.022
200	0.017
250	0.014
300	0.011

Table A.6. Percent change in MSE($\bar{v}$) (Table A.1) over MSE(v_c) (Table A.2)

m	n									
	2	3	4	5	6	7	8	10	25	50
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	52.073	25.903	16.884	13.289	9.206	8.055	6.425	6.544	2.664	0.815
3	68.193	35.431	20.908	15.835	13.093	10.186	7.706	7.619	2.964	1.380
4	75.557	35.226	24.624	18.797	12.587	12.943	9.253	9.351	1.935	0.722
5	84.482	41.662	27.244	20.039	15.126	13.957	10.916	9.449	2.340	1.101
6	89.940	40.801	28.364	17.146	19.364	14.511	10.580	9.076	3.501	0.988
7	84.869	45.270	28.358	25.440	17.214	10.842	11.563	10.332	4.594	1.879
8	85.890	48.805	30.047	22.209	16.430	14.215	12.691	10.746	3.352	2.133
9	89.357	43.492	31.116	20.060	18.264	14.980	11.704	9.735	3.795	1.924
10	88.726	44.652	30.283	23.127	16.977	16.780	11.599	9.772	2.714	1.527
11	90.284	47.833	31.328	20.932	17.115	15.487	12.731	10.912	3.942	1.449
12	87.522	43.637	33.993	23.695	18.684	14.152	13.556	10.073	2.775	1.944
13	92.710	43.511	26.581	26.125	19.541	13.470	14.819	9.337	2.875	2.287
14	85.025	44.385	32.747	22.168	17.461	14.061	12.865	9.206	3.346	1.852
15	92.841	51.279	34.034	21.948	19.869	16.353	13.868	9.921	3.877	0.501
16	88.865	48.512	31.479	24.520	17.975	15.630	12.666	9.818	3.403	2.457
17	93.852	49.103	30.251	22.397	18.714	15.667	13.909	10.301	4.064	1.888
18	90.367	49.454	35.143	23.567	19.157	12.699	13.300	9.184	2.637	2.173
19	93.210	46.948	33.387	23.265	16.551	17.442	14.703	10.343	3.472	2.220
20	98.648	51.753	31.433	24.018	19.463	16.540	12.286	11.093	3.394	1.117
25	92.231	50.113	32.890	24.573	19.684	15.712	15.426	10.608	3.755	1.647
30	101.498	47.193	30.602	20.961	19.572	14.935	12.462	10.124	3.892	2.687
50	99.336	49.047	32.104	25.351	19.752	17.257	14.353	11.368	3.573	1.625
75	104.021	42.990	31.257	27.672	19.255	14.066	13.853	12.540	4.104	2.447
100	103.253	48.012	32.163	27.019	20.347	15.836	14.674	9.487	4.611	2.196
150	99.622	49.317	32.110	25.578	19.767	14.875	13.895	12.096	2.737	2.173
200	99.086	48.312	33.452	27.408	20.831	18.136	14.511	11.227	4.837	1.806
250	98.234	50.075	28.355	24.681	18.559	14.259	14.870	11.612	3.553	1.313
300	95.180	48.556	33.210	26.037	20.520	16.140	14.207	10.015	4.702	2.797

Table A.7. Percent change in $\text{MSE}\left[\left(\bar{s}/c_4^*\right)^2\right]$ (Table A.4) over MSE($\bar{v}$) (Table A.1)

m	n									
	2	3	4	5	6	7	8	10	25	50
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	4.881	2.705	3.097	2.504	1.781	1.503	1.720	2.002	0.189	0.251
3	6.383	6.081	5.074	3.414	2.595	3.238	2.916	1.927	1.082	0.419
4	11.288	5.917	4.175	3.449	3.254	2.920	2.719	3.323	0.639	0.536
5	11.460	8.756	4.641	5.167	3.533	2.564	1.907	2.180	0.679	0.462
6	8.302	8.807	5.859	3.842	3.447	2.976	2.243	2.439	0.578	0.187
7	13.171	7.632	5.779	5.650	2.772	2.706	3.247	2.562	1.147	0.338
8	18.372	6.216	6.211	5.034	3.692	2.920	2.442	2.343	0.496	0.307
9	11.715	9.304	5.771	4.618	2.924	3.732	4.079	2.461	0.581	0.643
10	11.246	8.445	5.271	3.988	4.625	2.571	3.133	1.124	1.383	0.094
11	13.623	8.562	6.059	6.324	4.598	3.544	2.660	2.845	0.916	0.507
12	11.798	8.239	4.983	5.454	3.129	4.335	3.421	2.974	0.628	0.576
13	17.774	8.531	6.361	4.593	3.890	2.999	3.081	1.622	0.779	0.217
14	14.660	8.359	6.573	4.902	3.807	3.132	2.695	3.011	1.096	0.684
15	14.841	8.254	6.022	5.661	4.008	3.416	2.771	2.536	0.759	0.491
16	14.061	6.759	5.745	5.270	4.246	3.923	2.975	2.291	1.029	0.149
17	12.719	9.391	7.298	4.528	3.963	3.591	3.355	1.873	1.131	0.211
18	12.691	8.229	5.910	5.397	4.580	4.292	2.302	3.021	1.005	0.694
19	12.543	8.205	7.148	3.732	3.448	4.753	2.423	2.115	0.903	0.100
20	11.823	8.412	4.982	5.763	4.035	3.235	3.461	2.196	1.116	0.705
25	14.414	8.806	8.196	5.137	4.691	3.476	3.517	2.900	1.079	0.311
30	14.684	8.077	5.816	5.742	4.475	3.819	3.185	3.060	1.102	0.633
50	13.124	8.101	5.945	5.744	4.827	3.670	2.727	3.491	0.909	0.284
75	14.217	8.798	6.347	4.371	4.961	3.834	3.423	3.404	1.392	0.477
100	11.679	8.384	6.871	5.844	3.480	2.562	3.648	2.768	1.351	0.920
150	11.609	10.482	7.403	5.020	5.131	3.924	3.077	3.138	0.797	1.006
200	13.577	9.599	5.333	5.454	4.021	3.456	4.161	2.549	0.938	0.571
250	14.803	9.349	8.450	5.100	4.652	3.561	3.576	2.826	1.459	0.940
300	12.505	8.421	6.809	5.323	3.496	3.308	3.730	1.700	0.698	0.437

Table A.8. Percent change in $\mathrm{MSE}\left[\left(\overline{R}/d_2^*\right)^2\right]$ (Table A.3) over $\mathrm{MSE}\left[\left(\overline{s}/c_4^*\right)^2\right]$ (Table A.4)

m	n									
	2	3	4	5	6	7	8	10	25	50
1	0.001	1.215	3.480	8.627	10.131	13.936	15.067	22.798	55.428	113.664
2	-0.001	2.594	2.563	4.613	9.264	11.179	11.860	18.848	52.969	107.533
3	0.003	0.573	2.603	6.292	8.040	12.618	11.690	22.479	54.757	108.073
4	0.004	1.784	4.013	3.941	8.350	10.524	13.311	17.925	51.426	105.104
5	0.000	1.115	3.106	5.287	7.480	10.261	14.058	19.329	54.490	107.180
6	-0.001	1.379	3.138	6.893	7.902	9.331	12.450	20.489	52.802	109.893
7	0.002	0.942	3.337	5.896	8.455	9.545	12.367	18.134	55.828	102.097
8	0.001	1.123	2.172	3.487	7.175	11.319	13.533	19.499	53.487	105.157
9	0.002	1.121	2.672	6.139	7.143	10.362	14.281	17.587	47.735	103.686
10	-0.001	0.825	3.198	4.816	7.233	10.044	12.150	17.570	53.828	105.455
11	-0.004	1.292	3.164	5.275	6.896	9.743	13.373	18.760	57.409	100.255
12	0.001	0.775	1.920	4.957	7.255	9.834	11.513	19.786	49.903	106.588
13	-0.003	0.558	2.048	3.359	6.969	10.947	13.622	20.044	58.465	97.759
14	0.002	0.246	3.045	5.150	8.640	8.495	12.859	17.647	55.959	99.568
15	-0.003	0.448	2.745	5.215	8.112	8.094	12.096	17.428	57.369	107.824
16	-0.001	1.045	2.793	4.187	7.972	10.813	14.090	15.777	53.930	109.306
17	-0.001	0.356	2.518	4.946	7.901	8.322	12.944	18.603	53.725	107.640
18	0.002	0.776	3.205	6.324	7.750	11.144	13.361	19.721	57.326	105.909
19	0.001	0.745	2.892	3.913	7.910	8.911	10.971	17.007	56.335	107.476
20	-0.001	1.182	2.800	4.045	7.385	9.055	11.573	19.389	56.966	100.242
25	-0.002	0.191	2.892	4.748	8.011	9.757	13.276	17.238	57.042	98.729
30	0.001	0.959	2.271	5.115	6.917	9.134	13.342	17.543	51.071	99.970
50	0.003	0.887	2.441	5.394	5.941	10.419	12.549	19.035	53.174	111.775
75	0.004	1.585	2.236	4.956	6.513	10.382	13.174	18.330	57.105	96.958
100	0.003	0.592	2.768	5.044	7.278	10.218	10.040	16.489	52.726	100.217
150	0.001	0.667	1.895	4.526	6.441	8.006	12.889	16.511	51.540	106.852
200	-0.002	1.128	2.557	4.633	7.094	8.680	12.273	16.959	56.797	109.206
250	-0.004	1.189	2.730	4.952	5.964	8.974	12.622	19.308	54.066	106.406
300	-0.001	1.226	3.078	4.402	6.802	12.732	15.062	17.502	51.441	103.615

References

Bain, L. J. & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Belmont, CA: Duxbury Press.

David, H. A. (1951). Further applications of range to the analysis of variance. *Biometrika*, *38*, 393-409.

Duncan, A. J. (1974). *Quality control and industrial statistics* (4th ed.). Homewood, IL: Richard D. Irwin, Inc.

Elam, M. E. & Case, K. E. (2001). A computer program to calculate two-stage short-run control chart factors for $(\overline{X}, R)$ charts. *Quality Engineering*, *14(1)*, 77-102.

Elam, M. E. & Case, K. E. (2003). Two-stage short-run $(\overline{X}, v)$ and $(\overline{X}, \sqrt{v})$ control charts. *Quality Engineering*, *15(3)*, 441-448.

Elam, M. E. & Case, K. E. (2005a). Two-stage short-run $(\overline{X}, s)$ control charts. *Quality Engineering*, *17(1)*, 95-107.

Elam, M. E. & Case, K. E. (2005b). A computer program to calculate two-stage short-run control chart factors for $(\overline{X}, s)$ charts. *Quality Engineering*, *17(2)*, 259-277.

Elam, M. E. & Case, K. E. (2006a). Two-stage short-run (X, MR) control charts. *Journal of Modern Applied Statistical Methods*, *5*(2), in press.

Elam, M. E. & Case, K. E. (2006b). A computer program to calculate two-stage short-run control chart factors for (X, MR) charts. *Journal of Statistical Software, 15*(11).

FORTRAN PowerStation v4.0 in Microsoft Developer Studio (1994). Redmond, WA: Microsoft Corporation.

Harter, H. L. (1960). Tables of range and studentized range. *Annals of Mathematical Statistics*, *31*, 1122-1147.

Luko, S. N. (1996). Concerning the estimators $\overline{R}/d_2$ and $\overline{R}/d_2^*$ in estimating variability in a normal universe. *Quality Engineering*, *8(3)*, 481-487.

Marse, K. & Roberts, S. D. (1983). Implementing a portable FORTRAN uniform (0, 1) generator. *Simulation*, *41(4)*, 135-139.

Mathcad 2000 Professional (1999). Cambridge, MA: MathSoft, Inc.

Mead, R. (1966). A quick method of estimating the standard deviation. *Biometrika*, *53*, 559-564.

Minitab Release 14.1 Statistical Software (2003). State College, PA: Minitab Inc.

Numerical Recipes Extension Pack (1997). Cambridge, MA: MathSoft, Inc.

Palm, A. C. & Wheeler, D. J. (1990, Unpublished manuscript). Equivalent degrees of freedom for estimates of the process standard deviation based on Shewhart control charts.

Wheeler, D. J. (1995). *Advanced Topics in Statistical Process Control.* Knoxville, TN: SPC Press, Inc.

Appendix

Show: $\left(\bar{s}/c_4^*\right)^2$ is an unbiased point estimate of σ^2; i.e., show $E\left[\left(\bar{s}/c_4^*\right)^2\right] = \sigma^2$

$$E\left[\left(\frac{\bar{s}}{c_4^*}\right)^2\right] = \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot E\left[\left(\bar{s}\right)^2\right]$$

$$= \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot E\left[\left(\frac{\sum_{i=1}^{m} s_i}{m}\right)^2\right]$$

$$= \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot E\left[\left(\sum_{i=1}^{m} s_i\right)^2\right],$$

$$= \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot\left[Var\left(\sum_{i=1}^{m} s_i\right) + \left[E\left(\sum_{i=1}^{m} s_i\right)\right]^2\right]$$

$$= \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot\left[\sum_{i=1}^{m} Var\left(s_i\right) + \left(\sum_{i=1}^{m} E\left(s_i\right)\right)^2\right]$$

because the s_i's are independent.

$$\Rightarrow E\left[\left(\frac{\bar{s}}{c_4^*}\right)^2\right] = \left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot\left[\begin{array}{l}\sum_{i=1}^{m}\left(c_5^2\cdot\sigma^2\right)\\ +\left[\sum_{i=1}^{m}\left(c_4\cdot\sigma\right)\right]^2\end{array}\right],$$

because $Var(s) = c_5^2\cdot\sigma^2$ and $E(s) = c_4\cdot\sigma$ (by definition,

$$Var\left(\frac{s}{\sigma}\right) = c5^2$$

$$\Rightarrow \left(\frac{1}{\sigma^2}\right)\cdot Var(s) = c5^2 \Rightarrow Var(s) = c5^2\cdot\sigma^2\,;$$

by definition,

$$E\left(\frac{s}{\sigma}\right) = c4 \Rightarrow \left(\frac{1}{\sigma}\right)\cdot E(s) = c4 \Rightarrow E(s) = c4\cdot\sigma\,).$$

$$\Rightarrow E\left[\left(\frac{\bar{s}}{c_4^*}\right)^2\right]=\left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot\left[m\cdot c_5^2\cdot\sigma^2+(m\cdot c_4\cdot\sigma)^2\right]$$

$$=\left(\frac{1}{\left(c_4^*\right)^2}\right)\cdot\left(\frac{1}{m^2}\right)\cdot\left(m\cdot c_5^2\cdot\sigma^2+m^2\cdot c_4^2\cdot\sigma^2\right)$$

$$=\left(\frac{c_5^2\cdot\sigma^2}{m\cdot\left(c_4^*\right)^2}\right)+\left(\frac{c_4^2\cdot\sigma^2}{\left(c_4^*\right)^2}\right)=\sigma^2\cdot\left(\frac{c_4^2+\frac{c_5^2}{m}}{\left(c_4^*\right)^2}\right)$$

$$=\sigma^2\cdot\left(\frac{\left(c_4^*\right)^2}{\left(c_4^*\right)^2}\right),\text{ since }c_4^*=\left(c_4^2+\frac{c_5^2}{m}\right)^{0.5}$$

$$\Rightarrow E\left[\left(\frac{\bar{s}}{c_4^*}\right)^2\right]=\sigma^2\cdot(1)=\sigma^2$$

Show: $\left(\overline{MR}/d_2^*(MR)\right)^2$ is an unbiased estimate of σ^2; i.e., show $E\left[\left(\overline{MR}/d_2^*(MR)\right)^2\right]=\sigma^2$. One first needs to determine the variance of the distribution of the mean moving range $\overline{MR}/\sigma$.

$$\mathrm{Var}\left(\frac{\overline{MR}}{\sigma}\right)=\left(\frac{1}{\sigma^2}\right)\cdot\mathrm{Var}\left(\overline{MR}\right)$$

From Palm and Wheeler (1990),

$$\mathrm{Var}\left(\overline{MR}/d2\right)=\sigma^2\cdot r,$$

where

$$r=\frac{b\cdot(m-1)-c}{(m-1)^2}$$

with

$$b=\frac{2\cdot\pi}{3}-3+\sqrt{3}$$

and

$$c=\frac{\pi}{6}-2+\sqrt{3}$$

$$\Rightarrow r=\left(\frac{1}{\sigma^2}\right)\cdot\mathrm{Var}\left(\frac{\overline{MR}}{d2}\right)=\left(\frac{1}{\sigma^2}\right)\cdot\left(\frac{1}{d2^2}\right)\cdot\mathrm{Var}\left(\overline{MR}\right)$$

$$\Rightarrow d2^2\cdot r=\left(\frac{1}{\sigma^2}\right)\cdot\mathrm{Var}\left(\overline{MR}\right)$$

$$\Rightarrow \mathrm{Var}\left(\frac{\overline{MR}}{\sigma}\right)=d2^2\cdot r$$

$$E\left[\left(\frac{\overline{MR}}{d_2^*(MR)}\right)^2\right]=\left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot E\left[\left(\overline{MR}\right)^2\right]$$

$$=\left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[Var\left(\overline{MR}\right)+\left(E\left(\overline{MR}\right)\right)^2\right]$$

$$=\left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[d_2^2\cdot r\cdot\sigma^2+\left[E\left(\frac{\sum_{i=1}^{m-1}MR_i}{m-1}\right)\right]^2\right]$$

because

$$Var\left(\overline{MR}/\sigma\right)=d_2^2\cdot r\Rightarrow\left(1/\sigma^2\right)\cdot Var\left(\overline{MR}\right)$$

$$=d_2^2\cdot r\Rightarrow Var\left(\overline{MR}\right)\qquad.$$

$$=d_2^2\cdot r\cdot\sigma^2$$

$$\Rightarrow E\left[\left(\frac{\overline{MR}}{d_2^*(MR)}\right)^2\right]$$

$$=\left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[d_2^2\cdot r\cdot\sigma^2+\left(\frac{1}{(m-1)^2}\right)\cdot\left[E\left(\sum_{i=1}^{m-1}MR_i\right)\right]^2\right],$$

$$=\left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[d_2^2\cdot r\cdot\sigma^2+\left(\frac{1}{(m-1)^2}\right)\cdot\left(\sum_{i=1}^{m-1}E\left(MR_i\right)\right)^2\right]$$

$$= \left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[\begin{array}{l} d_2^2\cdot r\cdot\sigma^2 + \\ \left(\frac{1}{(m-1)^2}\right)\cdot\left[\sum_{i=1}^{m-1}\left(d_2\cdot\sigma\right)\right]^2 \end{array}\right]$$

because

$\mathrm{E(MR)} = \mathrm{d}_2\cdot\sigma$ (by definition,

$$\mathrm{E}\left(\frac{\mathrm{MR}}{\sigma}\right) = \mathrm{d2}$$

$$\Rightarrow \left(\frac{1}{\sigma}\right)\cdot \mathrm{E(MR)} = \mathrm{d2} \Rightarrow \mathrm{E(MR)} = \mathrm{d2}\cdot\sigma\text{).}$$

$$\Rightarrow E\left[\left(\frac{\overline{MR}}{d_2^*(MR)}\right)^2\right]$$

$$= \left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left[\begin{array}{l} d_2^2\cdot r\cdot\sigma^2 + \\ \left(\frac{1}{(m-1)^2}\right)\cdot\left((m-1)\cdot d_2\cdot\sigma\right)^2 \end{array}\right]$$

$$= \left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\left(d_2^2\cdot r\cdot\sigma^2 + d_2^2\cdot\sigma^2\right)$$

$$= \left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\sigma^2\cdot\left(d_2^2 + d_2^2\cdot r\right)$$

$$= \left(\frac{1}{\left(d_2^*(MR)\right)^2}\right)\cdot\sigma^2\cdot\left(d_2^*(MR)\right)^2$$

because

$$\mathrm{d}_2^*(\mathrm{MR}) = \left(\mathrm{d}_2^2 + \mathrm{d}_2^2\cdot\mathrm{r}\right)^{0.5}.$$

$$\Rightarrow \mathrm{E}\left[\left(\frac{\overline{\mathrm{MR}}}{\mathrm{d}_2^*(\mathrm{MR})}\right)^2\right] = \sigma^2\cdot(1) = \sigma^2$$

QED

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 217-229

1538 – 9472/06/$95.00

Variance Estimation and Construction of Confidence Intervals for GEE Estimator

Shenghai Zhang
Centre for Infectious Disease Prevention and Control
Public Health Agency of Canada

Mary E. Thompson
Department of Statistics & Actuarial Sciences
University of Waterloo

The sandwich estimator, also known as the robust covariance matrix estimator, has achieved increasing use in the statistical literature as well as with the growing popularity of generalized estimating equations (GEE). A modified sandwich variance estimator is proposed, and its consistency and efficiency are studied. It is compared with other variance estimators, such as a model based estimator, the sandwich estimator and a corrected sandwich estimator. Confidence intervals for regression parameters based on these estimators are discussed. Simulation studies using clustered data to compare the performance of variance estimators are reported.

Key words: Generalized estimating equation, sandwich estimator, bias corrected estimator, variance-covariance matrix

Introduction

Once the estimators of regression parameters are obtained from a generalized estimating equation (GEE) (see Diggle, Liang & Zeger,1994; Liang & Zeger,1986), one needs the variance estimator to conduct inferences about the parameters. The sandwich estimator, also known as the robust covariance matrix estimator, has been used to achieve this goal. Its virtue is that it provides consistent estimates of the covariance matrix for parameter estimates even if the correlation structure in the parametric model is mis-specified. However, the properties of the sandwich method, other than consistency, had been little discussed until Kauermann and Carroll (2001). Further discussion about the properties will be provided, as well as a new variance estimator. This will be compared with other variance estimators: (a) a model based estimator, (b) the sandwich estimator, and (c) a corrected sandwich estimator.

Estimation of $\text{cov}(Y_i)$ will be discussed first, where $Y_i = (y_{i1}, \cdots, y_{im})^T$ is a vector of repeated measurements taken on the ith subject; associated with each measurement y_{ij} is a vector of covariates $x_{ij} = (x_{ij1}, \cdots, x_{ijp})^T$ $(1 \le j \le m, \quad 1 \le i \le n)$. The mean of the marginal distribution of y_{ij} is denoted by μ_{ij}. It is assumed that Y_i and Y_k are independent vectors for all $i \ne k$. A bias reduced variance estimator will be provided next, and its consistency and efficiency will be discussed. Also, methods of constructing confidence intervals based on the variance estimators will be discussed. The simulation studies using clustered data to compare the performance of variance estimators will be reported.

Estimating Covariance

The main parameter of interest is $\beta = (\beta_1, \cdots, \beta_p)^T$, where β, covariates x_{ij} and the mean μ_{ij} of the marginal distribution are connected by a link function $h(.)$. The variance $\text{var}(y_{ij}) = \phi^{-1} v(\mu_{ij})$, where $v(.)$ is a known function, and where ϕ is a dispersion scalar that is either unknown or a known

Dr. Shenghai Zhang is Senior Biostatistician. Email: s.zhang@rogers.com. Mary E. Thompson is Professor of Statistics. The authors thank Professors J. F. Lawless, V. P. Godambe, P. Song and C. Wu at University of Waterloo for their valuable suggestions.

constant. Let $R(\alpha)$ be a $m \times m$ symmetric matrix which is a ‘working’ correlation matrix. The estimation of the nuisance parameter α will not be discussed and will be assumed to be known. The results could be generalized to the estimated $\hat{\alpha}$ of the α. Let

$$\eta_{ij} = x_{ij}^T \beta .$$

Then

$$\mu_{ij} = h(\eta_{ij}) ,$$

and

$$X_i = (x_{i1}, \cdots, x_{im})^T ;$$

$$A_i = diag(h'(\eta_{ij})) ;$$

$$\Gamma_i = diag(v(h(\eta_{ij}))) \tag{1}$$

are matrices with order $m \times p$, $m \times m$ and $m \times m$ respectively. It is well known that the general estimating function is defined as the following (Liang & Zeger, 1986):

$$g_n(\beta, \alpha) = \frac{1}{n} \sum_{i=1}^{n} [D_i]^T [V_i]^{-1} S_i , \tag{2}$$

where

$$V_i = V_i(\beta, \alpha) = \Gamma_i^{\frac{1}{2}} R_i(\alpha) \Gamma_i^{\frac{1}{2}} \;;$$

$$D_i = D_i(\beta) = -\frac{\partial S_i}{\partial \beta^T} = A_i X_i ;$$

and

$$S_i = Y_i - (h(x_{ij}{}^T \beta), \cdots, h(x_{ij}{}^T \beta))^T$$
$$= (S_{i1}, \cdots, S_{im})^T .$$

$\phi^{-1} V_i$ was used to replace the true covariance $\mathrm{cov}(Y_i)$ in the optimal estimating function linear in S_i. Because $\mathrm{cov}(Y_i)$ is usually unknown, the estimation of $\mathrm{cov}(Y_i)$ is first discussed. Typically, the residual estimator $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T$ is used to estimate $\mathrm{cov}(Y_i)$, where $\hat{\mu}_i = (\hat{\mu}_{i1}, \cdots, \hat{\mu}_{im})$ is the vector of fitted values based on the estimated parameters $\hat{\beta}_{GEE}$ obtained by solving equation $g_n(\beta, \alpha) = 0$. Because the fitted values tend to be closer to the observed values than the true values are, the residuals tend to be too small. Therefore, $\mathrm{cov}(Y_i)$ tends to be underestimated by this method. To reduce the bias in general, another estimator of $\mathrm{cov}(Y_i)$ will be proposed.

Considering a first-order Taylor series expansion of $\hat{\mu}_i = \mu_i(\hat{\beta}_{GEE})$ at the true parameter β_0, one has the following expressions:

$$\begin{aligned} S_i(\hat{\beta}_{GEE}) &\equiv Y_i - \hat{\mu}_i \\ &= Y_i - \mu_i(\beta_0) - \frac{\partial \mu_i}{\partial \beta^T}(\hat{\beta}_{GEE} - \beta_0) - O_p(n^{-1}) \\ &= S_i(\beta_0) - D_i(\beta_0)(\hat{\beta}_{GEE} - \beta_0) - O_p(n^{-1}). \end{aligned} \tag{3}$$

Based on an expansion for $\hat{\beta}_{GEE} - \beta_0$ (see Zhang , 2003),

$$\begin{aligned} &(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T = S_i(\beta_0)[S_i(\beta_0)]^T \\ &- S_i(\beta_0) H_i(\beta_0, \alpha) - [H_i(\beta_0, \alpha)]^T [S_i(\beta_0)]^T \\ &+ [H_i(\beta_0, \alpha)]^T H_i(\beta_0, \alpha) \\ &+ O_p(n^{-\frac{3}{2}}) \end{aligned} \tag{4}$$

where

$$\begin{aligned} &H_i(\beta_0, \alpha) \\ &= \frac{1}{n} \sum_{k=1}^{n} S_k{}^T V_k{}^{-1} D_k (\dot{g}_{n,0})^{-1} D_i{}^T \Big|_{(\beta_0, \alpha)} , \end{aligned} \tag{5}$$

and

$$\dot{g}_{n,0}(\beta,\alpha)=\frac{1}{n}\sum_{i=1}^{n}D_i^{\,T}[V_i]^{-1}D_i\Big|_{(\beta_0,\alpha)}. \tag{6}$$

$f\big|_{(\beta,\alpha)}$ is used to denote the value of a function f at (β,α). For example,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}D_i^{T}[V_i]^{-1}D_i\Big|_{(\beta_0,\alpha)}\\&\equiv\frac{1}{n}\sum_{i=1}^{n}[D_i(\beta_0)]^{T}[V_i(\beta_0,\alpha)]^{-1}D_i(\beta_0).\end{aligned}$$

Taking expectation on both sides of (4), under certain integral conditions,

$$\begin{aligned}&\mathrm{E}[(Y_i-\hat{\mu}_i)(Y_i-\hat{\mu}_i)^{T}]\\&=\{\mathrm{cov}(Y_i)-n^{-1}\,\mathrm{cov}(Y_i)h_{ii}-n^{-1}h_{ii}^{\;T}\,\mathrm{cov}(Y_i)\\&+n^{-2}h_{ii}^{\;T}\,\mathrm{cov}(Y_i)h_{ii}\\&+n^{-2}\sum_{k=1,k\neq i}^{n}h_{ki}^{\;T}\,\mathrm{cov}(Y_k)h_{ki}\}\Big|_{(\beta_0,\alpha)}+O(n^{-\frac{3}{2}})\\&=\{(I_i-n^{-1}h_{ii}^{\;T})\,\mathrm{cov}(Y_i)(I_i-n^{-1}h_{ii})\\&+n^{-2}\sum_{k=1,k\neq i}^{n}h_{ki}^{\;T}\,\mathrm{cov}(Y_k)h_{ki}\}\Big|_{(\beta_0,\alpha)}+O(n^{-\frac{3}{2}})\end{aligned}$$

where

$$h_{ki}=[V_k]^{-1}D_k[\dot{g}_{n,0}]^{-1}D_i^{\;T}, \tag{7}$$

for $i,k=1,\cdots,n$, and I_i is an identity matrix of the same dimension as that of h_{ii}. An alternative estimator for $\mathrm{cov}(Y_i)$ was proposed by Mancl and DeRouen (2001) that is intended to compensate for the bias of the residual estimator in hypothesis testing: $\mathrm{cov}(Y_i)$ could be estimated by

$$(I_i-n^{-1}\hat{h}_{ii}^{\;T})^{-1}\hat{S}_i[\hat{S}_i]^{T}(I_i-n^{-1}\hat{h}_{ii})^{-1}, \tag{8}$$

under the assumption that

$$n^{-2}\sum_{k=1,k\neq i}^{n}h_{ki}^{\;T}\,\mathrm{cov}(Y_k)h_{ki}$$

is negligible. Let $\hat{h}_{ii}=h_{ii}(\hat{\beta}_{GEE},\alpha)$ and $\hat{S}_i=S_i(\hat{\beta}_{GEE})$. It is hard to tell whether (8) is a good estimator, because the assumption is not always reasonable. If $R(\alpha)$ correctly specifies the correlation structure, the expectation of the estimator defined by (8) has the following expression:

$$\begin{aligned}&\mathrm{cov}(Y_i)\big|_{(\beta_0,\alpha)}+\{(I_i-n^{-1}h_{ii}^{\;T})^{-1}n^{-1}\\&D_i[\dot{g}_{n,0}]^{-1}D_i^{\;T}(I_i-n^{-1}h_{ii})^{-1}\}\big|_{(\beta_0,\alpha)}\\&+O(n^{-\frac{3}{2}})\end{aligned}$$

and the estimator is biased upwards with order $O(n^{-1})$. This makes it more conservative than the residual estimation. For the residual estimator of $\mathrm{cov}(Y_i)$,

$$\begin{aligned}&\mathrm{E}[(Y_i-\hat{\mu}_i)(Y_i-\hat{\mu}_i)^{T}]\\&=\mathrm{cov}(Y_i)\big|_{(\beta_0,\alpha)}-n^{-1}D_i[\dot{g}_{n,0}]^{-1}D_i^{T}\big|_{(\beta_0,\alpha)}.\\&+O_p(n^{-\frac{3}{2}})\end{aligned} \tag{9}$$

Because

$$n^{-1}D_i(\beta_0)[\dot{g}_{n,0}(\beta_0,\alpha)]^{-1}[D_i(\beta_0)]^{T}$$

is positively definite, the residual estimator appears to be biased downward with order $O(n^{-1})$.

If the parameter values were known, one could use the following covariance estimator of the $\mathrm{cov}(Y_i)$:

$$\text{co}\hat{\text{v}}(\mathrm{Y}_i)_c = \{(I_i - n^{-1}h_{ii}{}^T)^{-1} S_i{}^c (I_i - n^{-1}h_{ii})^{-1}\}\Big|_{(\beta_0,\alpha)} \qquad (10)$$

where

$$S_i{}^c = (Y_i - \mu_i)(Y_i - \mu_i)^T - n^{-2} \sum_{k=1,k\neq i}^{n} h_{ki}{}^T (Y_k - \mu_k)(Y_k - \mu_k)^T h_{ki}.$$

The notation $\text{co}\hat{\text{v}}(Y_i)$ in (10) means an estimation of the $\text{cov}(Y_i)$. In this case, the first order asymptotic bias disappears, because

$$\mathrm{E}[\text{co}\hat{\text{v}}(Y_i)_c] = \text{cov}(Y_i) + O(n^{-\frac{3}{2}}).$$

Therefore, if the covariance estimator (10) was able to be used, the first order bias reduction would hold even if the correlation structure were not correctly specified. In practice, plug-in estimates are proposed

$$\hat{h}_{ik} = h_{ik}(\hat{\beta}_{GEE}, \alpha)$$

and

$$\hat{\mu}_i = \mu_i(\hat{\beta}_{GEE})$$

to get $\text{co}\hat{\text{v}}(Y_i)_c$.

If there is a common correlation structure $R(\alpha) = R_i(\alpha) = corr(Y_i)$, observations are pooled across different clusters to estimate $R(\alpha)$ by

$$\hat{R} = \frac{\phi}{n}\sum_{i=1}^{n} \Gamma_i{}^{-\frac{1}{2}} \text{co}\hat{\text{v}}(Y_i)_c \Gamma_i{}^{-\frac{1}{2}}, \qquad (11)$$

where $\text{co}\hat{\text{v}}(Y_i)_c$ and Γ_i are the same as before. The estimator $\hat{R}$ is similar to Liang and Zeger's suggestion for estimation of correlation structure (see Zeger & Liang, 1992; Zhao & Prentice, 1990; Fahrmeir & Tutz, 2001). Once estimation of the correlation matrix R is obtained, then, the $\text{cov}(Y_i)$ may be estimated by another way (also see Pan, 2001):

$$\text{co}\hat{\text{v}}(Y_i)_{new} = \phi^{-1}\Gamma_i{}^{\frac{1}{2}}\hat{R}\Gamma_i{}^{\frac{1}{2}} = \Gamma_i{}^{\frac{1}{2}}[\frac{1}{n}\sum_{k=1}^{n}\Gamma_k{}^{-\frac{1}{2}} \text{co}\hat{\text{v}}(Y_k)_c \Gamma_k{}^{-\frac{1}{2}}]\Gamma_i{}^{\frac{1}{2}}. \qquad (12)$$

The $\text{co}\hat{\text{v}}(Y_i)_{new}$ is a consistent estimator of $\text{cov}(Y_i)$.

If there is not a common correlation structure $R(\alpha)$ across all clusters, one may classify clusters into several groups such that all subjects in the same group have the same correlation structure, and then apply (12) to obtain a correlation matrix for that group.

Estimating Covariance Matrix Of GEE Estimator

It is known that the covariance matrix of the estimator $\hat{\beta}_{GEE}$ has the following approximation:

$$\text{cov}(\hat{\beta}_{GEE}) \approx \frac{1}{n^2}\{[\dot{g}_{n,0}]^{-1}\sum_{i=1}^{n} D_i{}^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i [\dot{g}_{n,0}]^{-1}\}\Big|_{(\beta_0,\alpha)}. \qquad (13)$$

If the $R(\alpha)$ is correctly specified, that is, if

$$\text{cov}(Y_i) = \phi^{-1}\Gamma_i{}^{\frac{1}{2}} R_i(\alpha) \Gamma_i{}^{\frac{1}{2}}$$

then the first order approximation to $\text{cov}(\hat{\beta}_{GEE})$ is $n^{-1}\phi^{-1}[\dot{g}_{n,0}]^{-1}\Big|_{(\beta_0,\alpha)}$. So, one can estimate $\text{cov}(\hat{\beta}_{GEE})$ by

$$\text{co}\hat{\text{v}}(\hat{\beta}_{GEE})_{\text{model}} = n^{-1}\hat{\phi}^{-1}\{[\dot{g}_{n,0}]^{-1}\Big|_{(\hat{\beta}_{GEE},\alpha)}\}. \qquad (14)$$

The estimate $\hat{\phi}$ may be obtained by

$$\hat{\phi} = \frac{1}{nm} \sum \hat{Z}_i^{\,T} \hat{Z}_i$$

where $\hat{Z}_i = \Gamma_i^{-\frac{1}{2}} (Y_i - \hat{\mu}_i)$. It was suggested (see Chaganty, 1997) that the $\hat{\phi}$ can be replaced by $\hat{\phi}_{bc} = nm\hat{\phi}/(nm - p)$ if a bias-corrected estimate for ϕ is preferable. However, the correlation structure could be mis-specified, that is

$$\operatorname{cov}(Y_i) \neq \phi^{-1} \Gamma_i^{\frac{1}{2}} R_i(\alpha) \Gamma_i^{\frac{1}{2}},$$

because the correlation matrix may not be known in practice. In this case, it is well known that the variance $\operatorname{cov}(\hat{\beta}_{GEE})$ can be estimated consistently by the sandwich formula

$$\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{sand} = \{[\dot{g}_{n,0}]^{-1} \frac{1}{n^2} \sum_{i=1}^{n} D_i^{\,T} V_i^{-1} \varepsilon_i \varepsilon_i^{\,T} V_i^{-1} D_i [\dot{g}_{n,0}]^{-1}\}\Big|_{(\hat{\beta}_{GEE},\alpha)} \tag{15}$$

where $\varepsilon_i = (y_{i1} - \mu_{i1}, \cdots, y_{im} - \mu_{im})^T$ are the residuals. As previously discussed, estimating $\operatorname{cov}(Y_i)$ by fitted $\hat{\varepsilon}_i \hat{\varepsilon}_i^{\,T}$ ($\hat{\varepsilon}_i = \varepsilon_i \big|_{(\hat{\beta}_{GEE},\alpha)}$) could be biased downward. Thus, the sandwich estimate $\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{sand}$ will be biased downward for estimating $\operatorname{cov}(\hat{\beta}_{GEE})$. Recently, the bias corrected sandwich estimators have been provided by Mancl and DeRouen (2001) and Kauermann and Carroll (2001), where the estimation of $\operatorname{cov}(\hat{\beta}_{GEE})$ is obtained by replacing $\hat{\varepsilon}_i \hat{\varepsilon}_i^{\,T}$ by $\operatorname{co\hat{v}}(Y_i)_c$ defined by (10), that is

$$\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{sand_u} = [\dot{g}_{n,0}]^{-1} \frac{1}{n^2} \sum_{i=1}^{n} D_i^{\,T} V_i^{-1} \operatorname{co\hat{v}}(Y_i)_c V_i^{-1} D_i [\dot{g}_{n,0}]^{-1}\Big|_{(\hat{\beta}_{GEE},\alpha)} \tag{16}$$

Finally, if $\operatorname{co\hat{v}}(Y_i)_{new}$ is used, a more efficient sandwich estimator could be obtained:

$$\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{new} = \{[\dot{g}_{n,0}]^{-1} \frac{1}{n^2} \sum_{i=1}^{n} D_i^{\,T} V_i^{-1} \operatorname{co\hat{v}}(Y_i)_{new} V_i^{-1} D_i [\dot{g}_{n,0}]^{-1}\}\Big|_{(\hat{\beta}_{GEE},\alpha)} \tag{17}$$

Consider the following:

Theorem:

$$\operatorname{cov}(vec(\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{sand})) - \operatorname{cov}(vec(\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{new})) = \Omega_n - \delta_n$$

where Ω_n is nonnegative definite, δ_n has higher order than Ω_n and the operator "vec" is used to stack the columns of a matrix together to obtain a vector.

Proof: Because $\hat{\beta}_{GEE}$ is $\sqrt{n}$-consistent, expand $\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{new}$ and $\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{sand}$ at (β_0, α). Then, the following expansions are obtained:

$$\operatorname{co\hat{v}}(\hat{\beta}_{GEE})_{new} = \{[\dot{g}_{n,0}]^{-1} \frac{1}{n^2} \sum_{i=1}^{n} D_i^{\,T} V_i^{-1} \Gamma_i^{\frac{1}{2}} \frac{1}{n} \sum_{k=1}^{n} \Gamma_k^{-\frac{1}{2}} H_{ii}^{\,T} \varepsilon_k \varepsilon_k^{\,T} H_{ii} \Gamma_k^{-\frac{1}{2}} \Gamma_i^{\frac{1}{2}} V_i^{-1} D_i [\dot{g}_{n,0}]^{-1}\}\Big|_{(\beta_0,\alpha)} + O_p(n^{-\frac{3}{2}})$$

where $H_{ii} = (I - n^{-1} h_{ii})^{-1}$. Similarly,

$$\begin{aligned}
&\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand} \\
&= \{[\dot{g}_{n,0}]^{-1} \frac{1}{n^2} \sum_{i=1}^{n} D_i{}^T V_i{}^{-1} \varepsilon_i \\
&\varepsilon_i{}^T V_i{}^{-1} D_i [\dot{g}_{n,0}]^{-1}\}\Big|_{(\beta_0,\alpha)} + O_p(n^{-\frac{3}{2}})
\end{aligned}$$

By Theorem 7.16 in Schott (1997),

$$\begin{aligned}
&vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand}) \\
&\approx \frac{1}{n^2} \sum_{i=1}^{n} \{A_{i,n} vec(\varepsilon_i \varepsilon_i{}^T)\}\Big|_{(\beta_0,\alpha)}
\end{aligned} \tag{18}$$

and

$$\begin{aligned}
&vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new}) \\
&\approx \frac{1}{n^2} \sum_{i=1}^{n} \{A_{i,n} vec\{\Gamma_i^{\frac{1}{2}} \frac{1}{n} \sum_{k=1}^{n} \Gamma_k{}^{-\frac{1}{2}} H_{ii}{}^T \varepsilon_k \\
&\varepsilon_k{}^T H_{ii} \Gamma_k{}^{-\frac{1}{2}} \Gamma_i^{\frac{1}{2}}\}\}\Big|_{(\beta_0,\alpha)} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \{A_{i,n} \{\frac{1}{n} \sum_{k=1}^{n} B_{k,i} vec(\varepsilon_k \varepsilon_k{}^T)\}\}\Big|_{(\beta_0,\alpha)},
\end{aligned} \tag{19}$$

where

$$A_{i,n} = ([\dot{g}_{n,0}]^{-1} D_i V_i{}^{-1}) \otimes ([\dot{g}_{n,0}]^{-1} D_i V_i{}^{-1})$$

and

$$B_{k,i} = (\Gamma_i^{\frac{1}{2}} \Gamma_k{}^{-\frac{1}{2}} H_{ii}) \otimes (\Gamma_i^{\frac{1}{2}} \Gamma_k{}^{-\frac{1}{2}} H_{ii}) .$$

The covariance matrices of

$$vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new})$$

and

$$vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand})$$

can be obtained from (19) and (18):

$$\begin{aligned}
&\operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand})) \\
&\approx \frac{1}{n^4} \sum_{i=1}^{n} \{A_{i,n} \operatorname{cov}(vec(\varepsilon_i \varepsilon_i^T) A_{i,n}{}^T\}\Big|_{(\beta_0,\alpha)}
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new})) \\
&\approx \frac{1}{n^6} \sum_{i=1}^{n} \{\sum_{k=1}^{n} A_{i,n} B_{k,i} \operatorname{cov}(vec(\varepsilon_k \varepsilon_k{}^T) \\
&B_{k,i}{}^T A_{i,n}{}^T\}\Big|_{(\beta_0,\alpha)} .
\end{aligned}$$

Notice that $vec(\varepsilon_i \varepsilon_i{}^T)$ $(i = 1, \cdots, n)$ are independent and free of n. It is clear that $\|B_{k,i}\|$ is bounded when $n \to \infty$. Hence, under some regularity conditions (see details in Zhang, 2003), there is the following result:

$$\begin{aligned}
&\frac{1}{n^2} \sum_{k=1}^{n} \{B_{k,i} \operatorname{cov}(vec(\varepsilon_k \varepsilon_k^T)) B_{k,i}{}^T\}\Big|_{(\beta_0,\alpha)} \\
&= O(n^{-1}),
\end{aligned}$$

as $n \to \infty$. Finally,

$$\begin{aligned}
&\operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand})) \\
&- \operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new})) \\
&= \frac{1}{n^4} \sum_{i=1}^{n} A_{i,n} (\operatorname{cov}(vec(\varepsilon_i \varepsilon_i^T) \\
&- O_p(n^{-1})) A_{i,n}{}^T \Big|_{(\beta_0,\alpha)} \\
&= (\Omega_n - \delta_n)\Big|_{(\beta_0,\alpha)} .
\end{aligned}$$

where

$$\Omega_n = \frac{1}{n^4} \sum_{i=1}^{n} A_{i,n} \operatorname{cov}(vec(\varepsilon_i \varepsilon_i{}^T)) A_{i,n}{}^T$$

is a non-negative definite matrix and the δ_n has higher order of convergence to zero than Ω_n. Thus, it has been proven that

$$\begin{aligned}&\operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{sand}))\\&-\operatorname{cov}(vec(\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new}))\geq 0\end{aligned}$$

asymptotically. The proof of the Theorem is completed.

In summary, the estimator of the covariance matrix of regression parameters could gain some efficiency. Also it is expected that the method is more plausible for small sample sizes n than other estimators of the covariance.

For construction of confidence intervals, inference about $L^T\beta$ is of interest, where L^T is a $1\times p$ dimensional contrast vector of unit length, that is, $L^T L=1$. If the $R(\alpha)$ is correctly specified, then the first-order of approximation of $\operatorname{var}(L^T\hat{\beta}_{GEE})$ is $\phi^{-1}L^T[\dot{g}_{n,0}]^{-1}L\Big|_{(\beta_0,\alpha)}$. Thus, $\operatorname{var}(L^T\hat{\beta}_{GEE})$ can be estimated by

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{mod}el}&=\hat{\operatorname{var}}(L^T\hat{\beta}_{GEE})_{\mathrm{mod}el}\\&=\frac{1}{n}\hat{\phi}^{-1}L^T[\dot{g}_{n,0}]^{-1}L\Big|_{(\hat{\beta}_{GEE},\alpha)}\end{aligned}\tag{20}$$

Based on (20), a symmetric confidence interval is given by

$$(\hat{\theta}\pm z_q\hat{\sigma}_{\mathrm{mod}el}),\tag{21}$$

where z_q is the q quantile of the standard normal distribution and $\hat{\theta}=L^T\hat{\beta}_{GEE}$.

Corresponding to this estimate, another symmetric confidence interval is obtained

$$(\hat{\theta}\pm z_q\hat{\sigma}_{sand_u}).\tag{25}$$

Based on the estimation of the covariance matrix (17), if the $R_i(\alpha)$ is mis-specified, the variance $\operatorname{var}(L^T\hat{\beta}_{GEE})$ can be estimated consistently by the sandwich formula

$$\begin{aligned}\hat{\sigma}^2_{sand}&=\hat{\operatorname{var}}(L^T\hat{\beta}_{GEE})_{sand}\\&=L^T[\dot{g}_{n,0}]^{-1}\frac{1}{n^2}\{\sum_{i=1}^{n}D_i^TV_i^{-1}\varepsilon_i\\&\varepsilon_i^TV_i^{-1}D_i\}[\dot{g}_{n,0}]^{-1}L\Big|_{(\hat{\beta}_{GEE},\alpha)}\end{aligned}\tag{22}$$

where the ε_i's are the same as before. Then, based on (22), the symmetric confidence interval is given by

$$(\hat{\theta}\pm z_q\hat{\sigma}_{sand}).\tag{23}$$

It follows from the discussion that the sandwich estimate appears to be biased downward. Therefore, the bias corrected sandwich estimation of $\operatorname{var}(L^T\hat{\beta}_{GEE})$ can be obtained by replacing $\varepsilon_i\varepsilon_i^T$ by $\operatorname{cov}(Y_i)_c$ defined by (10). Thus, the bias reduced sandwich estimate of the variance $\operatorname{var}(L^T\hat{\beta}_{GEE})$ is obtained by

$$\begin{aligned}\hat{\sigma}^2_{sand_u}&=\hat{\operatorname{var}}(L^T\hat{\beta}_{GEE})_{sand_u}\\&=L^T[\dot{g}_{n,0}]^{-1}\frac{1}{n^2}\{\sum_{i=1}^{n}D_i^TV_i^{-1}\\&\operatorname{cov}(Y_i)_cV_i^{-1}D_i\}[\dot{g}_{n,0}]^{-1}L\Big|_{(\hat{\beta}_{GEE},\alpha)}\end{aligned}\tag{24}$$

$$\begin{aligned}\hat{\sigma}^2_{new}&=\hat{\operatorname{var}}(L^T\hat{\beta}_{GEE})_{new}\\&=L^T\hat{\operatorname{cov}}(\hat{\beta}_{GEE})_{new}L.\end{aligned}\tag{26}$$

Then, a confidence interval is obtained:

$$(\hat{\theta}\pm z_q\hat{\sigma}_{new}).\tag{27}$$

Simulation Study and Discussions

Suppose that y_{ij} has marginally a negative binomial distribution, that is, $y_{ij} \sim NB(1, \mu_{ij})$, i =1,...,n and j =1,...,m. The link function is *log*, i.e. $\log(\mu_{ij}) = x_{ij}^{T}\beta$, where $\beta = (\beta_0, \beta_1, \beta_2)^T$ and $x_{ij} = (1, x_{ij1}, x_{ij2})^T$ are the covariates: $x_{ij2} \sim N(0,1)$ and x_{ij1} are constants. The correlation structure among $y_{i1}, \cdots, y_{im}$ is assumed to be given as an AR(1) with ρ =0.8. Now, the procedures developed in the last two sections are applied to the model $\mathrm{E}(y_{ij}) = e^{x_{ij}^{T}\beta}$. The simulation study is completed for the number n of clusters as 10, 20, 30, $\cdots$, 90, 100 respectively.

A comparison of the performance of the estimators of the asymptotic variances is required. The estimators, $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_{GEE})$, $\hat{\sigma}^2_{sand}(\hat{\beta}_{GEE})$, $\hat{\sigma}^2_{sand_u}(\hat{\beta}_{GEE})$, and $\hat{\sigma}^2_{new}(\hat{\beta}_{GEE})$, are defined by taking the vector L in an appropriate form in (20), (22), (24) and (26). Each of these variance estimators is related to a specified correlation structure $R_i(\alpha)$.

First, the situation is observed, where the $R_i(\alpha)$ in the estimators of variances are correctly specified to a constant. Figure 1 shows the comparisons of $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_1)$, $\hat{\sigma}^2_{sand}(\hat{\beta}_1)$, $\hat{\sigma}^2_{sand_u}(\hat{\beta}_1)$, and $\hat{\sigma}^2_{new}(\hat{\beta}_1)$ and the true variance (empirical variance) $\mathrm{var}(\hat{\beta}_1)$ over 1000 simulations, when the regression parameters are estimated by the GEE estimator. From Figure 1, it is found that the estimator $\hat{\sigma}^2_{new}$ of the variance is better than other three, since the biases are smaller, even for the clusters with small sample size.

The curves shown in Figure 1 are consistent with the property that all four estimators are asymptotically unbiased. Notice that, in all these plots, the sandwich estimator $\hat{\sigma}^2_{sand}(\hat{\beta}_1)$ has the biggest bias when the sample size is small. It corresponds to the fact that the sandwich estimator would be expected to underestimate the variance of $\hat{\beta}_1$. It is not surprising that the model based estimator $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_1)$ performs better than the sandwich estimator because the model is correct (the $R_i(\alpha)$ is correctly specified except for the constant α).

When the model is mis-specified, for example, if $R_i(\alpha)$ is an identity matrix, the model based estimator $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_1)$ is the worst one. Figure 2 shows that (i) estimators $\hat{\sigma}^2_{sand}(\hat{\beta}_{GEE})$, $\hat{\sigma}^2_{sand_u}(\hat{\beta}_{GEE})$, and $\hat{\sigma}^2_{new}(\hat{\beta}_{GEE})$ are asymptotically unbiased; (ii) the $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_{GEE})$ is significantly biased; (iii) the new estimator $\hat{\sigma}^2_{new}(\hat{\beta}_1)$ of the variance is the best one to estimate the $\mathrm{var}(\hat{\beta}_1)$.

Now, the efficiency of the variance estimators is compared. For Figure 3, the study is based on 1000 simulations for each number of clusters being 10, 20, $\cdots$, 100 respectively. The variances are calculated by

$$\mathrm{var}(\hat{\sigma}^2_{estimator}) = s^2_{\ estimator} \ ,$$

where $s^2_{\ estimator}$ is sample variance of values of $\hat{\sigma}^2_{estimator}$ which is obtained from the formula in the last section for each simulation. The estimator can be “model”, “sand”, “$sand_u$ ” and “new” respectively. Figure 3 illustrates that the corrected sandwich variance estimator $\hat{\sigma}^2_{sand}(\hat{\beta}_1)$ has the biggest standard error even for large sample size.

When the correlation structure is correctly specified, the model based estimator $\hat{\sigma}^2_{\mathrm{mod}el}(\hat{\beta}_1)$ could be better than the corrected sandwich variance estimator, especially, when the sample size is small. When the number of clusters is greater than 30, the simulation shows that new variance estimator is the most stable one. It follows from Figure 4 that these facts still hold when the correlation structure is mis-specified in the variance estimators in the manner of the example. Of course, the model based variance estimator should not be used in this case because it is biased, although its variance is the smallest one. If the sample size is small, the sandwich estimator performs well.

With variance estimators at hand, confidence intervals could be constructed with different variance estimators. It will be seen that the confidence intervals obtained by the new variance estimator perform better than the other three in terms of coverage probability. The problem of testing a null hypothesis $H_0 : \beta \in \vartheta_0$ will be considered. Essentially, confidence intervals are closely related with tests. The aim is to compare CI's which are related to the various estimators introduced in the third sections of this article. In the simulation study, the CI for β_1 corresponds to a test that $H_0 : \beta_1 = \beta_{10}$. The test statistic could be $T_{new} = (\hat{\beta}_1 - \beta_{10}) / \hat{\sigma}_{new}(\hat{\beta}_1)$ or other ones obtained by different variance estimators. It follows from Figure 5 that the coverage percentages with the new variance estimator are bigger; therefore, the confidence interval based on the new variance estimator is accurate for smaller sample sizes than other ones with the variance estimators 'model', 'sand' or 'sand$_u$' .

It appears to be better to use the new variance estimator to construct confidence intervals, especially when the sample size is small. In the example of a mis-specified correlation structure in the variance estimators, the new and adjusted sandwich estimators both give accurate confidence intervals (see Figure 6). Again, the model based variance estimator should not be used in this case.

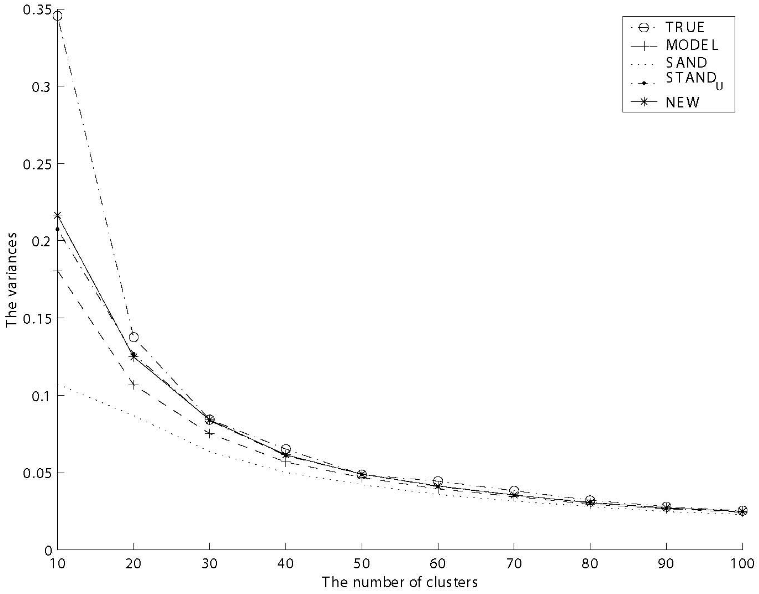


Figure 1

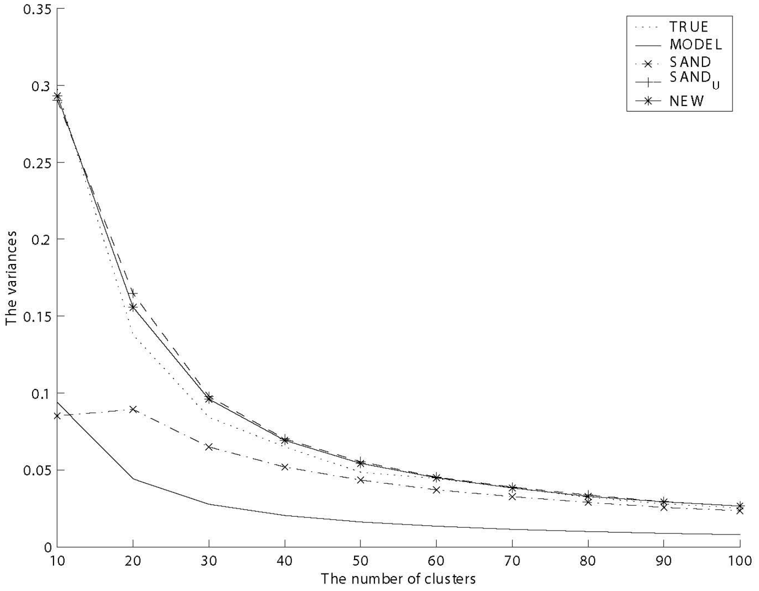


Figure 2

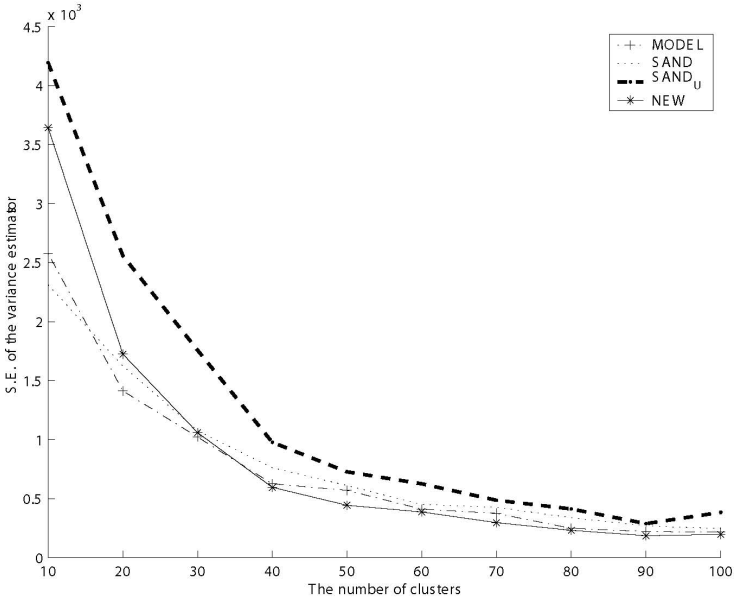


Figure 3

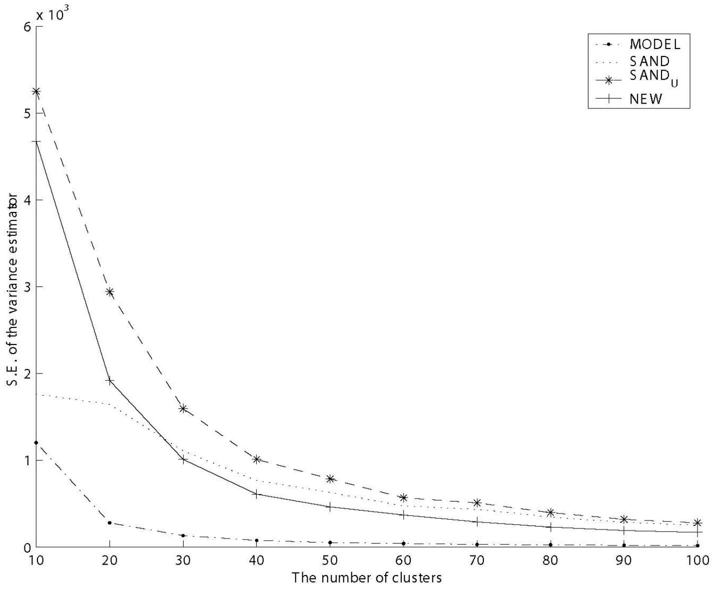


Figure 4

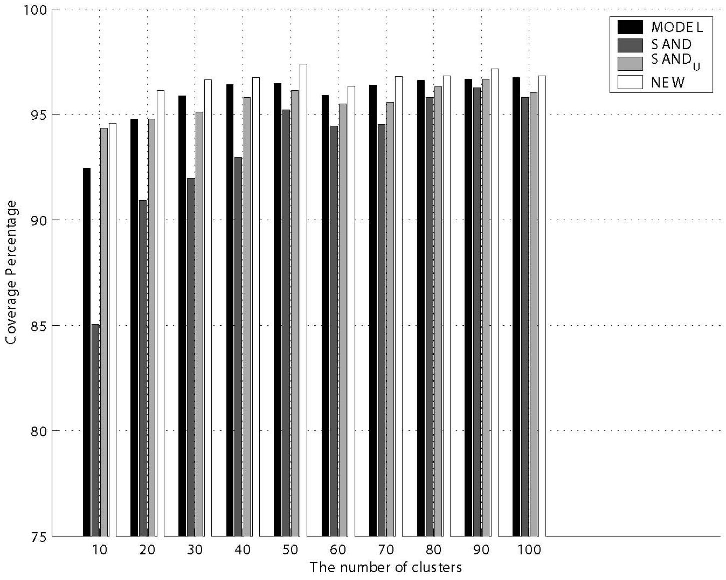


Figure 5

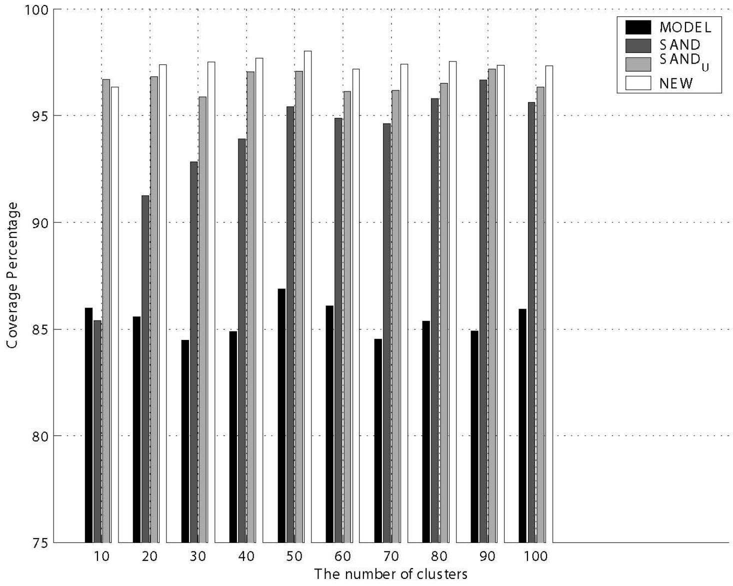


Figure 6

References

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistica Planning and Inference*, *63*, 39–54.

Diggle, P. J., Liang, K. Y., & Zeger, S. T. (1994). *Analysis of longitudinal data*. NY: Oxford Science Publications.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modeling based on gerealized linear models*. (2nd ed.). NY: Springer.

Kauermann, G., & Carroll, R. J. (2001). *Journal of the American Statistical Association*, *96*, 1387-1396.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 12-22.

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, *57*, 126-134.

Pan, W. (2001). On the robust variance estimator in generalized estimating equations. *Biometrika, 3*, 901-906.

Schott, J. R. (1997). *Matrix analysis for statistics*. NY: Wiley.

Zeger, S. L., & Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, *11*, 1825-1839.

Zhang, S. (2003). *Statistical analysis for clustered data*. Unpublished doctoral dissertation. Ontario, Canada: University of Waterloo.

Zhao, L. P., & Prentice, R. L. (1990). Correlated binary regression using a generalized quadratic models. *Biometrika*, *77,* 642-648.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 230-247

1538 – 9472/06/$95.00

The Use Of Hierarchical ANCOVA In Curriculum Studies

Show-Mann Liou
National Taiwan Normal University

Chao-Ying Joanne Peng
Indiana University-Bloomington

Many educational studies are carried out in intact settings, such as classrooms or groups in which individual data were collected before and after a treatment. Researchers advocate either the use of individual scores as the unit of analysis or class means. Both approaches suffer from conceptual and methodological limitations. In this article, the use of hierarchical ANCOVA for analyzing quasi-experimental data including baseline measures is designed and promoted. It is illustrated with a real-world data set collected from a curriculum study. Results showed that the hierarchical ANCOVA is a conceptually and methodologically sound approach, and is better than ANCOVA based on individual scores or ANCOVA based on class means. The potential of using hierarchical ANCOVA designs for curriculum studies is discussed in terms of statistical power and congruence with study plans.

Key words: Educational research methodology, hierarchical ANCOVA, Project Citizen, civic education, civic skills, civic dispositions, adolescent students

Introduction

Among educational research methods, true experiments are designed to investigate causes and consequences in behavior (Fraenkel & Wallen, 2000; McMillan & Schumacher, 2001). However, most circumstances in education prevent the possibility of random selection and random assignment of subjects into experimental and control conditions. Consequently, the use of true experiments is limited in educational research. Instead, quasi-experiments are much more prevalent.

Even with quasi-experiments, educational researchers are faced with another difficulty that weakens the internal validity of a study. Namely, students in the same classroom are often administered the same treatment by the same instructor making their performances not statistically independent. Consider a study in which a researcher is interested in studying the effectiveness of two instructional strategies on students' achievement in biology. To carry out this study, a researcher may randomly select intact classes and train teachers of these classes to implement the instructional strategies. Consequently, students in a classroom cannot be randomly assigned to learn from a particular strategy, nor can teachers teach students independently or in isolation. To account for the difference in students' achievement that already existed in the beginning of the study and to compensate for the lack of independence among students' performances, a researcher can administer a pretest to determine a baseline measure of the outcome (i.e., biology achievement in this case). A one-way analysis of covariance (ANCOVA) can be subsequently applied to posttest measures to test differences due to the two strategies while statistically controlling for pretest differences. The ANCOVA approach has been a method of choice since Lindquist (1940) brought to light

Show-Mann Liou, Associate Professor of Civic Education & Leadership. Research interests include citizenship education and experimental design. C.-Y. Joanne Peng, Professor of Inquiry Research Methodology. Research interests include logistic regression and missing data. Send correspondence to Show-Mann Liou, Department of Civic Education and Leadership, National Taiwan Normal University, Hoping E. Rd. Sec. 1 No. 162, Taipei, Taiwan 106, phone: 886-2-2369-8673 ex.31, fax: 886-2-2363-8821, email: t11033@ ntnu.edu.tw.

the issues with non-independence in subjects' responses in intact groups.

It is generally agreed that ANCOVA is an appropriate statistical technique for analyzing quasi-experimental data with baseline measures as long as its assumptions—linearity and independence between the covariate and the independent variable—are met (Buser, 1995; Henson, 1998; Hines & Foil, 2000; Loftin & Madison, 1991). There is, however, one issue remaining: what is the proper unit of analysis in quasi-experimental studies, class means or individual scores? (Barcikowski, 1981; Blair & Higgins, 1986; Hopkins, 1982; Morran, Robison, & Hulse-Killacky, 1990; Peckham, Glass, & Hopkins, 1969).

The issue has generated and received considerable attention in the literature ever since Lindquist (1940) presented an argument and rationale for using group means as the unit of analysis for data collected from intact groups. At the heart of the disagreement is: what is the most appropriate unit for data analysis and interpretation? With the use of individual scores, it is assumed that students in the same classroom are unrelated, as far as treatments are concerned, and therefore statistically independent. This assumption and its computational approach could lead to an overestimation of treatment effects with sufficiently large samples. Conversely, using group means as the unit of analysis ensures that the independence assumption is met, at the individual level, and the interpretation of the data has internal validity (Peckham, Glass, & Hopkins, 1969). However, this approach results in a great loss in sample size; hence, a decrease in statistical power (Barcikowski, 1981). Furthermore, the use of group means limits the generalizability of the findings only to classes, and results may not be informative to educators in general. It is evident from the brief summary that each approach has its own conceptual and methodological limitations.

This article addresses the limitations raised above regarding the use of these two traditional ANCOVAs, one based on individual's scores and the other on group means, and proposes a third approach. This approach applies the hierarchical ANCOVA to data collected from intact settings such as classrooms. It will be shown that the hierarchical ANCOVA is a conceptually and methodologically sound analytical approach that is well suited to educational research. Specifically, this approach isolates the nuisance variable of classes and incorporates the inherent hierarchical nature of the data structure into the analysis. Consequently, this approach not only takes into account the independence assumption required of individuals' scores but also makes valid and meaningful inferences at the individual's level.

The hierarchical ANCOVA is introduced and demonstrated using a real world data set (Liou, 2002). The Liou study was primarily interested in the effects of *We the People…Project Citizen* on civic skills and four dimensions of the civic dispositions of adolescent students. The study exemplified most educational research in which classrooms are randomly selected or even assigned to treatment conditions but students are not. Furthermore, students' levels of civic skills and civic dispositions were assessed both before and after the implementation of *Project Citizen*. Data were analyzed by three methods: ANCOVA based on individual scores, ANCOVA based on class means, and hierarchical ANCOVA based on individual scores. Results from the three methods were shown to be different; they were interpreted in terms of substantive implications and methodological considerations (i.e., statistical power, practical as well as statistical significance). Recommendations and implications for educational researchers are offered in light of the relative superiority of hierarchical ANCOVA over the other two methods.

Design Structures: Crossed and Nested (Hierarchical) Designs

To ensure the internal and external validities of statistical analysis of quasi-experiments, one should carefully plan two aspects of a study: the structure of the design and the unit of analysis. Specifically, two major structures are possible for a quasi-experimental design: crossed and nested (or hierarchical) (Peng, 2004). Likewise, two types of units of analysis need to be distinguished conceptually and computationally: the unit of research design and the unit of statistical analysis.

A crossed design employs all combinations of levels of two or more independent variables in a study. It is typically used to test differences in a dependent variable due to main effects of independent variables and their interactions. A nested design is a research design in which levels of one independent variable (say *B*) are hierarchically subsumed under (or nested within) levels of another independent variable (say *A*). As a result, assessing the complete combination of *A* and *B* levels is not possible in a nested design.

Nested design is alternatively called hierarchical design; it is used most often in quasi-experimental studies in which researchers have little or no control over random assignment of observations into treatment conditions. The design is popular, and sometimes necessary, among curriculum studies, clinical, sociological, and ethological research in which participants belong to intact groups (such as classes, therapeutic groups, etc.); these intact groups cannot be dismantled to allow for a random assignment of participants into different treatment conditions.

Many studies in education can be carried out only in nested designs. Consider the example mentioned earlier in which instructional strategies are administered in classroom settings. Even though students individually learn and are tested on their achievement in biology, their learning effects are to an extent dependent on the learning environment and dynamics of interactions among peers. Thus, students are nested within classrooms which in turn are nested within instructional strategies. In this case, a crossed design neglects the hierarchical nature of the data and produces incorrect interpretations of the results. According to Roberts (2000), neglecting a nested design leads to the following consequences:

> Neglecting a nested design when one actually exists will make the research: (1) wrongly attribute a main effect to an interaction effect when, in fact, no interaction exists; (2) divide by the wrong degrees of freedom when determining the mean square and *F*-value (and the statistical significance of the *F*-value); and (3) assume that a main effect has a smaller effect size (eta square) because the sum of squares for that effect is being partly attributed to the interaction effect. (Roberts, 2000, p. 6)

Unit of Research Design and Unit of Statistical Analysis

Another issue that should be taken into consideration when analyzing quasi-experiments is the unit of analysis. Valid statistical inferences from data depend on the compatibility between the unit of a research design and that of statistical analysis (Peckham, Glass & Hopkins, 1969; Glass & Stanley, 1970; Morran, Robinson, & Hulse-Killacky, 1990). Units of a research design refer to entities that are allocated to a condition of the independent variable, independently from other entities. Units of statistical analysis refer to entities whose measures or scores form the basis of statistical inferences. Clearly, a research design unit can be either individuals or classes. Even if classes are the research design units, students' scores can still be treated as units of statistical analyses.

When analyzing data in an ANOVA framework, educational researchers may, and frequently do, make an a priori decision to treat individuals as the unit of statistical analysis (Morran, Robinson, & Hulse-Killacky, 1990). Several reasons contribute to this decision. One is to ensure that the statistic, whether it is *F*, *q*, or *t*, is tested with the maximal *df* based on the sample. Another reason for regarding individuals as the unit of analysis is to retain the variability at the individual level, thus, maximizing information a research can glean from the data. This approach further affords researchers the opportunity to study the effects of certain organismic or demographic characteristics and their interactions with independent variables on the dependent variable (Hopkins, 1982; Morran, Robinson, & Hulse-Killacky, 1990; Peckham, Glass, & Hopkins, 1969). It is impossible to study these effects if group means are analyzed. Thus, the group means approach ignores the hierarchical nature of the data collected in typical educational settings and consequently impoverishes inferences that may be drawn at the individual level.

Yet, a few researchers advocate the use of group means on statistical grounds. They argue that participants studied in intact settings are not the appropriate unit of analysis since they fail to meet the independence assumption. The result of such a violation is deflated within-group variability, hence, inflated treatment effects. In a typical educational setting, the classroom provides a shared educational experience; thus, students are not entirely independent insofar as sampling errors are concerned. According to Peckham, Glass, and Hopkins (1969), "violating the assumption of independence of errors may substantially affect the validity of probability statements" (p.338). They concluded that the use of group means promotes "the greatest insurance that the independence assumption has been met" (p.344); and therefore statistical inferences from the result are valid. Some proponents went further in arguing that when the independent assumption is not tenable, treating individuals as the unit of statistical analysis leads to non-replicable findings.

As Hopkins (1982) showed that the recommendation of using class means proves to be restrictive, unnecessary, and less powerful than alternatives that are derived directly from individual data and proper statistical models. A better treatment of the inter-dependence among units of observation is to employ an efficient statistical modeling technique, such as the hierarchical ANCOVA, that adequately represents the condition under which data were collected and provides the greatest statistical power and external validity.

Hierarchical ANCOVA

In light of the issues raised in the preceding two sections, it is not without understanding that the two ordinary ANCOVA's – one based on class means and the other on individual scores – are unlikely to yield satisfactory interpretation of data collected from hierarchical settings that include pretests or baseline measures. In their places, researchers have proposed that nested or hierarchical ANCOVA be used in order to account for variances due to treatments, classes, and individual students nested within classrooms (Hopkins, 1982; Lindman, 1992; Morran, Robison & Hulse-Killacky, 1990; Robert, 2000). Hierarchical ANCOVA combines features from a hierarchical research design with those of analysis of covariance.

Assume that a researcher wishes to study the effect of Internet search strategies (Factor *A*) on college students' information seeking efficiency (the dependent variable). Six classes of freshmen English at a state college are randomly selected; three classes are assigned to the linear search condition and the other three to the nonlinear search condition. At the onset of the study, all freshmen are assessed in terms of their information seeking efficiency. These measures will be treated as covariates in analysis of covariance. Figure 1 illustrates the research design.

Because freshmen enrolled in these classes form intact groups, they cannot be randomly assigned to the two treatment conditions on an individual basis. Furthermore, their learning processes and behaviors are likely to be mutually dependent; differences in students' information seeking behavior among classes are embedded within each treatment condition. This restriction makes this design a nested design rather than a fully crossed design. In addition, the pretest measures taken from all participants can serve as a covariate in the hierarchical ANCOVA model presented below:

$$Y_{ijk} = \mu_y + beta(X_{ijk} - \mu_x) + \alpha_j + \beta_{k(j)} + e_{i(jk)}, \tag{1}$$

Where

i = 1, …,*n* (number of freshman in a class, say, 20);

j = 1, …,*p* (number of treatment condition=2 in this example);

k= 1,…,*q* (number of classes=3 in this example);

Factor *A*
Internet Search Strategy

		Treatment 1		Treatment 2	
		Linear		Nonlinear	
		Pretest	Posttest	Pretest	Posttest
Factor *B*	Class 1	$\overline{X}_{1(1)}$	$\overline{Y}_{1(1)}$		
	Class 2	$\overline{X}_{2(1)}$	$\overline{Y}_{2(1)}$		
Freshman English Class	Class 3	$\overline{X}_{3(1)}$	$\overline{Y}_{3(1)}$		
	Class 4			$\overline{X}_{4(2)}$	$\overline{Y}_{4(2)}$
	Class 5			$\overline{X}_{5(2)}$	$\overline{Y}_{5(2)}$
	Class 6			$\overline{X}_{6(2)}$	$\overline{Y}_{6(2)}$

Figure 1

Y_{ijk} is the dependent score of the *i*th participant in the *j*th level of Factor *A* and *k*th level of Factor *B*;

μ_y is the population mean of the dependent scores;

beta is the pooled within-group regression coefficient derived from regressing the covariate score, X_{ijk} on the dependent score Y_{ijk};

X_{ijk} is the covariate measure (such as the pretest score) of the *i*th participant in the *j*th level of Factor *A* and *k*th level of Factor *B*;

μ_x is the population mean of the covariate measures;

α_j is the effect of the *j*th treatment condition of Factor *A*; algebraically, it equals the deviation of the *j*th population mean (μ_{y_j}) from the grand mean (μ_y). It is a constant for all participants' dependent scores in the *j*th condition, subject to the restriction that all α_j sum to zero across all conditions.

$\beta_{k(j)}$ is the effect of the *k*th condition under Factor *B*, nested within the *j*th level of Factor *A*; algebraically, it equals the deviation of the population mean ($\mu_{y_{jk}}$) in the *k*th and *j*th combined level from the grand mean (μ_y). It is a constant for all observations' dependent scores in the *k*th condition, nested within Factor *A*'s *j*th condition. The effect is assumed to be normally distributed in its underlying population.

$e_{i(jk)}$ is the random sampling error associated with the *i*th participant in the *j*th condition of Factor *A* and *k*th condition of Factor *B*. It is a random variable that is normally distributed in the underlying population and is independent of $\beta_{k(j)}$. In comparison, the ordinary ANCOVA model based on individual scores does not examine nor acknowledge the nested effect, $\beta_{k(j)}$ in its model as follows:

$$Y_{ijk} = \mu_y + beta(X_{ijk} - \mu_x) + \alpha_j + e_{ijk}, \qquad (2)$$

where all terms are defined as previously, except that there is no $\beta_{k(j)}$ effect and no nested effect of classes within treatment conditions.

A third approach, i.e., the ordinary ANCOVA based on class means, follows the same model as model (2) except that data are aggregated over the entire class before they are analyzed by the ANCOVA model as stated below:

$$Y_{jk} = \mu_y + beta(X_{jk} - \mu_x) + \alpha_j + \varepsilon_{jk}, \tag{3}$$

where

$j =$ 1, …,p (number of treatment condition=2 in this example);

$k=$ 1,…,q (number of classes=3 in this example);

Y_{jk} is the average dependent score of the kth class in the jth level of Factor A;

μ_y is the population mean of average class dependent scores;

beta is the within-group regression coefficient derived from regressing the covariate score, X_{jk} on the dependent score Y_{jk};

X_{jk} is the average covariate measure (such as the pretest score) of the kth class in the jth level of Factor A;

μ_x is the population mean of average class covariate measures;

α_j is the effect of the jth treatment condition of Factor A; algebraically, it equals the deviation of the jth population mean (μ_{y_j}) from the grand mean (μ_y). It is a constant for all class average dependent scores in the jth condition, subject to the restriction that all α_j sum to zero across all conditions;

e_{jk} is the random sampling error associated with the kth class in the jth condition of Factor A. It is a random variable that is normally distributed in the underlying population.

Note in model (3), the i subscript is no longer present due to the fact that individuals are not the unit of analysis. Instead, class means are used; they are denoted by the k subscript.

Statistical Assumptions and Tests

The null hypothesis (H_0) for all the three models is identical, namely, the parameter α_j equals zero in the population for all conditions (or linear search and nonlinear search according to the present example). The alternative hypothesis (H_1) states that some of the α_j's do not equal zero. To test the null hypothesis according to models (1), (2), or (3), data are organized to form a ratio of mean squares treatment (MS_t) over mean squares error (MS_e). The ratio is distributed as a central F distribution under the null hypothesis but non-central F distribution under the alternative, provided that statistical assumptions are met. For all three models, it is assumed that random sampling errors [$e_{i(jk)}$, e_{ijk}, or e_{jk}] are normally distributed, homogeneous in variances, and independent from each other in the population. Furthermore, the covariate (pretest in the example) is assumed by three models to be linearly related with the dependent variable, independent of the independent variable, homogeneous in regression slopes and variances, and measured without errors. Finally, for Model (1) alone, it is assumed that the $\beta_{k(j)}$ effect is normally distributed in its underlying population, as stated earlier.

It might be asked why researchers need three models when any of the three can be used to test the null hypothesis. The answer lies in selecting a model that renders the greatest statistical power and the least bias. In terms of statistical power, the hierarchical ANCOVA model in (1) enables a researcher to separate the nuisance variable of classrooms that may affect the participant's performance on the dependent variable, from the sampling error. The inclusion of the nested effect $\beta_{k(j)}$ in Model (1) effectively removes a portion of the sum of squares due to this effect from the error sum of squares (or SS_e). Consequently, the magnitude of SS_e in Model (1) is smaller than that in Model (2). The reduction in SS_e is accompanied by a reduction in degrees of freedom for the error term as well. As it will be shown with real world data in the next section, if the reduction in SS_e is sizeable, it can offset the loss in degrees of freedom. Hence, the MS_e ($=SS_e/df_e$) is made smaller in Model (1) than in Model (2). A smaller MS_e in the denominator of an F-ratio inevitably leads to a greater F statistic and potentially more powerful F test. Compared with Models (1) and (2), Model (3) has the lowest statistical power because it aggregates data over all participants in a classroom. This approach reduces the sample size (in terms of number of classes, rather than number of individuals) and therefore the statistical power.

All three models employ a covariate to statistically adjust differences due to covariates in nonrandomized studies, or to provide a more precise estimation of the treatment effect (i.e., α_j) in randomized studies. Thus, three models are comparable in these regards. In the next section, the application of hierarchical ANCOVA is illustrated in a curriculum study. Results of this application will be contrasted with those obtained from two ordinary ANCOVA's based on individual scores and class means, respectively. The empirical evidence based on real data will support the recommendation for the hierarchical ANCOVA as a conceptually sound and analytically powerful method for interpreting data gathered from intact groups that also include a pretest or baseline measure.

An Illustration

To help illustrate the superiority of hierarchical ANCOVA modeling over two ordinary ANCOVA's, a real world data set with all three methods was analyzed. Results will be shown to be different. They are discussed in terms of interpretability, generalizability, and statistical power.

Data Set and Its Related Study

Data came from a curriculum study by Liou (2002), which was carried out in Taiwan. There were dramatic political changes in Taiwan in recent years. These political changes created a society that is becoming politically more open and democratic than ever before. In order to prepare citizens for future developments of a truly democratic society and the rule of law, the civic curricula in the Taiwanese educational system aim at cultivating in students the knowledge, skills, and dispositions indispensable for such developments and fostering a participatory perspective. However, civic education faces formidable barriers, most notably a gap between pedagogical theory and classroom practice, and a conventional emphasis on the acquisition of factual knowledge regarding the political system instead of actual civic participation. Consequently, the goal of adequately preparing democratic citizens through education is not being fulfilled.

Project Citizen is a civic education program for middle school students. The program actively engages students in learning how to monitor and influence public policy through an interactive and cooperative process. It is typically implemented as a class project. For the project, students work together to identify and study a public policy issue, eventually developing an action plan for implementing their policy solution. According to its developers, the goal of *Project Citizen* is to motivate and empower adolescents to exercise their rights and to accept the responsibilities of democratic citizenship through the intensive study of a local community problem. Specifically, *Project Citizen* is designed to help adolescents:

- learn how to monitor and influence public policy in their communities;
- learn the public policy-making process;
- develop concrete skills and the foundation needed to become responsible participating citizens;
- develop effective and creative communication skills; and
- develop more positive self-concepts and confidence in exercising the

> rights and responsibilities of citizenship. (Center for Civic Education, 2000)

In light of the goals of *Project Citizen* and problems facing Taiwan's civic education, it seems that *Project Citizen* can be used as a curriculum supplement to remedy some of the weaknesses of Taiwan's civic education and to help Taiwan prepare participatory citizens. Consequently, Liou conducted the study to evaluate the effects of *Project Citizen* on the civic skills and dispositions of adolescent students in Taiwan.

Research Design

For administrative reasons, it was deemed impractical to randomly assign students into different pedagogical conditions. Therefore, the study employed a pretest-posttest quasi-experimental design with one treatment and one comparison conditions. Twelve Taiwanese high school teachers, each teaching one experimental and one comparison class, participated in this research. Classes taught by the same teacher were randomly assigned to either the treatment or the comparison condition. In the fall of 2001, students in the experimental classes received instruction in *Project Citizen* as an adjunct to the traditional instruction of *Civics* or *Three Principles of the People*. The comparison students received traditional, discipline-based instruction that focused on the hierarchical model of knowledge acquisition. Liou collected data from 942 students on the pre- and post-treatment assessment of their civic skills and civic dispositions along with their demographic, experiences, teacher-related, and school-related information.

Measurements

To help illustrate the hierarchical ANCOVA approach, students' pre-test and post-test of the civic skills and four dimensions of civic dispositions as a function of their group (treatment versus comparison) information were analyzed; all extracted from Liou's study (2002). Civic skills are those intellectual and participatory capacities that enable active involvement in civic life (Vontz, et al., 2000). Civic dispositions are those traits of public and private character that contribute to both the political efficacy of the individual and the common good of society (Vontz, et al., 2000). Civic dispositions in the Liou study were operationalized by summing the mean scores derived from four subscales of Adolescent Student Civic dispositions Scale (ASCDS): Politic Interest, Propensity to Participate in Future Political Life, Commitment to Rights and Responsibilities of Citizenship, and Sense of Political Efficacy.

Means on the civic skills and dispositions ranged from 1 to 6; the higher the score, the better was the performance. Descriptive information about the pre-test and the post-test of civic skills and civic dispositions is presented in Table 1. The post-test means were adjusted for the pre-test scores using the ANCOVA approach based on individual scores. The group information was coded dichotomously, 1 for the experimental group (participated in *Project Citizen*) and 2 for the comparison group (did not participate in *Project Citizen*). There were equal numbers of students in each group.

Research Hypothesis and Data Analyses

The research hypothesis posted to data was: there was significant difference between experimental and comparison students in their civic skills and four dimensions of civic disposition, namely, political interest, propensity to participate, commitment of rights and responsibilities of citizenship, and sense of political efficacy due to the implementation of *Project Citizen*. To test this research hypothesis, three statistical procedures were applied to the data: ANCOVA based on individual scores, ANCOVA based on class means, and hierarchical ANCOVA based on individual scores. The statistical model underlying ANCOVA based on individual scores was Model (2); Model (3) underlay ANCOVA based on class means, and Model (1) for hierarchical ANCOVA based on individual scores. All three ANCOVA's treated the post-test scores of the five outcome variables as the dependent

Table 1. Descriptive Information about the Sample Data.

Outcome variables	Group	Pretest		Adjusted Posttest
		Mean	SD	Mean
Civic skills	Experiment	3.45	.85	3.62
	Comparison	3.60	.80	3.45
Political interest	Experiment	3.40	.87	3.47
	Comparison	3.55	.86	3.38
Propensity to participate	Experiment	3.61	.78	3.64
	Comparison	3.67	.72	3.56
Commitment of rights and responsibilities of citizenship	Experiment	5.22	.51	5.11
	Comparison	5.19	.53	4.97
Sense of political efficacy	Experiment	4.47	.84	4.49
	Comparison	4.41	.81	4.42

Note. Full sample: N=942. Females: n_f= 475 (50.4%). Males: n_m = 467 (49.6%). Experimental group: n_e = 471 (50%). Comparison group: n_c = 471 (50%).

variables and the pre-test scores as the covariate. The independent variable was the implementation (or lack of) of *Project Citizen* in civic education curriculum. Prior to analyses, statistical assumptions such as normality, equal variance, independence of errors, linearity between pretest (the covariate) and posttest scores, and common slope for all treatment conditions were examined. All assumptions associated with the three procedures were satisfactorily met. Appendix A lists SAS® programming codes for examining these assumptions.

Based on the rationale and previous research, it was hypothesized that *Project Citizen* would have a positive impact on adolescent's civic skills and civic dispositions. Hence, statistical tests pertaining to the research hypothesis were conducted as one-tailed at an alpha level of .025. It was also decided that univariate tests were preferred over multivariate tests of all five dependant variables because the objective of this article was to compare models, instead of accounting for underlying relationships among these dependant variables. The data were analyzed using SAS® version 8.2 (SAS Institute Inc., 1999) and SPSS® version 10 (SPSS Inc., 1999) in the Windows 2000 environment.

ANCOVA Results Based on Individual Scores

Data of the 942 observations were submitted to the GLM procedure in SPSS® version 10 to determine the effect of *Project Citizen* on the civic skills and dispositions of Taiwanese adolescents. Univariate ANCOVA results based on individual scores are shown in Table 2. The five *F*-tests were carried out using MS_{error} as the denominator. An examination of the results indicated that students participating in *Project Citizen* significantly outperformed students in the comparison group on civic skills and three dimensions of civic dispositions including political interest, propensity to participate, and commitment to rights and responsibilities of citizenship. The two groups were comparable on the fourth dimension of civic disposition, namely, sense of political efficacy.

ANCOVA Results Based on Class Means

The second ANCOVA procedure used class means instead of individual scores as the unit of statistical analysis. In order to perform ANCOVA based on class means, data were first aggregated by classes resulting in 24 classroom means (12 treatment class means with 471 students and 12 comparison class means with 471 students). ANCOVA was subsequently

Table 2. ANCOVA Results Of Civic Skills And Four Civic Dispositions Subscales Using Individual Scores As The Unit Of Analysis

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	7.93	1	7.93	19.89	< .001**
Error	374.352	939	.399		
Political interest					
Group	1.62	1	1.62	4.15	.011*
Error	365.45	939	.389		
Propensity to participate					
Group	1.17	1	1.17	4.29	.010*
Error	255.78	939	.272		
Commitment to rights and responsibilities of citizenship					
Group	4.98	1	4.98	17.12	< .001**
Error	273.26	939	.291		
Sense of political efficacy					
Group	1.22	1	1.22	2.44	NS [a]
Error	468.86	939	.499		

* p < .025 (one-tailed), **p < .01 (one-tailed).
[a] Not significant at α = .025.

applied to these 24 class means using the GLM procedure in SPSS® version 10. Results are shown in Table 3. According to Table 3, students participating in *Project Citizen* significantly outperformed students in the comparison group on civic skills. Furthermore, two dimensions of civic dispositions, namely, propensity to participate and commitment to rights and responsibilities of citizenship were also found to be significant with experimental students outperforming comparison students.

Table 3. ANCOVA Results of Civic Skills and Four Civic Dispositions Subscales with Class Means as The Unit Of Analysis.

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	.19	1	.19	10.77	.001**
Error	.37	21	.018		
Political interest					
Group	.037	1	.037	2.66	NS [a]
Error	.288	21	.014		
Propensity to participate					
Group	.039	1	.039	3.21	.022*
Error	.254	21	.012		
Commitment to rights and responsibilities of citizenship					
Group	.111	1	.111	5.40	.008*
Error	.431	21	.021		
Sense of political efficacy					
Group	.020	1	.020	1.07	NS [a]
Error	.393	21	.019		

* $p < .025$ (one-tailed), **$p < .01$ (one-tailed).
[a] Not significant at $\alpha = .025$.

Hierarchical ANCOVA Results

The results of the hierarchical ANCOVA are presented in Table 4 that treated intact classes as nested in the two experimental conditions and students nested in classes. As shown in Table 4, students participating in *Project Citizen* significantly outperformed students in the comparison group in civic skills and also in three dimensions of civic dispositions, namely, political interest, propensity to participate, and commitment to rights and responsibilities of citizenship.

SAS® programming codes for performing the hierarchical ANCOVA is provided in Appendix A for each of the dependent variables. Note that for each dependent variable (such as civic skills); two statistical procedures in SAS® were applied to data: PROC REG and PROC GLM, twice. The purpose of each statistical analysis is explained in the TITLE statement immediately preceding the RUN; statement. For example, the purpose of REG procedure was to test the linearity assumption regarding the linear relationship between the covariate and the dependent variable. The linear relationship was assumed within each condition as well as for the entire data set. The first GLM procedure was to apply the ANCOVA model to the data according to equation (1) presented earlier. The second GLM procedure was to test the equal slope assumption assumed by the ANCOVA model. This assumption was tested via the interaction between the covariate (i.e., pretest) and the independent variable (participating in *Project Citizen* or not). Non-significant *F* test results were obtained for all five dependent variables indicating that the equal slope assumption was met.

Table 4. Hierarchical ANCOVA Results for Civic Skills And Four Civic Dispositions Subscales Using Individual Scores as The Unit of Analysis

Source	SS	*df*	MS	*F*	*p*
Civic skills					
Group	7.37	1	7.37	10.89	< .001**
Class (Group)	14.90	22	.677	1.73	.0201
Error	359.46	417	.391		
Political interest					
Group	1.803	1	1.803	3.53	.019*
Class (Group)	11.233	22	.511	1.32	.1466
Error	354.219	917	.386		
Propensity to participate					
Group	1.280	1	1.280	2.81	.024*
Class (Group)	10.031	22	.156	1.70	.0232
Error					
Commitment to rights and responsibilities of citizenship					
Group	4.8855	1	4.885	6.03	.006*
Class (Group)	17.815	22	.810	2.91	< .001**
Error	255.441	917	.279		
Sense of political efficacy					
Group	1.062	1	1.062	1.43	NS [a]
Class (Group)	16.315	22	.742	1.50	.0643
Error	452.549	917	.494		

* $p < .025$ (one-tailed), **$p < .01$ (one-tailed).

[a] Not significant at $\alpha = .025$.

Comparison of Three Results

Results obtained from three statistical approaches regarding the research question are contrasted in Table 5. For civic skills, propensity to participate, commitment to rights and responsibilities of citizenship, and sense of political efficacy, there was agreement among the three approaches. For the political interest of Taiwanese adolescent students, ANCOVA based on class means yielded a non-significant result; this contrasted with a significant finding ($p < .025$) obtained from the hierarchical ANCOVA and ANCOVA based on individual scores. As stated earlier, ANCOVA based on class means aggregated scores into class means leading to great loss in units of analysis and therefore, statistical power, compared to the other two approaches. Further, findings from the means approach limit the interpretation and generalizability to class averages only—a result not useful or relevant to most educators or parents.

The hierarchical ANCOVA approach yielded results comparable to those obtained from ANCOVA based on individual scores. Yet, the hierarchical approach uncovered additional class differences that could not be found by ANCOVA based on individual scores due to its model configuration. As shown in Table 4 in gray areas, the 12 classes nested in each treatment condition exhibited statistically significant differences ($p < .05$, two tailed) on civic skills, propensity to participate, and commitment to rights and responsibilities of citizenship. On sense of political efficacy, class differences were significant at the $p < .10$ (two-tailed) level but not at .05.

Table 5. Comparison Of Three ANCOVA Results For Civic Skills And Four Civic Dispositions Subscales

Source	Hierarchical ANCOVA	ANCOVA (Individual Scores)	ANCOVA (Class Means)
	p	*p*	*p*
Civic skills	< .001**	< .001**	<.001**
Political interest	.019*	.011*	NS [a]
Propensity to participate	.024*	.010*	.022*
Commitment to rights and responsibilities of citizenship	.006*	< .001**	.008*
Sense of political efficacy	NS [a]	NS [a]	NS [a]

* *p* < .025 (one-tailed), ***p* < .01 (one-tailed).

[a] Not significant at α = .025.

These differences merited further investigation as to why and how these differences existed, as well as to what extent these differences were due to teacher-related, school-related, or student-related characteristics. Research into these class differences can be a worthy endeavor; findings may suggest curricula or cultural changes to schools or classes in order to bring about equality.

Implications for Educational Researchers

In this article, the application of hierarchical ANCOVA for analyzing quasi-experimental data including baseline measures is demonstrated. This procedure is illustrated with a real-world data set to investigate the effect of *Project Citizen* on Taiwan adolescent students' civic skills and four dimensions of civic dispositions, namely, political interest, propensity to participate, commitment of rights and responsibilities of citizenship, and sense of political efficacy. Results obtained from the hierarchical ANCOVA and ANCOVA based on individual scores were comparable. Both statistical approaches were shown to be more powerful than ANCOVA based on class means. Additional statistically significant differences among classes assigned to either the treatment or the comparison condition were uncovered by the hierarchical ANCOVA, but not by ANCOVA based on individual scores. On the basis of statistical power, interpretability, and generalizability, it was concluded that the hierarchical ANCOVA was superior to ANCOVA based on individual scores or class means. The latter two approaches suffered from conceptual and methodological limitations.

In accounting for effects associated with *Project Citizen*, the hierarchical ANCOVA approach incorporated the hierarchical (or nested) nature of Liou's (2002) quasi-experimental design into the analysis of covariance model. Consequently, data analysis was congruent with the way the study was actually carried out. It retained the maximum number of degrees of freedom for testing pertinent population parameters. It employed the pretest score as a covariate in order to control for pre-existing differences in students that were unrelated to *Project Citizen*. The hierarchical ANCOVA was shown in this article to be well suited to educational research in which data are collected from intact settings (such as classrooms) in quasi-experimental designs that also include one or more baseline measures.

To ensure credibility and to minimize, if not eliminate, potential bias in the findings reported in quasi-experimental research, it is necessary that educational researchers keep the following recommendations in mind.

First and the foremost, efforts should be exerted to randomly assign subjects to treatments. By so doing, educational researchers exclude the confounding issue of unit of analysis from their research and therefore, reduce bias and distortion in estimating population parameters or testing pertinent hypotheses. Researchers are advised to achieve random assignment whenever possible.

Second, data collected in intact groups deserve a rigorous examination. In educational research, it is possible to randomly assign subjects to treatment conditions and to establish circumstances in which the outcome measures are isolated from systematic carryover effects or threats to the independence assumption. Yet, it is often impossible or even undesirable to administer treatments individually in isolation. To account for the hierarchical nature of research designs and to maintain the interpretation of results at the individual level, an appropriate statistical model such as hierarchical ANCOVA should be employed.

Lastly, it should be noted that, even though the hierarchical ANCOVA has been proven to be a conceptually and methodologically sound procedure, this approach should be regarded as a viable approach that exercises only statistical control of biases. Moreover, the hierarchical ANCOVA is computationally more complex than an ordinary ANCOVA; it requires a set of restrictive statistical assumptions (Kirk, 1995). These assumptions must be met before valid inferences can be drawn from data analysis.

References

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6* (3), 267-285.

Blair, R. C., & Higgins, J. J. (1986). Comment on "Statistical Power with Group Mean as the Unit of Analysis”. *Journal of Educational Statistics, 11*(2), 161-169.

Buser, K. (1995, April). *Dangers in using ANCOVA to evaluate special education program effects*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 384 654).

Center for Civic Education. (2000). *We the People…Project Citizen: A professional development manual*. Calabasas, CA: Center for Civic Education.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). New York: McGraw Hill.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Henson, R. K. (1998, November). *ANCOVA with intact groups: Don’t do it*! Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 426 086).

Hines, J. L., & Foil, C. R. (2000, January). *Covariance corrections: What they are and what they are not.* Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service No. ED 445 076)

Hopkins, K. D. (1982). The units of analysis: Group means versus individual observations. *American Educational Research Journal*, *19* (1), 5-18.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Lindman, H. R. (1992). *Analysis of variance in experimental design*. New York: Springer-Verlag.

Lindquist, E. F. (1940). *Statistical analysis in educational research*. New York: Houghton Mifflin.

Liou, S.-M. (2002). *The effect of We the People…Project Citizen on the civic skills and dispositions of Taiwanese adolescent students*. Unpublished doctoral dissertation, Indiana University-Bloomington.

Loftin, L., & Madison, S. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological development* (Vol. 1, pp. 133-147.). Greenwich, CT: JAI Press Inc.

McMillan, J. H., & Schumacher, S. (2001). Research in education: A conceptual introduction (5th ed.). New York: Longman.

Morran, D. K., Robinson, F. F., & Hulse-Killacky, D. (1990). Group research and the unit of analysis problem: The use of ANOVA designs with nested factors. *The Journal for Specialists in Group Work, 15*(1), 10-14.

Peckham, P. D., Glass, G. V., & Hopkins, K. D. (1969). The experimental unit in statistical analysis. *The Journal of Special Education, 3* (4), 337-349.

Peng, C.-Y. J. (2004). Nested Design. In *The SAGE Encyclopedia of Social Science Research Methods, 2,* 717-719. Thousand Oaks, CA: Sage Publications.

Roberts, J. K. (2000). *Nested ANOVA vs. crossed ANCOVA: When and how to use which*. (ERIC Document Reproduction Service No. ED 440 112).

SAS Institute Inc. (1999). *SAS/STAT® user's guide, version 8, volume 2.* Cary, NC: SAS Institute Inc.

SPSS Inc. (1999). *SPSS base 10.0 user's guide*. Chicago: SPSS Inc.

Appendix A SAS® Programming Codes

```
*--------------------------------------------------Test of Civic Skills---------------------------------------------------------;
PROC    REG;
        MODEL q2_ski=q1_ski;
        PLOT q2_ski*q1_ski;
        BY q1_group;
TITLE   'TEST OF LINEARITY ASSUMPTION: Civic Skills';
RUN;

PROC    GLM;
        CLASS q1_group class;
        MODEL q2_ski=q1_ski q1_group class(q1_group)/SOLUTION;
        TEST H=q1_group E=class(q1_group);
        Means q1_group;
        LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE   'Hierarchical ANCOVA for Civic Skills';
RUN;

PROC    GLM;
        CLASS q1_group;
        MODEL q2_ski=q1_ski q1_group q1_ski*q1_group;
TITLE   'TEST OF EQUAL SLOPE ASSUMPTION: Civic Skills';
RUN;

*--------------------------------------------------Test of Political Interest-----------------------------------------------------;
PROC    REG;
        MODEL q2_int=q1_int;
        PLOT q2_int*q1_int;
        BY q1_group;
TITLE   'TEST OF LINEARITY ASSUMPTION: Political Interest';
RUN;

PROC    GLM;
        CLASS q1_group class;
        MODEL q2_int=q1_int q1_group class(q1_group)/SOLUTION;
        TEST H=q1_group E=class(q1_group);
        Means q1_group;
        LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE   'Hierarchical ANCOVA for Political Interest';
RUN;
PROC    GLM;
        CLASS q1_group;
        MODEL q2_int=q1_int q1_group q1_int*q1_group;
TITLE   'TEST OF EQUAL SLOPE ASSUMPTION: Political Interest';
RUN;
```

```
*-------------------------------------------Test of Propensity to Participate-----------------------------------;
PROC   REG;
       MODEL q2_par=q1_par;
       PLOT q2_par*q1_par;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Propensity to Participate';
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_par=q1_par q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Propensity to Participate';
RUN;

PROC   GLM;
       CLASS q1_group;
       MODEL q2_par=q1_par q1_group q1_par*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Propensity to Participate';
RUN;

*---------------------------Test of Commitment to Rights and Responsibilities--------------------------------;
PROC   REG;
       MODEL q2_right=q1_right;
       PLOT q2_right*q1_right;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Commitment to Rights and Responsibilities' ;
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_right=q1_right q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Commitment to Rights and Responsibilities';
RUN;

PROC   GLM;
       CLASS q1_group;
       MODEL q2_right=q1_right q1_group q1_right*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Commitment to Rights and Responsibilities';
RUN;
```

```
*--------------------------------------------------Test of Political Efficacy---------------------------------------------------
-;
PROC   REG;
       MODEL q2_effic=q1_effic;
       PLOT q2_effic*q1_effic;
       BY q1_group;
TITLE  'TEST OF LINEARITY ASSUMPTION: Political Efficacy';
RUN;

PROC   GLM;
       CLASS q1_group class;
       MODEL q2_effic=q1_effic q1_group class(q1_group)/SOLUTION;
       TEST H=q1_group E=class(q1_group);
       Means q1_group;
       LSMEANS q1_group/E=class(q1_group) ADJUST=BON E STDERR PDIFF;
TITLE  'Hierarchical ANCOVA for Political Efficacy';
RUN;

PROC   GLM;
       CLASS q1_group;
       MODEL q2_effic=q1_effic q1_group q1_effic*q1_group;
TITLE  'TEST OF EQUAL SLOPE ASSUMPTION: Political Efficacy';
RUN;
```

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 248-257

A Combined Individuals and Moving Range Control Chart

Michael B. C. Khoo S. H. Quah C. K. Ch'ng
School of Mathematical Sciences
Universiti Sains Malaysia

An individuals control chart is usually used to monitor shifts in the process mean when it is not possible to form subgroups. The moving range of two successive process measures is used as the basis for estimating the process variability. Similar to the case of the $\overline{X} - R$ and $\overline{X} - S$ charts, the individuals-moving range (I-MR) charts are used simultaneously in the monitoring of the process mean and variance respectively for individual observations, requiring maintaining two different charts. In this article, a new approach is suggested where the measurements of both the process mean and variance are plotted on one chart. It is referred to as the combined I-MR chart. An average run length (ARL) study is conducted to evaluate its performance with respect to shifts in the process mean and variance. Examples are provided.

Key words: Individuals charts; moving range charts; average run length (ARL); process mean; process variance

Introduction

There are many situations in which the sample size used for process monitoring is one (Montgomery, 2001). Some of these are in situations involving the use of automated inspection and measurement technology where every unit manufactured is analyzed. Situations where the production rate is slow and monitoring of the process is required before the time needed to form subgroups may also call for process monitoring involving individual observations. The monitoring of individual observations is also important in situations where repeat measurements on a process differ only because of laboratory or analysis error, as in many chemical processes.

Traditionally, individuals control charts are used in the monitoring of processes involving individual observations. For such cases, the moving range charts are employed in the monitoring of the process variability. Here, the moving range of two successive observations is defined as (Montgomery, 2001):

$$MR_i = \left| X_i - X_{i-1} \right|, \quad i = 2, 3, \ldots \qquad (1)$$

A moving range chart is established by plotting the moving ranges computed from eq. (1) based on the limits

$$UCL = D_4 \overline{MR} \qquad (2a)$$

$$CL = \overline{MR} \qquad (2b)$$

$$LCL = D_3 \overline{MR} \qquad (2c)$$

Michael B. C. Khoo is a Lecturer, with research interests in statistical process control and reliability analysis. He is a member of the Editorial Boards of *Quality Engineering* and *Journal of Modern Applied Statistical Methods*. Email him at mkbc@usm.my. S.H. Quah is a retired Professor of Statistics His research interests include quality management, statistical process control, and statistical design of experiments. Email him at shquah@cs.usm.my. C.K. Ch'ng is a doctoral student in Industrial Statistics. Email him at chngchuankim@yahoo.com.

where $\overline{MR}$ is the average of the moving range computed from a preliminary set of data. After establishing an in-control state for the process variability, the individuals chart is set up by plotting the individual observations, X_i, on a chart with limits (Montgomery, 2001):

$$UCL = \overline{X} + 3\frac{\overline{MR}}{d_2} \tag{3a}$$

$$CL = \overline{X} \tag{3b}$$

$$LCL = \overline{X} - 3\frac{\overline{MR}}{d_2} \tag{3c}$$

Note that in eqs. (2a), (2c), (3a) and (3c), D_3, D_4 and d_2 are control chart constants for $n = 2$ whose values are given in most quality control textbooks.

A Combined I-MR Chart

Let X_i, i = 1, 2, …, represent individual observations from a process for a quality characteristic of interest. It is assumed that $X_i \sim N(\mu + a\sigma, b^2\sigma^2)$, where a = 0 and b = 1 indicate that the process is in-control; otherwise, the process is out-of-control. Here, μ and σ denote the on-target mean and standard deviation. Define

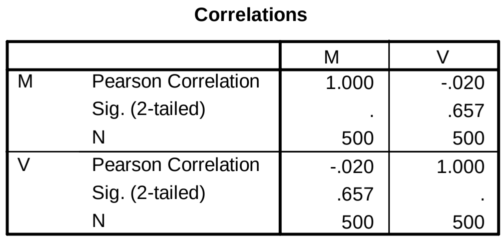
$$M_i = \frac{X_i - \mu}{\sigma} \sim N(0,1), \quad i = 1, 2, \ldots \tag{4}$$

and

$$V_i = \Phi^{-1}\left\{H_1\left[\frac{1}{2\sigma^2}(X_i - X_{i-1})^2\right]\right\} \sim N(0,1), \quad i = 2, 3, \ldots \tag{5}$$

where $\Phi^{-1}(\cdot)$ and $H_1(\cdot)$ are the inverse of the standard normal distribution function and the chi-square distribution function with one degree of freedom respectively. Because the value of X_i is unavailable when i = 0, V_1 is computed using $\Phi^{-1}\left\{H_1\left[\frac{1}{2\sigma^2}(X_1 - \mu)^2\right]\right\}$. It is found that V_i follows a standard normal distribution (Appendix).

Due to the transformation of the V_i statistic in (5), $\text{cov}(M_i, V_i)$ is intractable. Thus, in finding the correlation of M_i and V_i to determine the extent of the relationship between the two statistics, 500 individual observations from a N(0,1) distribution are generated, the M_i and V_i statistics computed and the sample correlation coefficient of M_i and V_i is calculated using the Pearson correlation procedure from SPSS version 11. The output is shown in Figure 1. Note that the individual observations can also be generated from other normal distributions. From Figure 1, the correlation of M_i and V_i is insignificant at the 1% significance level because its associated p-value is 0.657. Here, the sample correlation coefficient is –0.02. Based on this result, it can be concluded that the correlation of M_i and V_i is negligible if the underlying distribution of the individual observations is normal.



Correlations

		M	V
M	Pearson Correlation	1.000	-.020
	Sig. (2-tailed)	.	.657
	N	500	500
V	Pearson Correlation	-.020	1.000
	Sig. (2-tailed)	.657	.
	N	500	500

Figure 1. The Sample Correlation Coefficient of the M_i and V_i Statistics based on 500 Individual Observations (Output from SPSS)

M_i monitors the process mean while V_i the process variability. These two statistics are combined to form a new statistic given by

$$C_i = \max(|M_i|, |V_i|) \tag{6}$$

The statistic C_i will be large when the process mean has shifted away from its target value and/or when the process variance has increased or decreased.

Because the correlation of M_i and V_i is negligible, it is shown (Appendix) that the approximate density function of C_i for the in-control case is

$$f(c) = 4\phi(c)\{2\Phi(c) - 1\}, \quad c \geq 0 \tag{7}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and distribution functions of a standard normal random variable respectively. The combined I-MR chart only requires an upper control limit (*UCL*) because C_i is nonnegative. Suppose that the desired Type-I error set by management is α, then the *UCL* can be obtained from the following integral:

$$\int_{UCL}^{\infty} f(c)dc = \alpha \tag{8}$$

Steps for Implementing the Combined I-MR Chart

The following steps serve as guidelines in setting up a combined I-MR chart:

(i) If the process parameter(s) are unknown, then they are estimated as follow: The process mean, μ, is estimated from the formula, $\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m}$, where *m* is the number of observations in the stable preliminary data set used in the estimation. The process standard deviation, σ, is estimated using $\frac{\overline{MR}}{d_2}$, where

$$\overline{MR} = \frac{MR_2 + MR_3 + ... + MR_m}{m-1}$$

is the average of the moving ranges. Here, d_2 is the value of the control chart constant for sample size, *n* = 2.

(ii) Compute M_i, V_i and C_i for each observation.

(iii) Determine the *UCL* using eq. (8) based on a desired Type-I error.

(iv) When $C_i \leq UCL$, plot a dot at time *i*. When $C_i > UCL$, check both $|M_i|$ and $|V_i|$ against *UCL*. If only $|M_i|$ is greater than *UCL*, plot "*m*+" at time *i* when $M_i > 0$ to indicate the process mean has increased, and plot "*m*–" at time *i* when $M_i < 0$ to indicate the process mean has decreased.

Similarly, if $|V_i|$ alone is greater than *UCL*, plot "*v*+" at time *i* when $V_i > 0$ to indicate the process variability has increased, and plot "*v*–" at time *i* when $V_i < 0$ to indicate the process variability has decreased. For the case when both $|M_i|$ and $|V_i|$ are greater than the *UCL*, plot "++", "+–", "–+" or "––" if $M_i > 0$ and $V_i > 0$, $M_i > 0$ and $V_i < 0$, $M_i < 0$ and $V_i > 0$, or $M_i < 0$ and $V_i < 0$ respectively.

(v) Investigate the cause(s) for each out-of-control point so that appropriate corrective actions can be taken.

Plots for Determining the *UCL*

Figure 2 gives a plot for approximating the *UCL* based on a desired Type-I error. The plot is based on in-control ARLs (ARL_0s) between 100 and 1000. It is constructed from points (*UCL* , ARL_0) obtained using a simple Mathematica 4.0 program shown in Figure 3 based on eq. (8).

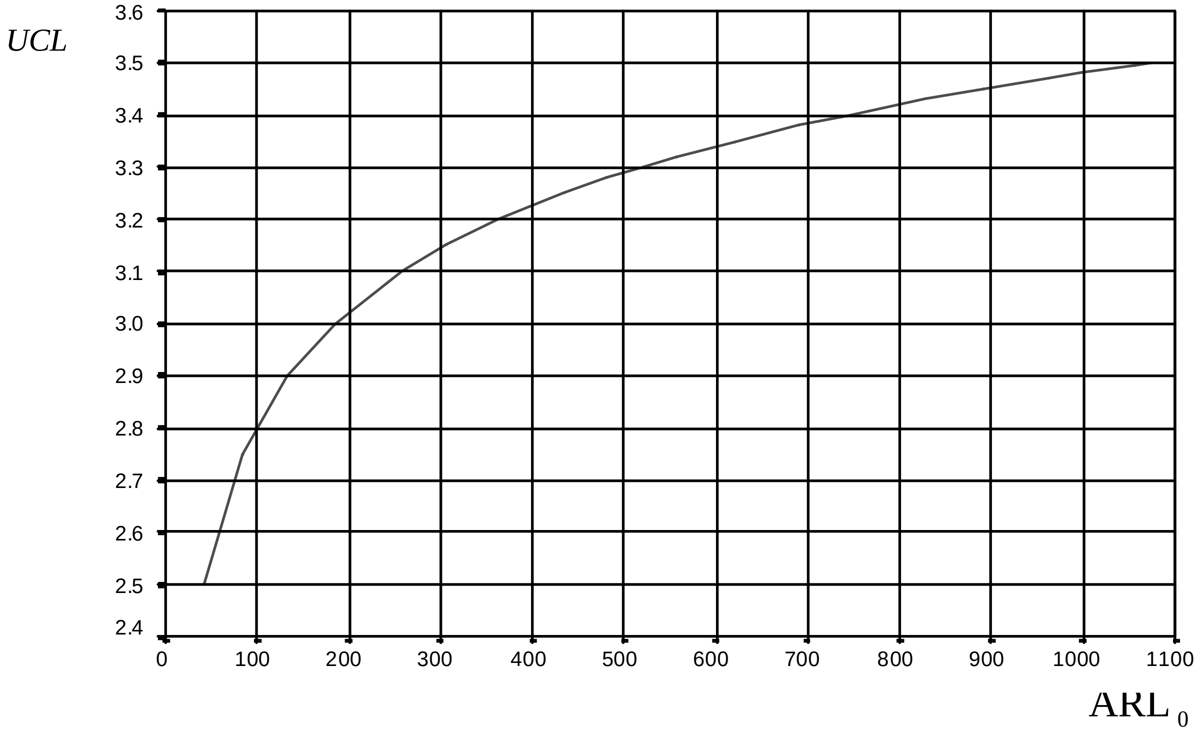


Figure 2. A Plot of *UCL* vs. ARL_0 for the Combined I-MR Chart

$$\mathbf{UCL} =$$

$$\mathbf{NIntegrate}\left[2\sqrt{\frac{2}{\pi}}\times\left(e^{-\frac{c^2}{2}}\right)\times\left(\sqrt{\frac{2}{\pi}}\ \mathbf{Integrate}\left[e^{-\frac{t^2}{2}},\{t,-\infty,c\}-1\right]\right),\{c,\mathbf{UCL},\infty\}\right]$$

Figure 3. A Mathematica 4.0 Program to Compute the *UCL* for a Combined I-MR Chart

A sensitivity analysis can be performed using the Mathematica 4.0 program in Figure 3 to obtain the exact *UCL* for a desired Type-I error. The following example shows how a sensitivity analysis is performed, assuming that the Type-I error is set at $\alpha = 0.004$. The corresponding in-control ARL is $\text{ARL}_0 = 250$. The value of *UCL* approximated from the plot in Figure 2 is 3.08. Values of α which correspond to values of *UCL*s close to the one approximated, i.e., 3.08 are computed using the program in Figure 3 and are tabulated in Table 1. From Table 1, it is noticed that the value of *UCL* which produces the closest Type-I error to $\alpha = \dfrac{1}{250} = 0.004$ is 3.09.

A Study on the Performance of the Combined I-MR Chart

A simulation study is conducted using SAS version 8 to compute the ARL values of the combined I-MR chart based on $\text{ARL}_0 = 250$ and 500. Each ARL reading is based on 5000 simulation trials. The *UCL*s are determined, using the approach discussed in the previous section, to be 3.09 and 3.29 for $\text{ARL}_0 = 250$ and 500 respectively. Shifts in both the process mean and variance are considered. The process mean shifts from μ to $\mu + a\sigma$ while the process variance from σ to $b\sigma$, where $a = 0$ and $b = 1$ represent the in-control case. The values of $a \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 3\}$ and $b \in \{1, 1.05, 1.1, 1.2, 1.25, 1.5, 2, 2.5, 3, 4, 5\}$ are considered. The simulation results for $\text{ARL}_0 =$ 250 and 500 are given in Tables 2 and 3 respectively.

Both Tables 2 and 3 show that as the magnitude of the shift (either in the process mean or variance or both) increases, the value of ARL decreases. For example, consider the case of $b = 1$ and $a \in \{0, 0.25, 0.5, \ldots, 3\}$ in Table 2, where only the process mean shifts. The ARL values for this case are 275.71, 234.75, 153.14, …, 2.13 for a = 0, 0.25, 0.5, …, 3 respectively, where the values show a declining trend as the magnitude of the shift in the mean increases. A similar trend is observed when only the process variance shifts. For example, from Table 2, when $a = 0$ and $b \in \{1, 1.05, 1.1, \ldots, 5\}$, the ARL values of 275.71, 192.88, 133.86, …, 1.84 show a declining trend as *b* increases, i.e., as the magnitude of the shift in the variance increases. Note that the ARL values will also show a decreasing trend if the magnitude of shifts in both the mean and variance increase simultaneously. It is shown in Tables 2 and 3 that the computed ARL_0 values are 275.71 and 546.38 respectively, where they differ only slightly from the desired values of 250 and 500. This shows that the *UCL* computed from the approximate density function, *f* (*c*) in eq. (7) is reliable, which indicates that the correlation between M_i and V_i is negligible. The difference in the estimated versus intended Type-I errors is very little, i.e., $\dfrac{1}{275.71} = 0.00363$ vs. $\dfrac{1}{250} = 0.004$ and $\dfrac{1}{546.38} = 0.00183$ vs. $\dfrac{1}{500} = 0.002$.

Table 1. Values of the Type-I Error (α) Computed from Corresponding *UCL*s

UCL	α
3.07	0.00427659
3.08	0.00413573
3.09	0.00399912

Table 2. ARL Profiles of the Combined I-MR Chart for $\text{ARL}_0 = 250$ with *UCL* = 3.09

	a								
b	0	0.25	0.5	0.75	1	1.25	1.5	2	3
1	275.71	234.75	153.14	87.97	50.43	29.15	17.39	7.16	2.13
1.05	192.88	163.40	108.28	67.63	40.59	24.28	15.00	6.59	2.12
1.1	133.86	116.37	81.44	52.36	32.71	20.65	13.13	6.13	2.12
1.2	71.61	64.18	49.05	34.37	23.12	15.33	10.62	5.44	2.11
1.25	55.41	50.93	39.89	28.31	19.80	13.52	9.82	5.21	2.10
1.5	20.70	19.46	16.98	13.70	10.90	8.51	6.59	4.23	2.08
2	7.21	7.04	6.72	6.13	5.51	4.84	4.26	3.27	2.04
2.5	4.29	4.28	4.12	3.94	3.72	3.47	3.21	2.75	1.99
3	3.10	3.10	3.07	3.01	2.93	2.80	2.65	2.39	1.92
4	2.22	2.21	2.19	2.17	2.13	2.10	2.06	1.97	1.77
5	1.84	1.84	1.83	1.82	1.81	1.79	1.78	1.74	1.64

Table 3. ARL Profiles of the Combined I-MR Chart for $\text{ARL}_0 = 500$ with *UCL* = 3.29

	a								
b	0	0.25	0.5	0.75	1	1.25	1.5	2	3
1	546.38	457.34	283.45	156.09	83.45	46.63	26.52	10.25	2.57
1.05	358.51	302.16	196.48	110.97	64.00	37.53	22.57	9.08	2.54
1.1	240.74	205.22	137.70	83.06	50.35	30.53	19.04	8.27	2.51
1.2	115.87	101.70	75.92	50.99	33.47	23.77	14.34	7.01	2.46
1.25	85.47	76.72	59.06	41.48	27.85	18.86	12.85	6.52	2.44
1.5	27.48	26.38	22.79	18.12	14.00	10.76	8.26	5.10	2.34
2	8.67	8.42	7.93	7.26	6.40	5.64	4.88	3.69	2.23
2.5	4.89	4.81	4.65	4.44	4.18	3.87	3.60	3.01	2.13
3	3.41	3.41	3.37	3.28	3.16	3.04	2.90	2.58	2.03
4	2.33	2.33	2.33	2.30	2.27	2.22	2.18	2.08	1.85
5	1.94	1.94	1.92	1.90	1.89	1.88	1.85	1.80	1.71

The combined I-MR chart has an advantage over the traditional individuals and moving range charts because the former allows practitioners to set the joint Type-I error of the two charts. Conversely, the Type-I error of the traditional moving range chart cannot be set by practitioners because it is based on fixed limits given in eqs. (2a) – (2c). Another advantage of using the combined chart is practitioners do not have to plot two charts separately, i.e, one each for individual measurements and moving ranges. Due to the advent of modern computers, the computation of the combined I-MR chart's statistics in eqs. (4), (5) and (6) is only a trivial problem.

Applications

Two examples will be given to illustrate how the combined I-MR chart is used in real situations. They are based on observations generated from SAS version 8. The in-control observations are assumed to follow a standard normal distribution. Out-of-control observations are generated from a $N(\mu + a\sigma, b^2\sigma^2)$ distribution with $a > 0$, $b > 1$, $\mu = 0$ and $\sigma = 1$. The first example deals with a shift in the process mean while the second a shift in the process variance. The Type-I error for the two examples is set as $\alpha = 0.004$ which corresponds to $\text{ARL}_0 = 250$. Thus, the *UCL* is determined to be 3.09.

Example 1

The first 5 observations are generated from a $N(0,1)$ distribution to represent the in-control situation. Observations 6 to 20 which represent the out-of-control situation involving a shift in the mean are generated from a $N(3,1)$ distribution. Here, the magnitude of the shift in the mean in multiples of standard deviation is $a = 3$. The individual observations generated, X_i, $i = 1, 2, \ldots, 20$ together with the computed M_i, V_i and C_i statistics are shown in Table 4. The combined I-MR chart is plotted in Figure 4.

The chart shows that out-of-control signals due to a shift in the mean are detected at observations 7, 9, 12, 13, 15, 19 and 20. Following the first out-of-control signal at observation 7, an investigation needs to be made to search for the assignable cause(s) so that the process can return to an in-control state again.

Example 2

This example involves a shift in the variance. The first 5 observations which represent the in-control case are generated from a standard normal distribution. This is followed by generating the next 15 observations from a $N(0,4)$ distribution, where the magnitude of the shift in the standard deviation is $b = 2$. Table 5 summarizes the values of the individual observations, X_i, and their corresponding computed M_i, V_i and C_i statistics. The combined I-MR chart is given in Figure 5.

Conclusion

The proposed combined I-MR chart extends the work of Chen, Cheng and Xie (2001) where they suggested a joint monitoring of the process mean and variance of subgrouped data with one EWMA chart. The combined I-MR chart enables a simultaneous monitoring of the process mean and variance involving individual measurements. It combines the usual individuals chart and the moving range chart into a single chart. The advantages of the combined chart discussed in an earlier section serve as basis for practitioners to use the chart in place of its traditional counterparts.

Table 4. An Example of Application for a Shift in the Process Mean based on a = 3 and UCL = 3.09

Obs. No., i	X_i	M_i	V_i	C_i
1	0.7508	0.7508	–0.2416	0.7508
2	0.7835	0.7835	–2.0870	2.0869
3	0.6009	0.6009	–1.2660	1.2660
4	0.1087	0.1087	–0.6063	0.6063
5	–0.1614	–0.1614	–1.0300	1.0300
6	2.4860	2.4860	1.5447	2.4860
7	4.2386	4.2386*	0.7884	4.2386*
8	2.9663	2.9664	0.3363	2.9664
9	3.2089	3.2089*	–1.0978	3.2089*
10	1.1256	1.1256	1.0771	1.1256
11	2.9149	2.9149	0.8211	2.9149
12	3.4370	3.4370*	–0.5592	3.4370*
13	3.2020	3.2020*	–1.1171	3.2020*
14	2.9880	2.9880	–1.1737	2.9880
15	4.3715	4.3715*	0.4456	4.3715*
16	3.0377	3.0377	0.3972	3.0377
17	2.6764	2.6764	–0.8357	2.6764
18	2.1498	2.1498	–0.5523	2.1498
19	4.6574	4.6574*	1.4311	4.6574*
20	3.2859	3.2859*	0.4340	3.2859*

Note: * indicates the out-of-control points

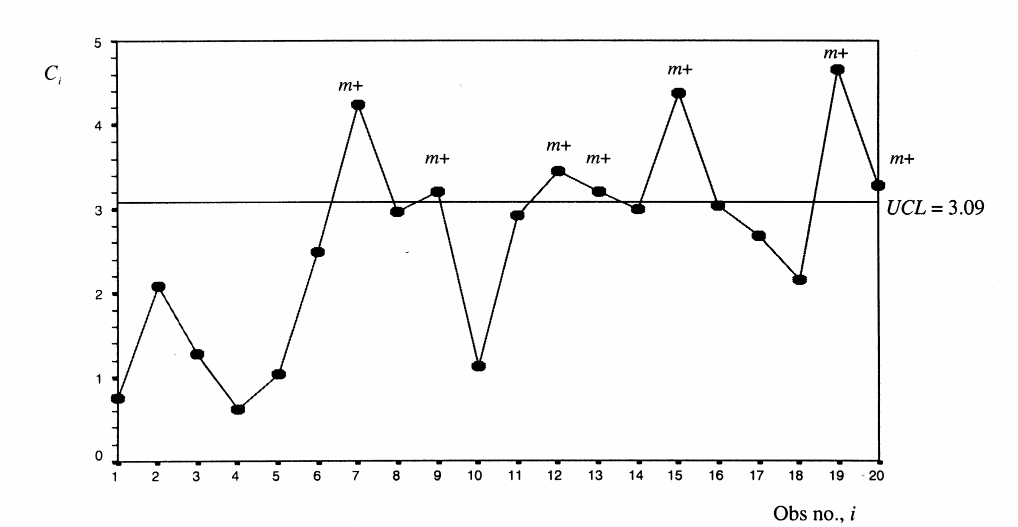


Figure 4. A Combined I-MR Chart for a Shift in the Process Mean

Table 5. An Example of Application for a Shift in the Process Variance based on $b = 2$ and $UCL = 3.09$

Obs. No., i	X_i	M_i	V_i	C_i
1	−0.3487	−0.3487	−0.8605	0.8605
2	−1.2907	−1.2907	−0.0134	1.2907
3	1.0317	1.0317	1.2784	1.2784
4	0.0442	0.0442	0.0376	0.0442
5	−0.1895	−0.1895	−1.1207	1.1207
6	−2.0778	−2.0778	0.9086	2.0778
7	−0.1000	−0.1000	0.9864	0.9864
8	0.4558	0.4558	−0.5081	0.5081
9	−0.3241	−0.3241	−0.2053	0.3241
10	3.0338	3.0338	2.1065	3.0338
11	0.4064	0.4064	1.5286	1.5286
12	1.8603	1.8603	0.5132	1.8603
13	2.3679	2.3679	−0.5818	2.3679
14	−2.7172	−2.7172	3.4111*	3.4111*
15	1.8373	1.8373	3.0162	3.0162
16	−1.4168	−1.4168	2.0258	2.0258
17	−0.7237	−0.7237	−0.3162	0.7237
18	0.9509	0.9509	0.7180	0.9509
19	−0.5085	−0.5085	0.5184	0.5184
20	−1.6768	−1.6768	0.2308	1.6768

Note: * indicates the out-of-control point

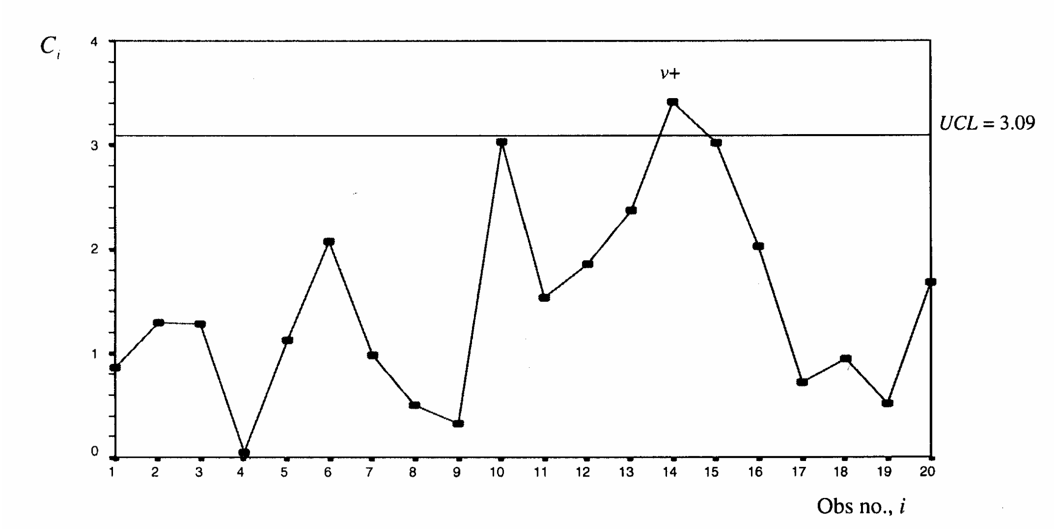


Figure 5. A Combined I-MR Chart for a Shift in the Process Variance

References

Montgomery, D. C. (2001). *Introduction to statistical quality control* (4th. ed.). New York: N.Y. John Wiley & Sons.

Chen, G., Cheng, S. W., & Xie, H. (2001). Monitoring process mean and variability with one EWMA chart. *Journal of Quality Technology, 33*, 223 – 233.

Hogg, R. V. & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: N.Y. Macmillan.

Appendix

If $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots$, then

$$M_i = \frac{X_i - \mu}{\sigma} \sim N(0,1), \quad i = 1, 2, \ldots . \qquad \text{(A1)}$$

Because

$$X_i - X_{i-1} \sim N(0, 2\sigma^2), \quad i = 2, 3, \ldots,$$

it follows that

$$\frac{1}{\sigma\sqrt{2}}(X_i - X_{i-1}) \sim N(0,1), \quad i = 2, 3, \ldots .$$

Since the square of a standard normal statistic is a chi-square statistic with one degree of freedom (Hogg & Craig, 1978), it follows that

$$\frac{1}{2\sigma^2}(X_i - X_{i-1})^2 \sim \chi^2(1), \quad i = 2, 3, \ldots .$$

Then,

$$H_1\left[\frac{1}{2\sigma^2}(X_i - X_{i-1})^2\right] \sim U(0,1), \quad i = 2, 3, \ldots$$

where $H_1(\cdot)$ is the chi-square distribution function with one degree of freedom. Let $\Phi^{-1}(\cdot)$ be the inverse of the standard normal distribution function so that

$$V_i = \Phi^{-1}\left\{H_1\left[\frac{1}{2\sigma^2}(X_i - X_{i-1})^2\right]\right\} \sim N(0,1), \quad i = 2, 3, \ldots \qquad \text{(A2)}$$

Note that V_1 is computed using

$$\Phi^{-1}\left\{H_1\left[\frac{1}{2\sigma^2}(X_1 - \mu)^2\right]\right\}$$

because X_0 is unavailable at time $i = 1$. Define

$$C_i = \max(|M_i|, |V_i|), \quad i = 1, 2, \ldots \qquad \text{(A3)}$$

so that

$$\begin{aligned} F(c) &= P(C_i \le c) \\ &= P(|M_i| \le c, |V_i| \le c) \end{aligned}$$

where $F(\cdot)$ is the distribution function of C_i. Since the correlation of M_i and V_i is negligible, $P(|M_i| \le c, |V_i| \le c)$ can be approximated by

$$\begin{aligned} P(|M_i| \le c, |V_i| \le c) &\approx P(-c \le M_i \le c) \times P(-c \le V_i \le c) \\ &= [P(-c \le Z_i \le c)]^2 \\ &= [2\Phi(c) - 1]^2 \end{aligned}$$

Thus, the approximate density function of C_i is

$$f(c) = 4\phi(c)\{2\Phi(c) - 1\}, \quad c \ge 0 .$$

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 258-265

1538 – 9472/06/$95.00

A Combined Standard Deviation Based Data Clustering Algorithm

Kuttiannan Thangavel
Department of Mathematics
Gandhigram Rural Institute, Deemed University

Durairaj Ashok Kumar
Department of Computer Science
Government Arts College

The clustering problem has been widely studied because it arises in many knowledge management oriented applications. It aims at identifying the distribution of patterns and intrinsic correlations in data sets by partitioning the data points into similarity clusters. Traditional clustering algorithms use distance functions to measure similarity centroid, which subside the influences of data points. Hence, in this article a novel non-distance based clustering algorithm is proposed which uses Combined Standard Deviation (CSD) as measure of similarity. The performance of CSD based K-means approach, called K-CSD clustering algorithm, is tested on synthetic data sets. It compared favorably to widely used K-means clustering algorithm.

Key words: Clustering algorithm; combined standard deviation.

Introduction

A fundamental problem that frequently arises in a great variety of fields, such as pattern recognition, image processing, machine learning and statistics in the clustering problem (Narasimha, Jain, & Flyinn, 1999). In its basic form, the clustering problem is defined as the problem of finding homogenous groups of data points in a given data set. Each of these groups is called a cluster and can be defined as a region in which the density of objects is locally higher than in other regions.

Clustering methods can be classified into two categories: Hierarchical and Non-Hierarchical. The hierarchical methods can be further divided into agglomerative methods is viewed as a cluster and at each level, some clusters are divided into smaller clusters. There are also many non-hierarchical methods, which divide the set into clusters. These methods are further divided into two: the partitioning method, in which the clusters are mutually exclusive and the clumping method, in which overlap is allowed.

The simplest form of clustering is partitional clustering which aims at partitioning a data set into disjoint subsets (clusters) so that specific clustering criteria are optimized. The most widely used criteria in this clustering is the error criterion, which for each point computes its squared distance from the corresponding cluster center and then takes the sum of these distances for all points in the data set. A popular clustering method that minimizes the clustering error is the K-means clustering algorithm. However, the k-means clustering algorithm is a local search procedure and it is well known that its performance heavily depends on the initial starting conditions and centroid computed based on that (Pena & Larranaga, 1999). To treat this problem, several other techniques have been developed that are based on stochastic global optimization methods (eg. Genetic algorithm

K. Thangavel is Head of the Department of Mathematics, Gandhigram Rural Institute–Deemed University, Gandhigram–624 302, Tamilnadu, India. His research includes optimization algorithms, pattern searching and recognition algorithms, and neural networks. E–mail: ktvel@rediffmail.com. D. Ashok Kumar, a doctoral candidate, Lecturer, and Head in the Department of Computer Science, Government Arts College, Udumalpet–642 126, Tamilnadu, India. Email: akudaiyar@rediffmail.com. His research interest includes pattern recognition, neural networks, and genetic algorithms

simulated annealing). However, it must be noted that these techniques have not gained wide acceptance and in many practical applications the clustering method that is used in the K-means clustering algorithm with multiple restarts (Maulik & Bandyopadhyay, 2000).

The K-CSD clustering algorithm is proposed, which constitutes an effective clustering for minimization of the clustering error. The basic idea underlying the proposed method is that an optimal solution for a clustering problem with K clusters can be obtained using combined standard deviation. At each step, instead of placing the data point by minimum distance between centroid and the data point, the minimum combined standard deviation is used which leads to optimal clusters. In addition to effectiveness, the method is deterministic and does not depend on centroid. These are significant advantages over all clustering approaches mentioned above.

Clustering

Clustering has been always a key task in the process of acquiring knowledge. The complexity and especially the diversity of phenomena have forced society to organize the things based on their similarities (Spath, 1989). One can say that the objective of the cluster analysis is to sort out the observations into groups such that the degree of natural association is high among members of the same group and low between members of different groups. And clustering is a technique, which is used to find groups of clusters that are somehow similar in characteristic from the given data set for which the real structure is unknown.

Clustering is often confused with classification, but there are some differences between the two. In classification, the data are assigned to predefined classes or clusters, whereas in clustering the classes or clusters are also to be defined and also when the only data available are unlabelled. The classification problems are, sometimes, referred to as unsupervised classification. Cluster analysis can be defined as a wide variety of procedures that can be used to create a classification. These procedures empirically form clusters of groups of highly similar entities. In other words, it can be said that cluster analysis defines group of cases through a number of procedures, which are more similar among them than all the others.

The clustering methods can be basically classified into two categories: Hierarchical and Nonhierarchical. The hierarchical methods can be further divided into the agglomerative methods and the divisive methods. The agglomerative methods merge together the most similar clusters at each level and the merged clusters will remain in the same cluster at all higher levels. In the divisive methods, initially, the set of all object is viewed as a cluster and at each level, some clusters are divided into smaller clusters. There are also many nonhierarchical methods which divide the dataset into clusters. These methods are further divided into two: the partitioning method, in which the clusters are mutually exclusive and the clumping method, in which overlap is allowed.

For years, many clustering techniques were proposed in partitional clustering and are now available in the literature (Narasimha, Jain, & Flyinn, 1999). The methods are Forgy's algorithm, Kmeans algorithm, ISODATA and its variants. The extensive studies (Tseng & Yang, 1999; Narashinha & Sridhar, 1991; Maulik & Bandyopadhyay, 2000) dealing with comparative analysis of different clustering methods suggests that there is no general strategy, which works equally well in the different problems domain. However, it has been found that it is usually beneficial to run schemes that are simpler, and execute them several times, rather than using schemes that are very complex but need to be run only once.

K-Means Clustering Algorithm

The aim of this study is a clustering technique that will not assume any particular underlying distribution of the data set being considered. As well, it should be conceptually simple like the K-means algorithm (Duda & Hart, 1973; Macqueen, 1967). The searching through algorithm is explored in order to search for appropriate cluster centers in the feature space such that a similarity metric of the resulting cluster is optimized.

In fact, to compare the performance or to check the optimality, one does not have the sufficient information regarding the structure of the data set. Thus, to determine the best clusters, a better algorithm is devised which is more valid. It can be established by ranking the utility of clustering results obtained from different clusters algorithms, with respect to certain application domains, where utility can be measured. As the cluster centers are updated in the K-means and proposed algorithms, the distance between the cluster centers and each of its points can be treated as a unique measure. Mathematically, the clustering metric μ for K clusters C_1, C_2, …, C_K

$$\mu(C_1, C_2,..., C_K) = \sum_{i=1}^{K} \sum_{xj \in Ci} ||xj - zi||$$

where Ci are clusters and zi are cluster centers.

The clustering algorithm searches for the appropriate cluster centers z_1, z_2, …, z_K such that the clustering metric μ is minimized. The K-means algorithm is briefly described below in the sequel:

Input: Set of sample patterns {x_1, x_2, …, x_m},
$x_i \in R^n$

Output: Set of Clusters { C_1, C_2, …, C_K }.

Step 1: Choose K initial cluster centers z_1, z_2, …, z_K randomly from the m patterns { x_1, x_2, …, x_m } where K < m.

Step 2: Assign pattern x_i to cluster C_j, where i = 1, 2, …, m and j ∈ {1, 2, …, K}, if and only if $||x_j - z_j|| < ||x_j - z_p||$, p = 1, 2, …, K and j ≠ p. Ties are resolved arbitrarily. Compute cluster centers for each point x_i as follows,
$z_i = (1/n_i) \sum x_j$, i = 1, 2 , … , K. $x_j \in C_i$ Where n_i is the number of elements belongs to cluster Ci.

Step 3: Assign each pattern x_i to cluster Cj, where i = 1, 2, …, m and j ∈ {1, 2, …, K} if and only if $||x_j - z_j|| < ||x_j - z_p||$, p = 1, 2, …, K and j ≠ p, where || • || is an Euclidean metric norm. Ties are resolved arbitrarily, without changing the cluster centers z_j, j = 1, 2, …, K

Step 4: Stop.

K-CSD Clustering Algorithm

In a nutshell, the clustering capability of proposed clustering technique using combined standard deviation (Gupta, 2001) is stated in the following steps:

Input: Set of sample patterns {x_1, x_2, …, x_m}, x_i ∈ R^n

Output: Set of clusters { C_1, C_2, …, C_K }.

Step 1: Choose K initial cluster points z_1, z_2, …, z_K randomly from the m patterns {x_1, x_2, …, x_m} (where K < m) for each cluster.

Step 2: Assign pattern x_i to cluster C_j, where i = 1, 2, …, m and j ∈ {1, 2, …, K}, if and only if CSD(x_j , C_j) < CSD(x_j, C_p), p = 1, 2, …, K and j ≠ p. Ties are resolved arbitrarily. The CSD(x_j , C_j) is obtained by including point x_i into Cluster C_j and find the Combined Standard Deviation of new cluster C_j .

Step 3: Compute cluster centers for each point x_i as follows, $z_i = (1/n_i)\sum x_j$, i = 1, 2 , … , K. $x_j \in C_i$ Where n_i is the number of elements belongs to cluster Ci.

Step 4: Assuming z_i are the new initial points to each cluster $C_{j.}$ Assign each pattern x_i to cluster C_j, where i = 1, 2, …, m and j ∈ {1, 2, …, K} if and only if CSD(x_j , C_j) < CSD(x_j, Cp), p = 1, 2, …, K and j ≠ p. Ties are resolved arbitrarily, without changing the cluster centers z_j, j = 1, 2, …, K

Step 5: Stop

Experimental Results

The experimental results are carried out to compare the Proposed Algorithm clustering algorithm with the K-means clustering algorithm using two synthetic data sets: Data1 and Data2. These are described below:

Data1: This is a non-overlapping two dimensional data set where the number of classes is three. It has several patterns which are selected from those classes by giving equal probabilities. The value of K is chosen to be 3 for this data set.

Class 1: [0, 20] X [40, 60]
Class 2: [40, 60] X [0, 20]
Class 3: [80,100] X [60, 80]

The results of K-means clustering algorithm and Proposed Algorithm clustering algorithm are shown in the following Tables: Table 1, Table 2, Table 3, and Table 4 for 30, 60, 90, and 120 patterns of Data 1 respectively for different configurations of data sets generated.

Table 1 : 30 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	186.17	3	115.69
2	3	145.12	3	131.74
3	3	156.12	3	130.42
4	3	186.05	3	235.82
5	3	77.52	3	129.23
Total	15	750.98	15	742.90
Average	3	150.196	3	148.58

Table 2 : 60 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	282.32	3	320.43
2	3	214.27	3	187.92
3	3	274.54	3	201.53
4	3	102.26	3	187.97
5	3	224.85	3	179.29
Total	15	1098.24	14	1077.14
Average	3	219.648	2.8	215.428

Table 3 : 90 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	µ – Euclidean metric	Number of Clusters	µ - Euclidean metric
1	3	264.46	3	216.52
2	3	282.80	3	250.27
3	3	187.65	3	140.41
4	3	338.13	3	344.81
5	3	128.46	3	128.94
Total	15	1201.50	15	1080.95
Average	3	240.30	3	216.19

Table 4 : 120 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	µ – Euclidean metric	Number of Clusters	µ - Euclidean metric
1	3	252.87	3	272.63
2	3	326.26	3	278.94
3	3	371.83	3	272.04
4	3	323.89	3	277.12
5	3	276.22	3	248.57
Total	15	1551.07	15	1349.30
Average	3	310.214	3	269.86

Data2: This is an overlapping two dimensional data set where the number of classes is three. It has several patterns which are selected from those classes by giving equal probabilities. In the K-means algorithms, the value of K is chosen to be 3 for this data set.

Class 1: [-3.3,-0.7] X [0.7, 3.3]
Class 2: [-1.3, 1.3] X [0.7, 3.3]
Class 3: [-3.3,-0.7] X [-1.3, 1.3]

The results of K-means clustering algorithm and the Proposed Algorithm clustering algorithm are shown in the following Tables: Table 5, Table 6, Table 7 and Table 8 for 30, 60, 90 and 120 patterns of Data 2 respectively for different configurations of data sets generated.

Table 5 : 30 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	µ Euclidean metric	Number of Clusters	µ Euclidean metric
1	3	10.22	3	14.33
2	3	13.55	3	9.40
3	3	8.17	3	9.82
4	3	14.27	3	14.21
5	3	16.22	3	9.88
Total	15	62.43	15	57.64
Average	3	12.486	3	11.528

Table 6 : 60 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	µ Euclidean metric	Number of Clusters	µ Euclidean metric
1	3	13.65	3	10.07
2	3	13.54	3	12.92
3	3	14.03	3	16.64
4	3	13.25	3	17.64
5	3	17.79	3	13.10
Total	15	72.26	15	70.37
Average	3	14.452	3	14.074

Table 7 : 90 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	µ Euclidean metric	Number of Clusters	µ Euclidean metric
1	3	26.38	3	15.29
2	3	21.22	3	27.18
3	3	23.83	3	17.03
4	3	20.83	3	16.55
5	3	17.19	3	16.63
Total	15	109.45	15	92.68
Average	3	21.88	3	18.536

Table 8 : 120 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	µ Euclidean metric	Number of Clusters	µ Euclidean metric
1	3	28.63	3	24.74
2	3	30.44	3	19.80
3	3	18.56	3	18.37
4	3	19.22	3	21.87
5	3	20.13	3	20.72
Total	15	116.98	15	105.5
Average	3	23.396	3	21.10

Table 9

Data	No. of Patterns	K-means		K-CSD	
		Number of Clusters	Average Euclidean metric - µ	Number of Clusters	Average Euclidean metric - µ
1	30	3	150.196	3	148.580
	60	3	219.648	3	215.428
	90	3	240.30	3	216.190
	120	3	310.214	3	269.860
2	30	3	12.486	3	11.528
	60	3	14.452	3	14.074
	90	3	21.88	3	18.536
	120	3	23.396	3	21.100
Total		24	992.572	24	915.296
Average		3	124.072	3	114.412

Conclusion

The implemented K-means and proposed K-CSD clustering algorithm is tested with two different synthetic datasets to optimize the clustering metric µ. The tested average metric measures of the Data 1 and Data 2 are tabulated in Table 9.

From the Table 9, it could be seen that the average metric is reduced in the proposed algorithm. Future work is planned to design and implement algorithms to cluster data sets with large amount of objects. Such algorithms are required in a number of data mining applications, such as partitioning very large heterogeneous sets of objects into a number of smaller and more manageable homogeneous subsets that can be more easily modeled and analyzed and detecting underrepresented concepts, e.g., fraud in a very large number of insurance claims.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York, N.Y.: Wiley.

Gupta S.P., *Statistical methods*, Sultan Chand & Sons, 2001.

Lin, Y. T. & Shiueng, B.Y (2001). A genetic apporach to automatic clustering problem. *Pattern Recognition*, *34*(2), 415-424.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate data. Proc. 5th Berkeley Symposium on probability and statistics. Berkeley, C.A.: University of California Press.

Narashinha, M. M. & Sridhar, V. (1991). A knowledge based clustering algorithm. *Pattern Recognition letters*, *12*, 511-517.

Narasimha, M. M., Jain, A. K., & Flyinn, P. J. (1999). Data clustering : A review. ACM Computing Surveys, ,*31*(3), 264-323.

Pena, J. A. L. J. M. & Larranaga, P. (1999). An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters, 20*, 1027-1040.

Spath, H. (1989). *Cluster analysis algorithms*. Chichester, U.K.: Ellis Horwod.

Ujiwal, M. & Sanghamitra, B. (2000). Genetic algorithm based clustering technique. *Pattern Recognition Letters, 33*, 1455-1465.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 266-272

JMASM22: A Convenient Way Of Generating Normal Random Variables Using Generalized Exponential Distribution

Debasis Kundu Anubhav Manglick
Department of Mathematics and Statistics
Indian Institute of Technology

Rameshwar D. Gupta
Department of Computer Science and Applied Statistics
University of New Brunswick

A convenient method to generate normal random variable using a generalized exponential distribution is proposed. The new method is compared with the other existing methods and it is observed that the proposed method is quite competitive with most of the existing methods in terms of the $K-S$ distances and the corresponding p-values.

Key words: Generalized exponential distribution; Kolmogorov-Smirnov distances; random number generator.

Introduction

Generating normal random numbers is an old and very important problem in the statistical literature. Several algorithms are available in the literature to generate normal random numbers like Box-Muller methods, Marsaglia-Bray method, Acceptance-Rejection method, Ahrens-Dieter method, etc. The book of Johnson, Kotz and Balakrishnan (1995) provided an extensive list of references of the different algorithms available today. Among the several methods the most popular ones are the Box-Muller transformation method or the improvement suggested by Marsagilia and Bray. Most of the statistical packages like, SAS, IMSL, SPSS, S-Plus, or Numerical Recipes use this method. In this article, a simple and convenient method of generating normal random numbers using generalized exponential distribution is proposed.

Generalized exponential (GE) distribution has been proposed and studied quite extensively recently by Gupta and Kundu (1999; 2001a; 2001b; 2002; 2003a). The readers may be referred to some of the related literature on (GE) distribution by Raqab (2002), Raqab and Ahsanullah (2001), and Zheng (2002). The two-parameter GE distribution has the following distribution function:

$$F_{GE}(x;\alpha,\lambda)=(1-e^{-\lambda x})^{\alpha};\ \alpha,\lambda>0 \qquad (1)$$

for $x>0$ and 0 otherwise. The corresponding density function is;

$$f_{GE}(x;\alpha,\lambda)=\alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x};\ \alpha,\lambda>0, \qquad (2)$$

for $x>0$ and 0 otherwise. Here α and λ are the shape and scale parameters respectively. When $\alpha=1$, it coincides with the exponential distribution. If $\alpha\le 1$, the density function of a

Debasis Kundu is a Professor of Statistics. His research interests include statistical signal processing, reliability analysis, statistical computing and competing risks models. Rameshwar D. Gupta is a Professor and Acting Associate Dean of Graduate Studies. His research interests are statistical inference, multivariate statistical analysis, and statistical analysis of reliability and life-testing models. Anubhav Manglick is a Master’s student. His major research interests include generalised exponential models, model discrimination and high frequency financial time series data. He has published in the *Journal of Statistical Planning and Inference* and *Naval Research Logistiscs*.

GE distribution is a strictly decreasing function and for $\alpha > 1$, it has a uni-modal density function. The shape of the density function of the GE distribution for different α can be found in Gupta and Kundu (2001a).

In a recent study by Kundu, Gupta and Manglick (2005), it was observed that in certain cases log-normal distribution can be approximated quite well by GE distribution and vice versa. In fact, for certain ranges of the shape parameters of the GE distributions the distance between the GE and log-normal distributions can be very small.

The main idea in this article is to use this particular property of a GE distribution to generate log-normal random variables and in turn generate normal random variables. It may be mentioned that the GE distribution function is an analytically invertible function, therefore, the generation of GE random variables is immediate using uniform random variables.

Methodology

The density function of a log-normal random variable with scale parameter θ and shape parameter σ is denoted as

$$f_{LN}(x;\theta,\sigma) = \frac{1}{\sqrt{2\pi}\, x\sigma} e^{-\frac{(\ln x - \ln\theta)}{2\sigma^2}} \;;\; \theta,\sigma > 0 \quad , (3)$$

for $x > 0$ and 0 otherwise. If X is a log-normal random variable with scale parameter θ and shape parameter σ, then

$$E(X) = \theta e^{\frac{\sigma^2}{2}}$$

and

$$V(X) = \theta^2 e^{\sigma^2}(e^{\sigma^2} - 1). \qquad (4)$$

Note that $\ln X$ is a normal random variable with mean $\ln\theta = \mu$ (say) and variance σ^2.

Similarly, if X is a generalized exponential random variable with the scale parameter λ and shape parameter α, then

$$E(X) = \frac{1}{\lambda}(\psi(\alpha+1) - \psi(1))$$

and

$$V(X) = \frac{1}{\lambda^2}(\psi'(1) - \psi'(\alpha+1)). \qquad (5)$$

It was observed by Kundu, Gupta and Manglick (2005) that a generalized exponential distribution can be approximated very well by a log-normal distribution for certain ranges of the shape parameters. The first two moments of the two distribution functions are equated to compute σ and θ from a given α and λ. Without loss of generality, $\lambda = 1$ is taken. For a given $\alpha = \alpha_0$, equating (4) and (5) one obtains

$$\theta e^{\frac{\sigma^2}{2}} = \psi(\alpha_0 + 1) - \psi(1) = A_0 \qquad (6)$$

$$\theta^2 e^{\sigma^2}(e^{\sigma^2} - 1) = \psi'(1) - \psi'(\alpha_0 + 1) = B_0 \qquad (7)$$

Therefore, solving (6) and (7), one obtains

$$\ln\theta_0 = \mu_0 = \ln\, A_0 - \frac{1}{2}\ln(1 + \frac{B_0}{A_0^2}), \qquad (8)$$

$$\sigma_0 = \sqrt{\ln(1 + \frac{B_0}{A_0^2})}\,. \qquad (9)$$

Using (8) and (9), standard normal random variable can be easily generated as follows:

Algorithm

Step 1: Generate U an uniform $(0,1)$ random variable.

Step 2: For a fixed α_0, generate $X = -\ln(1 - U^{\frac{1}{\sigma_0}})$. Note that X is a generalized exponential random variable with shape parameter α_0 and scale parameter 1.

Step 3: Compute $Z = \dfrac{\ln X - \mu_0}{\sigma_0}$. Here Z is the desired standard normal random variable.

An alternative approximation is also possible. Instead of equating the moments of the two distributions, one can equate the corresponding L-moments also. The L-moments of any distribution are analogous to the conventional moments, but they are based on the quantiles and they can be estimated by the linear combination of order statistics, i.e. by L-statistics (see Hosking, 1990, for details). It is observed by Gupta and Kundu (2003b) in a similar study of approximating gamma distribution by generalized exponential distribution that the L-moments perform better than the ordinary moments.

Let Z be any random variable having finite first moment and suppose $Z_{1:n} \leq .... \leq Z_{n:n}$ be the order statistics of a random sample of size n drawn from the distribution of Z. Then the L-moments are defined as follows:

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(Z_{r-k:r}); \; r = 1, 2, ... \quad (10)$$

The two $L-$moments of a log-normal distribution are

$$\lambda_1 = \theta e^{\frac{\sigma^2}{2}}$$

and

$$\lambda_2 = \theta e^{\frac{\sigma^2}{2}} erf(\frac{\sigma}{2}), \quad (11)$$

where $\text{erf}(x) = 2\phi(\sqrt{2}\ x) - 1$ and $\phi(x)$ is the distribution function of the standard normal distribution. Similarly, the two L-moments of a GE random variable are

$$\lambda_1 = \frac{1}{\lambda}(\psi(\alpha+1) - \psi(1))$$

and

$$\lambda_2 = \frac{1}{\lambda^2}(\psi(2\alpha+1) - \psi(\alpha+1)). \quad (12)$$

Therefore, as before equating the first two L-moments for a given $\alpha = \alpha_0$ and for $\lambda = 1$, one obtains

$$\theta e^{\frac{\sigma^2}{2}} = \psi(\alpha_0 + 1) - \psi(1) = A_0 \quad (13)$$

$$\theta e^{\frac{\sigma^2}{2}} erf(\frac{\sigma}{2}) = \psi(2\alpha_0 + 1) - \psi(\alpha_0 + 1) = B_1 \quad (14)$$

Solving (13) and (14), one obtains the solutions of θ and σ as

$$\ln \theta_1 = \mu_1 = \ln\ A_0 - \frac{\sigma_1^2}{2} \quad (15)$$

$$\sigma_1 = \sqrt{2}\, \phi^{-1}\ (\frac{1}{2}(1 + \frac{B_1}{A_0})), \quad (16)$$

where ϕ is the cumulative distribution function of standard normal distribution. Therefore, in the proposed algorithm, instead of using $(\mu_0, \sigma_0), (\mu_1, \sigma_1)$ also can be used.

Numerical Comparisons and Discussions

In this section, an attempt is made to determine the value of α_0, so that the distance between the generalized exponential distribution and the corresponding log-normal distribution is minimum. All the computations are performed using Pentium IV processor and the random number generation routines by Press et al. (1993). The distance function between the two distribution functions is considered as the Kolmogorv-Smirnov $(K-S)$ distance only. To be more precise, the $K-S$ distance between the GE is computed, with the shape and scale parameter as σ_0 and 1 respectively, and log-normal distribution with the corresponding shape and scale parameter as $\sigma_0(\sigma_1)$ and $\theta_0(\theta_1)$ respectively. It is believed that the distance function should not make much difference, any other distance function may be considered also. It is observed that as α_0 increases from 0, the $K-S$ distance first decreases, and then increases. When the moments (L-moments) equations have been used, the minimum $K-S$ distance occurred at $\alpha_0 = 12.9 (12.8)$. When $\alpha_0 = 12.9 (12.8)$, then from (8) and (9) ((15) and (16)), the corresponding μ_0 = 1.0820991 (μ_1 = 1.0792510) and σ_0 = 0.3807482 (σ_1 = 0.3820198) was obtained.

To compare the proposed method with the other existing methods, the $K-S$ statistics and the corresponding p-values were mainly used. The method can be described as follows. The standard normal random variables for different sample sizes namely n = 10, 20, 30, 40, 50 and 100 by using Box-Muller (BM) method, Marsaglia-Bray (MB) method, Acceptance-Rejection (AR) method, Ahren-Dieter (AD) method were generated, using moments equations (MM) and using L-moments equations (LM). In each case, the $K-S$ distance and the corresponding p-value between the empirical distribution function and the standard normal distribution function was computed. The process was replicated 10,000 times and the average $K-S$ distances, the average p-values and the corresponding standard deviations were computed. The results are reported in Table 1. In each case the standard deviations are reported within bracket below the average values. From the table values it is quite clear that, based on the $K-S$ distances and p values the proposed methods work quite well. Also, an effort is made to compute $(Z \le z)$ using the proposed approximation, where Z denotes the standard normal random variable. Note that

$$P(Z \le z) \approx (1 - e^{-e^{z\sigma_0 + \mu_0}})^{12.9}$$

or

$$P(Z \le z) \approx (1 - e^{-e^{z\sigma_1 + \mu_1}})^{12.8}. \qquad (17)$$

The results are reported in Table 2. It is clear from Table 2 that using μ_0 and σ_0 the maximum error can be 0.0005, where as using μ_1 and σ_1 the maximum error can be 0.0003. From Table 2, it is clear that L-moments approximations work better than the moments approximations.

Conclusions

A simple and convenient method of generating normal random variables is provided. Even simple scientific calculator can be used to generate normal random number from the uniform generator very quickly. It can be implemented very easily by using a one line program. It is also observed that the standard normal distribution function can be approximated at least up to three decimal places using the simple approximations.

Table 1. The average K-S distances and the corresponding p-values for different methods based on 10,000 replications. The standard deviations are reported within brackets in each case below the average values.

N		BM	MB	AR	AD	MM	LM
10	K-S	0.2587	0.2587	0.2597	0.2591	0.2586	0.2587
		(0.0796)	(0.0796)	(0.0809)	(0.0804)	(0.0794)	(0.0795)
	p	0.5127	0.5128	0.5109	0.5114	0.5135	0.5132
		(0.2938)	(0.2938)	(0.2970)	(0.2955)	(0.2930)	(0.2931)
20	K-S	0.1851	0.1851	0.1871	0.1860	0.1866	0.1867
		(0.0571)	(0.0571)	(0.0575)	(0.0578)	(0.0571)	(0.0572)
	p	0.5178	0.5178	0.5068	0.5135	0.5089	0.5085
		(0.2934)	(0.2934)	(0.2934)	(0.2957)	(0.2927)	(0.2928)
30	K-S	0.1532	0.1532	0.1533	0.1537	0.1524	0.1525
		(0.0467)	(0.0467)	(0.0466)	(0.0477)	(0.0465)	(0.0465)
	p	0.5094	0.5094	0.5086	0.5088	0.5150	0.5145
		(0.2937)	(0.2937)	(0.2923)	(0.2953)	(0.2930)	(0.2930)
40	K-S	0.1331	0.1331	0.1331	0.1335	0.1334	0.1334
		(0.0409)	(0.0488)	(0.0410)	(0.0412)	(0.0410	(0.0410)
	p	0.5111	0.5111	0.5121	0.5094	0.5097	0.5092
		(0.2923	(0.2923)	(0.2926)	(0.2945)	(0.2927)	(0.2928)
50	K-S	0.1191	0.1191	0.1197	0.1193	0.1199	0.1200
		(0.0370)	(0.0370)	(0.0364)	(0.0368)	(0.0366)	(0.0366)
	p	0.5140	0.5140	0.5071	0.5120	0.5058	0.5053
		(0.2931)	(0.2931)	(0.2924)	(0.2923)	(0.2927)	(0.2927)
100	K-S	0.0852	0.0852	0.0851	0.0854	0.0851	0.0852
		(0.0257)	(0.0257)	(0.0262)	(0.0257)	(0.0259)	(0.0259)
	P	0.5059	0.5059	0.5096	0.5043	0.5082	0.5077
		(0.2914)	(0.2914)	(0.2932)	(0.2895)	(0.2912)	(0.2912)

Table 2. The exact value of $\phi(z)$ and the two approximate values are reported.

Z	L-Moment	Exact	Moment
0.0	0.49984	0.50000	0.50014
0.1	0.53981	0.53983	0.54006
0.2	0.57935	0.57926	0.57955
0.3	0.61808	0.61791	0.61824
0.4	0.65564	0.65541	0.65574
0.5	0.69168	0.69145	0.69174
0.6	0.72594	0.72572	0.72595
0.7	0.75818	0.75800	0.75815
0.8	0.78822	0.78810	0.78814
0.9	0.81593	0.81588	0.81582
1.0	0.84125	0.84127	0.84112
1.1	0.86416	0.86424	0.86400
1.2	0.88469	0.88482	0.88452
1.3	0.90292	0.90308	0.90273
1.4	0.91893	0.91911	0.91875
1.5	0.93288	0.93305	0.93269
1.6	0.94490	0.94505	0.94472
1.7	0.95517	0.95528	0.95500
1.8	0.96385	0.96392	0.96369
1.9	0.97112	0.97114	0.97097
2.0	0.97714	0.97711	0.97701
2.1	0.98209	0.98200	0.98197
2.2	0.98610	0.98597	0.98600
2.3	0.98933	0.98916	0.98924
2.4	0.99189	0.99170	0.99181
2.5	0.99390	0.99370	0.99384
2.6	0.99547	0.99526	0.99542
2.7	0.99667	0.99647	0.99663
2.8	0.99759	0.99739	0.99755
2.9	0.99827	0.99809	0.99825
3.0	0.99878	0.99861	0.99876
3.5	0.99983	0.99976	0.99982
4.0	0.99998	0.99997	0.99998

References

Gupta, R. D. & Kundu, D. (1999). Generalized exponential distributions. *Australian and New Zealand Journal of Statistics, 41*(2), 173-188.

Gupta, R. D. & Kundu, D. (2001). Exponentiated exponential distribution: An alternative to gamma and Weibull distributions. *Biometrical Journal, 43*(1), 117-130.

Gupta, R. D. & Kundu, D. (2001). Generalized exponential distributions: Different methods of estimations. *Journal of Statistical Computations and Simulations*, *69*(4), 315-338.

Gupta, R. D. & Kundu, D. (2002). Generalized exponential distributions: Statistical inferences. *Journal of Statistical Theory and Applications, 1*(2), 101-118, 2002.

Gupta, R. D. & Kundu, D. (2003). Discriminating between Weibull and generalized exponential distributions. *Computational Statistics and Data Analysis, 43*, 179-196.

Gupta, R. D. & Kundu, D. (2003). Closeness of gamma and generalized exponential distributions. *Communications in Statistics - Theory and Methods*, *32*(4), 705-721.

Hosking, J. R. M. (1990). L-moment: Analysis and estimation of distribution using linear combination of order statistics. *Journal of Royal Statistical Society, 52*(B), 105-124.

Johnson, N., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distribution* (vol. 1). New York, N.Y.: John Willey and Sons.

Kundu, D., Gupta, R. D., & Manglick, A. (2005). Discriminating between log-normal and generalized exponential distribution. *Journal of Statistical Planning and Inference, 127*, 213-227.

Press et al. (1993). *Numerical recipes*. Cambridge: Cambridge University Press.

Raqab, M. Z. (2002). Inference for generalized exponential distribution based on record statistics. *Journal of Statistical Planning and Inference*, *104*(2), 339-350.

Raqab, M. Z. & Ahsanullah, M. (2001). Estimation of the location and scale parameters of the generalized exponential distribution based on order statistics. *Journal of Statistical Computations and Simulations, 69*(2), 109-124.

Zheng, G. (2002). On the Fisher information matrix in type-II censored data from the exponentiated exponential family. *Biometrical Journal, 44*(3), 353-357.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 273-280

1538 – 9472/06/$95.00

JMASM23: Cluster Analysis In Epidemiological Data (Matlab)

Andrés M. Alonso
Department of Statistics
Universidad Carlos III de Madrid

Matlab functions for testing the existence of time, space and time-space clusters of disease occurrences are presented. The classical scan test, the Ederer, Myers and Mantel's test, the Ohno, Aoki and Aoki's test, and the Knox's test are considered.

Key words: Time cluster, space cluster, time-space cluster, epidemiology, Monte Carlo.

Introduction

The concept of groups or clusters of disease occurrences is enough well-known and intuitive. A cluster is defined as an unusual, real or perceived group of health events that are grouped in the time and/or in the space. Many triumphs in the control of infectious diseases have been the result of the epidemiological study of clusters of cases, for instance, the epidemic of cholera in London in the 1850s and the investigation of cases of pneumonia in Philadelphia in 1976 (legionary disease). The investigation of clusters of non-infectious diseases also has remarkable examples: dermatitis in people who use rings made with contaminated gold and vaginal carcinomas in women whose mothers who consumed diethylstilbestrol (see CDC, 1990).

The investigation of perceived clusters of health events requires the knowledge of some statistical instruments for determining if the observed group is real, taking into account the circumstances under study (the data type, the availability of comparison data, etc.). In this article, the aim is to describe some of the statistical techniques used to investigate clusters of health events and to provide Matlab routines that implement these techniques.

Detection of Time Clusters

A time cluster is defined as a non-uniform distribution of the cases in the time interval for a given population under study. The objectives of these studies are:

1. To identify secular tendencies of the frequency of diseases in the populations.
2. To identify cyclical fluctuations in the occurrence of a disease.
3. To identify local epidemics of a disease.

Attention is focused on the methods related to the detection of local epidemics.

Scan test

The scan test is used to determine if the cases that appear in a geographic area are significantly near in time. The test statistics are the maximum number of events that happen in a time interval of fixed size t. This value is obtained by scanning in all the intervals of length t in the period under study. The critical values for this test are provided in the tables calculated by Naus (1965, 1966) and Wallenstein (1980).

It is assumed that T is the complete observational interval and t is the duration time of one epidemic. Let be $r = t/T$, N the number of cases that happened in time T, and $p = \Pr(n, N, r)$ is the probability that a maximum number of cases in any interval of length t exceeds or is equal to n. This probability is calculated under the hypothesis that the N events are uniformly distributed in the interval T. The problem consists of estimating p. Wallenstein (1980) proposed the following algorithm: If the observed interval is a multiple of 12, 24, 36, 48 or 60 months, and if the duration of the epidemic is a multiple from 2 to 4

Andrés Alonso is a *Juan de La Cierva* Researcher at the Department of Statistics. His areas of research interest are statistical computing, resampling methods and biostatistics. E-mail: andres.alonso@uc3m.es.

or 6 months, many quotients $r=t/T$ can be reduced to the fraction $1/L$ with $L = 4, 6, 8, 12, 15$ or the 24. If N is greater than 10 and smaller than 100, then tables in Wallenstein (1980) give the critical values of the distribution of n.

Example 1: The following table shows the number of cases of trisomia and spontaneous abortion in the city of New York between July/1975 and June/1977 (see Bailar et al., 1970).

```
function p = ProbabilityOfScanTest(n, N, t, T, B)

% Inputs:
% -------
% n : Maximum number of cases observed in t periods.
% N : Number of cases observed in T periods.
% t : Epidemic duration time.
% T : Total observation time.
% B : Number of replications.
%
% Output:
% -------
% p : Probability of having a value bigger or equal to n.

% Cases are B independents replicas of a uniform distribution
% of N cases in T periods.
Cases = zeros(T, B);
for b = 1:B
   X = rand(N, 1);
   for ii = 1:N
      for tt = 1:T
         if ((tt-1)/T < X(ii, 1) & X(ii, 1) < tt/T)
            Cases(tt, b) = Cases(tt, b) + 1;
         end
      end
   end
end

% Calculating the scan statistics using the B generates replicas
% stored in variable Cases.
ScanStatistics = zeros(B, 1);
for b = 1:B
   for tt = 1:T-t+1
      if (ScanStatistics(b, 1) < sum(Cases(tt:tt+t-1, b)))
         ScanStatistics(b, 1) = sum(Cases(tt:tt+t-1, b));
      end
   end
end

% Estimating the probability of having a scan statistics bigger
% or equal to the observed value, n.
p = sum(ScanStatistics >= n)/B;
```

Figure 1. Matlab Function `p`

Month / Year	Cases
07/1975 – 12/1975	0, 4, 1, 2, 1, 3
01/1976 – 06/1976	1, 3, 2, 2, 3, 4
07/1976 – 12/1976	1, 1, 1, 2, 4, 7
01/1976 – 06/1976	7, 2, 2, 6, 1, 2

Therefore, N = 62, T = 24 months and the epidemic duration is fixed to t=2 months. Then n=14 and Pr(14,62,1/12) can be calculated. The Matlab function in Figure 1 obtain the probability p = Pr(n, N, r) by a Monte Carlo simulation procedure. The results of the above function for the data in Example 1 is Pr(14,62,2,24)= 0.0113. It supports the conclusion of a time cluster.

Test of Ederer, Myers and Mantel

The period under study is divided in k disjoints intervals. Under the null hypothesis of no grouping, the n cases will have to be distributed uniformly in the k intervals. The test statistics, m, is the maximum number of cases in an interval. Mantel et al. (1976) calculated tables for the expectation and variance of m under the null hypothesis of no group and for selected values of k and n. In the following table, the approximated estimators of E (m) and Var(m) are shown when the number of cases is greater than 100 (see Mantel et al., 1976).

Number of intervals, k	$E(m)$	$Var(m)$
2	$n/2 + 0.3989 * n^{\frac{1}{2}}$	$0.09084 * n$
3	$n/3 + 0.4886 * n^{\frac{1}{2}}$	$0.07538 * n$
4	$n/4 + 0.5147 * n^{\frac{1}{2}}$	$0.06043 * n$
5	$n/5 + 0.5201 * n^{\frac{1}{2}}$	$0.04951 * n$

Example 2: Assume that the number of children with congenital malformations born in the same year is as follows: 1st trimester: 100 cases, 2nd trimester: 50 cases, 3rd trimester: 50 cases and 4th trimester: 70 cases. If k=4 and n=270, then one can use the estimators of the previous table: $E(m)$= 270/4+0.5147*√270 ≈ 75.95 and $Var(m)$ =0.06043*270 ≈ 16.32. The following statistic is calculated,

$$\chi = \frac{(m - E(m))^2}{Var(m)} = \frac{(100 - 75.95)^2}{16.32} \approx 35.44\,,$$

and it may be concluded that it exists a time cluster.

```
function [E, V] = EdererMyersMantelTest(m, n, k, B)

% Inputs:
% -------
% m : Maximum number of cases observed in one interval.
% n : Number of cases observed in the period under study.
% k : Number of intervals.
% B : Number of replications.
%
% Output:
% -------
% E : Expected value of m.
% V : Variance of m.

% Cases are B independents replicas of a uniform distribution
% of n cases in k intervals.
Cases = zeros(k, B);
for b = 1:B
   X = rand(n, 1);
   for ii = 1:n
      for tt = 1:k
         if ((tt-1)/k < X(ii, 1) & X(ii, 1) < tt/k)
            Cases(tt, b) = Cases(tt, b) + 1;
         end
      end
   end
end

% Calculating the maximum m using the B generated replicas
% stored in variable Cases.
mStatistics = max(Cases);

% Estimating the mean and the variance of m.
E = mean(mStatistics);
V = var(mStatistics);
```

Figure 2. Matlab function `[E, V]`

The Matlab function in Figure 2 obtains the estimators of $E(m)$ and $Var(m)$ by a Monte Carlo simulation procedure. The results of this function for the data in Example 2 is $E(m) = 76.07$ and $Var(m) = 17.52$.

Detection of Space Clusters

A space cluster is defined as a non-uniform distribution of the cases in the area under study relative to the distribution of the population under study. The presence of clusters suggests a possible environmental etiology. The simplest analysis of space cluster is the comparison of the incidence or the prevalence of a particular disease in different geopolitical areas.

Test of Ohno, Aoki and Aoki

The test proposed by Ohno et al. (1979) determines if the obtained geographic pattern is different from the expected geographic pattern under the assumption of a uniform random distribution of the cases in the area under study. The procedure is as follows:

1. Define $k > 2$ disjoint categories of the incidence rates.

2. Identify the adjacent geographic areas in a map of the area under study.
3. Count the number of concordant area pairs.
4. Calculate the expected number of concordant adjacent pairs for each category: Let be N the number of areas and N_i the number of areas in the i-th category, then the number of concordant pairs in category i is $N_i(N_i\text{-}1)/2$. Let A be the number of adjacent pairs of regions, then the expected number of adjacent pairs with the i-th category is

$$E(C_i) = \frac{A}{N(N-1)} N_i(N_i - 1) .$$

5. Calculate the expected number of concordant adjacent pairs:

$$E(C) = \sum_{i=1}^{k} E(C_i) .$$

Finally a χ^2 test statistics, $\chi^2 = \frac{(C - E(C))^2}{E(C)}$, is calculated.

Example 3: The mortality rates of vesicle and esophagus cancer in Japan (1967-71) is categorized according to the following criterion:

Category 1. Rate≥140 by 10000 inhabitants.
Category 2. 120≤Rate≤139.9 by 10000 inhabitants.
Category 3. 80≤Rate≤119.9 by 10000 inhabitants.
Category 4. 60≤Rate≤79.9 by 10000 inhabitants.
Category 5. Rate≤60 by inhabitants.

In 1970, Japan had N = 1,123 cities and towns, without counting the prefecture of Okinawa, with A=2840 adjacent pairs of regions. The number of regions by category was: N_1 = 293, N_2 = 78, N_3 = 256, N_4 = 116 and N_5 = 380. In the following table, the calculation required for Ohno, Aoki and Aoki's test is presented.

Concordant pairs	Observed, C_i	Expected, $E(C_i)$	χ^2
(1,1)	201	192.84	0.35
(2,2)	17	13.54	0.89
(3,3)	170	147.14	3.55
(4,4)	25	30.07	0.85
(5,5)	315	324.61	0.28
Total	728	708.20	0.55

Finally, χ^2=0.55 and it is concluded that evidence does not exist for the geographic association of the vesicle and esophagus cancer in men for these years in Japan. The following Matlab function obtain the value of Ohno, Aoki and Aoki's test statistics given N, A, C and the N_i.

```
function OAAtest = OhnoAokiAokiTest(N, A, Ni, C)

% Inputs:
% -------
% N : Total number of regions.
% A : Number of adjacent regions.
% Ni : Number of regions in the ith category (k x 1 vector).
% C : Observed number of concordant adjacent regions.
%
% Output:
% -------
% OOAAtest : Ohno, Aoki and Aoki test statistics.

% Numbers of categories.
k = length(Ni);

% Expected number of adjacent regions in the ith category.
ECi = A*Ni.*(Ni-1)/(N*(N-1));

% Expected number of concordant adjacent regions.
EC = sum(ECi);

% Ohno, Aoki and Aoki test statistics.
OAAtest = (C-EC)^2/EC;
```

Figure 3. Matlab Function `OAAtest`

Detection of Space-Time Clusters

A space-time cluster is defined as a non-uniform distribution of the cases in space and time, simultaneously. In general, the test of space-time cluster of health events needs a more a more sophisticated elaboration because one needs to prove that if the cases are associated in space they are also significantly near in the time, and vice versa (see, e.g., Kleinbaum et al., 1982).

Test of Knox

The test proposed by Knox (1964) is used to determine if there exists a significant interaction between the sites and the moments of appearance of the disease. It divides the dimensions in space-time into two parts, for which the critical distance in space, *E*, and the critical distance in time, *T*, must be defined. In a contingency table, each pair of cases is classified in one of the following categories: (i) near only in space, (ii) near only in time, (iii) near in space-time, and (iv) distant both in space and in time. The procedure is as follows:

1. Let be *n* the number of cases. For each case, one knows its position in the space and in the time, then there are $N = n(n-1)/2$ possible pairs of cases.
2. Determine the distances in space, *e*, and in time, *t*, for each pair of cases.
3. Classify the *N* pairs according to the following criterion:

   (a) A pair is near in space if $e<E$.

   (b) A pair is near in time if $t<T$.

   (c) A pair is near in space-time if it fulfills (a) and (b), simultaneously.

   (d) When a pair satisfies neither (a) nor (b), then we say that it is not near nor in space nor in time.

4. Construct the following table:

	Space		
Time	Near	Non-Near	Total
Near	X	N_t - X	N_t
Non-Near	N_e - X	N - N_t - N_e+X	N - N_t
Total	N_e	N - N_e	N

where N_e is the number of pairs near in the space, N_t the near ones in the time, and X the near pairs in space-time.

5. The test statistic is the observed number of pairs near in space-time, X. In Knox (1964) it is assumed that X distributes as a Poisson, therefore,

$$p = \Pr(X \geq x) = \sum_{i=x}^{N} \frac{e^{-\lambda}\lambda^{i}}{i!},$$

where λ=N_eN_t/N.

Example 4: The following table shows the results of the method of Knox for 5 cases of meningococcal disease in a territory given in a period of one year, it takes like critical distance in space 500 meters and in time 5 days.

	Space		
Time	Near	Non-Near	Total
Near	X=4	0	N_t= 4
Non-Near	1	5	6
Total	N_e= 5	5	N=10

Therefore, λ=5*4/10 = 2 and Pr(X≥4) = 0.142. The Matlab function in Figure 4 obtains the value of above *p*-value given X, N_e, N_t and N.

```
function pKtest = KnoxTest(X, Ne, Nt, N)

% Inputs:
% -------
% X : Number of pairs near in space-time.
% Ns : Number of pairs near in space.
% Nt : Number of pairs near in time.
% N : Total number of pairs.
%
% Output:
% -------
% pKtest : Pvalue of Knox test statistics.

% Parameter of the Poisson distribution.
lambda = Ne*Nt/N;

% p-value of Knox test statistics.
pKtest = 0;
for i = X:N
   pKtest = pKtest + exp(-lambda)*lambda^i/factorial(i);
end
```

Figure 4. Matlab Function `pKtest`

References

Bailar, J. C., Eisenberg, H. & Mantel, N. (1970). Time between pairs of leukemia cases. *Cancer*, *25*, 1301-1303.

CDC: Center of Disease Control (1990). Guidelines for investigating clusters of health events, *MMWR*, *39*.

Kleinbaum, D. G., Kupper, L. L. & Morgenstern, H. (1982). *Epidemiologic research, principles and quantitative methods*. New York, N.Y.: Van Nostrand Reinhol Company.

Knox, G. (1964). Detection of space-time interactions. *Applied Statistics*, *13*, 25-30.

Mantel, N., Kryscio, R. J. & Myers, M. H. (1976). Tables and formulas for extended use of Ederer-Myers-Mantel disease-clustering procedure. *American Journal of Epidemiology*, *104*, 576-588.

Naus, J. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, *60*, 532-538.

Naus J. I. (1966). Some probabilities, expectations and variances for the size of the largest clusters and smallest intervals. *Journal of the American Statistical Association*, *61*, 1191-1199.

Ohno, Y., Aoki, K. & Aoki, N. (1979). A test of significance for geographic clustering of disease. *International Journal of Epidemiology*, *8*, 273-281.

Wallenstein, S. (1980). A test for detection of clustering over time. *American Journal of Epidemiology*, *111*, 367-372.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 281

1538 – 9472/06/$95.00

ERRATA
Confidence Intervals On Subsets May Be Misleading

Juliet Popper Shaffer
University of California

This errata pertains to Shaffer (2004, "Confidence intervals on subsets may be misleading", *Journal of Modern Applied Statistical Methods*, *3*(2), 261-270). The section entitled "Conditioning when significant results in one direction only are noted" (p. 267-269) has some errors, and the associated Table 3 has an incorrect heading.

(a): The last sentence should be changed to: If the true value is in the direction that is reported, the values in Table 1 are underestimates of the probabilities that the reported intervals cover the true values. Table 4 below gives the probabilities in this case.

(b): The second sentence should be changed to: If the favored direction happens to be the true one, the confidence interval coverage will be greater than the nominal .95 coverage, changing from .97 at the origin (effect size 0) to .95 as the effect size increases.

(c): The correct heading of Table 3 is:

Table 3: Probability that the nominal .95 confidence interval covers the correct value when the results are not significant in the true direction, for a two-sample z test (values in parentheses are probabilities that the intervals are reported; dividing the entries by these probabilities gives the conditional coverage of the intervals, given that they are the only ones reported)

Table 4: True conditional probability that the nominal .95 confidence interval based on the *z* test covers the correct value, given rejection of the null hypothesis in the correct direction (values in parentheses are probabilities of rejection in the correct direction)

	Effect size				
Sample size	.1	.2	.3	.4	.5
5	.30(.05)	.50(.06)	.64(.07)	.73(.10)	.79 (.12)
10	.39(.05)	.62(.07)	.75(.10)	.83(.15)	.87(.20)
20	.50(.06)	.73(.09)	.84(.16)	.90(.24)	.93(.35)
30	.57(.07)	.79(.11)	.88(.21)	.93(.34)	.95(.48)
40	.62(.07)	.83(.14)	.91(.26)	.94(.43)	.96(.60)
50	.65(.08)	.85(.17)	.92(.32)	.95(.51)	.96(.70)

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 282-283

1538 – 9472/06/$95.00

Statistical Pronouncements V

"I commenced a deliberate system of time-killing, which united some profit with a cheering-up of the heavy hours. As soon as I came on deck and took my place and regular walk, I began with repeating over to myself a string of matters which I had in my memory, in regular order. First, the multiplication table" – Richard Henry Dana (1841, *Two years before the mast: A personal narrative of life at sea*).

"I had been to school most all the time and could spell and read and write just a little, and could say the multiplication table up to six times seven is thirty-five, and I don't reckon I could ever get any further than that if I was to live forever. I don't take no stock in mathematics, anyway" – Mark Twain (Samuel L. Clemens) (1884, *The adventures of Huckleberry Finn*).

"He said he was repeating the multiplication table over and over to steady his nerves and for pity's sake not to interrupt him, because if he stopped for a moment he got frightened and forgot everything he ever knew, but the multiplication table kept all his facts firmly in their proper place!" – Lucy Maud Montgomery (1908, *Anne of Green Gables*).

"It is a capital mistake to theorize before one has data" – Arthur Conan Doyle (1891, "*A scandal in Bohemia: The adventures of Sherlock Holmes*").

"Other things are *not* always equal" – Edward L. Thorndike (1922, *The psychology of arithmetic*, NY: MacMillian, p. 12).

"There is never a quantity which does not measure some quality, and never an existing quality that in non-quantitative. Even our halos vary in diameter" – William A. McCall (1922, *How to measure in education*, NY: Macmillian, p. 4).

"At least a half a dozen scales now exist by which it would have been possible to measure the quality of the Handwriting on the Wall" – William A. McCall (*ibid*).

"Poincaré confesses that he is a rather poor numerical calculator, and so am I" – Jacques Salomon Hadamard (1945, "*An essay on the psychology of invention in the mathematical field*", Princeton: Princeton University Press, p. 58).

"It is of utmost importance… that the third kind of error in statistical consulting be emphasized…*the error committed by giving the right answer to the wrong problem*" – A. W. Kimball (1957, *Journal of the American Statistical Association*, *52*, p. 134).

"An incident from Pearson's infancy which Julia Bell once related to me… She had asked him what was the first thing he could remember… 'Well,' he said,… 'I was sitting in a high chair and I was sucking my thumb. Someone told me to stop sucking it and said that unless I did so the thumb would wither away. I put my two thumbs together and looked at them for a long time. 'They look alike to me' I said to myself, 'I can't see that the thumb I suck is any smaller than the other. I wonder if she could be lying to me' " – Helen M. Walker (1958, The contributions of Karl Pearson, *Journal of the American Statistical Association*, *53*, p. 13)

"Pearson was a prodigious and compulsive worker. I remember asking him once how he had time to write so much and compute so much… he replied… 'I never answer a telephone or attend a committee meeting' " – Samuel A. Stouffer (1958, Karl Pearson – An appreciation on the 100th anniversary of his birth, *ibid*, p. 23)

"We have become accustomed to-day to a standard of published mathematical proof which can hide rather than reveal the actual process by which discoveries are made" – B. L. Welch (1958, *ibid*, p. 786)

"Scientists are rarely given ladies and cups of tea to experiment with" – N. T. Gridgeman (1959, Book Review, *Journal of the American Statistical Association*, *54*, p. 778).

"There is often great temptation to *assume*… two independent runs… will inevitably be in 'reasonable agreement,' and hence that there is no need of repeating the measurement process. This is one of the most hazardous assumptions which can be made in any field of science" – Samuel Stanley Wilks (1961, Some aspects of quantification, *Quantification*, (Harry Woolf, Ed.), Indianapolis: Bobbs-Merrill, p. 6-7).

"It is a genius that leaps ahead of the facts, leaving the rather different talent of the experimentalist and the instrumentalist to catch up" – Thomas S. Kuhn (1961, Measurement in modern physical science, *ibid*, p. 42).

"The American Statistical Association is, I am told, the second oldest learned society [in America], the American Philosophical Society being the oldest. This news usually shocks our colleagues in economics, whose American Economic Association was founded forty-six years later; in science, whose American Association for the Advancement of Science came along nine years later; in modern languages, forty-four years later; in physics, sixty years later; in chemistry, thirty-seven years later, and so forth. Statistics is somehow still regarded by some as a new and youthful subject, one which is by now perhaps beyond hope of ever maturing" – W. Allen Wallis (1966, Economic statistics and economic policy, *Journal of the American Statistical Association*, *61*, p. 2)

"Neither statisticians nor philosophers build bombs, automate production, cure cancer, meet payrolls, or carry precincts" – W. Allen Wallis (*ibid*, p. 2-3)

"The complaint that statistics is never the star of the show is not unlike the complaint that a lineman on a football team rarely scores any points. One who is not temperamentally suited to being a lineman ought not to take up statistics." – W. Allen Wallis (1966, *ibid*, p. 3)

"The Wilcoxon rank-sum test…show[s] only slight losses in both large and small sample efficiency relative to the t-test in the normal case, while in many non-normal cases, efficiency exceeds 100%" – Duane Meeter (1967, Book Review, *Journal of the American Statistical Association*, *62*, p. 1505)

"If your experiment needs statistics, you ought to have done a better experiment" – Ernest Rutherford (1871-1937, cited in N. T. J. Bailey, 1967, *The mathematical approach to biology and medicine*, NY: Wiley).

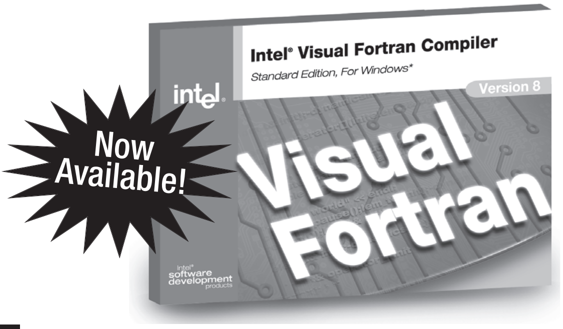
"I fear that the first act of most social scientists upon seeing a contingency table is to compute a chi-square for it" – Frederick Mosteller (1968, Association and estimation in contingency tables, *Journal of the American Statistical Association*, *63*, p. 1).

"Any sensible analysis would reject this theory - even a Bayesian t-test using an informationless prior" – Irwin D. J. Bross (1969, Applications of probability: Science vs. pseudoscience, *Journal of the American Statistical Association*, *64*, p. 52)

"The acid test of a good scientist is how he behaves when a favorite theory is refuted by incontrovertible facts" – Irwin D. J. Bross (*ibid*, p. 52).

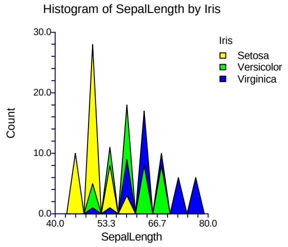
"All of us are unable to see any virtue in criticisms of our work but in this dimension of personality Fisher undoubtedly excelled" – Oscar Kempthorne (1970, Book Review, *Journal of the American Statistical Association*, *65*, p. 456.)

"During my 18 years," Mantle said, "I came to bat almost 10,000 times. I struck out about 1,700 times and walked maybe 1,800 times. You figure a ballplayer will average about 500 at bats a season. That means I played seven years without ever hitting a ball" – Mickey Mantle (1970, *San Francisco Chronicle*).



Now Available!
Intel® Visual Fortran Compiler
Standard Edition, For Windows*
intel
Version 8
Visual Fortran
intel® software development products

NCSS
329 North 1000 East
Kaysville, Utah 84037



Announcing NCSS 2004

Seventeen New Procedures

NCSS 2004 is a new edition of our popular statistical **NCSS** package that adds seventeen new procedures.

New Procedures

Two Independent Proportions
Two Correlated Proportions
One-Sample Binary Diagnostic Tests
Two-Sample Binary Diagnostic Tests
Paired-Sample Binary Diagnostic Tests
Cluster Sample Binary Diagnostic Tests
Meta-Analysis of Proportions
Meta-Analysis of Correlated Proportions
Meta-Analysis of Means
Meta-Analysis of Hazard Ratios
Curve Fitting
Tolerance Intervals
Comparative Histograms
ROC Curves
Elapsed Time Calculator
T-Test from Means and SD's
Hybrid Appraisal (Feedback) Model

Documentation

The printed, 330-page manual, called ***NCSS User's Guide V***, is available for $29.95. An electronic (pdf) version of the manual is included on the distribution CD and in the Help system.

Two Proportions

Several new exact and asymptotic techniques were added for hypothesis testing (null, noninferiority, equivalence) and calculating confidence intervals for the difference, ratio, and odds ratio. Designs may be independent or paired. Methods include: Farrington & Manning, Gart & Nam, Conditional & Unconditional Exact, Wilson's Score, Miettinen & Nurminen, and Chen.

Meta-Analysis

Procedures for combining studies measuring paired proportions, means, independent proportions, and hazard ratios are available. Plots include the forest plot, radial plot, and L'Abbe plot. Both fixed and random effects models are available for combining the results.

Curve Fitting

This procedure combines several of our curve fitting programs into one module. It adds many new models such as Michaelis-Menten. It analyzes curves from several groups. It compares fitted models across groups using computer-intensive randomization tests. It computes bootstrap confidence intervals.

Tolerance Intervals

This procedure calculates one and two sided tolerance intervals using both distribution-free (nonparametric) methods and normal distribution (parametric) methods. Tolerance intervals are bounds between which a given percentage of a population falls.

Comparative Histogram

This procedure displays a comparative histogram created by interspersing or overlaying the individual histograms of two or more groups or variables. This allows the direct comparison of the distributions of several groups.

Random Number Generator

Matsumoto's Mersenne Twister random number generator (cycle length > 10**6000) has been implemented.

Binary Diagnostic Tests

Four new procedures provide the specialized analysis necessary for diagnostic testing with binary outcome data. These provide appropriate specificity and sensitivity output. Four experimental designs can be analyzed including independent or paired groups, comparison with a gold standard, and cluster randomized.

ROC Curves

This procedure generates both binormal and empirical (nonparametric) ROC curves. It computes comparative measures such as the whole, and partial, area under the ROC curve. It provides statistical tests comparing the AUC's and partial AUC's for paired and independent sample designs.

Hybrid (Feedback) Model

This new edition of our hybrid appraisal model fitting program includes several new optimization methods for calibrating parameters including a new genetic algorithm. Model specification is easier. Binary variables are automatically generated from class variables.

Statistical Innovations Products

Through a *special arrangement* with Statistical Innovations (S.I.), NCSS customers will receive $100 discounts on:

Latent GOLDÒ - latent class modeling
SI-CHAIDÒ - segmentation trees
GOLDMineRÒ - ordinal regression

For demos and other info visit:
www.statisticalinnovations.com

Please rush me the following products:

Qty

___ **NCSS 2004 CD upgrade from NCSS 2001**, $149.95 .... $_____

___ **NCSS 2004 User's Guide V**, $29.95... $_____

___ **NCSS 2004 CD, upgrade from earlier versions**, $249.95........... $_____

___ **NCSS 2004 Deluxe (CD and Printed Manuals),** $599.95........... $_____

___ **PASS 2002 Deluxe**, $499.95 .... $_____

___ **Latent Gold® from S.I.,** $995 - $100 NCSS Discount = $895..... $_____

___ **GoldMineR® from S.I.**, $695 - $100 NCSS Discount = $595 ..... $_____

___ **CHAID® Plus from S.I.,** $695 - $100 NCSS Discount = $595.... $_____

Approximate shipping--depends on which manuals are ordered (U.S: $10 ground, $18 2-day, or $33 overnight) (Canada $24) (All other countries $10) (Add $5 U.S. or $40 International for any S.I. product) .... $_____

Total.......... $_____

TO PLACE YOUR ORDER
CALL: (800) 898-6109 **FAX:** (801) 546-3907
ONLINE: www.ncss.com
MAIL: NCSS, 329 North 1000 East, Kaysville, UT 84037

My Payment Option:

___ Check enclosed
___ Please charge my: __VISA __ MasterCard ___Amex
___ Purchase order attached________________________

Card Number ________________________________Exp ________

Signature____

Telephone:

( ) ____

Email:

____

Ship to:

NAME ____

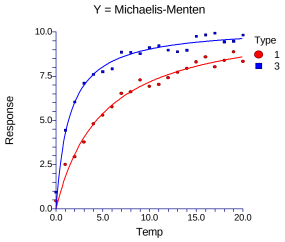
ADDRESS ____

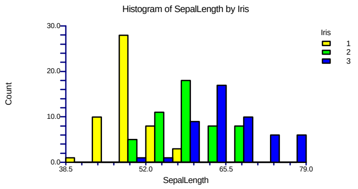
ADDRESS ____

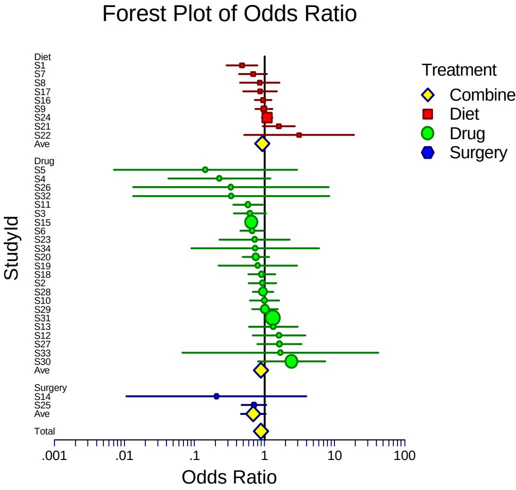
ADDRESS ____

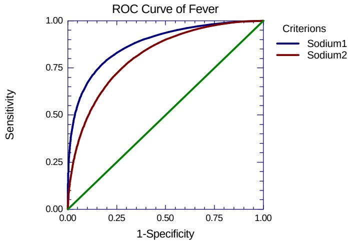
CITY ___________________________ STATE _______________

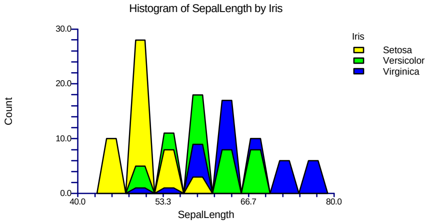
ZIP/POSTAL CODE ____________________COUNTRY _____________











Statistical and Graphics Procedures Available in NCSS 2004

Analysis of Variance / T-Tests

Analysis of Covariance
Analysis of Variance
Barlett Variance Test
Crossover Design Analysis
Factorial Design Analysis
Friedman Test
Geiser-Greenhouse Correction
General Linear Models
Mann-Whitney Test
MANOVA
Multiple Comparison Tests
One-Way ANOVA
Paired T-Tests
Power Calculations
Repeated Measures ANOVA
T-Tests – One or Two Groups
T-Tests – From Means & SD's
Wilcoxon Test

Time Series Analysis

ARIMA / Box - Jenkins
Decomposition
Exponential Smoothing
Harmonic Analysis
Holt - Winters
Seasonal Analysis
Spectral Analysis
Trend Analysis

***New Edition in 2004**

Plots / Graphs

Bar Charts
Box Plots
Contour Plot
Dot Plots
Error Bar Charts
Histograms
Histograms: Combined*
Percentile Plots
Pie Charts
Probability Plots
ROC Curves*
Scatter Plots
Scatter Plot Matrix
Surface Plots
Violin Plots

Experimental Designs

Balanced Inc. Block
Box-Behnken
Central Composite
D-Optimal Designs
Fractional Factorial
Latin Squares
Placket-Burman
Response Surface
Screening
Taguchi

Regression / Correlation

All-Possible Search
Canonical Correlation
Correlation Matrices
Cox Regression
Kendall's Tau Correlation
Linear Regression
Logistic Regression
Multiple Regression
Nonlinear Regression
PC Regression
Poisson Regression
Response-Surface
Ridge Regression
Robust Regression
Stepwise Regression
Spearman Correlation
Variable Selection

Quality Control

Xbar-R Chart
C, P, NP, U Charts
Capability Analysis
Cusum, EWMA Chart
Individuals Chart
Moving Average Chart
Pareto Chart
R & R Studies

Survival / Reliability

Accelerated Life Tests
Cox Regression
Cumulative Incidence
Exponential Fitting
Extreme-Value Fitting
Hazard Rates
Kaplan-Meier Curves
Life-Table Analysis
Lognormal Fitting
Log-Rank Tests
Probit Analysis
Proportional-Hazards
Reliability Analysis
Survival Distributions
Time Calculator*
Weibull Analysis

Multivariate Analysis

Cluster Analysis
Correspondence Analysis
Discriminant Analysis
Factor Analysis
Hotelling's T-Squared
Item Analysis
Item Response Analysis
Loglinear Models
MANOVA
Multi-Way Tables
Multidimensional Scaling
Principal Components

Curve Fitting

Bootstrap C.I.'s*
Built-In Models
Group Fitting and Testing*
Model Searching
Nonlinear Regression
Randomization Tests*
Ratio of Polynomials
User-Specified Models

Miscellaneous

Area Under Curve
Bootstrapping
Chi-Square Test
Confidence Limits
Cross Tabulation
Data Screening
Fisher's Exact Test
Frequency Distributions
Mantel-Haenszel Test
Nonparametric Tests
Normality Tests
Probability Calculator
Proportion Tests
Randomization Tests
Tables of Means, Etc.
Trimmed Means
Univariate Statistics

Meta-Analysis*

Independent Proportions*
Correlated Proportions*
Hazard Ratios*
Means*

Binary Diagnostic Tests*

One Sample*
Two Samples*
Paired Samples*
Clustered Samples*

Proportions

Tolerance Intervals*
Two Independent*
Two Correlated*
Exact Tests*
Exact Confidence Intervals*
Farrington-Manning*
Fisher Exact Test
Gart-Nam* Method
McNemar Test
Miettinen-Nurminen*
Wilson's Score* Method
Equivalence Tests*
Noninferiority Tests*

Mass Appraisal

Comparables Reports
Hybrid (Feedback) Model*
Nonlinear Regression
Sales Ratios

PASS 2002

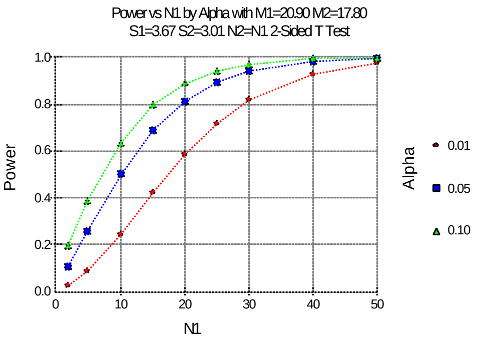
Power Analysis and Sample Size Software from NCSS

PASS performs power analysis and calculates sample sizes. Use it before you begin a study to calculate an appropriate sample size (it meets the requirements of government agencies that want technical justification of the sample size you have used). Use it after a study to determine if your sample size was large enough. ***PASS*** calculates the sample sizes necessary to perform all of the statistical tests listed below.

A power analysis usually involves several "what if" questions. ***PASS*** lets you solve for power, sample size, effect size, and alpha level. It automatically creates appropriate tables and charts of the results.

PASS is accurate. It has been extensively verified using books and reference articles. Proof of the accuracy of each procedure is included in the extensive documentation.

PASS is a standalone system. Although it is integrated with ***NCSS***, you do not have to own ***NCSS*** to run it. You can use it with any statistical software you want.



PASS comes with two manuals that contain tutorials, examples, annotated output, references, formulas, verification, and complete instructions on each procedure. And, if you cannot find an answer in the manual, our free technical support staff (which includes a PhD statistician) is available.

System Requirements

PASS runs on Windows 95/98/ME/NT/2000/XP with at least 32 megs of RAM and 30 megs of hard disk space.

PASS sells for as little as **$449.95**.

***PASS* Beats the Competition!**

No other program calculates sample sizes and power for as many different statistical procedures as does *PASS*. Specifying your input is easy, especially with the online help and manual.

PASS automatically displays charts and graphs along with numeric tables and text summaries in a portable format that is cut and paste compatible with all word processors so you can easily include the results in your proposal.

Choose ***PASS***. It's more comprehensive, easier-to-use, accurate, and less expensive than any other sample size program on the market.

Trial Copy Available

You can try out ***PASS*** by downloading it from our website. This trial copy is good for 30 days. We are sure you will agree that it is the easiest and most comprehensive power analysis and sample size program available.

Analysis of Variance
Factorial AOV
Fixed Effects AOV
Geisser-Greenhouse
MANOVA*
Multiple Comparisons*
One-Way AOV
Planned Comparisons
Randomized Block AOV
New Repeated Measures AOV*

Regression / Correlation
Correlations (one or two)
Cox Regression*
Logistic Regression
Multiple Regression
Poisson Regression*
Intraclass Correlation
Linear Regression

Proportions
Chi-Square Test
Confidence Interval
Equivalence of McNemar*
Equivalence of Proportions
Fisher's Exact Test
Group Sequential Proportions
Matched Case-Control
McNemar Test
Odds Ratio Estimator
One-Stage Designs*
Proportions – 1 or 2
Two Stage Designs (Simon's)
Three-Stage Designs*

Miscellaneous Tests
Exponential Means – 1 or 2*
ROC Curves – 1 or 2*
Variances – 1 or 2

T Tests
Cluster Randomization
Confidence Intervals
Equivalence T Tests
Hotelling's T-Squared*
Group Sequential T Tests
Mann-Whitney Test
One-Sample T-Tests
Paired T-Tests
Standard Deviation Estimator
Two-Sample T-Tests
Wilcoxon Test

Survival Analysis
Cox Regression*
Logrank Survival -Simple
Logrank Survival - Advanced*
Group Sequential - Survival
Post-Marketing Surveillance
ROC Curves – 1 or 2*

Group Sequential Tests
Alpha Spending Functions
Lan-DeMets Approach
Means
Proportions
Survival Curves

Equivalence
Means
Proportions
Correlated Proportions*

Miscellaneous Features
Automatic Graphics
Finite Population Corrections
Solves for any parameter
Text Summary
Unequal N's

***New in PASS 2002**

NCSS Statistical Software • 329 North 1000 East • Kaysville, Utah 84037
Internet (download free demo version): http://www.ncss.com • Email: sales@ncss.com
Toll Free: (800) 898-6109 • Tel: (801) 546-0445 • Fax: (801) 546-3907

PASS 2002 adds power analysis and sample size to your statistical toolbox

WHAT'S NEW IN PASS 2002?

Thirteen new procedures have been added to ***PASS*** as well as a new home-base window and a new Guide Me facility.

MANY NEW PROCEDURES

The new procedures include a new multi-factor repeated measures program that includes multivariate tests, Cox proportional hazards regression, Poisson regression, MANOVA, equivalence testing when proportions are correlated, multiple comparisons, ROC curves, and Hotelling's T-squared.

TEXT STATEMENTS

The text output translates the numeric output into easy-to-understand sentences. These statements may be transferred directly into your grant proposals and reports.

GRAPHICS

The creation of charts and graphs is easy in ***PASS***. These charts are easily transferred into other programs such as MS PowerPoint and MS Word.

NEW USER'S GUIDE II

A new, 250-page manual describes each new procedure in detail. Each chapter contains explanations, formulas, examples, and accuracy verification.

The complete manual is stored in PDF format on the CD so that you can read and printout your own copy.

GUIDE ME

The new *Guide Me* facility makes it easy for first time users to enter parameter values. The program literally steps you through those options that are necessary for the sample size calculation.

NEW HOME BASE

A new home base window has been added just for PASS users. This window helps you select the appropriate program module.

COX REGRESSION

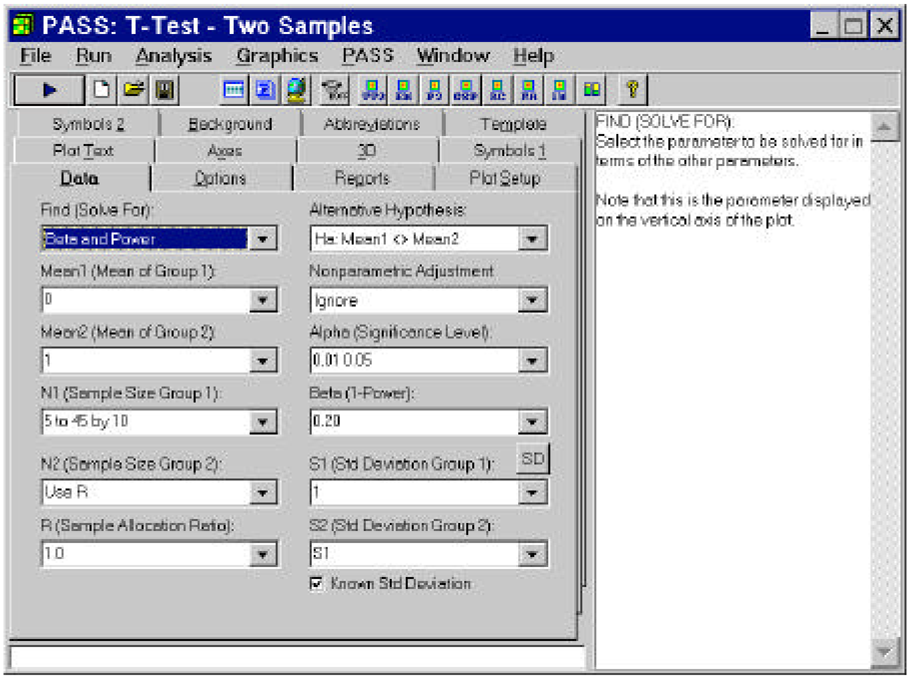
A new Cox regression procedure has been added to perform power analysis and sample size calculation for this important statistical technique.

REPEATED MEASURES

A new repeated-measures analysis module has been added that lets you analyze designs with up to three grouping factors and up to three repeated factors. The analysis includes both the univariate F test and three common multivariate tests including Wilks Lambda.

RECENT REVIEW

In a recent review, 17 of 19 reviewers selected ***PASS*** as the program they would recommend to their colleagues.



Find (Solve For): Beta and Power

Alternative Hypothesis: Ha: Mean1 <> Mean2

Mean1 (Mean of Group 1): 0

Nonparametric Adjustment: Ignore

Mean2 (Mean of Group 2): 1

Alpha (Significance Level): 0.01 0.05

N1 (Sample Size Group 1): 5 to 45 by 10

Beta (1-Power): 0.20

N2 (Sample Size Group 2): Use R

S1 (Std Deviation Group 1): 1

R (Sample Allocation Ratio): 1.0

S2 (Std Deviation Group 2): S1

☑ Known Std Deviation

FIND (SOLVE FOR):
Select the parameter to be solved for in terms of the other parameters.

Note that this is the parameter displayed on the vertical axis of the plot.

PASS calculates sample sizes for...

Please rush me my own personal license of ***PASS 2002.***

Qty

___ **PASS 2002 Deluxe** (CD and User's Guide)**:** $499.95..............$ ______

___ **PASS 2002 CD (electronic documentation):** $449.95..........$ ______

___ **PASS 2002 5-User Pack (CD & 5 licenses):** $1495.00........$ ______

___ **PASS 2002 25-User Pack (CD & 25 licenses):** $3995.00....$ ______

___ **PASS 2002 User's Guide II (printed manual):** $30.00.........$ ______

___ **PASS 2002 Upgrade CD** for *PASS 2000* users**:** $149.95 ....$ ______

Typical Shipping & Handling: USA: $9 regular, $22 2-day, $33 overnight. Canada: $19 Mail. Europe: $50 Fedex......................$ ______

Total: ....$ ______

FOR FASTEST DELIVERY, ORDER ONLINE AT

WWW.NCSS.COM

Email your order to sales@ncss.com
Fax your order to (801) 546-3907
NCSS, 329 North 1000 East, Kaysville, UT 84037
(800) 898-6109 or (801) 546-0445

My Payment Options:

____ Check enclosed
___ Please charge my: __VISA __MasterCard __Amex
___ Purchase order enclosed

Card Number

____Expires______

Signature____
Please provide daytime phone:

( )____

Ship my *PASS 2002* to:

NAME

COMPANY

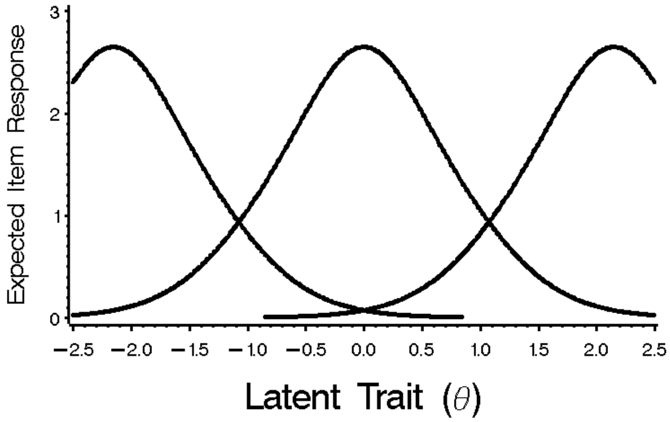
ADDRESS

CITY/STATE/ZIP

COUNTRY (IF OTHER THAN U.S.)

Introducing GGUM2004

Item Response Theory Models for Unfolding



The new GGUM2004 software system estimates parameters in a family of item response theory (IRT) models that unfold polytomous responses to questionnaire items. These models assume that persons and items can be jointly represented as locations on a latent unidimensional continuum. A single-peaked, nonmonotonic response function is the key feature that distinguishes unfolding IRT models from traditional, "cumulative" IRT models. This response function suggests that a higher item score is more likely to the extent that an individual is located close to a given item on the underlying continuum. Such single-peaked functions are appropriate in many situations including attitude measurement with Likert or Thurstone scales, and preference measurement with stimulus rating scales. This family of models can also be used to determine the locations of respondents in particular developmental processes that occur in stages.

The GGUM2004 system estimates item parameters using marginal maximum likelihood, and person parameters are estimated using an expected *a posteriori* (EAP) technique. The program allows for up to 100 items with 2-10 response categories per item, and up to 2000 respondents. GGUM2004 is compatible with computers running updated versions of Windows 98 SE, Windows 2000, and Windows XP. The software is accompanied by a detailed technical reference manual and a new Windows user's guide. ***GGUM2004 is free*** and can be downloaded from:

http://www.education.umd.edu/EDMS/tutorials

GGUM2004 improves upon its predecessor (GGUM2000) in several important ways:

- It has a user-friendly graphical interface for running commands and displaying output.
- It offers real-time graphics that characterize the performance of a given model.
- It provides new item fit indices with desirable statistical characteristics.
- It allows for missing item responses assuming the data are missing at random.
- It allows the number of response categories to vary across items.
- It estimates model parameters more quickly.

Start putting the power of unfolding IRT models to work in your attitude and preference measurement endeavors. Download your free copy of GGUM2004 today!

JOIN DIVISION 5 OF APA!

The Division of Evaluation, Measurement, and Statistics of the American Psychological Association draws together individuals whose professional activities and/or interests include assessment, evaluation, measurement, and statistics. The disciplinary affiliation of division membership reaches well beyond psychology, includes both members and non-members of APA, and welcomes graduate students.

Benefits of membership include:

- subscription to *Psychological Methods* or *Psychological Assessment* (student members, who pay a reduced fee, do not automatically receive a journal, but may do so for an additional $18)
- *The Score* – the division's quarterly newsletter
- Division's Listservs, which provide an opportunity for substantive discussions as well as the dissemination of important information (e.g., job openings, grant information, workshops)

Cost of membership: $38 (**APA membership not required**); student membership is only $8

For further information, please contact the Division's Membership Chair, Yossef Ben-Porath (ybenpora@kent.edu) or check out the Division's website:

http://www.apa.org/divisions/div5/

ARE YOU INTERESTED IN AN ORGANIZATION DEVOTED TO EDUCATIONAL AND BEHAVIORAL STATISTICS?

Become a member of the **Special Interest Group - Educational Statisticians** of the American Educational Research Association (SIG-ES of AERA)!

The mission of SIG-ES is to increase the interaction among educational researchers interested in the theory, applications, and teaching of statistics in the social sciences.

Each Spring, as part of the overall AERA annual meeting, there are seven sessions sponsored by SIG-ES devoted to educational statistics and statistics education.
We also publish a twice-yearly electronic newsletter.

Past issues of the SIG-ES newsletter and other information regarding SIG-ES can be found at http://orme.uark.edu/edstatsig.htm

To join SIG-ES you must be a member of AERA. Dues are $5.00 per year.

For more information, contact Joan Garfield, President of the SIG-ES, at jbg@umn.edu.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2-282

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.

STATISTICIANS

HAVE YOU VISITED THE

Mathematics Genealogy Project?

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

Please visit our web site

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

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

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

harry.coonce@ndsu.nodak.edu

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

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